REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID - TLEMCEN

THÈSE

Présentée à :

FACULTE DES SCIENCES – DEPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

DOCTORAT EN SCIENCES

Spécialité : INFORMATIQUE

Par :

Mr GAOUAR Adil

Sur le thème

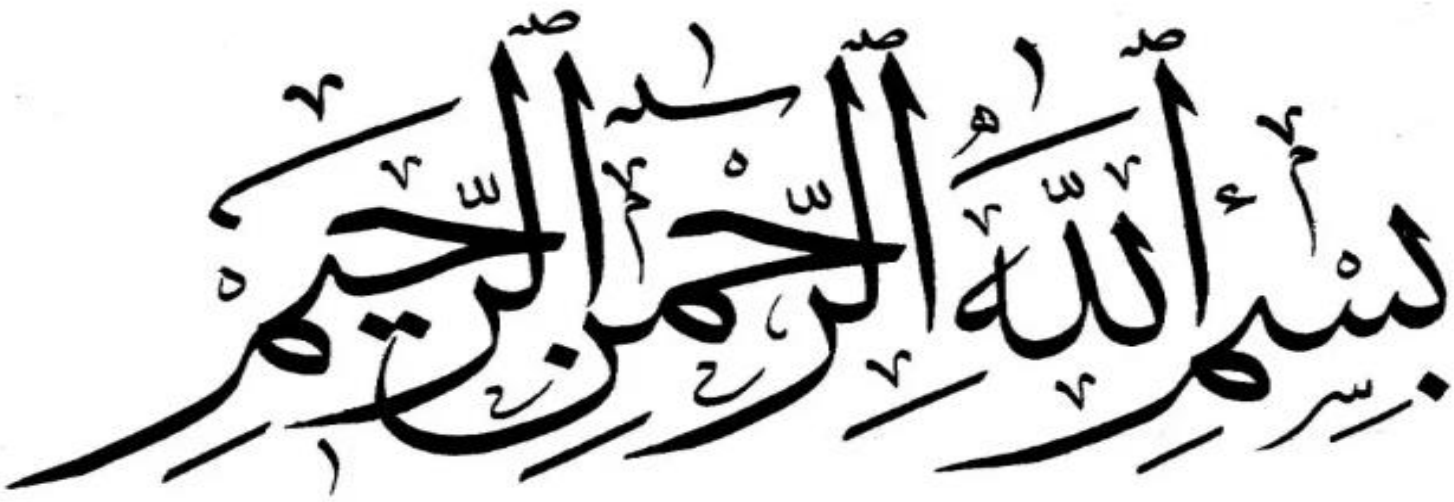
Explicabilité des modèles intelligents basés sur l'apprentissage profond (Deep Learning) : Détection automatique du paludisme

---

Soutenue publiquement le 13 décembre 2025 à Tlemcen devant le jury composé de :

Mr CHIKH Azeddine	Professeur	Université de Tlemcen	Président
Mr RAHMOUN Abdelatif	Professeur	ESI - Sidi Bel Abbés	Directeur de thèse
Mr BOUKLI HACENE Ismail	Professeur	Université de Tlemcen	Examinateur
Mr MALIKI Fouad	Maître de Conférences A	ESSA - Tlemcen	Examinateur
Mr MEGNAFI Hicham	Maître de Conférences A	ESSA – Tlemcen	Examinateur
Mr MERZOUG MOHAMMED	Maître de Conférences A	Université de Tlemcen	Examinateur
Mme HAMZA-CHERIF Souaad	Maître de Conférences A	Université de Tlemcen	Invitée

Année Universitaire 2025-2026


بسم الله الرحمن الرحيم

Remerciements

Louange à ALLAH qui m'a donné la force et la détermination pour finaliser ce modeste travail

Je remercie en premier lieu le Professeur Abdellatif RAHMOUN, mon encadrant et Directeur de thèse ; professionnellement, j'ai beaucoup appris avec lui tout au long de ces années d'étude pendant lesquelles, à maintes reprises, son expérience et ses conseils m'ont été d'une grande utilité et d'un apport inestimable pour ma formation post-graduée. Je lui suis reconnaissant de la confiance qu'il m'a témoignée pour mener à bien ce modeste travail.

Je tiens aussi à remercier les membres du jury d'avoir accepté de juger notre travail, ainsi que pour le temps et l'effort qu'ils ont consacrés à l'examen de cette thèse.

Une mention spéciale va à notre cher Professeur Azeddine Chikh pour son aide, ses encouragements, son soutien indéfectible et surtout pour l'enseignant qu'il a été pour nous.

Mes vifs remerciements vont au Professeur Boukli Hacéne Ismail du département de GBM Tlemcen, qui a toujours su m'encourager et me motiver. Je lui suis énormément reconnaissant pour ses conseils judicieux et pour son amitié.

Mes vifs remerciements vont à Mr Fouad Maliki, Maître de Conférences à l'École Supérieure en Sciences Appliquées de Tlemcen, d'avoir accepté de juger ce travail. Vos remarques et suggestions Monsieur seront de grandes valeurs pour enrichir cette thèse.

C'est avec un grand plaisir que je remercie Mr Hicham Megnafi, Maitre de Conférences à l'École Supérieure en Sciences Appliquées de Tlemcen, pour avoir accepté de faire partie de mon jury. Vos conseils et suggestions, tant scientifiques que techniques, enrichiront ce travail.

Mes chaleureux remerciements vont aussi à Mr Merzoug Mohammed, Maître de Conférences à l'Université de Tlemcen, pour avoir accepté de faire partie de ce jury et examiner nos travaux de recherches. Vos remarques pertinentes, suggestions et conseils ne feront qu'enrichir nos travaux.

Mes sincères remerciements vont aussi à notre Doyen de la Faculté de Technologie, Monsieur le Professeur Amine Chikh, pour l'enseignant, le père et le frère qu'il représente pour moi. Son expérience et ses conseils m'ont été d'une grande utilité et d'un apport inestimable pour ma formation post graduée et ma carrière d'enseignant.

Je tiens à remercier très sincèrement Madame la Professeure Narjès Bellamine Ben Saoud, Directrice de L'ENSI de l'Université de Manouba (Tunisie), ainsi que notre cher frère et ami le Professeur Cherif Ameur, Vice-Recteur de l'Université de Manouba, pour m'avoir accueilli et ouvert les portes de leur Université, ainsi que pour tout le temps qu'ils m'ont consacré durant mon stage de préparation du Doctorat. Ce fut un grand plaisir de travailler avec eux.

Mes plus sincères remerciements vont à ma collègue, cousine et amie Mme Souad Hamza Cherif, pour l'aide incommensurable et les encouragements qu'elle m'a apporté ; ainsi qu'à notre docteure préférée Mlle ELAOUABER Zineb Aziza pour son aide et son dévouement.

Une mention spéciale et mes plus chaleureux remerciements vont à mon cher frère et ami Mr El Habib Daho Mostafa dit "Si Djaâfar", que je ne saurai remercier pour tout ce qu'il a fait pour moi. Son aide, ses encouragements et son soutien indéfectible m'ont plus que porté pour accomplir ces travaux.

Mes remerciements les plus sincères vont à mes collègues et amies du Département d'Informatique, Mme Zeyneb El Yebdri et Mme Iles Nawel, pour leur aide et leurs encouragements, tout au long de mon parcours en tant qu'étudiant et au cours de ma post- graduation.

Mes collègues, ami(e)s, frères et sœurs Hamza-Cherif Souaad, Lamia Kazi, Bensmail Ilhem, Belaidi Asmaa, Belarouci sara, Rerbal Souhila, Baakek Yettou Nour El Houda, Bouckli Hacène Ismail, Lazouni Amine, Messadi Mahammed, Hamza-Cherif Lotfi, Dib Nabil, Taleb Tariq, Kholkhal Mourad, Benali Redouane, Kerai Salim, Taouli Sidi Ahmed, Behadada Omar, Djebbari Abdelghani, Ouchdi Mohamed, Baba Ahmed Ilyes, Youbi Redha, ainsi qu'à tous et à toutes mes collègues du département de GBM, je vous dis à tous et à chacun MERCI pour votre amitié et vos encouragements.

Mes vifs remerciements vont aussi aux doyens de nos enseignants du département de GBM, Mr Debbale Sidi Mohammed, Mr Berekci Reguig Fethi, Mr Rahmoun Fethi, Mr Bechar Hassane, Mr Hadj Slimane Zine eddine et Mr Bessaid Abdelhafid.

Un grand merci aussi à mes amis du service de scolarité Bettouaf Hicham, Dahaoui Mourad et El Hadi, Bensmaine Yassir.

Je ne saurais oublier toutes les autres personnes, mes collègues, ami(e)s et les secrétaires Houcine et Lamia et hayatte du Département de GBM.

Dédicaces

Je dédie ce modeste travail à….

À *mes très chers parents, puisse ALLAH les garder et les protéger, ils m'ont soutenu tout au long de mes études. À Ma très chère mère, que je ne saurais assez remercier pour m'avoir encouragé et boosté afin de poursuivre mes études et ce, malgré les responsabilités et les aléas de la vie, ainsi que pour avoir été un pilier inébranlable dans ma vie. À Mon très cher père qui a été toujours présent pour moi avec sa sagesse et ses conseils inestimables.*

À *ma femme pour ses encouragements et ses sacrifices durant la réalisation de ce travail….*

À *mes filles Lilya, Lina et Sarah pour leur amour et leur support affectif*

À *mes chers frères et sœurs qui m'ont aidé et porté par leur Amour et leur affection tout au long de ma vie, ainsi qu'à mes beaux-frères, Tariq Boursali et Amine Yadi pour leurs encouragements et leur gentillesse.*

À *mes beaux-parents pour leurs encouragements et dévouement ainsi qu'à mes beaux-frères Adil, Nassim et leurs femmes.*

À *mes chers neveux et mes chères nièces.*

À *mon cher et regretté collègue et ami Mr Bouallem Attou Slimani puisse ALLAH l'accueillir dans son vaste paradis.*

Citation choisie : Proverbe indien

Celui qui ne sait pas et qui sait qu'il ne sait pas instruis-le,
Celui qui sait et qui ne sait pas qu'il sait initie-le,
Celui qui ne sait pas et qui ne sait pas qu'il ne sait pas fuis-le,
Celui qui sait et qui sait qu'il sait suis-le...

ملخص

لا تزال الملاريا تمثل تحديًا عالميًا للصحة العامة، حيث تؤثر على أكثر من 247 مليون شخص وتسبب 619,000 حالة وفاة في جميع أنحاء العالم في عام 2024 (وفقًا لمنظمة الصحة العالمية) .التشخيص السريع ضروري للعلاج الفعال ولتحسين فرص بقاء المرضى على قيد الحياة.

في هذه الدراسة، نقترح إطار تعلم عميق قابل للتفسير لتشخيص الملاريا بدقة باستخدام صور مسحات الدم .كما نقوم بتقييم ومقارنة عدة نماذج أساسية للتعلم العميق (DL) (الأساسيات)، مثل VGG-16 و VGG-19المخصصتين، بالإضافة إلى نماذج DL الأحدث مثل Vision Transformer (ViT) وMobileNet، ولأول مرة، شبكة الذاكرة الطويلة القصيرة المكدسة -stacked) (LSTMمع آلية الانتباه للكشف التلقائي عن الملاريا من صور مسحات الدم .تم تدريب هذه النماذج والتحقق من صحتها على مجموعة بيانات متاحة للجمهور تحتوي على أكثر من 27,000 صورة ملطخة بالدم مصنفة .أظهرت الدراسة المقارنة والإحصائية التي أجريت في هذا البحث أن نموذج Stacked-LSTM المقترح مع آلية الانتباه تفوق على جميع الأساليب الأخرى، محققًا دقة تصنيف (0.9912)، وحساسية، ونوعية، ودقة، ودرجة(0.9911) F1 ، ومساحة تحت المنحنى (AUC) تفوق جميع النماذج الأخرى .على الرغم من أدائها القوي، غالبًا ما تُعتبر هذه النماذج "صناديق سوداء" بسبب نقص الشفافية في عملية اتخاذ القرار، مما يشكل تحديات كبيرة في التطبيقات الطبية والمجالات التي تكون فيها حياة الإنسان على المحك .لمعالجة ذلك، قمنا بدمج تقنيات الذكاء الاصطناعي القابل للتفسير(XAI) ، وهي Grad-CAM وLIME، لتحسين قابلية تفسير النموذج .تظهر نتائجنا القيمة التكميلية لدمج نماذج التعلم العميق عالية الأداء مع تقنيات الذكاء الاصطناعي القابل للتفسير (XAI) لتعزيز الثقة واليقين في التشخيص الطبي المدعوم بالذكاء الاصطناعي، مما يشير إلى أن نموذجنا يمكن أن يدعم الكشف المبكر والقابل للتفسير عن الملاريا في البيئات السريرية.

الكلمات المفتاحية: كشف الملاريا، التعلم العميق(DL) ، LSTM، الذكاء الاصطناعي القابل للتفسير(XAI) ، القابلية للتفسير، التفسير، LIME، تكنولوجيا الصحة، تطبيق ويب.

Abstract

Malaria remains a global public health challenge, affecting more than 247 million people and causing 619,000 deaths worldwide in 2024 (according to WHO). Rapid diagnosis is essential for effective treatment and to improve patients' chances of survival.

In this study, we propose an interpretable deep learning framework for accurate malaria diagnosis using blood smear images. Also, We evaluate and compare several baseline deep learning (DL) models (fundamentals), customized VGG-16 and VGG-19, as well as newer DL models such as Vision Transformer (ViT) and MobileNet, and, for the first time, a stacked long-short-term memory network (Stacked-LSTM) with an attention mechanism for automatic detection of malaria from blood smear images. These models were trained and validated on a publicly available dataset of over 27.000 labeled blood smear images. The comparative and statistical study conducted in this research showed us that the proposed Stacked-LSTM model with attention mechanism outperformed all other approaches, achieving a classification accuracy (0.9912), sensitivity, specificity, precision, F1 score (0.9911), and area under the curve (AUC) superior to all other models. Despite their solid performance, these models are often considered "black boxes" due to their lack of transparency in the decision-making process, which poses significant challenges in medical applications and fields where human life is at stake. To address this, we have integrated explainable AI (XAI) techniques, namely Grad-CAM and LIME, to improve the model's interpretability. Our results demonstrate the complementary value of combining high-performance deep learning models with XAI methods to enhance trust and certainty in AI-assisted medical diagnosis, suggesting that our model can support early and interpretable malaria detection in clinical environments.

Keywords: Malaria detection, deep learning (DL), LSTM, explainable AI (XAI), interpretability, explainability, LIME, health technology, web application

Résumé

Le paludisme reste un défi mondial de santé publique, touchant plus de 247 millions de personnes et causant 619 000 décès dans le monde en 2024 (selon l'OMS). Le diagnostic rapide est essentiel pour un traitement efficace et pour améliorer les chances de survie des patients.
Dans cette thèse, nous proposons un cadre d'apprentissage profond interprétable pour un diagnostic précis du paludisme à l'aide d'images de frottis sanguins. Nous évaluons également et comparons plusieurs modèles de deep learning (DL) de base (fondamentaux), les modèles personnalisés VGG-16 et VGG-19, ainsi que des modèles DL plus récents tels que Vision Transformer (ViT) et MobileNet, et, pour la première fois, un réseau de mémoire à long et court terme empilé (Stacked-LSTM) avec un mécanisme d'attention pour la détection automatique du paludisme à partir d'images de frottis sanguins. Ces modèles ont été entraînés et validés sur un ensemble de données publiques de plus de 27 000 images de frottis sanguins annotées. L'étude comparative et statistique menée dans cette recherche nous a montré que le modèle Stacked-LSTM proposé avec mécanisme d'attention a surpassé toutes les autres approches, atteignant une précision de classification (0,9912), une sensibilité, une spécificité, une précision, un score F1 (0,9911) et une aire sous la courbe (AUC) supérieurs à tous les autres modèles. Malgré leurs performances solides, ces modèles sont souvent considérés comme des « boîtes noires » en raison de leur manque de transparence dans le processus de prise de décision, ce qui pose des défis importants dans les applications médicales et les domaines où la vie humaine est en jeu. Pour y remédier, nous avons intégré des techniques d'IA explicable (XAI), à savoir Grad-CAM et LIME, afin d'améliorer l'interprétabilité du modèle. Nos résultats démontrent la valeur complémentaire de la combinaison de modèles de deep learning haute performance avec des méthodes XAI pour renforcer la confiance et la certitude dans le diagnostic médical assisté par IA, suggérant que notre modèle peut soutenir la détection précoce et interprétable du paludisme dans les environnements cliniques.

Mots Clés : Détection du paludisme, apprentissage profond (DL), LSTM, IA explicable (XAI), Interprétabilité, Explicabilité, LIME, technologie de la santé, application web.

Table des matières

Liste des Figures

Liste des tableaux

Liste des acronymes

ACRONYME	DESIGNATION
ACC	Accuracy (Exactitude)
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
ANN	Artificial Neural Network (Réseau de neurones artificiel)
AUC	Area Under the Curve (Aire sous la courbe)
BD	Bases de données
BDD	Bases de données
BN	Batch Normalization
BNN	Biological Neural Network
BSERESU-NET	Before-activation Squeeze-and-Excitation ResU-Net
BFCN	Butterfly-FCN
BI-LSTM	Bi-directionnal Long Short Term Memory
CAD	Computer-Aided Diagnostic
CAM	Class Activation Map
CART	Classification And Regression Tree
CNN	Convolutional Neural Network (Réseaux de Neurones Convolutifs)
CONV	Couche de convolution
CO	Cup Optique
CPU	Central Processing Unit
DCA	Decision Curve Analysis
DCNN	Deep Convolutional Neural Network
DLA	Deep Learning Algorithm (Algorithme d'apprentissage profond)
DL	Deep Learning
D-MNET	Deformable convolutional M-shaped Network
DNN	Deep Neural Network
DTDNN	Distributed Time Delay Neural Network
ELM	Extreme Learning Machine
FC	Couche entièrement connectée
FCN	Fully Connected Network (Réseau entièrement connecté)
FCM	Fuzzy C-Means
FDA	Food and Drug Administration
FDT	Technologie de Doublement De Fréquence
FFBPNN	Feed Forward Back Propagation Neural Networks
FOV	Field Of View
GAM	General Additive Models
GAP	Global Average Pooling

GD	Gradient Descent
GRAD-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
IOU	Intersection over Union
IRM	Imagerie par Résonance Magnétique
ISNT	Inférieur, Supérieur, Nasal, Temporel
ISODATA	Iterative Self-Organizing Data Analysis Technique
K-NN	k-Nearest Neighbor
LDA	Linear Discriminant Analysis (Analyse discriminante linéaire)
LEAKY RELU	Leaky Rectified Linear Unit
LIME	Local Interpretable Model-agnostic Explanations
LCMS	Lightweight Convolution Modules
LSTM	Long Short Term Memory
ML	Machine Learning
MLP	Multi Layer Perceptron (Perceptron multicouche)
MOBILE-NET	Mobile Network
MPS-NET	Multi-Path Scale Network
MSE	Mean Squared Error
NAG	Nestrov Accelerated Gradient
NB	Naive Bayes
NUMPY	Numerical Python
OMS	Organisation Mondiale de la Santé
PRC	Precision Recall Curve
POOL	Couche de pooling
RELU	Rectified Linear Unit
RESNET	Residual Network
RF	Random Forest
RGPD	Règlement Général sur la Protection des Données
RMSPROP	Root Mean Square Propagation
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SCIPY	Scientific Python
SGD	Stochastic Gradient Descent
SGDM	Stochastic Gradient Descent with Momentum
SHAP	SHapley Additive exPlanations
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
STACKED-LSTM	Stacked Long Short Term Memory
TANH	Tangente hyperbolique
TDM	Tomodensitométrie
TPU	Tensor Processing Unit

T-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VGG	Visual Geometry Group
VIT	Vision Transformer
XAI	eXplainable Artificial Intelligence

INTRODUCTION GÉNÉRALE

Introduction générale

Sommaire

I. Contexte

Le paludisme, causé par des parasites du genre **Plasmodium**, constitue un problème majeur de santé publique, affectant des millions de personnes chaque année. Selon l'Organisation Mondiale de la Santé (OMS), en 2023, plus de 247 millions de cas ont été signalés dans le monde, avec une majorité des décès survenant en Afrique subsaharienne. Cette maladie, pourtant évitable et traitable, continue de prospérer et de faire des ravages dans les régions où l'accès aux diagnostics et aux traitements est limité **(OMS, 2024)**.

Le diagnostic précoce et précis joue un rôle essentiel dans la lutte contre le paludisme ainsi que pour réduire considérablement le taux de mortalité (OMS, 2024). En effet, plus la détection du plasmodium dans le sang, est rapide, plus les chances de survie des patients augmentent. Les frottis sanguins, analysés au microscope, représentent le standard de référence pour l'identification des globules rouges infectés. Cependant, cette méthode est lente, dépend fortement de l'expertise et de la qualification des microscopistes, et peut être imprécise dans des environnements où les ressources médicales sont limitées **(OMS, 2024)**.

Dans ce contexte, au cours de cette dernière décennie, l'**intelligence artificielle (IA)** et plus particulièrement l'**apprentissage profond (DL)** (Deep Learning en Anglais), ont révolutionné le monde dans de nombreux domaines tels que la médecine, dont un mauvais diagnostic peut engager le pronostic vital d'une personne ; offrant ainsi des opportunités uniques pour automatiser et améliorer le diagnostic médical. En effet, Les modèles de DL, notamment les réseaux neuronaux profonds (DNN), les réseaux neuronaux convolutifs (CNN) et les réseaux neuronaux récurrents (RNN) dont font partie les réseaux à mémoire à long et à court terme (Long-Short Terme Memory LSTM), ont démontré une efficacité et une capacité jusqu'à lors inégalées dans l'analyse d'images médicales, fournissant une aide au diagnostic médical très importante. Dans cette optique, nous souhaitons tirer profit de toute la puissance de l'IA pour permettre une détection précoce du paludisme (avant que le parasite plasmodium n'atteigne son apogée) et une aide à la décision médicale significative pour déterminer quel individu est infecté et qui ne l'est pas.

Cela-dit, l'intégration de l'IA dans le processus décisionnel médical, et par conséquent dans la détection du paludisme, reste un véritable challenge, car il est quasiment impossible de

faire confiance aveuglément à quelque chose que l'on ne comprend pas. Comment l'IA procède pour avoir des résultats ? Sur quelles bases elle prend une décision ? Quelles sont les paramètres présents dans les images qui lui permettent d'affirmer avec certitude qu'un tel patient est infecté ou pas ? Autant de questions qui restent sans réponses et qui empêche l'utilisation massive de l'IA dans les secteurs dits sensibles tels que la médecine.

II. Problématique et motivations

Malgré leurs exploits, les modèles intelligents peinent à être largement utilisés par les professionnels de la santé et ce, par méfiance et par manque de confiance dans les résultats qu'ils obtiennent. En effet, déléguer des décisions critiques à des systèmes qui ne peuvent pas être interprétés ou qui ne fournissent pas d'explications sur leurs processus de raisonnement et de leur façon d'obtenir des résultats peut être dangereux. Effectivement, ces modèles sont perçus comme des **« boîtes noires »**, limitant leur adoption dans des environnements critiques comme la médecine.

D'autre-part, si on se réfère au droit et aux exigences éthiques de la santé, comprendre une décision médicale pour un patient, n'est pas juste souhaitable, mais c'est un droit fondamental dans certains pays comme le mentionne le Règlement Général sur la Protection des Données (RGPD) dans l'Union Européenne (UE).

En réponse à toute cette agitation autour de l'IA, et pour pouvoir tirer profit de sa puissance dans tous les domaines, lui permettant un déploiement à la hauteur de ses exploits, les scientifiques du domaine se sont alloués la lourde tâche d'expliquer cette IA, créant ainsi un vaste champ de recherches dénommé **L'IA eXplicable (XAI)** de l'Anglais **eXplainable Artificial Intelligence.**

Cet axe de recherche en plein essor, combiné à notre désir ardent de proposer une solution effective et très satisfaisante pour la détection précoce du paludisme, tentent de répondre à une problématique qu'on peut résumer ainsi :

1. **Confiance et adoption par les utilisateurs :** Les cliniciens et autres professionnels de la santé doivent comprendre et faire confiance aux recommandations des systèmes d'IA pour les intégrer dans leurs pratiques.
2. **Comment rendre les explications suffisamment claires et utiles ?**
3. **Balance entre performance et explicabilité :** Les modèles performants, comme les réseaux de neurones profonds, sont souvent des "boîtes noires". Comment concilier leur performance avec des exigences d'explicabilité ?
4. **Contraintes réglementaires et éthiques :** Les réglementations comme le RGPD imposent la transparence et le droit à l'explication. Comment s'assurer que les systèmes XAI respectent ces contraintes ?
5. **Complexité des données médicales :** Les données médicales sont souvent hétérogènes, volumineuses et sensibles. Comment adapter les approches XAI à ces spécificités pour garantir des explications fiables et compréhensibles ?

6. **Amélioration de la prise de décision :** L'aptitude de fournir une aide à la décision médicale précise, interprétable par le personnel médical, les explications fournies par les systèmes XAI améliorent-elles réellement la qualité des décisions cliniques ?

7. Enfin, pouvons-nous **détecter précocement le paludisme** par l'IA ? pouvons-nous proposer un modèle intelligent, à la fois puissant, précis et explicable ?

Autant de questions, auxquelles nous allons tenter d'y répondre tout au long de cette thèse.

III. Contribution

Nos travaux de recherche menés dans le cadre de cette thèse visant à pallier aux limitations actuelles, pour améliorer la transparence et la confiance dans les systèmes de détection du paludisme. Nos expérimentations intègrent l'analyse des performances des modèles avec l'explicabilité dans le contexte de la détection automatique précoce du paludisme. Les principales contributions de ce travail incluent :

- **Proposition d'un modèle de classification robuste basé sur un réseau Stacked-LSTM muni d'un mécanisme d'attention** ;
    - Une approche novatrice, exploitant les capacités des Stacked-LSTM à générer les informations contextuelles dans le processus d'extraction des caractéristiques, à partir de séquences temporelles ou spatiales extraites des images de frottis sanguin, pour détecter efficacement les globules rouges infectés, permettant une classification hautement précise.
    - Une optimisation des paramètres pour garantir une robustesse face aux variations des conditions d'acquisition des images.
    - Démontrer l'efficacité d'un modèle Stacked-LSTM avec un mécanisme d'attention pour la classification d'images médicales et les tâches complexes de vision par ordinateur.
- **Intégration d'un mécanisme d'explicabilité par les techniques XAI** :
    - Ajout d'un mécanisme d'explicabilité après notre classifieur basé sur les réseaux Stacked-LSTM pour concentrer le modèle sur les régions d'images les plus pertinentes, permettant d'extraire et d'expliquer les motifs ou les patterns responsables de la décision de notre modèle intelligent.
    - Évaluation de deux méthodes XAI, la carte d'activation de classe pondérée par le gradient (Gradient-weighted Class Activation Mapping **Grad-CAM**) et les explications locales interprétables indépendamment du modèle (Local Interpretable Model-agnostic Explanations **LIME**), pour interpréter et expliquer les prédictions faites par le classificateur Stacked-LSTM. Ces techniques d'explicabilité nous permettent d'analyser comment le modèle prend des décisions et d'identifier les régions d'intérêt influençant ses

prédictions. Cette approche vise à améliorer l'interprétabilité des modèles d'IA et à garantir leur application fiable dans les contextes critiques de soins de santé.

- **Détection précoce du paludisme :**

La "détection précoce" fait référence à l'identification de l'infection par le paludisme aux premiers stades de la présentation des symptômes, généralement avant que des manifestations graves ne se développent. Notre modèle est conçu pour analyser les échantillons de laboratoire (images de frottis sanguins) avec une grande sensibilité, permettant un diagnostic plus précoce par rapport à une suspicion clinique tardive ou à la microscopie traditionnelle.

- **Étude comparative avancée entre plusieurs modèles de DL :**

Une évaluation comparative de cinq modèles d'apprentissage profond—VGG-16, VGG-19, Stacked-LSTM, Vision Transformer (ViT) et MobileNetV2—pour la détection du paludisme à l'aide d'images de frottis sanguins.

- **Développement d'une plateforme web applicative** :
  - Une interface interactive permettant d'intégrer le diagnostic automatisé avec une visualisation des explications générées par le modèle.
  - Un système léger et déployable dans des environnements à faibles ressources.

Ces contributions combinent précision, transparence et accessibilité, ouvrant ainsi la voie à une adoption plus large de l'IA dans le diagnostic médical.

IV. Organisation du document

Notre mémoire est organisé en deux parties principales : la première, intitulée « **Cadre Théorique** », présente les notions théoriques liées au contexte médical portant sur le paludisme et à l'explicabilité de l'IA par les techniques XAI ; ce qui nous permet d'introduire le contexte applicatif dans lequel s'inscrivent les travaux menés dans le cadre de cette thèse, et de montrer un état de l'art lié à la problématique dans ce même contexte. La seconde partie du document intitulée « **Contributions Pratiques** », présente nos contributions à travers l'approche proposée, d'un point de vu conceptuel dans un premier temps, puis l'implémentation et le déploiement de notre approche par la suite.

Le cadre théorique quant à lui inclut trois chapitres :

- **Chapitre 1 : intelligence artificielle et explicabilité**

Ce chapitre, lui-même se scinde en deux parties distinctes. La première, introduit les fondements de l'IA, plus précisément l'apprentissage automatique (Machine Learning en Anglais **ML**) ; notamment les paradigmes principaux (apprentissage supervisé, non supervisé, apprentissage par renforcement, etc.), et nous permet de présenter les différentes architectures des modèles profonds. La seconde partie quant à elle, nous permet de jeter les bases de l'XAI et d'élucider pourquoi l'explicabilité est devenue cruciale et quels sont leurs défis d'un point de vue éthique et sociétal.

- **Chapitre 2 : Le paludisme**

Ce chapitre est consacré à la présentation du contexte médical de cette thèse tout en fournissant des définitions et des présentations simples et concises liées au paludisme et des statistiques. Enfin, nous présenterons les différentes techniques d'acquisition des images de frottis sanguin ou de gouttes de sang et leur importance, en ce sens où, elles constituent la base sur laquelle repose le diagnostic médical, qu'il soit réalisé par l'humain ou par l'IA.

- **Chapitre 3 : Travaux connexes : Approches de Classification et explicabilité dans la détection du paludisme.**

Dans ce chapitre, nous présentons les principaux travaux récents traitant de notre problématique, nous permettant de dresser un paysage et une vue d'ensemble portant sur les principales avancées technologiques, les nouvelles méthodes utilisées ainsi que leurs limites. Par la suite, une étude comparative de ces derniers avec nos travaux de recherche est proposée afin de positionner nos travaux par rapport aux autres travaux.

La seconde partie de ce mémoire, relative aux contributions pratiques, regroupe deux chapitres :

- **Chapitre 4 : Détection automatique du paludisme : Méthodes, résultats et discussion.**

La mise en œuvre conceptuelle de notre solution est présentée au cours de ce chapitre. Nous y exposerons notre contribution par le biais d'une approche novatrice, visant à répondre aux différents défis soulevés lors de l'expression de la problématique (section II), tout en mettant l'accent sur les différentes méthodes et techniques utilisées afin d'y parvenir. Aussi, nous présenterons et discuterons les résultats obtenus par notre approche.

- **Conclusion Générale et Perspectives.**

Nous clôturons nos investigations par une synthèse de nos principales contributions et du travail effectué tout au long de cette thèse, puis nous développerons les perspectives possibles ainsi que les futurs objectifs et travaux.

CHAPITRE I

INTELLIGENGE ARTIFICIELLE & EXPLICABILITÉ

CHAPITRE I

Intelligence Artificielle et Explicabilité

Sommaire

I. Introduction

Les modèles intelligents en général et ceux basés sur le Deep Learning en particulier ont révolutionné de nombreux domaines de l'intelligence artificielle, offrant des avancées significatives dans des tâches complexes telles que la reconnaissance d'images, la compréhension du langage naturel et la prédiction des résultats. Ces modèles ont démontré des performances exceptionnelles dans des domaines critiques comme la santé, la finance, la sécurité et la justice.

Malgré toutes leurs performances, ces modèles rencontrent des difficultés à être adoptés par les professionnels en raison de la méfiance et du manque de confiance dans les résultats qu'ils produisent. En effet, confier des décisions cruciales à des systèmes non interprétables, dépourvus d'explications sur leurs processus de raisonnement et d'obtention de résultats, peut s'avérer périlleux, surtout dans les domaines ou la vie des individus peut être compromise.

Ce manque de transparence est dû au fait que, de par la nature complexe des mécanismes de décision de ces modèles, ils sont souvent considérés comme des « **boîtes noires »** par les utilisateurs finaux, car ils fonctionnent de manière opaque. Cette opacité soulève des préoccupations importantes concernant les résultats et les prédictions présentés, en particulier dans des domaines où les décisions peuvent avoir des conséquences néfastes sur la vie des individus.

À partir de là, et si l'on considère le besoin viscéral qu'éprouve l'être humain à comprendre et à expliquer tout ce qui l'entoure, « l'interprétabilité » et « l'explicabilité » ont émergé ces dernières années comme un domaine de recherche crucial qui vise à rendre les modèles intelligents, plus transparents et compréhensibles par l'être humain.

Au cours de ce chapitre, nous décrirons les réseaux de neurones convolutifs (CNN) ainsi que les différents modèles de deep learning employés pour la classification des images. Dans la deuxième partie, nous allons présenter de façon assez exhaustive les méthodes et les

techniques se rapportant aux notions **« Interprétabilité »** et **« Explicabilité »** et discuterons des défis généraux les concernant, ainsi que de leurs avantages et limitations individuels.

II. Intelligence artificielle

L'Intelligence Artificielle (IA) constitue une branche de l'informatique qui englobe l'application d'algorithmes avancés et des modèles mathématiques en vue de concevoir des systèmes aptes à reproduire certaines manifestations de l'intelligence humaine.

Il s'agit d'une discipline scientifique visant à identifier des méthodes permettant de résoudre des problèmes présentant une complexité logique ou algorithmique élevée, et ce, dans un délai minimal tout en garantissant une grande précision.

Ces techniques ont pour objectif d'analyser les données et de prendre des décisions de manière autonome, ce qui pourrait engendrer des implications significatives dans divers domaines, en particulier dans le secteur médical.

1. Apprentissage automatique

L'apprentissage automatique, également désigné par l'acronyme ML (Machine Learning en anglais), représente une discipline au sein de l'intelligence artificielle. Il englobe un ensemble de techniques mathématiques qui permettent aux machines d'apprendre à partir de données numériques, généralement sous forme tabulaire, dans le but d'exécuter des tâches spécifiques. Sa popularité a considérablement augmenté en raison de l'accroissement des volumes de données accessibles et de l'amélioration des capacités de calcul. Cette approche s'avère particulièrement pertinente dans des contextes où les méthodes algorithmiques traditionnelles, ainsi que l'intervention humaine, se révèlent inadaptées ou insuffisantes. L'apprentissage automatique confère aux systèmes informatiques la capacité d'acquérir des connaissances, de s'améliorer de façon continue et, dans certains cas, de dépasser les performances humaines **(Murphy, 2012).**

L'apprentissage automatique se répartit principalement en trois catégories : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage semi-supervisé **(Mathieu-Dupas, 2010)**. Chaque catégorie aborde des types de données particuliers et regroupe des algorithmes élaborés pour résoudre des problématiques complexes.

1.1. L’apprentissage automatique supervisé

Le modèle le plus courant et le plus aisément interprétable s'applique dans le cas où les données d'entrée sont annotées, c'est-à-dire qu'elles possèdent des sorties ou des classes établies par des experts. Le processus se divise en deux phases principales ; la première consiste en l'entraînement, durant lequel un modèle acquiert des connaissances à partir d'une base de données annotée. Par la suite, la phase de test a pour objectif de prédire les classes de nouvelles données, en s'appuyant sur le modèle préalablement formé, dans le but d'évaluer ses performances.

1.2. Apprentissage non supervisé

Employée dans le cas où les données ne sont pas annotées, cette méthode a pour objectif d'organiser les points en groupes ou en classes selon un critère de similarité préétabli. L'objectif consiste à identifier des regroupements naturels au sein des données, sans avoir de connaissances préalables concernant les relations entre les différents éléments. Une méthode de

regroupement efficace favorise une grande homogénéité au sein des classes et une faible similarité entre les différentes classes. Ce mode d'apprentissage est fréquemment utilisé dans des applications telles que la segmentation d'images, où les pixels sont classés en régions homogènes selon un critère de proximité.

1.3. Apprentissage automatique semi-supervisé

Cette méthode intègre les principes de l'apprentissage supervisé ainsi que ceux de l'apprentissage non supervisé. Elle s'avère particulièrement bénéfique lorsque la base de données renferme une faible proportion d'éléments annotés et une vaste quantité de données non annotées. Les algorithmes semi-supervisés tirent parti des données annotées afin de former le modèle, puis utilisent les données non annotées en s'appuyant sur le modèle préalablement entraîné.

2. Apprentissage profond

L'apprentissage profond, également désigné par l'acronyme DL en anglais, constitue une sous-discipline de l'intelligence artificielle, dérivant de l'apprentissage automatique. Il se distingue par sa capacité à permettre à la machine d'apprendre directement à partir de données brutes, sans nécessiter d'intervention humaine pour identifier ou sélectionner les caractéristiques fondamentales. À cet égard, il se distingue de l'apprentissage automatique traditionnel, qui requiert une phase manuelle de préparation des données ainsi qu'une sélection des attributs pertinents. La **Figure 1** représente les relations hiérarchiques entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond.

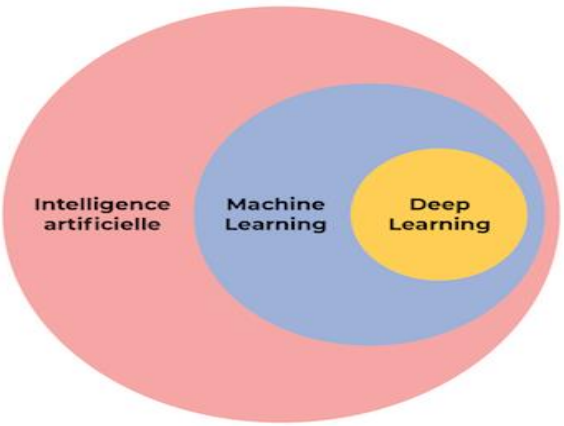


Figure 1. Relation entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond **(McClelland, 2017)**

Les algorithmes d'apprentissage profond utilisent directement les images en tant qu'entrées, sans requérir une extraction manuelle des caractéristiques, contrairement aux approches traditionnelles. Ces algorithmes ont pour fonction d'identifier et d'apprendre de manière autonome des caractéristiques diverses, allant des plus élémentaires aux plus sophistiquées, dans le but de concevoir et d'entraîner un modèle performant **(Hilali, 2009)**.

Au cours des dernières années, les techniques d'apprentissage profond ont considérablement révolutionné le domaine de la vision par ordinateur, engendrant des applications dans des secteurs variés tels que la reconnaissance faciale, la détection d'objets, ainsi que le traitement d'images, de textes et de la parole. Dans le secteur médical, elles facilitent

la résolution de défis complexes qui s'avèrent souvent difficiles à aborder par des méthodes conventionnelles.

Les avancées observées dans le domaine de l'apprentissage profond sont principalement imputables à l'amélioration des capacités de calcul ainsi qu'à l'accès à d'importantes bases de données annotées. À l'heure actuelle, les réseaux de neurones convolutifs (CNN) figurent parmi les modèles les plus performants et les plus largement employés, notamment pour des tâches telles que la classification des images médicales.

2.1. Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (Convolutional Neural Networks, CNN) constituent des modèles sophistiqués des réseaux de neurones artificiels, généralement caractérisés par la présence de plus de cinq couches cachées. Ils se composent d'un ensemble de couches interconnectées, chacune analysant les données de manière autonome afin d'acquérir des représentations à divers niveaux d'abstraction. Les couches inférieures examinent des attributs fondamentaux des données d'entrée, tandis que les couches supérieures synthétisent ces attributs afin de produire des représentations plus élaborées et abstraites. En recourant à des algorithmes de rétropropagation, les réseaux de neurones convolutifs (CNN) modifient leurs paramètres internes afin de déterminer la représentation de chaque couche à partir de celle qui la précède **(Castelvecchi, 2016)**.

Un CNN est organisé sous la forme d'un empilement de couches, parmi lesquelles les deux types principaux sont la couche de convolution et la couche de pooling. La couche de convolution, qui est à l'origine du nom du réseau de neurones convolutifs (CNN), applique des filtres de dimensions réduites afin d'extraire les caractéristiques significatives de l'image d'entrée. Elle est fréquemment accompagnée d'une couche de pooling, laquelle diminue la taille des données tout en préservant les informations essentielles par le biais de leur agrégation. Ces deux catégories de couches sont disposées de manière alternée afin d'établir une hiérarchie de caractéristiques visuelles : les couches initiales identifient des éléments simples tels que des contours ou des formes rectilignes, tandis que les couches plus profondes sont chargées de détecter des motifs ou des formes complexes. Après l'extraction des caractéristiques, la sortie de la dernière couche de convolution est transformée en un vecteur de caractéristiques aplati.

Ce vecteur est ensuite utilisé comme entrée pour une ou plusieurs couches entièrement connectées, qui sont fréquemment employées dans des tâches telles que la classification ou la segmentation. La couche finale est généralement une couche de classification, laquelle applique une fonction d'activation afin de générer des probabilités de classe, en fonction des catégories que le réseau a été formé à identifier. L'entraînement des réseaux de neurones convolutifs (CNN) s'appuie sur l'ajustement des poids des connexions neuronales par le biais de la rétropropagation du gradient, dans le but de minimiser la fonction de perte. La **Figure 2** présente une illustration des diverses couches d'un CNN.

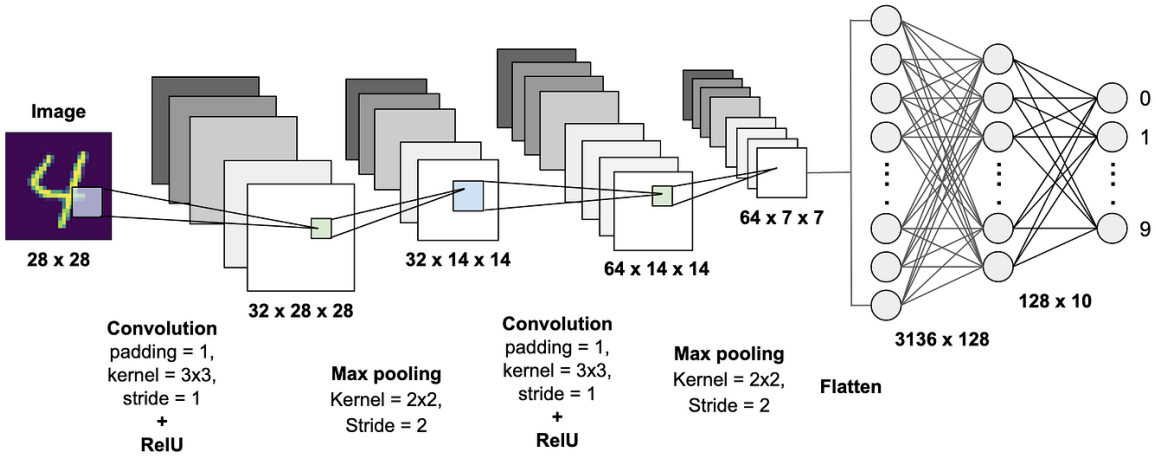


Figure 2. Exemple d'un CNN pour la classification des images[1]

Dans cette section, nous décrirons brièvement les rôles des différentes couches qui composent un réseau de neurones convolutif (CNN).

2.1.1. Couches des réseaux neuronaux convolutifs

Les couches d'un réseau de neurones convolutif (CNN) constituent des éléments fondamentaux qui permettent au modèle de traiter de manière efficace des tâches complexes et sophistiquées relatives aux images. Chaque catégorie de couche exerce une fonction particulière au sein de la chaîne de traitement des données. Ces couches peuvent être classées selon leur fonction.

En règle générale, un réseau de neurones convolutifs (CNN) se compose de quatre types de couches principales : la couche de convolution, la couche d'activation (également désignée sous le terme de correction), la couche de pooling, ainsi que la couche entièrement connectée, comme l'illustre la Figure 2. D'autres couches, telles que la couche de normalisation et la couche de régularisation, peuvent également être intégrées dans le but de réduire le risque de surapprentissage.

Dans les sections qui suivent, nous exposerons les caractéristiques ainsi que les fonctions des diverses couches employées dans la conception des réseaux de neurones convolutifs (CNN).

A. Couche de convolution (CONV)

La couche de convolution constitue l'élément essentiel des réseaux de neurones convolutifs (CNN). Sa fonction principale consiste à extraire des caractéristiques locales des images d'entrée en recourant à des filtres (ou noyaux) de dimensions réduites, généralement de

[1] https://becominghuman.ai/building-a-convolutional-neural-network-cnn-model-for-image-classification-116f77a7a236

3x3 ou 5x5. Au cours du processus de convolution, chaque filtre se déplace à travers l'image en effectuant une opération mathématique : les valeurs des pixels sont multipliées par les coefficients du filtre, puis additionnées afin de produire une nouvelle valeur. Ces valeurs sont consolidées au sein d'une nouvelle représentation, désignée sous le terme de "carte de caractéristiques" (ou feature map), comme le montre la **Figure 3**.

Les premières couches de convolution sont chargées de détecter des caractéristiques élémentaires telles que les contours ou les textures, tandis que les couches plus profondes se consacrent à l'identification de structures plus complexes. Ces cartes de caractéristiques constituent par la suite une entrée pour les couches subséquentes, facilitant ainsi une extraction hiérarchique des informations.

Les filtres employés sont caractérisés par des matrices de poids qui ont été initialisées de manière aléatoire. Au cours de l'entraînement du modèle, ces poids sont modifiés afin d'optimiser l'extraction des caractéristiques en fonction des données d'apprentissage **(Lipton, 2018)**.

1) Hyper-paramètres de la couche de convolution

Quatre hyper-paramètres clés doivent être configurés pour chaque couche de convolution :

1. **Taille et nombre de filtres :** Les dimensions des filtres doivent être inférieures à celles de l'image d'entrée. L'accumulation de plusieurs couches dotées de petits filtres s'avère souvent plus efficace que l'emploi de quelques couches équipées de filtres de dimensions plus importantes. Néanmoins, une augmentation des dimensions des filtres engendre une prolongation du temps d'entraînement. Le nombre de filtres, qui détermine la quantité de cartes de caractéristiques générées, est en adéquation avec la profondeur du volume de sortie. Divers filtres permettent d'extraire des caractéristiques variées et de fournir une représentation plus exhaustive de l'entrée.
2. **Pas (stride) :** Ce paramètre détermine le déplacement du filtre à chaque étape. Une augmentation du pas diminue la dimension de la sortie et restreint le chevauchement entre les segments d'image examinés.
3. **Remplissage (zero-padding) :** Cette technique implique l'ajout de zéros en périphérie de l'image d'entrée afin de préserver ses dimensions après l'application de la convolution.

Il est impératif de sélectionner ces paramètres avec une attention particulière, car certaines configurations peuvent optimiser les performances tout en diminuant le temps de calcul.

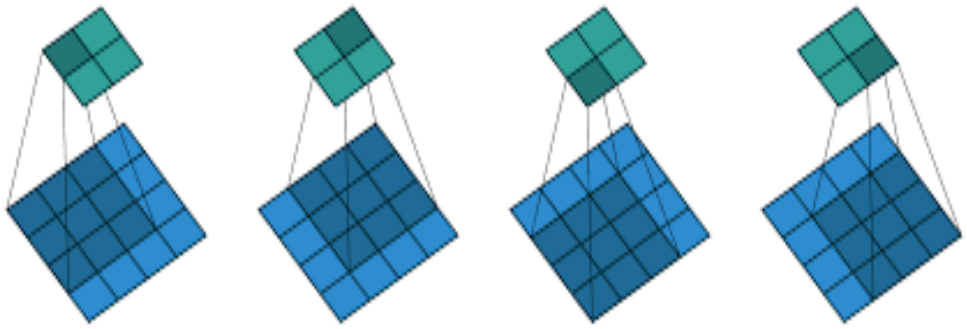


Figure 3. Principe de convolution pour une entrée de taille 4x4 avec un filtre de taille 3x3 et un pas de 1 [2]

1. Fonctions d'activation

Après chaque couche de convolution, une fonction d'activation est appliquée aux valeurs de la carte des caractéristiques. Cette étape introduit une non-linéarité dans le modèle, ce qui permet de modéliser des relations complexes entre les variables d'entrée et de sortie.

Parmi les fonctions d'activation les plus couramment utilisées : **sigmoïde**, **tanh**, **ReLU**, **Leaky ReLU**, et **Softmax (Lipton, 2018).**

d) Fonction sigmoïde :

Elle convertit une valeur réelle en une valeur située entre 0 et 1. Cette fonction en forme de S est fréquemment employée pour normaliser la sortie d'un neurone. L'équation de la fonction sigmoïde f est exprimée par:

$$Sigmoide(x) = 1/(1 + exp^{-x}) \quad \textbf{(1)}$$

Où x et la valeur d'entrée. Si x est positif, la fonction *f* tend vers 1 sinon elle tend vers 0 pour une valeur négative de x.

a) Fonction tanh (tangente hyperbolique) :

Est similaire à la sigmoïde, mais avec une plage de sortie plus large. Elle transforme une valeur réelle en une valeur entre -1 et 1. Tanh est définie comme :

$$tanh\ (x) = (exp^{x} - exp^{-x})/(exp^{x} + exp^{-x}) \quad \textbf{(2)}$$

b) Fonction Unité linéaire rectifiée (Rectifd Linear Unit ReLU) :

Elle constitue la fonction d'activation par défaut au sein des architectures d'apprentissage profond, se distinguant par sa simplicité, sa rapidité de calcul et son efficacité pour un grand nombre de tâches. ReLU fait référence à la fonction réelle non linéaire définie par :

[2] https://medium.com/inveterate-learner/deep-learning-book-chapter-9-convolutional-networks-45e43bfc718d

$$\text{ReLU}(x)=\max(0, x) \quad \textbf{(3)}$$

Cette fonction substitue toutes les valeurs négatives fournies en entrée par un zéro « 0 » (**Figure 4**), ce qui peut avoir un impact sur les performances du réseau de neurones, en diminuant le nombre de neurones actifs et en entraînant une perte de précision. Afin de remédier à ce problème, diverses variantes de la fonction ReLU ont été suggérées, parmi lesquelles figure la fonction Leaky ReLU.

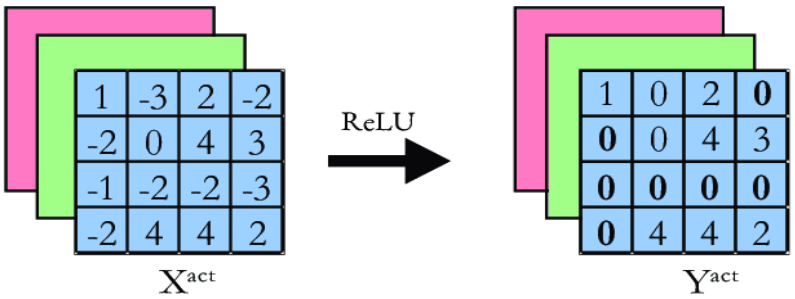


Figure 4. Le principe de fonctionnement du ReLU **(Abdelouahab, 2018)**

c) Fonction Leaky ReLU (Leaky Rectifd Linear Unit) :

Est une version modifiée de ReLU développée pour réduire le problème de ReLU. Elle présente une petite pente négative au lieu d'une pente plate. Mathématiquement, elle se définit comme :

$$Leaky\ ReLU(x) = \begin{cases} 0.01x, & si\ x < 0 \\ x, & ailleurs \end{cases} \quad \textbf{(4)}$$

La **Figure 5** ci-dessous illustre les tracés des quatre fonctions d'activation décrites précédemment : sigmoïde, tanh, ReLU et Leaky ReLU.

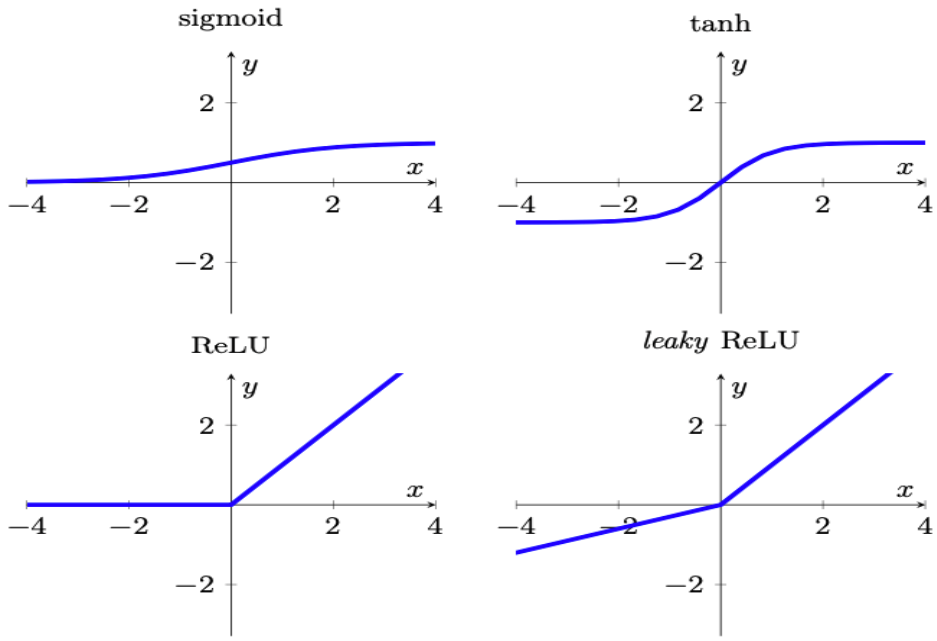


Figure 5. Les courbes des quatre fonctions d'activation

d) Fonction Softmax :

Fréquemment employée dans les réseaux de neurones, cette méthode permet de convertir les résultats en une distribution de probabilités normalisées sur plusieurs catégories. Elle est

fréquemment utilisée en tant que couche finale dans les réseaux de neurones convolutifs pour la classification multi-classes.

La fonction Softmax reçoit en entrée un vecteur de valeurs (z) et génère en sortie un vecteur de probabilités normalisées pour chacune des classes. Chaque probabilité se situe entre 0 et 1, et la somme de l'ensemble des probabilités est égale à 1. La classe présentant la probabilité la plus élevée est sélectionnée en tant que prédiction finale.

Cette fonction est définie comme l'exponentielle de chaque valeur (zi), rapportée à la somme des exponentielles de l'ensemble des valeurs (z_j), où j=1, 2,…k et k est le nombre de classes. Chaque valeur de la sortie de cette fonction est calculée comme suit :

$$Softmax(z) = e^{z_i} / \sum_{j=1}^{k} e^{z_j} \quad \textbf{(5)}$$

B. Couche de pooling (POOL)

Le pooling constitue une technique essentielle au sein des réseaux de neurones convolutifs (CNN). Il s'agit d'une méthode de sous-échantillonnage ou de réduction de dimensions, employée dans le but de minimiser la taille spatiale d'une image intermédiaire, tout en préservant les informations les plus significatives, afin de diminuer le nombre de paramètres et le temps de calcul au sein du réseau. Il régule le surapprentissage et renforce la robustesse du modèle. Le principe de pooling consiste à segmenter l'image d'entrée en petites fenêtres de dimensions n*n pixels (généralement 2*2 ou 3*3), sans chevauchement, et à extraire une seule valeur de chaque fenêtre en fonction du type de pooling sélectionné.

On distingue deux catégories de pooling : Le max-pooling est largement employé dans les réseaux de neurones convolutionnels (CNN), et consiste à retenir la valeur maximale de chaque fenêtre au sein de la carte des caractéristiques. Le pooling moyen (average pooling) permet de préserver la valeur moyenne de chaque fenêtre, comme illustré dans la Figure 6 **(Lipton, 2018)**.

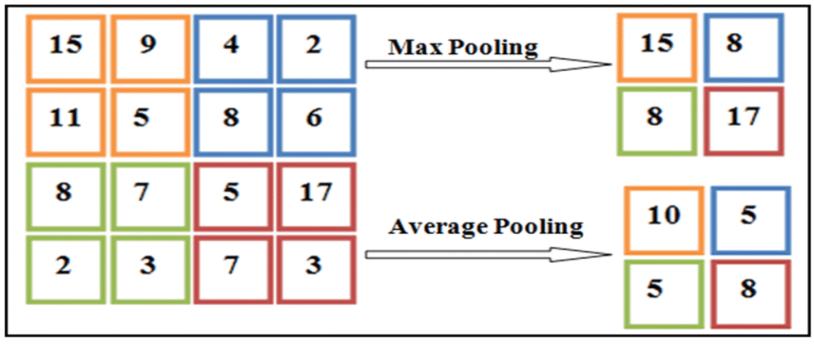


Figure 6. Processus de max et average pooling avec fenêtre de taille 2x2 et pas 2

C. Couche entièrement connectée (FC)

Cette couche représente la phase finale d'un réseau de neurones convolutif (CNN). Elle reçoit en entrée les caractéristiques extraites par les couches antérieures et les convertit en un seul vecteur de sortie, permettant ainsi de générer des résultats appropriés pour des tâches telles que la classification et la régression. À l'instar d'un réseau de neurones artificiels (ANN), chaque neurone de la couche entièrement connectée (FC) est relié à l'ensemble des neurones de la couche antérieure. La quantité de neurones au sein de cette couche est généralement établie en

fonction de l'ampleur de la tâche de classification ou de prédiction à réaliser. Elle est fréquemment accompagnée de la fonction Softmax, qui a pour objectif de normaliser les sorties et de générer une distribution de probabilités pour chacune des classes envisageables. La classe présentant la probabilité la plus élevée est sélectionnée en tant que prédiction finale du modèle **(Lipton, 2018).**

D. Couche de normalisation :

La normalisation par lot, également connue sous le terme anglais de "batch normalization", constitue une méthode de régularisation largement employée dans le domaine de l'apprentissage profond. Elle vise à accélérer la convergence de l'apprentissage et à optimiser les performances d'un modèle. Elle procède à la normalisation des entrées (activations) de la couche précédente pour chaque mini-lot, plutôt que pour chaque échantillon individuel. Cela contribue à stabiliser le processus d'apprentissage et diminue de manière significative le nombre d'époques de formation requises pour élaborer des réseaux profonds **(Plamen et al., 2021).**

E. Couche de Dropout :

Le dropout constitue une méthode de régularisation visant à lutter contre le surapprentissage (over-fitting) et à optimiser l'apprentissage d'un réseau de neurones convolutifs. Elle implique la désactivation temporaire et aléatoire, à chaque époque, d'un certain nombre de neurones (c'est-à-dire leur mise à zéro) au sein du modèle, ainsi que de toutes les connexions d'entrée et de sortie associées. Le nombre de neurones à désactiver est déterminé en fonction d'une probabilité préalablement établie, généralement fixée à p=0,5 **(Gunning, 2017).**

2.1.2. Les architectures CNN pour la classification des images

La classification des images à l'aide des modèles d'apprentissage profond a enregistré un succès considérable au cours des dernières années et a démontré sa robustesse en comparaison avec les méthodes traditionnelles. Dans cette section, nous procéderons à l'évocation des modèles les plus renommés dans la littérature en matière de classification, notamment VGG 16 et VGG 19.

A. Architectures VGG16 et VGG19

Les architectures VGG16 **(Figure 7)** et VGG19 constituent deux modèles de réseaux de neurones convolutifs (CNN) élaborés par le Visual Geometry Group (VGG) de l'université d'Oxford en 2014. Elles sont réputées pour leur simplicité et leur efficacité en matière de classification d'images et de détection d'objets. Le modèle VGG16 se compose de 16 couches dotées de poids entraînables, dont 13 sont des couches de convolution et 3 des couches entièrement connectées. En revanche, le modèle VGG19 comprend 19 couches, réparties en 16 couches de convolution et 3 couches entièrement connectées. Les deux architectures emploient de manière systématique des filtres de convolution de dimensions 3×3, suivis de couches de max-pooling de 2×2, afin d'extraire progressivement les caractéristiques de l'image tout en diminuant les dimensions spatiales. À la conclusion du réseau, des couches entièrement connectées convertissent les caractéristiques en vecteurs destinés à la classification, généralement en utilisant une fonction d'activation Softmax. Ces réseaux sont constitués de plus d'un million d'images provenant de la base de données ImageNet et ont la capacité de classifier les images en 1000 catégories d'objets **(Adadi et Berrada, 2018).**

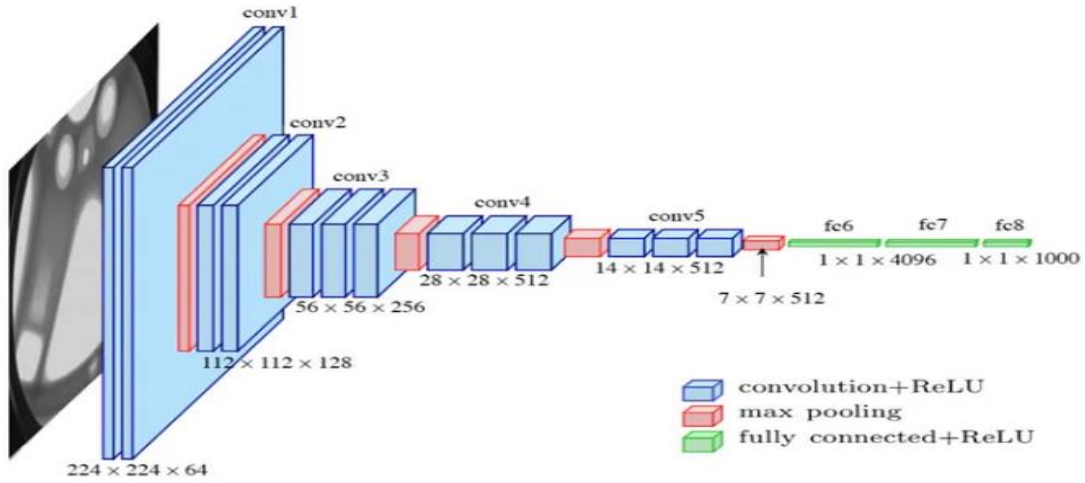


Figure 7. Architectures VGG16 **(Garcia-Garcia et al., 2017)**

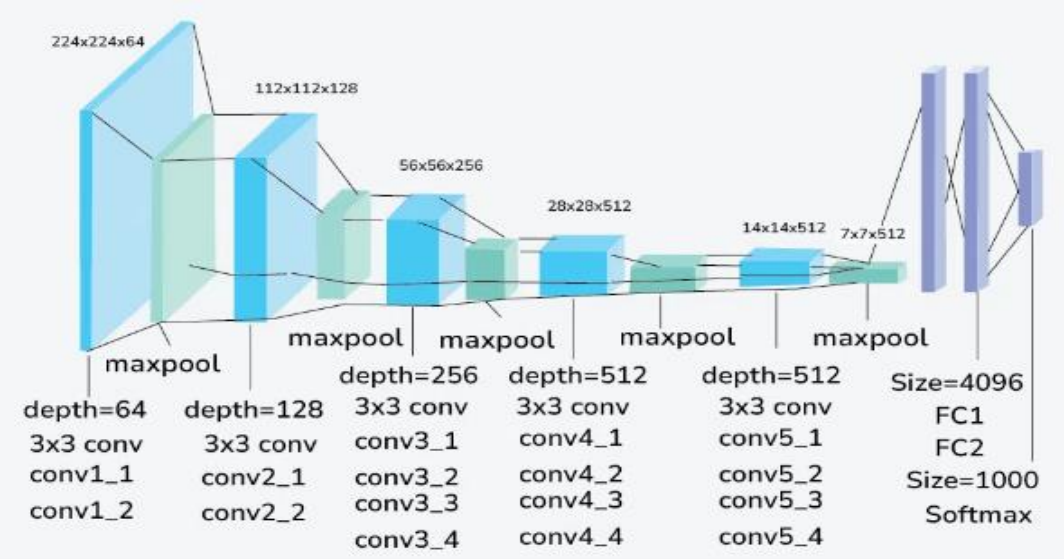


Figure 8. Architecture VGG19 **(Garcia-Garcia et al., 2017)**

III. Explicabilité de l'IA

1. Explicabilité vs Interprétabilité

Avant de poursuivre plus loin dans ce chapitre, il convient d'abord d'établir un point de compréhension commun en ce qui concerne les termes **explicabilité** et **interprétabilité** dans le contexte de l'IA et plus précisément en ML. Pour cela, nous devons nous arrêter sur les nombreuses définitions qui ont été données concernant ces concepts (**Quoi ?**), pour expliquer pour quelles raisons ces deux notions représentent un problème important en IA et en ML (**Pourquoi ? à-quoi-ça-sert ?**) et d'aborder la classification générale des méthodes et techniques d'explicabilité XAI qui nous guidera tout au long de ce chapitre (**Comment ?**).

L'interprétabilité et l'explicabilité sont étroitement liées pour aider les humains à comprendre le raisonnement derrière les prédictions d'un modèle intelligent. Bien qu'ils soient souvent utilisés de manière abusive et interchangeable dans la littérature, ce ne sont pas des concepts monolithiques et il existe une différence notable entre eux. Cependant, leurs définitions précises et formelles restent subjectives dans la littérature spécialisée. En effet,

jusqu'à l'heure actuelle, aucune spécification consensuelle sur ce qu'est un algorithme interprétable ou sur une manière appropriée de générer et d'évaluer des explications n'a été atteinte **(Lipton, 2018)**.

1.1. Définitions et terminologie (Quoi ?)

Bien qu'il existe des divergences en ce qui concerne les termes **« interprétabilité »** et **« explicabilité »**, nous pouvons dire que :

- **L'interprétabilité** fait référence à une caractéristique **passive** d'un modèle se référant à l'aptitude d'un modèle à donner du sens pour un humain. Cette caractéristique se traduit également par la ***transparence*** **(Arrieta et al., 2020)**.

- Une autre définition présente l'interprétabilité comme étant la capacité à présenter quelque chose de manière compréhensible. Dans le contexte du ML, un modèle interprétable permet aux utilisateurs d'observer les sorties, d'étudier l'architecture du modèle et de comprendre comment les données d'entrée sont mappées aux sorties, de manière mathématique et logique. Cela signifie que l'interprétabilité est vue comme un élément passif, indiquant dans quelle mesure on peut extraire du sens d'un domaine avec des informations abstraites **(Doshi-Velez and Kim, 2017).**

- **L'explicabilité** quant à elle, peut être considérée comme une caractéristique **active** d'un modèle, désignant toute action ou procédure effectuée par un modèle dans le but de clarifier ou de détailler ses fonctions internes **(Arrieta et al., 2020).**

- Par analogie à cette définition, nous pouvons dire que l'explicabilité est l'aptitude d'un système à être compris par un humain étant donné un contexte. En effet, l'explication peut prendre diverses formes et s'adapter aux parties prenantes. Les informations brutes comme le code source d'un algorithme ou la structure d'un modèle ne sont pas des explications mais de la transparence. Celle-ci ne suffit pas à rendre un modèle explicable, surtout si ce dernier est complexe **(Doshi-Velez and Kim, 2017).**

- **La compréhensivité** (ou également, **l'intelligibilité**) **(Understandability)** désigne la caractéristique d'un modèle à permettre à un humain de comprendre son fonctionnement, sans pour autant avoir besoin d'expliquer sa structure interne ou les moyens algorithmiques par lesquels le modèle traite les données en interne **(Montavon et al., 2018).**

- **La compréhensibilité (Comprehensibility),** lorsqu'elle est conçue pour les modèles de ML, fait référence à la capacité d'un algorithme d'apprentissage à représenter ses connaissances acquises de manière compréhensible pour un humain **(Gleicher, 2016 ; Doshi-Velez and Kim, 2017 ; Fernandez et al., 2019).**

- **La transparence** d'un modèle est la capacité de ce dernier à être compréhensible par lui-même **(Arrieta et al., 2020).**

A partir de cette terminologie, nous pouvons déduire que :

- La compréhensivité émerge comme le concept le plus essentiel en XAI.

- La transparence et l'interprétabilité sont toutes deux fortement liées à ce concept. Effectivement, tandis que la transparence se réfère à la caractéristique d'un modèle à être

compréhensible, par lui-même, pour un humain, la compréhensivité mesure le degré de compréhension d'une décision prise par un modèle par un humain.

- La compréhensibilité est également liée à compréhensivité en ce sens qu'elle repose sur la capacité du public à comprendre les connaissances contenues dans le modèle.

- Dans l'ensemble, la compréhensivité est une question à double facette : la compréhensivité liée au modèle et la compréhensivité humaine. C'est la raison pour laquelle la définition de l'IA explicable abordée plus loin dans ce chapitre fait référence au concept de public, car les compétences cognitives et l'objectif poursuivi par les utilisateurs du modèle doivent être pris en compte conjointement avec l'intelligibilité et la compréhensibilité du modèle utilisé.

- Ce rôle prépondérant accordé à la compréhensivité fait du concept d'audience la pierre angulaire de l'IA explicable (XAI), comme nous l'expliquons plus en détail ci-après.

- Enfin, bien qu'il n'existe pas de définition consensuelle en ce qui concerne l'interprétabilité et l'explicabilité, en se basant sur de nombreuses contributions dans la littérature, nous pouvons affirmer qu'on parle de « la réalisation d'un modèle interprétable » et de « techniques d'explicabilité ou de techniques favorisant l'explicabilité ».

À présent, abordons la définition de **« Intelligence Artificielle Explicable XAI».** Commençant par la définition qui revient le plus souvent dans la littérature, celle donnée par **Gunning (2017)** : « *L'IA explicable (XAI) créera un ensemble de techniques d'apprentissage automatique qui permettra aux utilisateurs humains de comprendre, de faire confiance de manière appropriée et de gérer efficacement la génération émergente de partenaires artificiellement intelligents*. ». Cette définition bien que très intéressante, prenant en charge deux notions importantes liées aux XAI, à savoir **« la compréhension »** et **« la confiance »**, reste incomplète, ne prenant pas en charge d'autres objectifs importants relatifs aux modèles d'IA interprétables et explicables, tels que **« la causalité, la transférabilité, l'informativité et l'équité** » **(Vellido et al., 2012 ; Doran, et al., 2017 ; Doshi-Velez and Kim, 2017 ; Lipton, 2018)**.

En cherchant la complétude pour définir le XAI, nous nous sommes rendu compte que cela serait impossible sans une définition exacte de « **l'explication** ». Le Cambridge Dictionary of English Language, définit l'explication comme (traduit de l'Anglais) « *les détails ou les raisons qu'une personne donne pour rendre quelque chose clair ou facile à comprendre* » **(Walter, 2008 ; Arrieta et al., 2020).** Par analogie au contexte d'un modèle de ML, cette dernière peut être reformulée comme : « les détails ou les raisons qu'un modèle donne pour rendre son fonctionnement clair ou facile à comprendre ». À partir de là, soulignons les deux notions les plus importantes à savoir ; *les raisons* et *la compréhension* toutes deux fortement dépendantes du public cible, et c'est pourquoi nous avons mentionné plus haut que l'audience constitue la pierre angulaire de l'IA explicable.

Finalement, la définition donnée par Alejandro Barredo Arrieta dans **(Arrieta et al., 2020),** nous semble très satisfaisante dans le sens où elle nous semble la plus complète et la plus appropriée dans le cadre de nos travaux : ***« Étant donné un public, une Intelligence Artificielle explicable est celle qui produit des détails ou des raisons pour rendre son fonctionnement clair ou facile à comprendre ».***

1.2. Pourquoi ?

Comme nous l'avons cité dans l'introduction, l'explicabilité est l'un des principaux obstacles auxquels fait face l'IA de nos jours, concernant sa mise en œuvre pratique et son large déploiement dans les divers secteurs d'activités.

L'incapacité d'expliquer ou de comprendre pleinement les raisons pour lesquelles les algorithmes du ML fonctionnent aussi bien, est un problème qui trouve ses racines dans deux causes différentes mais d'égale importance **(Arrieta et al., 2020)** :

- Premièrement, l'IA est confrontée à une forte résistance et une réticence au sein des secteurs d'activité fortement réglementés tels que les banques, les finances et la santé entre autres, ce qui crée un gouffre entre les chercheurs scientifiques et les acteurs de ces secteurs d'activité empêchant l'épanouissement et le déploiement de nouveaux modèles du ML.

- Deuxièmement, dans le monde de la recherche scientifique tel qu'il est aujourd'hui, où les métriques de performances dans les secteurs de l'IA et du ML en particulier, représentent la seule valeur ajoutée dans les travaux de recherches, des notions aussi importantes que l'interprétabilité et l'explicabilité peinent à trouver preneurs. D'autant plus, que comme nous l'expliquerons un peu plus loin, la notion de performance est **inversement proportionnelle** aux notions d'interprétabilité et d'explicabilité. Par contre, si on se réfère à la connaissance sociétale, ces dernières ont bien plus de poids que les métriques de performances.

De plus, la nécessité d'expliquer le comportement des algorithmes de ML non interprétables qui peuvent affecter la vie des gens n'est pas seulement une propriété souhaitable, mais aussi une exigence légale dans certains pays.

Par exemple, l'Union Européenne (UE) a réglementé le droit à l'explication dans son Règlement Général sur la Protection des Données (RGPD), intégrant des lois sur les décisions prises par des algorithmes, pour réduire les effets négatifs des systèmes informatiques sur la société **(EU Regulation, 2023).** Par ailleurs, le RGPD exige le droit aux individus d'accéder aux informations sur les décisions prises par des systèmes automatiques. Cela signifie que les responsables doivent expliquer aux gens comment ces décisions sont prises et leur fournir des détails si nécessaire **(Guidotti et al., 2019 ; Amparore et al., 2021)**. Le RGPD a engagé des discussions officielles sur de nouvelles règles, pour s'assurer que l'utilisation de l'intelligence artificielle soit conforme à la législation.

Aux USA, La Food and Drug Administration (FDA) a suggéré des lois pour les dispositifs médicaux utilisant l'IA et/ou le ML (FDA, 2024). Le cadre définit la nécessité de soumettre à l'appréciation et l'évaluation de la FDA, lorsque les algorithmes d'IA provoquent des changements qui affectent de manière significative la performance d'un dispositif médical. Cependant, la mise en œuvre de ces exigences pour le développement de produits reste un problème ouvert **(Maier-Hein, 2022)**.

1.3. A-quoi-ça-sert ?

Les objectifs autour de la réalisation d'une intelligence artificielle explicable sont nombreux et divers ; mais presque aucun des articles examinés ne s'accorde complètement sur ces derniers, bien qu'ils soient nécessaires pour décrire ce qu'un modèle explicable devrait exiger et/ou apporter. Cependant, tous ces différents objectifs pourraient aider à discriminer et à identifier précisément le but final pour lequel une technique d'explicabilité en ML est réalisée.

Malgré le fait que nous n'avons pas trouvé un document qui recense tous les objectifs liés à l'explicabilité (vu que les auteurs ne sont pas tous sur la même longueur d'onde), nous avons essayé de les recueillir à partir des divers travaux de recherche pour les présenter dans notre thèse. En effet, à partir des travaux menés dans **(Z. C. Lipton, 2018** ; **D. Doran, S. Schulz and T. R. Besold, 2017 ; A. Holzinger et al., 2017)**, nous avons pu recenser les objectifs, que nous avons jugé être les plus pertinents, afin de définir quel est le but ou bien la finalité recherchée derrière le XAI.

1.3.1. Les objectifs principaux du XAI

Dans cette section, nous allons présenter les différents objectifs et les qualités que devrait présenter une technique XAI.

a. La fiabilité : est relative à la confiance que les utilisateurs peuvent avoir dans les résultats fournis par le système, en sachant que ces résultats sont basés sur des processus transparents et justifiables. Elle est aussi définie comme la capacité d'un modèle à agir comme prévu lorsqu'il est confronté à un problème donné. Pour la majorité des auteurs, cet objectif est le plus important, bien qu'il n'existe aucun moyen de quantifier le degré de confiance.

b. La causalité : se réfère à la capacité d'un système d'IA à identifier et à expliquer les relations causales entre les variables qui influencent ses décisions ou ses comportements. En d'autres termes, la causalité dans le XAI implique que le système peut fournir des explications qui décrivent les mécanismes sous-jacents qui ont conduit à une décision ou à un résultat particulier. Prouver une relation de-cause-à-effet entre les données en entrée et les sorties, est d'autant plus intéressant dans le sens où cela nous permettrait de répondre à des questions comme : Qu'est-ce qui a causé cette décision ? Quels sont les caractéristiques qui ont contribué à ce résultat ? ou bien, Quelle serait l'issue si telle ou telle caractéristique était modifiée ?

c. La transférabilité : est la capacité d'un modèle ou d'un système d'IA à appliquer ses connaissances ou ses compétences acquises dans un domaine ou un contexte à un autre domaine, ou à un contexte différent. En d'autres termes, il s'agit de la capacité d'un modèle à généraliser ses apprentissages ou ses explications à de nouvelles situations ou à de nouveaux ensembles de données, encourageant ainsi les développeurs à la réutilisabilité des solutions déjà développées, ce qui augmente forcément la productivité et diminue les dépenses de ressources inutiles. Une technique XAI qui peut démontrer un bon niveau de transférabilité est souvent considérée comme robuste, car elle peut s'adapter à des variations dans les données ou dans les conditions d'utilisation.

d. Informativité : Les modèles du ML sont utilisés avec l'intention ultime de soutenir la prise de décision **(Huysmans, 2011)**. Par conséquent, une grande quantité d'informations est nécessaire pour pouvoir relier la décision de l'utilisateur à la solution donnée par le modèle, et pour éviter de tomber dans les pièges des fausses idées. À cette fin, les modèles de ML explicables devraient fournir des informations sur le problème traité. La plupart des raisons fournies dans les articles examinés concernent l'extraction d'informations sur les relations internes d'un modèle. Presque toutes les techniques d'extraction de règles justifient leur approche par la recherche d'une compréhension plus simple de ce que le modèle fait en interne, en affirmant que la connaissance (l'information) peut être exprimée par cette simplification, qu'ils considèrent comme expliquant le modèle antécédent.

e. La confiance : Est à notre avis, la clé de voûte de la réussite et de l'acceptabilité d'une technique d'explicabilité, en ce sens où elle est directement liée à des notions très importantes pour un système telles que la robustesse, la fiabilité et la stabilité. Comme mentionné par **(Ruppert, 1987 ; Yu et al., 2013 ; Basu, 2018)**, la stabilité est essentielle pour comprendre un modèle. On ne devrait pas utiliser des modèles instables pour faire des interprétations fiables. Ainsi, un modèle clair doit donner des informations sur la confiance de son propre fonctionnement.

f. L'équité : Les modèles d'IA peuvent parfois apprendre des biais présents dans les données d'entraînement. L'équité implique la nécessité de détecter ces biais et de les corriger pour garantir que les décisions prises par le modèle soient justes et équitables, permettant ainsi une analyse de l'équité ou éthique du modèle en question **(Chouldechova, 2017 ; Goodman and Flaxman, 2017)**.

g. L'accessibilité : Parmi les articles que nous avons étudiés, certains d'entre eux affirment que l'explicabilité est la propriété qui permet aux utilisateurs finaux de s'impliquer davantage dans l'amélioration d'un modèle de ML **(Chander et al., 2018 ; Miller et al., 2020)**. Il est clair que les modèles explicables faciliteront la tâche aux personnes qui ne sont pas férues de technologie ou expertes pour comprendre des algorithmes, qui au premier abord n'ont pas de sens. Cette accessibilité est d'autant plus importante, car elle est considérée comme le troisième objectif le plus pris en considération dans la littérature.

h. L'interactivité : Certains travaux de recherche **(Harbers, et al., 2010 ; Langley et al., 2017)** disent que l'un des objectifs d'un modèle de ML explicable est que le modèle puisse interagir avec l'utilisateur. Encore une fois, cet objectif concerne des domaines où les utilisateurs finaux sont très importants, et que leur habilité à modifier et à interagir avec ces modèles est ce qui garantit le succès.

i. Conscience de la vie privée : Concept très important, rendu possible grâce aux modèles de ML explicable par leur capacité à évaluer la vie privée. Les modèles de ML peuvent avoir des représentations complexes de leurs patterns appris. Le fait de ne pas comprendre précisément ce qui a été enregistré et stocké par le modèle dans sa représentation interne peut engendrer une violation de la vie privée **(Castelvecchi, 2016)**.

1.4. Comment ?

Dans la littérature on trouve une distinction claire entre les modèles **interprétables par conception** et ceux qui peuvent être expliqués par des **techniques XAI externes**. De même, une classification très largement adoptée par la communauté scientifique spécialisée dans le domaine, est celle des **modèles transparents (explicable ante-hoc),** désignant les modèles interprétables par conception et de **l'explicabilité post-hoc,** lorsqu'il s'agit des modèles intelligents opaques, expliqués par des techniques XAI.

Nous allons présenter plus en détail ces deux classes dans les sections suivantes.

IV. Les Modèles Transparents

Les modèles de régression linéaire, les arbres de décision, les modèles des k-plus proches voisins, l'apprentissage basés sur des règles, les modèles additifs généralisés (GAM) et les réseaux bayésiens sont autant de modèles, considérés couramment comme modèles transparents

(Arrieta et al., 2020 ; Belle and Papantonis, 2021), à condition qu'ils ne soient pas trop volumineux. En effet, seuls les petits modèles peuvent conserver leur compréhensibilité. Dans le cas de modèles avec de nombreuses règles ou paramètres, ceux-ci peuvent également devenir des boîtes noires.

L'utilisation réduite de ces modèles pour tous les problèmes d'apprentissage automatique, est due à leur manque de performance : comme nous le montre la **Figure 8**, ces modèles n'atteignent généralement pas une précision comparable à celle des modèles opaques, tels que les réseaux de neurones profonds (DNN). Ceci est bien connu sous le nom de compromis performance-interprétabilité **(Plamen and Angelov, 2021).**

L'utilisation réduite de ces modèles pour tous les problèmes d'apprentissage automatique, est due à leur manque de performance : comme nous le montre la **Figure 8**, ces modèles n'atteignent généralement pas une précision comparable à celle des modèles opaques, tels que les réseaux de neurones profonds (DNN). Ceci est bien connu sous le nom de compromis performance-interprétabilité

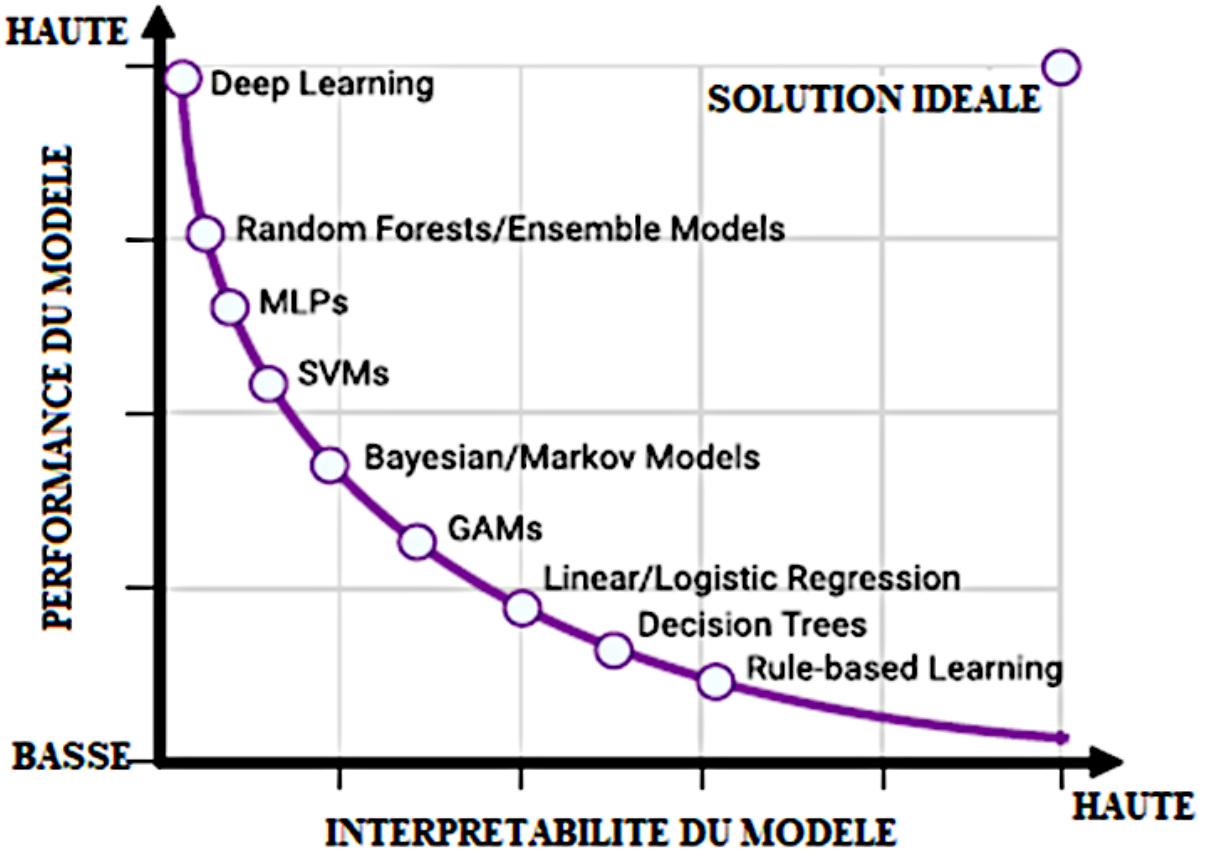


Figure 9. Compromis entre interprétabilité et performance prédictive des principaux modèles d'apprentissage. Les modèles à hautes performances sont complexes et offrent une faible interprétabilité de leurs décisions, tandis que les modèles plus transparents présentent de faibles performances prédictives **(Plamen and Angelov, 2021)**

1- Les différents niveaux de transparence dans le ML

Comme nous l'avons mentionné précédemment, les modèles transparents sont à un certain degré, interprétables par eux-mêmes. Les modèles appartenant à cette catégorie peuvent également être abordés en termes du domaine dans lequel ils sont interprétables, qui est aussi relatif à leur degré de transparence, à savoir, ***la transparence algorithmique, la décomposabilité*** et la ***simulabilité*** **(Guidotti, 2019)**.

1) ***La simulabilité :*** désigne la capacité d'un modèle à être simulé, pensé ou raisonné par un humain.

2) ***La transparence Algorithmique :*** est la capacité pour l'utilisateur de comprendre comment le modèle fonctionne pour donner un résultat à partir des données en entrée. En d'autres termes, un modèle linéaire est clair car on peut comprendre et analyser facilement ses erreurs. Selon **James (2013)**, cela permet à l'utilisateur de savoir comment le modèle réagira dans chaque situation rencontrée. Cependant, il est difficile de le comprendre dans les réseaux profonds, car le paysage de perte est vraisemblablement opaque **(Datta et al., 2016 ; Kawaguchi, 2016)** ; puisqu'il ne pas être entièrement saisi, on doit trouver une solution en utilisant des méthodes d'approximation, comme la descente de gradient stochastique. La principale difficulté pour ces modèles est qu'ils doivent être complètement explorés par des méthodes d'analyses et des méthodes mathématiques.

3) ***La décomposabilité :*** est relative à la capacité d'interpréter facilement chaque entrée et expliquer chacune des parties d'un modèle (entrées, paramètres et calculs). Cela peut être considéré comme de l'intelligibilité comme indiqué par **Lou (2012)** ; cependant, comme c'est le cas avec la transparence algorithmique, cette propriété n'est pas acquise par tous les modèles. De plus, pour qu'un modèle algorithmiquement transparent devienne décomposable, il faut que chaque partie du modèle soit compréhensible par l'humain sans avoir besoin d'outils supplémentaires.

2- Les modèles transparents en ML

Les modèles que nous allons présenter dans cette section sont un ensemble de modèles transparents qui peuvent intégrer un ou tous les niveaux de transparence des modèles décrits précédemment (à savoir, la simulabilité, la décomposabilité et la transparence algorithmique). Nous présenterons aussi dans la **Figure 9** une illustration graphique qui montre ces différents niveaux de transparence pour chacun des modèles de ML exposés.

2.1. Régression Linéaire/Logistique

La régression logistique est un modèle de classification binaire des données pour prévoir une catégorie. Cependant, si la variable dépendante est continue, on parle de régression linéaire. Ce modèle suppose qu'il y a une relation linéaire entre les facteurs et les résultats, ce qui rend l'adaptation aux données difficile. Ce manque de flexibilité est justement ce qui confère à ces modèles leur transparence et les maintient sous l'égide des méthodes transparentes. Cependant, comme nous l'avons cité plus haut, l'explicabilité est fortement liée au public récepteur selon qu'il soit expert ou non, ce qui fait que ces méthodes peuvent être partagées entre les deux catégories d'explication ante-hoc et post-hoc. Par conséquent, bien que les modèles de régression linéaire et de régression logistique intègrent clairement les trois niveaux de transparence précédemment définis, ils peuvent avoir besoin des techniques d'explicabilité post-hoc (principalement, la visualisation que nous verrons plus loin, dans la partie dédiée à la classification des méthodes XAI), lorsque le modèle doit être expliqué à un public non-expert **(Speith. 2022)**.

Il convient de souligner que, pour qu'un modèle de régression logistique ou de régression linéaire conserve sa capacité à être décomposable et simulable, il est essentiel de le maintenir à une taille restreinte et de veiller à ce que les variables employées demeurent compréhensibles pour les utilisateurs. Tel que stipulé auparavant, lorsque les données

alimentant le modèle sont d'une spécificité et d'une technicité élevée, difficile à appréhender, ledit modèle se trouvera bien loin de la simplicité lui permettant d'être décomposable. De plus, si le modèle dépasse tellement l'entendement humain au point qu'il ne peut être appréhendé dans sa totalité, sa capacité à être simulable sera remise en question.

2.2. Arbres de Décision

Selon **Quinlan (1987)**, ce sont des structures hiérarchiques dédiés à la prise de décision dans les problèmes de régression et de classification en ML. Comme pour les modèles précédents, les arbres de décisions véhiculent très bien les trois niveaux de transparence précédemment cités. En effet, dans leurs versions les plus simples, les arbres de décision sont des modèles simulables. Par contre, leurs configurations peuvent les rendre décomposables ou algorithmiquement transparents, faisant d'eux des modèles oscillants sur les différentes catégories de modèles transparents. Cependant, en raison de leurs caractéristiques individuelles, ces modèles sont généralement catégorisés parmi les modèles algorithmiquement transparents.

Un arbre de décision est dit simulable, s'il peut être géré par un utilisateur humain. Cela signifie que sa taille est relativement petite et que le nombre de caractéristiques et leur signification sont facilement compréhensibles. Par contre si sa taille augmente, le modèle devient décomposable. Enfin, en augmentant davantage sa taille et en utilisant des relations de caractéristiques complexes, le modèle deviendra algorithmiquement transparent, perdant ainsi les caractéristiques précédentes **(Quinlan, 1987).**

2.3. K-Nearest Neighbors KNN

Une autre méthode qui relève des modèles transparents est celle des KNN qui traite les problèmes de classification de manière simple et méthodologique : pour résoudre les problèmes de classification, une étiquette de classe est attribuée à travers un vote de majorité : c'est l'étiquette la plus fréquemment représentée autour d'un point de données qui est utilisée.

Lorsqu'il est utilisé dans le contexte des problèmes de régression, le concept est similaire à la classification, à la seule différence qu'ici, c'est la moyenne des k plus proches voisins qui est utilisée pour faire une prédiction. La principale distinction entre les deux types de problèmes, réside dans le fait que la classification concerne des valeurs discrètes, tandis que la régression porte sur des valeurs continues. En matière d'interprétabilité du modèle, sa méthode de prédiction évoque davantage un processus de prise de décision humaine s'appuyant sur l'expérience, relative aux issues de situations similaires antérieures. Il est important de se rappeler que la classe de transparence du KNN varie en fonction des caractéristiques, du nombre de voisins et de la fonction de distance choisie pour évaluer la similitude entre les données. Un niveau de complexité K trop élevé perturbe la capacité d'un utilisateur humain à simuler pleinement la performance du modèle. De même, l'emploi de caractéristiques complexes et/ou des méthodes de mesures de distance compliquées nuirait à la capacité du modèle à être décomposable, limitant ainsi son interprétabilité à la transparence de ses processus algorithmiques.

La **Figure 10** ci-après résume graphiquement les différents niveaux de transparence de chaque modèle susmentionné.

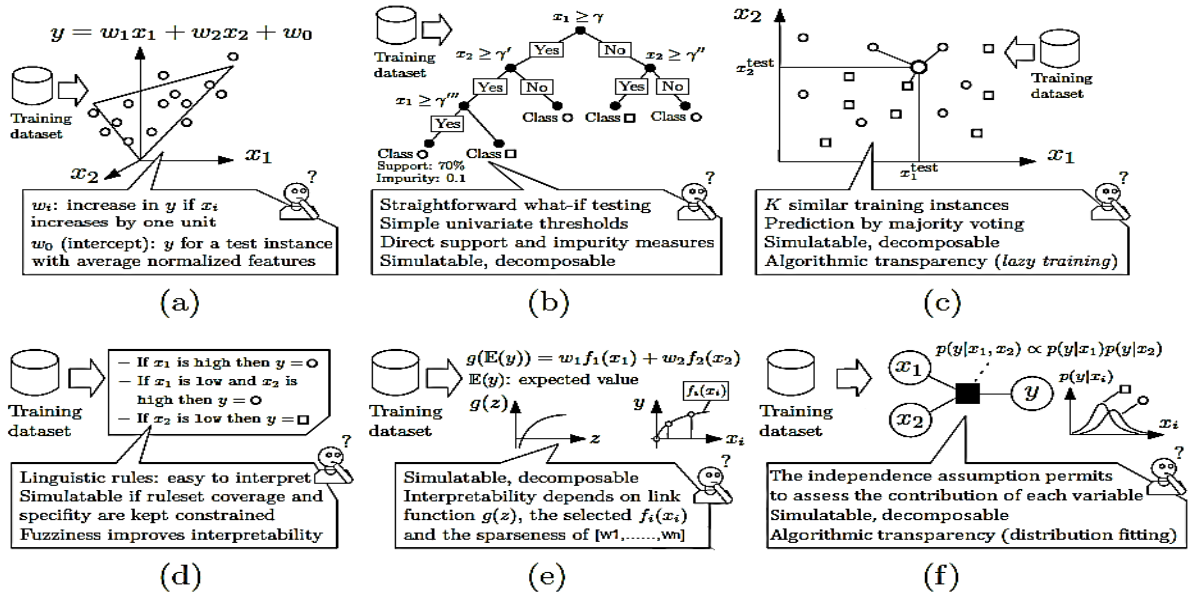


Figure 10. Illustration graphique des niveaux de transparence des différents modèles d'apprentissage automatique considérés dans cette vue d'ensemble : (a) Régression linéaire ; (b) Arbres de décision ; (c) K-Plus Proches Voisins ; (d) Apprenants basés sur des règles ; (e) Modèles Additifs Généralisés ; (f) Modèles Bayésiens **(Arrieta et al., 2020)**

2.4. Apprentissage basé sur les règles

Tout modèle capable de générer des règles pour caractériser les données à partir desquelles il est censé apprendre, fait partie de cette catégorie de modèle de ML. La complexité des règles conditionnelles peut varier de la forme la plus simple « si-condition-alors-résultat », à la forme la plus complexe dans laquelle une combinaison de règles simples est formée pour caractériser les données d'apprentissage.

Une autre variante de cette famille de modèles, est celle basée sur des règles floues, qui est conçue pour un champ d'action plus large, permettant la définition de règles formulées verbalement sur des domaines imprécis. Dans le cadre de nos travaux de recherche, les systèmes flous améliorent deux axes principaux et pertinents. Premièrement, ils permettent de réaliser des modèles plus compréhensibles par l'humain puisqu'ils fonctionnent en termes linguistiques. Deuxièmement, dans des contextes ayant un certain degré d'incertitude, ils fonctionnent mieux que les systèmes à base de règles classiques **(Nuñez et al., 2002 ; 2006).**

Les apprenants basés sur des règles se présentent comme des modèles transparents, fréquemment utilisés pour expliquer des modèles complexes, en émettant des règles qui justifient leurs prédictions. Cependant, un problème central demeure avec ces approches quant à la génération des règles se rapportant à la couverture (quantité) et à la spécificité (longueur) des règles générées. En réalité, lorsqu'on élabore une base de règles, un objectif de conception, souvent souhaité par l'utilisateur, est de pouvoir disséquer et appréhender le modèle. Ajouter un grand nombre de règles à un modèle augmentera indéniablement ses performances, mais au prix de sa transparence et de sa compréhension. De même, la spécificité des règles nuit également à l'interprétabilité, car une règle avec un grand nombre de variables en entrée et/ou de conséquences peut devenir difficile à interpréter. Plus la couverture ou la spécificité est grande, plus le modèle sera proche d'être uniquement algorithmiquement transparent. Parfois, la raison pour laquelle on passe des règles classiques aux règles floues est de s'affranchir des contraintes liées à la taille des règles **(Arrieta et al., 2020).**

2.5. Modèles Additifs Généralisés (General Additive Models) GAM

Ils permettent de modéliser une variable à expliquer avec des fonctions de lissage non-linéaires des prédicteurs. Dans ce contexte l'interprétabilité est facile à atteindre, car les GAM permettent à l'utilisateur de vérifier l'importance de chaque variable (à travers sa fonction correspondante), en vérifiant à quel point elle affecte la sortie prédite. Des méthodes de visualisation sont souvent utilisées pour faciliter davantage l'interprétation du modèle. Les GAM peuvent également être considérés comme des modèles simulables et décomposables, pour peu que leurs propriétés initiales soient conservées.

2.6. Les modèles Bayésiens

Un modèle bayésien est généralement représenté par un graphe acyclique, dirigé et probabiliste, dont les liens représentent les dépendances conditionnelles entre un ensemble de variables. Par exemple, un réseau bayésien pourrait représenter les relations probabilistes entre les maladies et les symptômes. Étant donné les symptômes, le réseau peut être utilisé pour calculer les probabilités de présence de diverses maladies. Transmettant une représentation claire des relations entre les caractéristiques et la cible qui, dans ce cas, sont données explicitement par les connexions reliant les variables entre elles. Ces modèles sont considérés comme transparents du fait qu'ils sont simulables, décomposables et algorithmiquement transparents. Cependant, il convient de noter que dans certains cas où les variables sont trop complexes ou trop grandes, ces modèles peuvent perdre leurs deux premières propriétés (simulables et décomposables).

V. Les Techniques d'Explicabilité Post-hoc : XAI

1- XAI : Besoins et Défis

Les modèles d'apprentissage profond (Deep Learning DL) et d'apprentissage par ensemble ont des mécanismes internes complexes et intriqués, qui sont pratiquement impossibles à interpréter par l'humain. De plus, les raisons menant à une décision ne peuvent pas être comprises, ce qui obscurcit les tâches de vérification qui tentent d'évaluer la logique derrière les prédictions **(Ribeiro et al., 2016a)**.

Comme nous l'avons précisé précédemment, Les modèles opaques sont comme des boîtes noires : on met des données d'un côté et on reçoit des prédictions de l'autre, mais on ne sait pas comment cela fonctionne. En haut de la Figure 10, nous montrons un exemple courant de l'apprentissage supervisé. Chaque modèle d'apprentissage a ses propres compétences, et chaque type de données peut demander des compétences différentes. Donc, des modèles différents peuvent donner des résultats différents sur les mêmes ensembles de données. Dans ce cas, les mesures de performance guident les data scientists à choisir le modèle le plus précis pour chaque situation.

Après l'entraînement, les tests et l'analyse des résultats, le modèle peut être utilisé pour classer des données non étiquetées. À ce moment-là, il sera dans un espace différent et utilisera tout ce qu'il a appris pendant l'entraînement (Patterns appris) sur les nouvelles données, à partir de là, il est difficile de surveiller le modèle à cause des diverses façons dont les données sont reliées aux espaces mappés **(Lundberg et al., 2020)**. Il est primordial de s'assurer que le modèle opérationnel parvient à catégoriser correctement les nouvelles données, sans se

contenter de performances apparemment satisfaisantes masquées par un biais lors de l'entraînement ou à cause d'un problème de définition. Néanmoins, sur le terrain, lorsque le modèle opère sur des données non annotées, il peut produire des classifications inexactes ou des classifications motivées par des raisons erronées, ce qui peut entraîner des problèmes dans les applications d'apprentissage artificiel. En effet, des perturbations indésirables, tels que le bruit parasite ou le biais pourraient ainsi demeurer dissimulés au sein du processus décisionnel.

Bien que les modèles soient essentiellement des fonctions mathématiques, les éléments internes des modèles d'apprentissage automatique sont couramment illustrés graphiquement par souci de lisibilité.

Malgré les illustrations graphiques qui tentent d'éclairer les mystères architecturaux des modèles, percer le secret du fonctionnement interne de ces derniers reste un défi de taille. Toutefois, pour que les applications basées sur l'intelligence artificielle gagnent notre confiance, il ne suffit pas qu'elles donnent l'impression de produire de bons résultats ; elles doivent être en même temps véritablement correctes et équitables. C'est à ce moment que le XAI entre en jeu, offrant des explications permettant de vérifier la validité des prédictions du modèle pour des raisons valables et légitimes. Ces explications apportent des garanties de conformité qui éclairent les processus décisionnels des systèmes opaques.

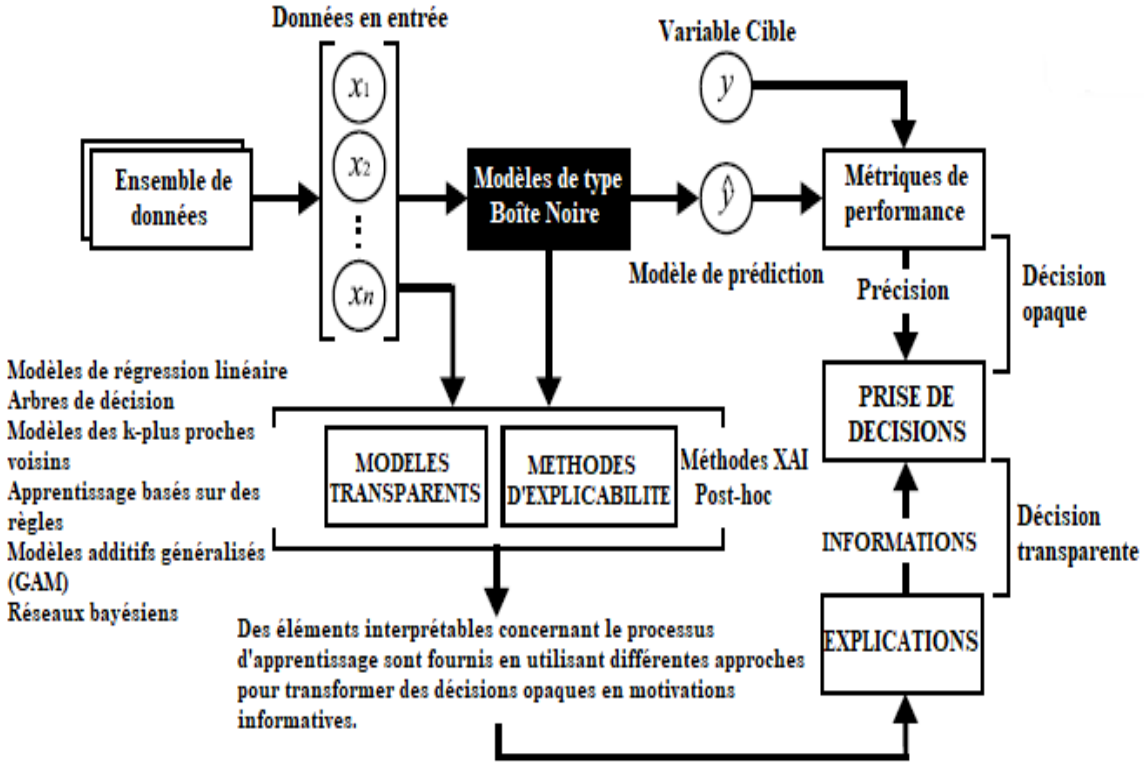


Figure 11. Complémentarité entre le ML et le XAI (inspiré de **Arrieta et al., 2020)**

Dans la **Figure 11**, l'explicabilité est positionnée comme un complément au ML. Les modèles complexes fonctionnent comme des boîtes noires et les techniques XAI expliquent les décisions des boîtes noires en termes interprétables, permettant aux praticiens de prendre des décisions basées sur des informations plus transparentes

Bien que de nouvelles architectures d'apprentissage basées sur le DL soient constamment développées, pour atteindre de meilleures performances dans les domaines les plus variés, leur compréhension a été principalement ignorée **(Datta et al., 2016)**. Des recherches ont mis en

lumière les faiblesses des modèles de pointe, révélant que la perfection n'est pas toujours au rendez-vous, même lorsque l'apprentissage artificiel peut atteindre des performances remarquables. Des biais ethniques ont été repérés dans un algorithme évaluant les risques de récidive criminelle, tandis qu'un programme d'Amazon a écarté les minorités ethniques lors de la sélection des régions aux USA, éligibles à des promotions **(Guidotti et al., 2019)**. En outre, un modèle entraîné pour prédire la probabilité de décès dû à une pneumonie, a attribué un risque plus faible aux patients asthmatiques **(Ba and Caruana, 2014)**.

2- Classification et Catégorisations des Approches et des Méthodes XAI

L'explicabilité post-hoc cherche à rendre compréhensibles des modèles initialement complexes, en utilisant diverses approches pour en faciliter l'interprétation, telles que les explications textuelles, les explications visuelles, les explications locales, les explications par exemple, les explications par simplification et les techniques d'explications de la pertinence des caractéristiques. Chacune de ces techniques représente l'une des façons les plus répandues par lesquelles les êtres humains tentent d'élucider les systèmes et les processus qui les entourent.

La classification que nous allons présenter, se veut d'être la plus exhaustive possible, présentant un véritable arsenal de méthodes et de techniques XAI, dans le but de guider tout chercheur dans sa quête de la méthode idéale, en parfaite adéquation avec ses compétences. Cette classification prend également en compte le domaine spécifique dans lequel les techniques ont été mises en œuvre.

Pour pouvoir recenser toutes les classes et les techniques XAI, nous avons dû examiner attentivement plusieurs articles spécialisés dans la littérature, car aucune d'entre-elles ne les présentait toutes à la fois. Néanmoins, dans nos travaux de recherche, nous avons pu dresser un paysage très large, les prenant toutes en considération. À partir de là, nous pouvons présenter les approches XAI, suivantes :

- ***Approche basée sur le modèle ;***
- ***Approche basée sur la granularité ;***
- ***Approche basée sur le fonctionnement ;***
- ***Approche basée sur les résultats.***

Nous allons présenter tour à tour, chacune d'elles dans la section qui suit. Notons tout de même le fait que certaines de ces approches partagent certaines classes de méthodes d'explicabilité XAI comme le montre la **Figure 12** ci-après.

2.1. Approche basée sur le modèle

Dans cette catégorie, les techniques XAI sont organisées en fonction du modèle pour lequel elles ont été conçues. On peut recenser principalement deux catégories de méthodes XAI, à savoir : ***Les méthodes spécifiques au modèle*** et ***les méthodes agnostiques au modèle (Indépendantes du modèle)*** **(Arrieta et al., 2020 ; Ortigossa et al., 2024).**

Les méthodes spécifiques au modèle, sont indiquées lorsque l'objectif d'explicabilité est de déchiffrer la logique d'un classifieur spécifique, ou lorsqu'un avantage peut être tiré de l'architecture du modèle à expliquer.

Ces méthodes peuvent conduire à des performances particulièrement élevées, en tirant profit des particularités fonctionnelles de la catégorie des modèles à expliquer. Néanmoins, leur champ d'action reste restreint. En effet, étant spécifiquement adaptés à une catégorie précise de modèles d'apprentissage, ils risquent de manquer de flexibilité pour s'adapter à d'autres types de modèles, en dehors de leur domaine initial **(Arrieta et al., 2020 ; Ortigossa et al., 2024)**.

Ce manque de flexibilité peut poser un réel problème lorsqu'on doit expliquer une série de modèles hétérogènes, où la sortie d'un modèle représente l'entrée d'un autre. Par exemple on peut utiliser un algorithme de GP (Genetic Programming) pour l'extraction des caractéristiques et faire une classification avec un RNN (Recurrent Neural Networks**)(avijeet, 2023)** ; si on devait expliquer les résultats des deux modèles avec une méthode XAI spécifique au modèle, cela serait probablement problématique, car ils ne fonctionnent pas de la même façon.

À contrario, **les méthodes agnostiques au modèle** peuvent être utilisées « THEORIQUEMENT » avec n'importe quel modèle d'apprentissage. Elles ne dépendent pas de l'architecture ni de l'algorithme du modèle qu'on désire expliquer, car elles ne tiennent pas compte de ses spécificités. En revanche, elles visent à comprendre le raisonnement derrière les prédictions, en utilisant des simplifications, des estimations de pertinence ou des visualisations sans entrer dans la logique interne du classificateur, ce qui font d'elles des méthodes précieuses dans le sens où elles permettent de comparer différents modèles par une même technique d'explicabilité **(Arrieta et al., 2020 ; Ortigossa et al., 2024)**.

2.2. Approche basée sur la granularité

Une autre catégorisation dans le XAI est liée à la granularité des explications. On recense principalement deux classes de méthodes en ce qui concerne la granularité, à savoir : ***Explicabilité Globale*** et ***Explicabilité Locale*** **(Barredo Arrieta et al., 2020 ; Ortigossa et al., 2024).**

Les méthodes d'explicabilité globale sont utilisées pour comprendre comment le modèle se comporte par rapport aux variables qui ont le plus d'impact. Ces méthodes donnent un aperçu général de la manière dont un modèle se comporte, sur l'ensemble des données ou sur une partie importante de celles-ci. La stratégie est souvent utilisée pour comparer l'importance globale d'une variable et voir quelles autres variables sont plus importantes dans les décisions concernant la population. Par exemple, évaluer le comportement général dans des situations comme le changement climatique ou l'utilisation des médicaments est plus utile que d'expliquer chaque détail des modèles possibles.

Les méthodes d'explicabilité locale sont utiles quand il s'agit de donner des détails précis sur une prédiction, pour un cas particulier. Avec des connaissances avancées, le but est de saisir pourquoi une prédiction a été faite. Les explications locales sont utiles pour les modèles complexes qui réagissent différemment selon les différentes entrées.

Durant l'étape de modélisation, on soumet le modèle généré à partir des données d'entraînement à des métriques d'évaluation, afin de simuler les interactions du monde réel. Cependant, les données d'entraînement et la réalité sont comme deux mondes parallèles qui ne se croisent que rarement. L'explicabilité globale permet d'explorer si un modèle se comporte conformément à ce qui est attendu de sa conception. Il peut s'avérer ardu, voire peu instructif, de saisir d'un seul coup d'œil l'ensemble des correspondances apprises, surtout dans les modèles

comportant de nombreux attributs. En effet, cela requiert que la méthode d'explication parvienne à découvrir un optimum permettant de déceler toute relation fonctionnelle entre les données d'entrée et les sorties, ce qui peut poser un défi NP-difficile dans l'ensemble **(Wojtas and Chen, 2020).**

Au-delà des métriques d'évaluation, chaque prédiction doit être vérifiée individuellement par une explicabilité locale, surtout si une erreur peut avoir de graves conséquences (comme un mauvais diagnostic médical). Ces explications individuelles peuvent montrer quelles sont les caractéristiques d'une instance de données, qui mènent à une décision précise quand le modèle dans son ensemble (explication globale) n'est pas assez descriptif.

D'autre part, la précision locale ne signifie pas la fidélité globale, car des caractéristiques globalement importantes peuvent ne pas être localement importantes, et vice-versa **(Wojtas and Chen, 2020)**. Une explication globale totalement fidèle ne peut être obtenue sans une description complète de l'ensemble du modèle. En effet, selon Wojtas et Chen (**Wojtas and Chen, 2020)**, une simple collection d'explications d'instances peut ne pas fonctionner pour la caractérisation au niveau de la population, car les explications locales sont spécifiques au niveau de l'instance et souvent incohérentes avec les explications globales. Par conséquent, identifier des explications globalement fidèles et interprétables reste un défi **(Ribeiro et al., 2016a)**.

2.3. Approche basée sur le fonctionnement

Dans cette catégorie de méthodes d'explicabilité, on retrouve principalement cinq classes différentes de modèles XAI. Les trois premières ont été proposées par **Samek et Müller (2019)** et les deux dernières par **Arrieta et al. (2020)** (**Figure 12**) :

- **Explicabilité basée sur les Perturbations Locales ;**
- **Explicabilité par Exploitation des Structures ;**
- **Explicabilité basée sur les Méta-Explications ;**
- **Explicabilité basée sur la Modification de l'Architecture ;**
- **Explicabilité basée sur les Exemples.**

2.3.1. Explicabilité basée sur les Perturbations Locales

Les méthodes d'explicabilité qui appartiennent à cette catégorie perturbent légèrement les entrées d'un modèle afin de déterminer l'importance de certaines caractéristiques sur la prédiction du modèle.

2.3.2. Explicabilité par Exploitation des Structures

Les méthodes de cette catégorie utilisent des caractéristiques spécifiques des modèles d'apprentissage automatique pour créer l'explication. Dans les DNN, une façon courante d'utiliser la structure est d'observer les gradients ; ces derniers sont une généralisation multivariée des dérivées, ils montrent à quel point chaque valeur d'entrée est importante. Ces méthodes d'explicabilité donnent souvent des résultats sur l'importance des caractéristiques, comme celles qui utilisent des perturbations locales.

2.3.3. Explicabilité basée sur les Méta-Explications

Les méthodes d'explicabilité de cette classe ne s'appliquent pas directement à un modèle d'apprentissage automatique, mais aux explications de ce modèle, produites par d'autres techniques d'explicabilité. Ces explications sont rassemblées et comparées entre-elles, pour donner une meilleure explication que chacune des méthodes prises séparément. Encore une fois, ces méthodes aboutissent souvent à des attributions d'importance des caractéristiques en entrée. On peut spécifier que les méta-explications ne peuvent pas être considérées comme un mode de fonctionnement, mais plutôt comme un résultat, car il faut les créer. Cependant, puisque la formation des méta-explications aboutit souvent à l'attribution de l'importance à certaines caractéristiques, cette méthode a été considérée comme une approche fonctionnelle pour mieux faire la distinction.

2.3.4. Explicabilité basée sur la Modification de l'Architecture

Ces méthodes essaient de rendre les modèles complexes plus simples, en modifiant leur structure. Dans les réseaux de neurones convolutionnels par exemple, on peut remplacer les couches de convolutions par des couches de max-pooling. Modifier l'architecture peut rendre les explications apportées par d'autres méthodes d'explicabilité plus claires, ou même aboutir à des modèles explicables ante-hoc.

2.3.5. Explicabilité basée sur les Exemples

Cette démarche repose sur l'extraction d'exemples de données qui se rapportent au résultat généré par un modèle, permettant ainsi de mieux comprendre le modèle lui-même. De la même manière que les humains se comportent lorsqu'ils tentent d'expliquer un processus donné, les explications par exemple sont principalement centrées sur l'extraction d'exemples représentatifs, qui saisissent les relations internes et les corrélations trouvées par le modèle analysé.

2.4. Approche basée sur les résultats

Cette approche considère le résultat d'une méthode d'explicabilité comme le constituant essentiel pour sa classification, composée de trois classes différentes :

- **Explicabilité basée sur l'Importance des Caractéristiques ;**
- **Explicabilité basée sur les Modèles de Substitution ;**
- **Explicabilité basée sur les Exemples** (voir section 2.3.5).

2.4.1. Explicabilité basée sur l'Importance des Caractéristiques

Dans la littérature sur XAI, on parle souvent de l'effet ou de l'impact des caractéristiques, de la contribution des variables et de leur interprétation. Ces termes expliquent comment et dans quelle mesure chaque caractéristique d'entrée influence la prédiction du modèle, c'est-à-dire **l'importance des caractéristiques**. **Breiman en (2001)** a suggéré l'une des premières méthodes pour repérer les caractéristiques importantes. Sa méthode consiste à permuter chaque caractéristique, en mélangeant au hasard les valeurs des caractéristiques, pour pouvoir examiner et évaluer leurs contributions individuelles. Plus clairement, permuter les valeurs des caractéristiques, a pour effet de déconnecter la caractéristique de la variable cible, ce qui fait chuter les performances de prédiction si la caractéristique est importante. Donc, le niveau de

perte de performance montre à quel point le modèle dépend de cette caractéristique. Néanmoins, la méthode de Breiman est spécifique aux Random Forests entraînées.

Expliquer les prédictions par des interprétations au niveau des caractéristiques est un objectif commun des approches XAI et plusieurs chercheurs classifient les méthodes XAI selon d'autres éléments ou mécanismes appliqués, pour accomplir la tâche de l'importance des caractéristiques. Nous comptons parmi elles :

A. Attribution des caractéristiques

Cette technique mesure les contributions des caractéristiques d'entrée individuelles aux performances d'un modèle d'apprentissage supervisé, en répartissant équitablement les valeurs prédites entre les variables d'entrée pour quantifier la pertinence de chaque variable **(Wojtas and Chen, 2020)**.

B. Importance Additive

Cette technique se base sur le calcul de la somme de toutes les importances des caractéristiques qui devrait approcher la valeur prédite originale **(Lundberg and Lee, 2017).**

C. Sensibilité

Elle mesure comment la performance prédictive d'un modèle d'apprentissage varie (augmente ou diminue), en perturbant chaque caractéristique d'entrée **(Mishra et al., 2021)**. Du point de vue de l'analyse de sensibilité, plus les variables sont importantes, plus leur contribution à la performance prédictive est significative.

D. Basé sur le gradient

Un cas particulier de l'approche de sensibilité qui évalue le comportement du modèle d'apprentissage automatique à travers des perturbations de taille très réduites **(Bhatt et al., 2021)**.

E. Sélection des caractéristiques

En partant d'un ensemble de données de **n** caractéristiques, cette méthode identifie une combinaison ou un sous-ensemble de **p** caractéristiques importantes, qui entraînent un modèle avec une perte minimale de précision. En pratique, $\mathbf{p \ll n}$ pour la plupart des tâches de sélection de caractéristiques **(Das et al. 2022).**

À présent, faisons la distinction entre **la sélection des caractéristiques** et **l'extraction de caractéristiques**. Bien que les deux méthodologies visent à améliorer la performance des modèles basés sur les données, en réduisant l'espace des caractéristiques d'origines, les méthodes pour extraire des caractéristiques sont liées à la réduction de la taille des données. Elles créent un nouvel ensemble de caractéristiques à partir des données originales, en utilisant des transformations simples ou complexes. Ces transformations permettent de réduire les données de grande taille tout en gardant les informations importantes **(Van Der Maaten and Hinton, 2008**).

Les méthodes de sélection des caractéristiques, bien que visant également à réduire la dimensionnalité, effectuent la sélection des caractéristiques, en supprimant les axes de données basés sur des projections canoniques au lieu d'apprendre des cartographies.

2.4.2. Explicabilité basée sur les Modèles de Substitution

Les méthodes XAI, qui construisent des modèles de substitution, essaient d'approximer (une partie spécifique) le modèle original avec un modèle plus simple et explicable à priori. Les modèles de substitution peuvent être créés de nombreuses manières, par exemple en sondant le modèle original via des perturbations locales, ou en exploitant sa structure. Par conséquent, les modèles de substitution peuvent être le résultat de la plupart des modes de fonctionnement.

Dans ce contexte, nous pouvons affirmer que la classe la plus importante de cette catégorie est :

A. Explicabilité par simplification

Elle est sans doute la technique la plus largement utilisée dans la catégorie des méthodes post-hoc indépendantes du modèle. Les explications locales sont également présentes, car parfois, les modèles simplifiés ne sont représentatifs que de certaines sections d'un modèle. Presque toutes les techniques empruntant cette voie pour la simplification des modèles sont basées sur des techniques d'extraction de règles (**Figure 12**).

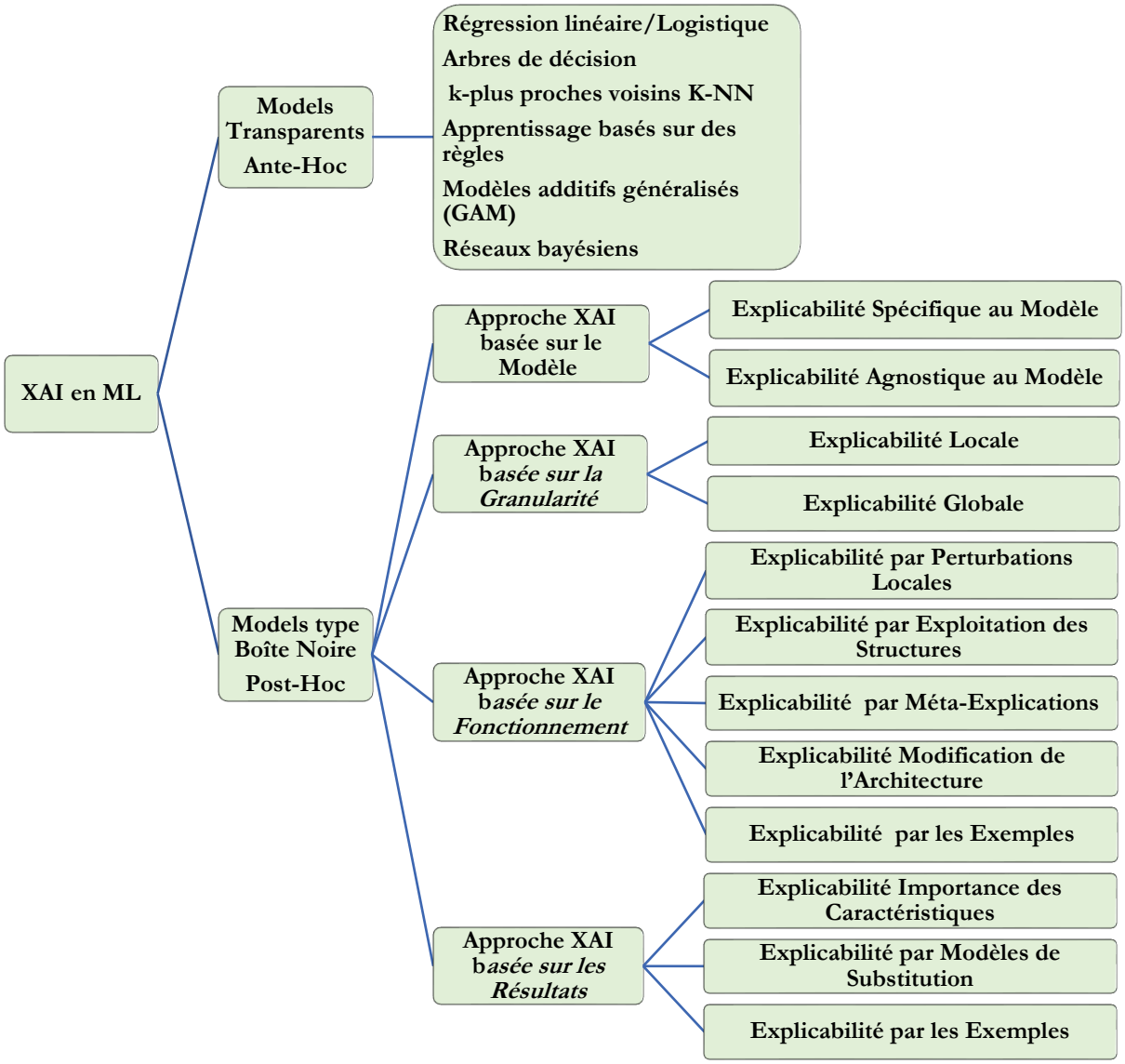


Figure 12. Catégorisation des approches et méthodes d'explicabilité XAI dans le ML

3- Techniques XAI

À partir des méthodes et des catégories précédemment citées, nous pouvons décrire les techniques les plus utilisées dans le domaine du XAI tout en essayant d'exposer leurs avantages, leurs inconvénients ainsi que leurs limitations.

3.1. Techniques XAI basées sur les Approximations

Selon **Hamilton et al. (2022)**, l'explicabilité peut être réalisée par des algorithmes intrinsèquement interprétables (Modèles type boîtes blanches (voir section 2)) afin d'atteindre la performance prédictive du modèle boîte noire d'origine. Un modèle boîte noire peut servir de « **Entraîneur** » pour transférer des connaissances à un modèle plus transparent et interprétable, qui approximera et expliquera les résultats du prédicteur original, un processus également désigné sous le terme de **distillation de modèle (Hinton et al., 2015)**.

Les arbres de décision, la régression linéaire/logistique (RL) ou encore la classification basée sur les règles, sont les modèles les plus utilisés pour fournir les explications sur les modèles types boîtes noires connus pour être intrinsèquement complexes et non linéaires.

Les arbres de décision, sont considérés par la majorité des chercheurs comme une norme d'interprétabilité, car la séquence logique d'un arbre de décision peut être interprétée intuitivement par un analyste humain **(Amann et al. 2022)**. Cependant, les arbres de décision ont tendance à avoir une faible généralisabilité en plus d'être sujets au surapprentissage ce qui conduit inéluctablement à des explications ayants une faible précision.

Dans le même registre de l'explicabilité, les modèles d'explicabilité, basés sur la RL partent du principe que les données d'entrées sont linéairement séparables, ce qui est rarement le cas dans la réalité rendant leur mise en œuvre difficile voire impossible dans certains cas.

Une solution élaborée par **Guidotti et al. (2018)** repose sur des classificateurs basés sur des règles, utilisant un algorithme génétique pour l'échantillonnage du voisinage d'une instance spécifique, entraîner un arbre de décision, puis produire une explication. Malgré leur réputation de transparence, les approches basées sur des règles rencontrent des défis de scalabilité, tout comme les modèles linéaires. Dans certaines situations, il peut être indispensable de créer un grand nombre de règles afin d'atteindre des performances de classification satisfaisantes, ce qui peut rendre l'analyse impraticable. Les modèles basés sur des règles sont particulièrement appropriés pour les approximations dans des domaines restreints, tandis que les modèles plus simples et transparents se révèlent inefficaces pour traiter des données de haute dimension comportant des relations complexes.

Outre ces modèles, les Modèles Additifs Généralisés (GAM) se sont présentés comme une alternative interprétable aux modèles de régressions complexes. Ces modèles s'appuient sur le lissage linéaire et décomposent une fonction prédictive en une agrégation de composants unidimensionnels définis pour chaque variable prédictive. Ainsi, il est possible de saisir les relations non linéaires spécifiques entre les variables lors de la modélisation pour construire des explications **(Lou et al., 2012 ; Caruana et al.,2015)**. Toutefois, les générateurs d'explications automatiques basés sur les GAM, se heurtent à une limitation en termes de capacité à fournir des explications pour les modèles les plus complexes.

Tan et al. (2023) ont présenté une étude comparative indiquant que les explications générées par des modèles transparents distillés obtiennent des résultats précis dans des contextes d'explications additives. Malheureusement, il est parfois impossible de construire un

modèle interprétable qui agit comme une approximation de meilleur ajustement imitant des modèles complexes type boîte noire. D'après **Linardatos et al. (2020)**, il est extrêmement complexe de mettre en place un modèle compétitif et transparent dans des domaines spécifiques tels que le traitement du langage et la vision par ordinateur, en raison de l'écart de performance insurmontable par rapport aux modèles reposant sur l'apprentissage profond.

3.2. Techniques XAI basées sur la visualisation des informations

La visualisation de l'information permet de cartographier les données sous des formats graphiques, facilitant ainsi leur représentation. Ce procédé aide les analystes à identifier visuellement des tendances, des motifs et des caractéristiques **(Alexandrina et al., 2019).**

Dans ce contexte, **Marcílio-Jr et al. (2021)** ont élaboré un outil indépendant du modèle, fondé sur des perceptions coordonnées, afin de visualiser la similarité entre les classes. Il évalue l'importance des caractéristiques en recourant à l'optimisation afin d'introduire des perturbations au sein des caractéristiques individuelles, dans le but de minimiser simultanément la performance du modèle ainsi que l'ampleur des perturbations.

Les graphiques de dépendance partielle **(Zhao and Hastie, 2021)** constituent des outils graphiques employés pour expliquer les modèles d'apprentissage supervisé, en permettant de visualiser l'effet marginal moyen (valeurs de dépendance partielle) entre les variables d'entrée et les prédictions. Bien qu'elle a montré une bonne précision d'explication, et une bonne aptitude a capté les relations monotoniques, cette technique, présente néanmoins un inconvénient majeur relatif au fait d'obscurcir les effets hétérogènes et les relations complexes résultant des interactions entre les caractéristiques.

Les courbes d'Expectation Conditionnelle Individuelle, proposées par **Goldstein et al. (2015)** et améliorée par **Casalicchio et al. (2018)**, abordent cette problématique en décomposant la sortie de dépendance partielle et en illustrant dans quelle mesure la prédiction d'une instance individuelle varie en fonction de la modification de la valeur d'une caractéristique sélectionnée.

Il existe d'autres techniques fondées sur la visualisation des modèles XAI qui emploient activement les techniques de réduction de dimensionnalité ou à des projections multidimensionnelles **(Nonato and Aupetit, 2019 ; Ortigossa et al., 2022)** dans le but de produire des représentations interprétables des espaces de caractéristiques des modèles d'apprentissage, notamment en ce qui concerne les relations entre les neurones et leur impact sur les données **(Hohman et al., 2019)**.

Cantareira et al. (2020) ont élaboré une technique qui illustre le flux de données d'activation au sein des couches cachées des réseaux de neurones. Les données facilitent la vérification de l'évolution du réseau au cours du processus d'entraînement, et des représentations sont générées lorsque l'information est transmise d'une couche à la couche suivante. **Rauber et al. (2017)** ont présenté une méthode analogue qui examine de manière visuelle la façon dont les neurones artificiels modifient les données d'entrée au fur et à mesure de leur passage à travers les couches cachées des réseaux de neurones profonds.

UMAP (Uniform Manifold Approximation and Projection) **(McInnes et al., 2018)** et t-SNE (t-distributed Stochastic Neighbor Embedding) **(Van Der Maaten and Hinton, 2008)**

constituent deux techniques de projection multidimensionnelle robustes, couramment employées dans les approches d'explicabilité des modèles (XAI) reposant sur la visualisation.

Néanmoins, les méthodes de réduction de dimensionnalité présentent des contraintes d'utilisabilité concernant le nombre de points pouvant être visualisés simultanément. L'explicabilité par le biais de la visualisation de l'information se heurte à des défis de scalabilité associés à la gestion d'un grand nombre d'éléments, tout en s'assurant de décrire de manière appropriée leurs relations.

3.3. Techniques XAI basées sur les Frontières de Décisions

L'analyse du comportement des modèles d'apprentissage automatique à travers l'étude de leurs frontières de décision constitue une démarche encore peu investiguée dans la littérature. **Karimi et al. (2019)** ont élaboré DeepDIG, une méthode fondée sur la génération d'échantillons adverses (Adversarial Sampling en Anglais) qui se situent à une distance suffisamment réduite de la frontière de décision, c'est-à-dire des instances synthétiques se trouvant entre deux classes distinctes. Plus précisément, DeepDIG opère sur un Réseau de Neurones Profond (DNN) préalablement entraîné et produit des échantillons d'instances de frontière avec des probabilités de classification pour deux classes distinctes, en veillant à ce qu'elles soient aussi proches que possible, ce qui engendre une incertitude dans la classification. Ces instances frontières synthétiques sont par la suite employées pour évaluer la complexité ou la non-convexité de la frontière de décision acquise par le réseau de neurones profond (DNN) entraîné.

Une des caractéristiques essentielles des réseaux de neurones profonds réside dans leur capacité exceptionnelle à généraliser, acquise par le biais de combinaisons élaborées de transformations non linéaires. En conséquence, les réseaux de neurones profonds (DNN) sont en mesure de cartographier des données présentant des relations complexes et de haute dimension, ce qui soulève la problématique de savoir si la complexité des données dans l'espace d'entrée est effectivement reflétée dans l'espace transformé (appris) du réseau.

Afin d'explorer cette problématique, **Karimi et al. (2019)** ont élaboré deux métriques destinées à caractériser la complexité des frontières de décision : l'une pour l'espace d'origine (données d'entrée) et l'autre pour l'espace transformé. Les auteurs ont par la suite examiné l'hypothèse formulée par **Li et al. (2018)** relative à la frontière de décision émanant de la dernière couche d'un réseau de neurones profond (DNN) entraîné par rétropropagation, laquelle converge vers la solution d'un SVM linéaire formé sur les données prétraitées.

3.4. Techniques XAI basées sur des Exemples Contrastifs et Contrefactuels

Le concept de contrastivité est emprunté aux sciences sociales, lesquelles affirment que les explications humaines découlent principalement des processus contrastifs **(Jacovi et al., 2021)**. Les explications contrastives permettent de clarifier les raisons pour lesquelles un événement s'est produit en opposition à un autre. En conséquence, la caractéristique de contrastivité stipule qu'une explication doit s'efforcer de répondre aux interrogations relatives aux raisons pour lesquelles un événement s'est produit, en tenant compte de ses causes potentielles (alternatives hypothétiques). À titre d'illustration, une explication « raisonnable » à une interrogation telle que « pourquoi l'événement A s'est-il manifesté plutôt que l'événement B ? » présenterait les motifs causaux ayant orienté le modèle vers l'événement A **(Stepin et al., 2021)**.

Dans le domaine de l'XAI, les méthodes contrastives fournissent des éclairages sur les motifs ayant conduit un modèle à effectuer une prédiction particulière, en soulignant les caractéristiques ayant influencé cette prédiction et en les opposant à des caractéristiques susceptibles de générer des résultats alternatifs.

De surcroît, divers scénarios peuvent être élaborés pour une prédiction spécifique en cas de légères altérations des données d'entrée, permettant ainsi d'expliquer les conséquences potentielles « contraires aux faits » de ces modifications. Les explications contrefactuelles possèdent une riche tradition dans les domaines de la philosophie et de la psychologie, car, en ce qui concerne les explications humaines, elles s'inscrivent dans des schémas de dépendance contrefactuels **(Verma et al., 2020)**. De même les techniques XAI faisant partie des méthodes XAI contrefactuelles, élaborent des instances ou des scénarios similaires à l'entrée initiale, pour lesquels la sortie du classificateur subit une modification **(Bhatt et al., 2021)**. Ces méthodes décrivent les caractéristiques susceptibles d'évoluer dans la prédiction en cas de perturbation, de suppression ou d'ajout de valeurs au sein des caractéristiques prédictives **(Liao et al., 2020)**.

Les explications contrefactuelles ne répondent pas de manière explicite à la question « pourquoi » un modèle émet une prédiction ; en revanche, leur objectif principal consiste à établir un lien entre ce qui aurait pu se produire si une certaine entrée avait été modifiée d'une manière spécifique, offrant ainsi des indications en direction de la prédiction recherchée **(Verma et al., 2020 ; Mishra et al., 2021)**.

3.5. Techniques XAI basées sur l'explication de l'Apprentissage Automatique des Graphes (GML) (Graph Machine Learning)

Les réseaux de neurones graphiques (GNN) constituent une catégorie puissante au sein des GML, exploités pour la génération des prédictions concernant des données qui sont liées à une structure de graphe sous-jacente. De nombreuses applications dans le monde réel se manifestent de manière intrinsèque sous la forme de modèles de graphes, telles que les réseaux sociaux, la détection de fraude, les graphes de connaissances, la bioinformatique, la modélisation des molécules en chimie, les cartes routières, la citation de documents, ainsi que l'optimisation des infrastructures **(Luo et al., 2020 ; Jia et al., 2023).**

L'ensemble des différentes approches spécifiques impliquées dans le GML est vaste, car les vecteurs de caractéristiques peuvent être associés aux nœuds du graphe (par exemple, le contenu d'un document), aux arêtes du graphe (par exemple, les messages entre utilisateurs dans un réseau social) et/ou à l'ensemble du graphe (par exemple, la toxicité d'une molécule).

La tâche d'apprentissage automatique, peut également consister en une prédiction :

- ✓ Au niveau des nœuds (par exemple, prédire la classe à laquelle appartiennent les documents),
- ✓ Au niveau des arêtes (par exemple, prévoir le flux de circulation dans les rues d'une ville),
- ✓ Au niveau du graphe (par exemple, prévoir la solubilité d'une molécule),
- ✓ Une prédiction de liens (par exemple, recommander des utilisateurs qui pourraient se suivre mutuellement),
- ✓ Une interaction graphe-à-graphe (par exemple, prédire les effets secondaires de la prise simultanée de deux médicaments).

Prendre en compte non seulement les caractéristiques associées aux éléments du graphe, mais aussi les interactions complexes définies par sa structure peut rendre l'explicabilité des GNN difficile, conduisant à une littérature moins étendue par rapport à un scénario XAI non basé sur les graphes.

Cependant, des progrès récents et notables ont été réalisés dans la recherche en XAI pour fournir des explications aux modèles complexes de GML. **Ying et al. (2019)** ont proposé GNNExplainer, une méthode basée sur la perturbation qui vise à expliquer les prédictions individuelles faites par des GNN entraînés. Elle met en évidence les nœuds et les arêtes les plus influents dans le graphique d'entrée en calculant les gradients de la prédiction d'intégration des nœuds. Si elle fournit des informations précieuses, la méthode présente toutefois certaines limites, notamment une sensibilité à l'intégration initiale des nœuds, le recours à des approximations pour les grands graphes et la difficulté d'expliquer les interactions complexes.

3.6. XAI basé sur l'Interprétation des Modèles d'Attention

Dans les modèles de séquence classiques, tels que les RNN ou les LSTM, les unités d'information sont transmises de manière séquentielle d'une étape à la suivante. Une telle nature séquentielle restreint la capacité à saisir le contexte, et ce, même en utilisant des RNN et des LSTM qui possèdent des architectures spécifiquement élaborées pour maintenir l'information sur une période prolongée **(Jozefowicz et al., 2016 ; Kim et al., 2017).**

En revanche, les méthodes fondées sur l'attention ne traitent pas les entrées de façon séquentielle. Le mécanisme d'attention confère au modèle la capacité d'attribuer une importance relative à diverses sections de la séquence d'entrée et de se concentrer sur certaines d'entre elles lors de l'élaboration des prédictions. Chaque jeton d'information est traité de manière simultanée en parallèle avec l'ensemble des autres jetons, ce qui permet d'effectuer des calculs plus efficaces et évolutifs **(Vaswani et al., 2017).** Par ailleurs, le mécanisme d'attention facilite une concentration sélective sur diverses sections de la séquence d'entrée, ce qui contribue à la conservation du contexte. Cette approche s'avère particulièrement efficace pour les applications nécessitant le traitement de données séquentielles, notamment dans le domaine du traitement du langage naturel (NLP) **(Biderman et al., 2023a)**.

Vaswani et al., (2017) ont présenté les Transformers, une catégorie d'architecture de réseau de neurones qui s'appuie sur des mécanismes d'attention (ou attention par produit scalaire) afin de saisir les relations entre divers mots ou tokens au sein d'une séquence. À un niveau avancé, le mécanisme d'auto-attention permet à un jeton au sein d'une séquence de se focaliser sur d'autres jetons présents dans cette même séquence, attribuant ainsi des niveaux d'importances variés à chaque jeton. L'approche transformer ne se limite pas aux contextes de taille fixe, permettant ainsi aux tokens d'exercer une influence directe les uns sur les autres, indépendamment de leur distance au sein de la séquence.

Les Transformers ont révélé des performances exceptionnelles dans de nombreuses tâches de traitement du langage naturel, et leur adoption s'est largement étendue à divers domaines. Bien que **Vaswani et al., (2017)** aient soutenu que les mécanismes d'attention employés dans les Transformers pouvaient engendrer des modèles plus interprétables, cette transparence demeure sujette à débat **(Jain and Wallace, 2019)**. En effet, l'interprétation du fonctionnement interne d'un modèle de transformateur peut s'avérer complexe et nécessite une compréhension approfondie **(Biderman et al., 2023a, 2023b)**.

Vig (2019) a succinctement abordé les recherches ayant conduit à l'élaboration d'outils permettant de visualiser l'attention au sein des modèles de traitement du langage naturel, allant des cartes de chaleur aux représentations graphiques. L'auteur a proposé une version optimisée de **BertViz (2019)**, un outil de visualisation structuré en trois perspectives selon le modèle des petits multiples, permettant d'explorer les modèles de transformateurs aux niveaux de tête d'attention, de modèle et de neurone. De plus, il a illustré un cas pertinent dans lequel un modèle fondé sur l'attention a intégré des biais de genre. Néanmoins, BertViz présente certaines restrictions. Il peut présenter une performance réduite lors du traitement d'entrées volumineuses ou de modèles de grande taille, et seuls quelques modèles fondés sur des transformateurs y sont intégrés. La présentation des cartes thermiques des poids d'attention peut prêter à confusion, conduisant par conséquent à des interprétations ambiguës. En outre, les expériences contrefactuelles ont la capacité de générer des cartes thermiques alternatives susceptibles de produire des prédictions équivalentes **(Jain and Wallace, 2019)**. Cependant, **Wiegreffe et Pinter (2019)** ont soutenu que la présence d'une explication alternative n'implique pas que l'explication fournie soit dénuée de sens ou erronée.

Le système Pythia **(Biderman et al., 2023a),** un cadre de référence destiné à l'évaluation des grands modèles de langage (LLM) (Large Language Models), comprend plusieurs modèles basés sur des transformateurs, accessibles librement et pré-entraînés, englobant une vaste gamme d'échelles atteignant jusqu'à 12 milliards de paramètres. L'étude souligne également l'importance capitale de la taille du modèle dans l'efficacité de la modélisation linguistique et propose des analyses concernant les biais de genre ainsi que la mémorisation. Bien que la mémorisation au sein des LLM soit devenue une préoccupation significative, peu d'outils sont à la disposition des data scientists pour en détecter et en prévenir les occurrences. **Biderman et al., (2023b)** ont présenté une synthèse de la mémorisation et ont suggéré des indicateurs pour en appréhender et en anticiper les mécanismes.

Garde et al. **(Garde et al., 2023)** ont présenté DeepDecipher, une interface interactive destinée à la visualisation et à l'interprétation des neurones au sein des couches MLP des modèles de transformateurs. Il propose des informations concernant le comportement des neurones afin de comprendre les circonstances et les raisons de l'activation d'un neurone MLP, en s'appuyant sur une base de données préétablie de séquences ainsi qu'une méthode permettant de générer un graphe de tokens **(Foote et al., 2023)**. Néanmoins, une perspective neuronale peut ne pas refléter de manière adéquate son comportement global, et DeepDecipher ne propose pas de méthode d'explication véritablement innovante.

Étant donné que les modèles fondés sur l'attention, qui sont à la fois vastes et complexes, ont acquis une influence croissante dans les applications intelligentes, il est impératif de leur conférer une interprétabilité. Des centaines de nouvelles études ont été récemment publiées et, bien que leur succès soit indéniable, il est impératif d'approfondir la compréhension des modèles de transformateurs **(Biderman et al., 2023b)**. Un grand nombre de solutions d'explicabilité des modèles d'attention mettent en œuvre des outils de visualisation. La visualisation des poids d'attention éclaire une partie du processus prédictif, toutefois elle ne garantit pas nécessairement une explication satisfaisante **(Jain and Wallace, 2019). Chefer et al. (2021)** ont présenté une approche fondée sur le gradient en vue de déterminer les scores de pertinence pour les modèles de transformateurs. Nous examinerons l'approche par gradient dans la section suivante.

L'étude des mécanismes par lesquels un modèle de transformateur spécifique acquiert et représente les données pourrait avoir un impact significatif sur le développement de la prochaine génération de logiciels **(Evandro et al., 2024).**

3.7. XAI basé sur les Gradients et la Décomposition du Signal

Les méthodes basées sur les gradients utilisent les dérivées partielles des modèles d'apprentissage pour expliquer leurs prédictions. Elles attribuent de l'importance aux caractéristiques d'entrée en analysant l'impact de petites perturbations de ces caractéristiques sur la sortie du modèle **(Ortigossa et al., 2024).** De plus, le calcul des gradients de la sortie par rapport à l'entrée est analogue à la vérification des coefficients d'un modèle de réseau de neurones **(Sundararajan et al., 2017)**, procédant à une généralisation de la procédure de reconstruction du réseau déconvulutionnel **(Simonyan et al., 2013 ; 2014)**. Les premières applications basées sur les gradients se sont concentrées sur la détermination des entrées maximisant l'activité neuronale des architectures de réseaux non **supervisés (Erhan et al., 2009)** et la génération de visualisations pour les couches convolutionnelles des réseaux profonds **(Zeiler and Fergus, 2014)**.

Simonyan et al. (2013 ; 2014) ont introduit l'utilisation des gradients pour générer des cartes de saillance pour les modèles supervisés (Une **carte de saillance** (ou **saliency map** en anglais) est une représentation visuelle utilisée pour expliquer ou interpréter les prédictions des modèles d'apprentissage automatique, en particulier des réseaux neuronaux profonds. Elle met en évidence les zones ou les caractéristiques des données d'entrée qui ont le plus influencé la décision du modèle.)– une telle approche est appelée **Vanilla Gradient** par la communauté XAI **(Agarwal et al., 2022 ; Krishna et al., 2022 ; Ortigossa et al., 2024)**. Il calcule directement les gradients de sortie du modèle à l'aide d'une expansion de Taylor du premier ordre **(Montavon et al., 2018)** autour d'une instance perturbée et d'un terme de biais. Le produit du gradient et des valeurs des caractéristiques d'entrée (sans modifications) est interprété comme une attribution de l'importance de ces dernières. Malgré sa simplicité, l'approche manque de sensibilité fine et est sujette au bruit au sein des gradients. De plus, ni la procédure de perturbation, ni le terme de biais n'ont été spécifiés de manière adéquate **(Bach et al., 2015)**.

De la même manière, T-Explainer une technique proposée par Ortigossa et al. (2024), s'appuie sur les développements des séries de Taylor afin d'approximer le comportement local des modèles de type boîte noire et de réaliser des attributions de caractéristiques. Néanmoins, la méthode procède au calcul des gradients par le biais de perturbations d'entrée dans le cadre d'une procédure d'optimisation fondée sur les différences finies, et ce, de manière indépendante de l'architecture du modèle. T-Explainer opère avec des données tabulaires, bien qu'il présente certaines limitations en ce qui concerne les caractéristiques catégorielles.

Les techniques XAI basées sur les gradients et la décomposition du signal sont très nombreuses **(Bach et al., 2015 ; Montavon et al., 2018 ; Lapuschkin et al., 2019 ; Kohlbrenner et al., 2020 …)** ; mais celle qui nous intéresse le plus, c'est la technique **Grad-CAM** (Gradient-weighted Class Activation Mapping) proposée par **Selvaraju et al. (2017)** (**Figure 13**).

Les couches convolutionnelles des architectures CNN, appliquent des filtres spécialisés sur les images d'entrée pour apprendre des motifs visuels complexes, tels que les informations spatiales et les sémantiques de haut niveau. Le modèle Grad-CAM **(Selvaraju et al., 2017)** génère des explications pour tout modèle CNN en attribuant des scores d'importance à chaque

neurone de la couche finale. Le processus d'attribution utilise des informations de gradient spécifiques à la classe **(Zhou, B. et al., 2016)** provenant du passage arrière de la rétropropagation pour produire une carte de localisation, qui met en évidence les régions les plus influentes de l'image d'entrée sur lesquelles le modèle se base pour prendre sa décision. Spécifiquement, Grad-CAM, calcule une matrice de scores d'importance W_C^K ; générant une carte de localisation $L^C_{Grad-CAM} \in \mathbb{R}^{m\,x\,n}$ où m et n représentent respectivement la largeur et la hauteur d'une image d'entrée, appartenant à une classe cible y^C, basée sur les gradients des poids des neurones sur chaque carte de caractéristiques M^K de la dernière couche de convolution, qui est calculée avant l'application de la fonction SoftMax $\frac{\partial y^C}{\partial M^k}$, en passant en arrière sur les dimensions m et n comme suit : $\boldsymbol{\omega_c^k = \frac{1}{z} \sum_i \sum_j \frac{\partial y^C}{M_{ij}^k}}$ **(6)**

Où : Z représente le nombre de pixels dans la carte de caractéristiques, qui est utilisé pour la normalisation de la sortie. Les scores d'importance W_C^K représentent une linéarisation partielle d'un CNN et décrivent l'importance de la carte de caractéristiques k pour la classe c. Enfin, les scores d'importance sont combinés linéairement en les moyennant globalement avec leurs cartes de caractéristiques correspondantes, en les passant dans une couche ReLU et en traçant la carte des scores finaux dans une carte de chaleur (Heatmap) :

$$\boldsymbol{L^C_{Grad-CAM} = ReLU(\sum_k W_C^K\, M^k)} \quad \textbf{(7)}$$

Grad-CAM et ses variantes récentes comme Grad-CAM++ **(Chattopadhay et al., 2018)** et LayerCAM **(Jiang et al., 2021)** génèrent des visualisations interprétables en superposant la carte thermique des scores sur l'image d'entrée originale, fournissant des informations visuelles permettant d'identifier les régions de l'image les plus influentes dans le processus de décision. Grad-CAM ne nécessite pas de modifications architecturales ni de réentraînement ; bien qu'il soit agnostique concernant les différents modèles de CNN, il est limité à ce type de modèle. Il dépend également de l'activation d'une couche ReLU pour une sensibilité correcte du gradient, ne permet pas de déterminer avec précision la couverture des régions de classe et est sujet à des instabilités lors de la localisation de plusieurs instances d'un objet dans une image **(Linardatos et al., 2020)**.

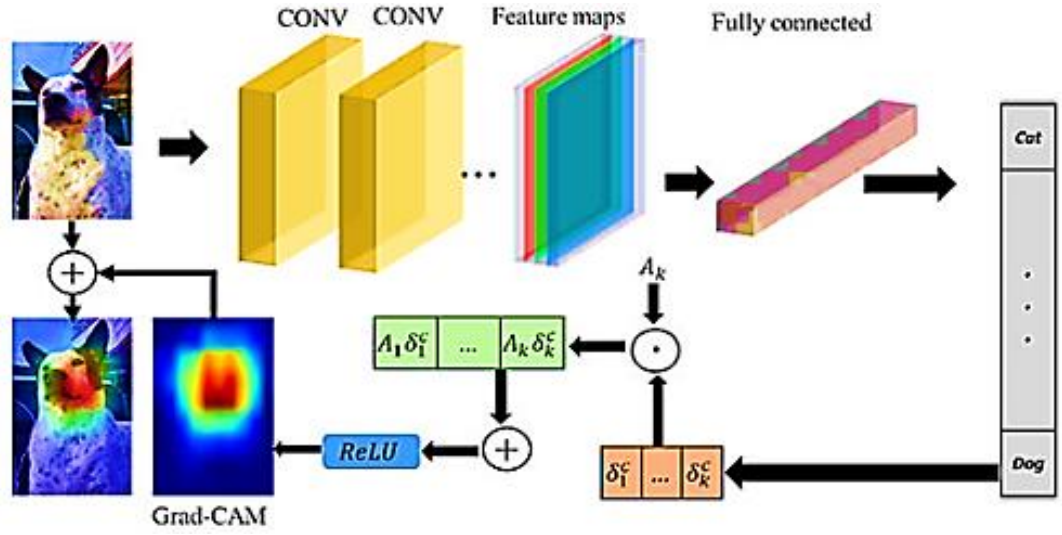


Figure 13. Architecture du modèle Grad-CAM **(Selvaraju et al., 2017)**

3.8. XAI basé sur des Simplifications

Sans nuls doutes, l'explicabilité par simplification, est l'une des techniques les plus importantes dans le paysage XAI. Elle comprend des techniques dans lesquelles un nouveau modèle explicatif est construit à partir d'un modèle entraîné à expliquer **(Arrieta et al., 2020)**. L'objectif du modèle simplifié est de reproduire un comportement similaire à celui du modèle original avec moins de complexité. Autrement dit, il doit conserver une performance prédictive similaire à celle du modèle original, mais basée sur des structures plus transparentes. **Thiagarajan et al. (2016)** ont développé TreeView, un outil qui interprète visuellement des modèles complexes. Il identifie les facteurs discriminatoires à travers les classes de données en utilisant une élimination séquentielle par une partition hiérarchique de l'espace des caractéristiques, regroupant les instances en clusters pour chaque facteur, où les associations indésirables sont écartées.

LIME **(Ribeiro et al., 2016a)** est l'une des techniques d'explicabilité les plus connues et largement appliquées. Elle consiste à déterminer un modèle linéaire interprétable qui approxime localement un modèle original. LIME génère un voisinage d'échantillons synthétiques autour de l'instance à expliquer en perturbant les instances du jeu de données original. Ceux-ci sont ensuite classés par le modèle d'apprentissage original, qui les pèse en appliquant un noyau de pondération selon leur proximité avec le point à expliquer. LIME détermine ensuite un modèle linéaire sur le voisinage, en minimisant une fonction de non-fidélité, et les prédictions sont expliquées par l'interprétation du modèle linéaire.

Plus précisément, soit f le modèle entraîné, g ∈ G un modèle interprétable, et G une classe de modèles potentiellement interprétables, tels que la régression linéaire ou les arbres de décision. Pour expliquer une instance n-dimensionnelle X = (x_1, . . . , x_n), un modèle interprétable g est déterminé en minimisant la fonction de perte $\mathcal{L}$ selon :

$$\mathcal{E}(X) = arg_{g \in G}\, min\ \mathcal{L}\,(f\,, g, \pi_{\mathbf{x}}) + \Omega(g) \qquad \textbf{(8)}$$

Où π_x définit le noyau de pondération centré sur x responsable de maintenir la fidélité locale de l'explication, et Ω est un terme de complexité (qui doit être maintenu *faible*) appliqué à g.

Certains auteurs ont également classé LIME comme un modèle de substitution local, défini comme la classe de méthodes qui expliquent les prédictions individuelles à travers un modèle de substitution entraîné localement, appelé « fidélité locale ». LIME dispose d'une interface graphique simple et informative. La **Figure 14** montre un exemple d'explication générée par LIME pour une instance du célèbre jeu de données Iris1 qui, pour simplifier, a été adapté pour ne contenir que deux classes. Concernant une tâche de classification binaire, LIME utilise un motif de deux couleurs (orange et bleu, dans ce cas). La **Figure 14a** montre les probabilités de classification de l'instance en cours d'examen, prédites par le modèle boîte noire comme appartenant à la classe Virginica. La **Figure 14b** affiche les attributs les plus pertinents par ordre d'importance pour la prédiction, avec les valeurs flottantes sur les barres horizontales indiquant l'importance attribuée à ces caractéristiques par LIME. La **Figure 13c** offre un aperçu de l'instance en cours d'examen, avec les valeurs originales de chaque caractéristique. Les couleurs sont réparties en fonction des contributions, c'est-à-dire que les attributs en orange ont contribué à la classe Virginica et ceux en bleu à la classe versicolor. Le code couleur est cohérent sur tous les graphiques. Cependant, la manière dont chaque attribut contribue

positivement ou négativement aux résultats de LIME n'est pas claire. Les auteurs ont également développé deux extensions de LIME **(Ribeiro et al., 2016b ; 2016c)** y compris une version améliorée qui fournit des explications textuelles plus claires basées sur des règles **(Ribeiro et al., 2018).**

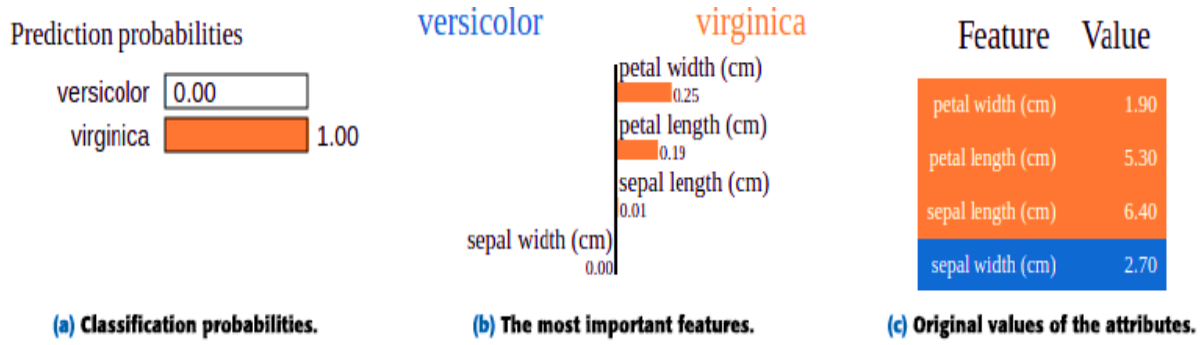


Figure 14. Explication LIME pour une instance de l'ensemble de données Iris, classée comme appartenant à la classe Virginica. LIME fournit des outils de visualisation avec différentes informations sur la classification du modèle, les valeurs d'importance attribuées localement et l'instance d'entrée elle-même. **(Ortigossa et al., 2024)**

En résumé, l'approche actuelle que LIME utilise pour approximer la solution de l'équation précédente (8) et produire une explication pour l'instance y, est :

1. Générer N échantillons "perturbés" de la version interprétable de l'instance pour expliquer y'. Soit $\{z_i' \in X' \mid i=1,\ldots,N\}$ l'ensemble de ces observations.
2. Récupérer les observations « perturbées » dans l'espace de caractéristiques original au moyen de la fonction de mappage. Soit $\{z_i \equiv h^y(z_i') \in X \mid i=1,\ldots,N\}$ l'ensemble dans la représentation originale.
3. Laissons le modèle boîte noire prédire le résultat de chaque observation « perturbée ». Soit $\{f(z_i) \in \mathbb{R} \mid i=1,\ldots,N\}$ l'ensemble des réponses, et soit $\mathfrak{Z}=\{(z'_i, f(z_i)) \in X' \times \mathbb{R} \mid i=1,\ldots,N\}$ l'ensemble de données des échantillons "perturbés" avec leurs réponses.
4. Calculer le poids de chaque observation « perturbée ». Soit $\{w^y(z_i) \in \mathbb{R}^+ \mid i=1,\ldots,N\}$ l'ensemble des poids.
5. Sélectionner K caractéristiques décrivant le mieux le résultat du modèle boîte noire à partir du jeu de données perturbé $\mathfrak{Z}$.
6. Ajuster un modèle de régression linéaire pondérée (en réalité, les implémentations actuelles de LIME ajustent une régression ridge (régression de crête) pondérée avec le paramètre de régularisation fixé à l'étape 1, à un ensemble de données réduit en caractéristiques composé des K caractéristiques sélectionnées à l'étape 5. Si le modèle boîte noire est un régresseur, le modèle linéaire prédira directement la sortie du modèle boîte noire. Si le modèle boîte noire est un classificateur, le modèle linéaire prédira la probabilité de la classe choisie.
7. Extraire les coefficients du modèle linéaire et les utiliser comme explications du comportement local du modèle boîte noire.

Un exemple de cette procédure pour les images est montré dans la **Figure 15**. Notez qu'en plus d'un modèle boîte noire (classificateur ou régresseur) **f** et d'une instance à expliquer **y** (ainsi que sa représentation interprétable y'), la procédure précédente nécessite de définir à l'avance le nombre d'échantillons **N**, la largeur du noyau **σ** et la longueur de l'explication **K**.

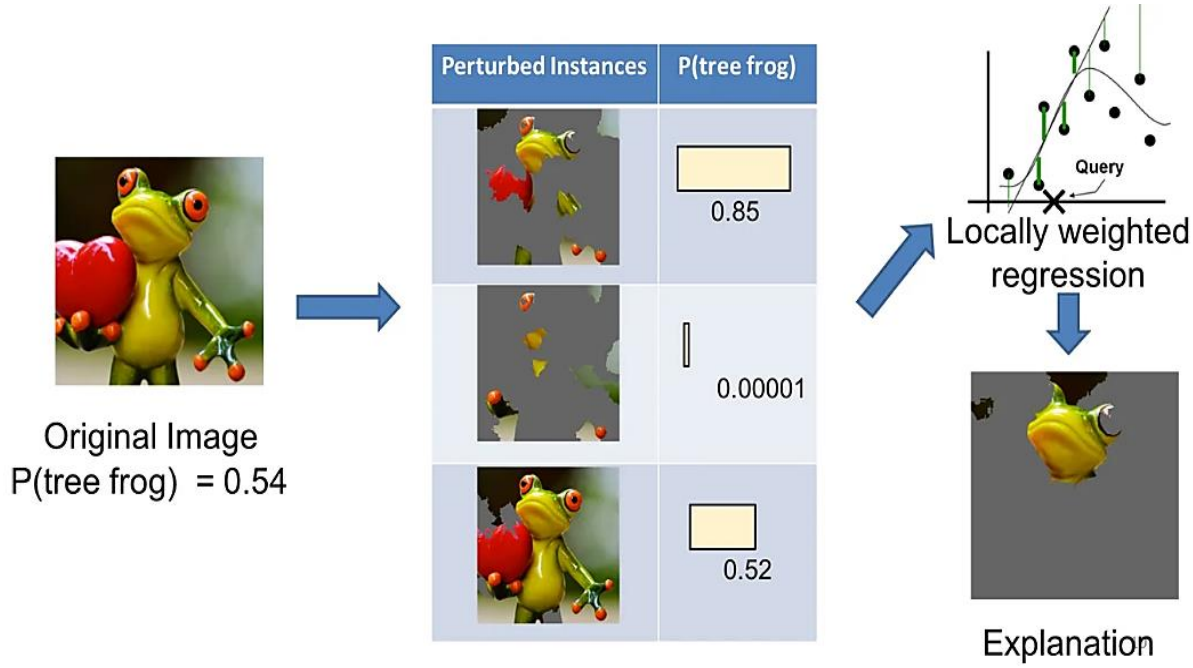


Figure 15. Expliquer la prédiction d'un classificateur avec LIME **(Ribeiro et al., 2016a)**

La **Figure 16**, montre un diagramme résumant le processus d'explicabilité avec LIME en quatre étapes. Pour générer des explications pour une classification, l'image donnée a d'abord été divisée en superpixels. Une distribution d'images perturbées a été générée et passée à travers le modèle de prédiction original pour calculer les probabilités de classification. Ces probabilités et images perturbées ont été présentées à un modèle de régression qui a estimé la contribution positive ou négative de chaque superpixel à la classification. Les poids de régression ont ensuite été tracés sur une carte de couleurs bleu-rouge.

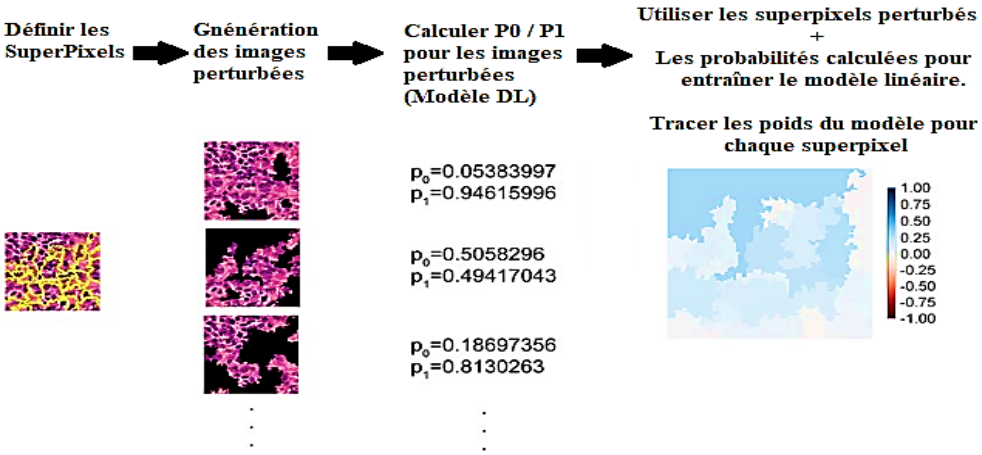


Figure 16. Diagramme de l'algorithme d'explicabilité avec LIME **(Ribeiro et al., 2016a)**

Aas et al. (2021) ont soutenu que LIME ne garantit pas une distribution parfaite des effets entre les variables. De plus, différents modèles peuvent s'adapter aux données échantillonnées, LIME en sélectionnant au hasard un modèle parmi ceux qui conviennent, sans garantir qu'il s'agit en fait de la meilleure approximation locale. Aucune garantie théorique solide n'indique qu'un modèle de substitution local simplifié représente adéquatement des modèles plus complexes, c'est-à-dire que les modèles originaux et de substitution produisent toujours des comportements prédictifs similaires.

Définir un voisinage significatif autour d'une instance d'intérêt est complexe. LIME surmonte une telle difficulté en construisant un voisinage autour du point à expliquer à l'aide du centre de gravité des données d'entraînement. Cette stratégie peut toutefois contribuer à l'instabilité de la technique en générant des échantillons considérablement différents de l'instance d'intérêt, malgré l'augmentation de la probabilité que LIME apprenne au moins une explication **(Molnar et al., 2019)**. LIME repose également sur des hypothèses simplistes concernant les frontières de décision des modèles d'apprentissage, en supposant qu'elles sont localement linéaires. Cependant, les frontières de décision des modèles tels que les réseaux de neurones peuvent être hautement non linéaires, même localement, et une approximation linéaire dans ce contexte pourrait mener à des explications instables **(Lou et al., 2013).**

Les valeurs de sortie de LIME manquent de signification comparative ; il n'est pas évident de comprendre la signification des valeurs attribuées à chaque caractéristique d'entrée (**Figure 14b**) ou la relation entre ces valeurs et la prédiction du modèle. De plus, l'augmentation du poids linéaire de LIME influence les échantillons non perturbés **(Hamilton et al., 2022)**. Aucune méthode raisonnable n'est capable d'estimer le noyau de pondération ou même de choisir une largeur appropriée.

LIME choisit ensuite des paramètres critiques, tels que le noyau de pondération, la taille du voisinage et le terme de complexité de manière heuristique, ce qui conduit à des comportements incohérents pouvant affecter la fidélité **locale (Lundberg and Lee, 2017 ; Amparore et al., 2021)**.

Des versions déterministes de LIME **(Zafar and Khan, 2021)**, ainsi que des stratégies d'optimisation **(Turner, 2016 ; Lakkaraju et al., 2020)** et basées sur l'apprentissage ont été proposées pour réduire l'instabilité ; cependant, ces alternatives ont le défaut d'augmenter le nombre de paramètres à régler.

3.9. Techniques XAI basées sur les valeurs de Shapley

Dérivées de la modélisation classique de la théorie des jeux, les valeurs de Shapley **(Shapley, 1953)** décrivent une manière de répartir les gains/coûts totaux d'un jeu coopératif entre les joueurs, tout en respectant des critères d'équité. Déterminer les valeurs de Shapley est donc un problème de répartition des coûts. Selon **Friedman and Moulin, (1999),** les problèmes de partage des coûts sont centraux dans plusieurs domaines où il est nécessaire de diviser les coûts communs et d'allouer proportionnellement leurs parts à chaque contributeur individuel.

L'électricité est un exemple de service public dont la production est longue et complexe. Elle commence aux unités de production d'énergie, passe par les transmetteurs et les distributeurs, et atteint enfin le consommateur final. Déterminer combien le consommateur final paiera et répartir équitablement cette valeur entre chaque maillon de la chaîne de production est un problème typique de partage des coûts.

Une valeur de Shapley représente la contribution marginale moyenne d'un joueur évaluée sur toutes les combinaisons possibles de joueurs, c'est-à-dire qu'il s'agit d'une moyenne pondérée des contributions individuelles liées à toutes les compositions possibles d'individus **(Molnar, 2019)**. Un aspect des valeurs de Shapley réside dans leur solide fondement théorique, qui garantit de manière axiomatique une distribution équitable des gains et des coûts entre les participants d'un jeu collaboratif. Selon **Kumar et al. (2020b)**, un jeu collaboratif comprend un ensemble de n joueurs et une fonction caractéristique v, qui associe des sous-ensembles $S \subseteq \{1,\ldots\ldots n\}$, à des valeurs réelles v(S), satisfaisant à la propriété $v(\emptyset) = 0$. La fonction

caractéristique décrit la manière dont le gain final peut être attribué aux joueurs individuels qui coopèrent en tant qu'équipe dans le jeu. Par conséquent, les valeurs de Shapley représentent une méthode de distribution de la valeur totale de la coopération, v({1, . . . , n}), entre n individus.

Considérons v(i) la fonction caractéristique appliquée à l'attribut i (un joueur) d'un sous-ensemble S d'attributs, c'est-à-dire i ∈ S. La valeur de Shapley peut être calculée comme une moyenne pondérée des contributions marginales de l'attribut i concernant chaque sous-ensemble possible d'attributs S ⊆ {1, . . . , n} et le nombre de permutations de S :

$$\emptyset_{\boldsymbol{v}}(\boldsymbol{i}) = \sum_{\boldsymbol{S}\subseteq \boldsymbol{n}(\boldsymbol{i})} \frac{|\boldsymbol{s}|!(\boldsymbol{n}-|\boldsymbol{s}|-\boldsymbol{1})!}{\boldsymbol{n}!} (\boldsymbol{v}(\boldsymbol{s}\cup\{\boldsymbol{i}\}) - \boldsymbol{v}(\boldsymbol{s})) \qquad \textbf{(9)}$$

Où $\emptyset_v(i)$ est la valeur de Shapley du i-ème attribut, v(S) est la valeur attendue de la fonction caractéristique conditionnelle aux sous-ensembles S ⊆ {1, . . . , n), c'est-à-dire la valeur attendue de la fonction caractéristique, $\mathbb{E}[v(S)]$, où n représente le nombre total d'attributs et |.| désigne la cardinalité **(Kumar and Chandran, 2021)**. Notez que v(S∪{i}) − v(S) décrit la contribution marginale d'un joueur à une combinaison de joueurs S, c'est-à-dire la variation de la valeur marginale $\Delta_v(i, S)$ générée lorsque i est inclus dans S **(Kumar et al. 2020b)**.

Dans le contexte de l'apprentissage automatique, les valeurs de Shapley quantifient la valeur modifiée dans la prédiction attendue lorsque les modèles d'apprentissage sont conditionnés par des combinaisons de cet attribut **(Lundberg and Lee, 2017)**. Les modèles appliqués étaient équivalents en termes d'hyperparamètres et de données d'entraînement (l'ensemble de données complet). La différence réside dans la combinaison des attributs inclus dans chaque modèle.

La modélisation de la valeur de Shapley a une longue histoire d'application, **Lipovetsky et Conklin (2001)** ont utilisé les valeurs de Shapley pour analyser l'importance globale des attributs dans les modèles de régression linéaire, et **Štrumbelj et Kononenko (2010)** ont mesuré les effets des caractéristiques dans les tâches de classification.

Plus précisément, une valeur de Shapley attribue une valeur d'importance à un attribut d'entrée, qui représente sa contribution à la prédiction finale lors de son inclusion dans le modèle. Le modèle d'apprentissage automatique en cours d'explication est pris comme fonction caractéristique pour le calcul des attributions d'importance selon **l'équation 9**. Il est appliqué à des sous-ensembles avec et sans l'attribut d'intérêt, ce qui permet d'estimer la contribution marginale de l'attribut, et d'obtenir ainsi une moyenne pondérée des contributions **(Lipovetsky and Conklin, 2001)**. Toutefois, comme il faut calculer toutes les combinaisons d'attributs pour obtenir les valeurs de Shapley de tous les attributs du jeu de données, le coût computationnel augmente de manière exponentielle à mesure que le nombre d'attributs augmente, ce qui entrave les modélisations opérationnelles sur des données de haute dimension **(Molnar, 2019).**

Pour pallier cette limite, **Štrumbelj et Kononenko (2014)** ont développé une version approximative basée sur l'échantillonnage de Monte Carlo. L'estimation de **l'équation 9** repose sur l'échantillonnage et suppose que les prédictions sont générées à partir d'instances (échantillonnées aléatoirement) contenant des permutations d'attributs sélectionnées aléatoirement (à l'exception de l'attribut en cours d'examen**) (Breiman, 2001 ; Casalicchio et al., 2018)**, plutôt que de toutes les n caractéristiques d'entrée originales. Par conséquent, la valeur de Shapley d'un attribut est estimée de manière itérative à partir des échantillons sélectionnés aléatoirement à chaque itération. Les variations des prédictions sont pondérées pour chaque échantillon selon la distribution de probabilité des données, puis le résultat est

calculé en tant que moyenne. La procédure est répétée pour chaque attribut afin d'estimer toutes les valeurs de Shapley **(Molnar, 2019)**.

Lundberg et Lee (2017) ont formulé SHAP (SHapley Additive exPlanations), l'une des approches les plus réussies basées sur les valeurs de Shapley. Il les estime en approchant le modèle d'apprentissage original à travers une fonction d'espérance conditionnelle sur des vecteurs avec une permutation d'attributs simplifiés. Il mesure ensuite le gain ou la perte de chaque prédiction en simulant la présence et l'absence d'attributs en échantillonnant les valeurs de la distribution marginale de chaque attribut. Notons que l'espérance conditionnelle est l'estimateur habituel qui résume la distribution de probabilité dans les applications de prédiction. Cependant, SHAP présume l'indépendance des caractéristiques et utilise une distribution marginale pour remplacer la distribution conditionnelle **(Aas et al., 2021),** ce qui permet à l'approximation de l'espérance conditionnelle d'estimer directement les valeurs de Shapley à l'aide d'un modèle d'attribution additive des caractéristiques.

En raison de la formulation linéaire de LIME, KernelSHAP est en mesure d'estimer les valeurs de Shapley en recourant à des solutions fondées sur la régression, ce qui s'avère être plus efficace sur le plan computationnel que le calcul direct de l'équation classique des valeurs de Shapley. Observation SHAP et LIME (dans leur version originale) offrent des explications distinctes concernant les prédictions. LIME met en évidence la caractéristique la plus significative pour une prédiction, tandis que SHAP précise la contribution de chaque caractéristique à cette même prédiction. Bien que les deux méthodes procèdent à une comparaison des prédictions en se fondant sur une explication dotée d'une probabilité moyenne, la méthode SHAP examine la disparité entre les valeurs prédites et celles attendues par rapport à la prédiction moyenne globale. Parallèlement, LIME clarifie la distinction entre la prédiction et la prédiction moyenne locale, laquelle est obtenue par le biais de l'échantillonnage de voisinage **(Aas et al., 2021).**

Néanmoins, le calcul précis des valeurs de Shapley requiert une quantité significative de ressources. **Aas et al. (2021)** ont souligné son incapacité à traiter les ensembles de données comprenant plus de dix variables. Dans la majorité des situations, seules des solutions approximatives peuvent être mises en œuvre **(Molnar, 2019). Hooker et al. (2018)** ont observé que SHAP manifeste un comportement déterministe lorsqu'il est appliqué à des données de faible dimension. En revanche, dans le cadre de données de haute dimension, il recourt à des méthodes d'échantillonnage statistique fondées sur l'intégration de Monte Carlo, ce qui conduit à des explications susceptibles d'être instables.

VI. Limites du XAI

Quelles sont les caractéristiques d'une explication de qualité ? **Miller (2019)** a caractérisé une explication adéquate comme étant celle qui se révèle véridique dans le monde réel. La réalité en matière d'apprentissage automatique se limite à la « vérité » acquise au cours du processus d'entraînement, ce qui peut dissimuler des biais non identifiés.

Une manière de favoriser la confiance consiste à accroître la transparence des applications intelligentes. Un aspect fondamental de l'augmentation de la transparence réside dans l'implémentation de l'explicabilité **(Amann et al., 2022)**, laquelle a la capacité de déchiffrer les mécanismes des algorithmes d'apprentissage sophistiqués, facilitant ainsi l'intégration de davantage d'éléments de confiance au sein des systèmes d'assistance qui recourent activement à des modèles intelligents. Néanmoins, des échanges substantiels ont eu lieu concernant les

limitations de l'XAI ainsi que les préoccupations précédemment évoquées relatives à l'insuffisance d'évaluation. Bien que cette étude ait abordé divers concepts, besoins, défis et méthodes relatifs à l'explicabilité de l'intelligence artificielle (XAI), **Kaur et al. (2020)** ont démontré que l'ensemble des data scientistes ne possède pas nécessairement les compétences requises pour appliquer de manière appropriée le XAI au sein des pipelines d'apprentissage automatique.

Krishna et al. (2022) ont examiné la fréquence à laquelle les explications générées par les méthodes de pointe présentent des divergences. Le phénomène de désaccord peut être attribué à une absence d'objectifs communs au sein des différentes méthodes d'explicabilité **(Han et al., 2022)**. Les auteurs ont également réalisé une étude auprès des data scientistes concernant les solutions envisageables face à de tels désaccords dans les explications.

Han et al. (2022) ont consolidé les méthodes couramment utilisées pour évaluer l'importance des caractéristiques locales au sein d'un cadre unifié, et ont mis en évidence qu'aucune de ces méthodes n'était en mesure de produire des explications optimales pour l'ensemble des sous-ensembles de données. Leurs résultats ont révélé que des divergences peuvent se manifester en raison du fait que divers explicateurs abordent le modèle en recourant à des quartiers et à des fonctions de perte distincts. Les auteurs ont également formulé des recommandations pour sélectionner les méthodes d'explicabilité de l'intelligence artificielle (XAI) en fonction de leur conformité au modèle.

Les techniques de permutation constituent l'une des principales approches en matière d'explicabilité des modèles d'intelligence artificielle (XAI) et se caractérisent par leur simplicité de description, de développement et d'utilisation. Ils présentent des résultats séduisants en générant des « caractéristiques nulles », ce qui rompt le lien entre les caractéristiques et les variables cibles **(Breiman, 2001 ; Casalicchio, 2018)**. Bien que les méthodes de permutation puissent s'avérer efficaces en présence d'un nul global, elles peuvent ne pas offrir d'explications précises dans des situations qui s'écartent de ce cadre **(Barber and Candès, 2015)**.

Hooker et al. (2021) ont mis en évidence la facilité avec laquelle il est possible de produire des exemples dans lesquels les explications fondées sur les permutations peuvent s'avérer trompeuses ou biaisées.

L'analyse des relations causales a rarement été examinée dans la littérature relative à l'Explicabilité des Systèmes d'Intelligence Artificielle (XAI). L'exploration de la causalité dans le domaine de l'apprentissage automatique s'avère être une tâche complexe, bien qu'elle soit reconnue comme un objectif majeur en matière d'explicabilité **(Bhatt et al., 2021)**.

Les variables non numériques constituent une autre contrainte qui peut être perçue comme un défi pour les méthodes d'explicabilité de l'intelligence artificielle (XAI), et la gestion adéquate des variables catégorielles représente également une difficulté dans le cadre des algorithmes d'apprentissage. La solution couramment employée dans ce contexte est l'encodage one-hot, une méthode élémentaire issue des circuits numériques, qui convertit les caractéristiques non numériques en matrices binaires. Cependant, dans le cadre des vastes ensembles de données comportant de nombreux attributs catégoriels, chacun présentant un nombre élevé de catégories distinctes, l'encodage one-hot entraîne une augmentation significative du degré de parcimonie des données. Cette situation, à son tour, accroît la dimensionnalité des données, la majorité des valeurs encodées étant intégrées sous forme de colonnes additionnelles, dont l'importance individuelle est souvent limitée. Une alternative à

l'encodage one-hot est l'encodage par cible **(Banachewicz et al., 2022)**, qui transforme chaque valeur d'un attribut catégorique en sa valeur attendue associée. La transformation obtenue n'introduit pas de colonnes additionnelles, ce qui permet d'éviter la conversion du jeu de données en un ensemble à haute dimension plus clairsemé. **Aas et al. (2021)** ont proposé des approches alternatives tirées de la littérature sur le clustering, lesquelles décrivent des fonctions de distribution permettant de traiter des données non numériques (Huang, 1998), ainsi que des généralisations de la distance de Mahalanobis applicables à des mélanges d'attributs nominaux, ordonnés et continus **(De Leon and Carrière, 2005).**

VII. Conclusion

Ce chapitre nous a permis d'identifier clairement les notions importantes relatives à l'IA, l'apprentissage automatique, l'apprentissage profond, l'interprétabilité et à l'explicabilité (XAI) des modèles de ML, ainsi que de leur importance dans la quête d'un déploiement plus important de l'IA, dans différents domaines, et ce par le rétablissement de la confiance entre l'IA et l'être humain.

Aussi, nous avons pu dresser un paysage assez large et exhaustif des approches et des méthodes relatives à l'explicabilité des modèles de MLn permettant d'expliquer des modèles considérés jusqu'à lors comme des **« Boîtes noires »,** dans le but d'aller vers une IA explicable et responsable, répondant aux attentes des différentes parties prenantes.

Le chapitre suivant nous permettra de présenter nos expérimentations, travaux et résultats relatifs au XAI ; nous discuterons ces derniers par rapport aux travaux similaires dans la littérature.

CHAPITRE II

PALUDISME

CHAPITRE II

Paludisme

Sommaire

I. Introduction

La malaria, surnommée fléau des marais, est une maladie infectieuse propagée par des parasites du genre Plasmodium, véhiculés par les piqûres des moustiques femelles appartenant au genre Anophèles.

La malaria, redoutable fléau, sévit principalement dans les contrées tropicales, menaçant la vie de milliers, voire de millions de personnes qui lui sont exposées. Cette parasitose, bien que curable, peut se métamorphoser en une version plus sinistre si elle n'est pas détectée promptement et soignée efficacement : le paludisme simple peut alors se transformer en une variante mortelle si elle reste sans traitement.

II. Données épidémiologiques mondiales

D'après le récent rapport de l'Organisation Mondiale de la Santé concernant la situation du paludisme à l'échelle mondiale, le nombre de cas de paludisme s'élevait à 263 millions en 2023, en comparaison avec les 249 millions enregistrés en 2022. En 2023, le nombre présumé de décès attribuables au paludisme s'élevait à 597 000, contre 608 000 en 2022 **(Figure 17)**. Selon l'OMS en 2024, près de la moitié de la population mondiale est exposée au risque de paludisme.

La région africaine demeure confrontée à une part significative et disproportionnée du fardeau mondial de la malaria. En 2023, près de 94 % des cas de paludisme et 95 % des décès liés à cette maladie ont été recensés dans cette zone géographique ; les enfants âgés de moins de cinq ans constituaient environ 76 % des décès attribués à ce fléau.

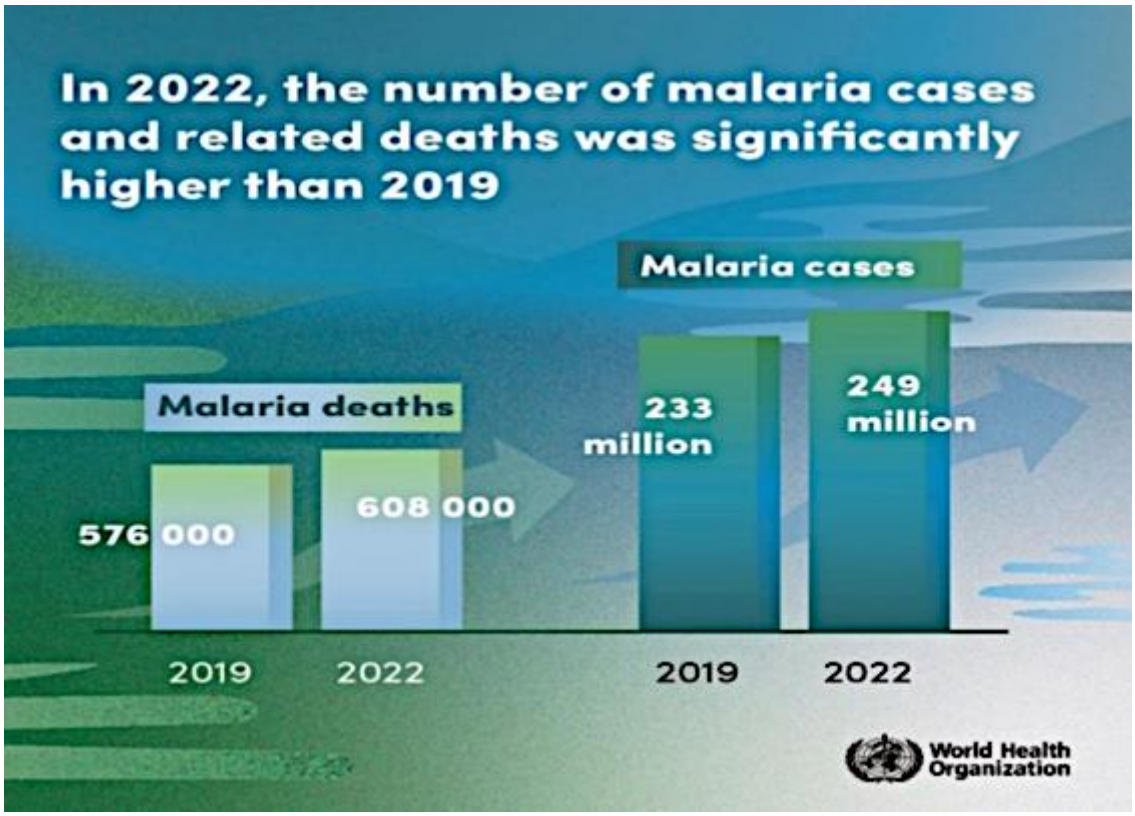


Figure 17. Statistiques mondiales sur le paludisme **(OMS, 2023)**

Certains individus présentent une prédisposition plus élevée à développer une forme sévère de malaria par rapport à d'autres. Les nourrissons, les enfants de moins de cinq ans, les femmes enceintes et les individus vivant avec le VIH/sida sont particulièrement vulnérables. Parmi les populations vulnérables, on peut citer les individus se déplaçant vers des régions à forte transmission, sans avoir acquis une immunité partielle, après une exposition prolongée ou sans suivre de traitement chimio-préventif, tels que les migrants, les populations nomades et les voyageurs **(OMS, 2023).**

- En 2022, la Région a enregistré 94 % des cas de paludisme (233 millions) et 95 % des décès liés à cette maladie (580 000).
- Les décès liés au paludisme dans la région étaient majoritairement attribuables aux enfants âgés de moins de cinq ans, représentant 80 % des cas.
- En 2024,
    - 2,5 milliards de personnes y ont été exposées ;
    - 247 millions de cas et 619 000 décès ;
    - En Afrique, 481 600 enfants de moins de 5 ans sont concernés ;
    - Répartis dans 85 pays ou territoires touchés.

Selon les statistiques de l'OMS, 90 % des cas de malaria se manifestent dans les régions tropicales de l'Afrique subsaharienne. Cependant, il est envisageable d'acquérir la maladie dans des pays spécifiques d'Asie du sud-est, d'Amérique du Sud et du Moyen-Orient **(Figure 18)**. Dans les pays

européens, les cas de paludisme sont principalement des cas d'importation, concernant des voyageurs ayant séjourné dans des régions où la maladie est endémique.

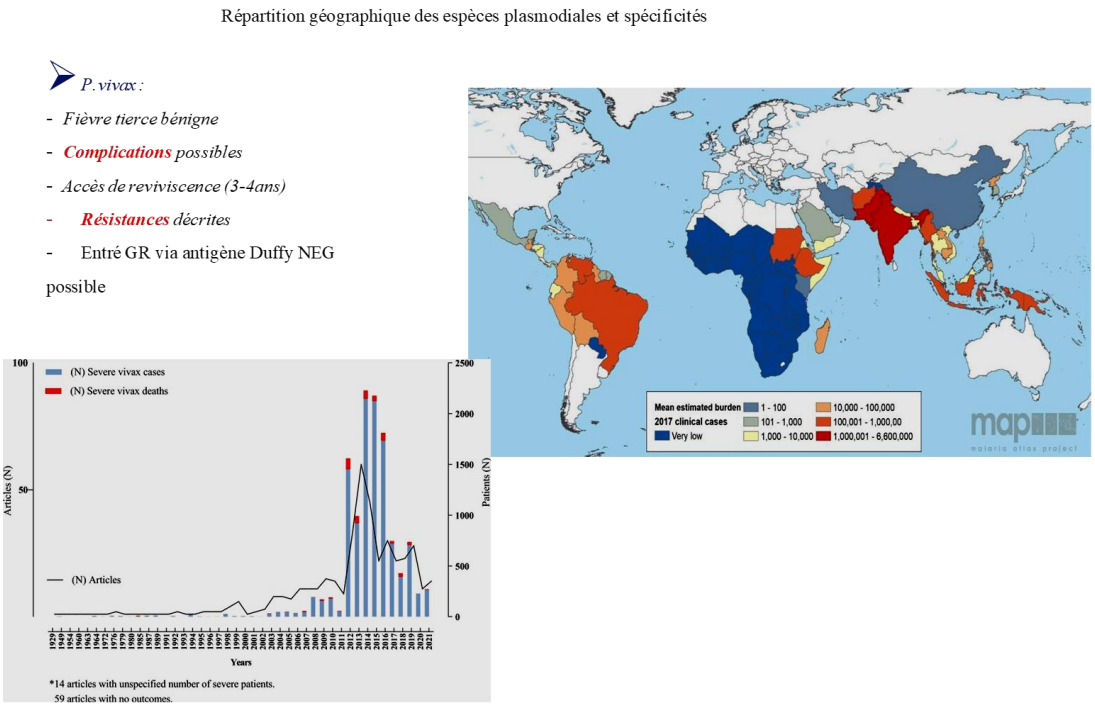


Figure 18. Répartition géographique des espèces plasmodiales et spécificités **(OMS, 2023)**

- ***P. falciparum*** (le plus répandu, **forme clinique mortelle**)
  - **Neuropaludisme+++**
  - Transmission > 18°C
  - Rechutes tardives rares
  - **90% accès dans les 2 mois**

En Algérie, le 27 septembre 2024, le ministère de la Santé a diffusé un communiqué de presse déclarant des mesures d'urgence appliquées dans les wilayas du sud du pays (Tamanrasset, In Guezzam et Bordj Badji Mokhtar) en réponse à la dégradation des conditions sanitaires de ces régions.

D'après le communiqué, des cas de malaria importés ont été signalés dans certaines wilayas du sud. Le ministère a confirmé que les personnes infectées ont été prises en charge en respectant les protocoles médicaux établis, soulignant que la situation épidémiologique est surveillée de manière continue au niveau central et local.

Le communiqué indique que le ministère a constaté une augmentation significative des cas graves en raison de la propagation rapide de l'infection et du déficit en personnel médical indispensable pour endiguer la progression de la maladie.

Devant cette situation, le ministère de la santé a déclaré que toutes les provinces du sud ont été désignées comme zones sinistrées. Par conséquent, l'accès à ces zones sera restreint, avec la fermeture de toutes les voies d'accès, et il sera formellement prohibé aux résidents d'y pénétrer ou d'en sortir pendant une durée de deux mois à partir de la date de la proclamation officielle **(Flutrackers, Le Matin d'Algérie, 2024)**.

III. Agents pathogènes

1. Parasites responsables du paludisme: les Plasmodium

Quatre espèces de *Plasmodium* sont impliquées dans la propagation du paludisme, comme indiqué dans le **Tableau 01**. *Plasmodium falciparum*, illustré dans la **Figure 19**, est l'espèce la plus pathogène, responsable des cas mortels de paludisme, est prédominante en Afrique. *Plasmodium vivax,* en coexistence avec *P. falciparum,* a été identifié dans des zones tempérées. En Afrique de l'Ouest, on peut observer la présence de *P. ovale*, un parasite qui n'est pas mortel mais qui entraîne des rechutes environ quatre à cinq ans après la première infection. Finalement, à l'échelle mondiale (avec une répartition inégale), on observe la présence de *P. Malariae* qui, bien qu'il ne soit pas mortel, comporte un risque de rechute pouvant survenir jusqu'à vingt ans après la première infection. Les rechutes sont associées à la présence persistante du parasite dans le foie, à l'état latent (hypnozoïte). Les cas de transmission directe de cette maladie entre individus sont peu fréquents. Ils peuvent se manifester, par exemple, lorsqu'il y a un partage de seringues contaminées ou par transmission transplacentaire pendant la grossesse **(Garrido-Cardenas et al., 2019).**

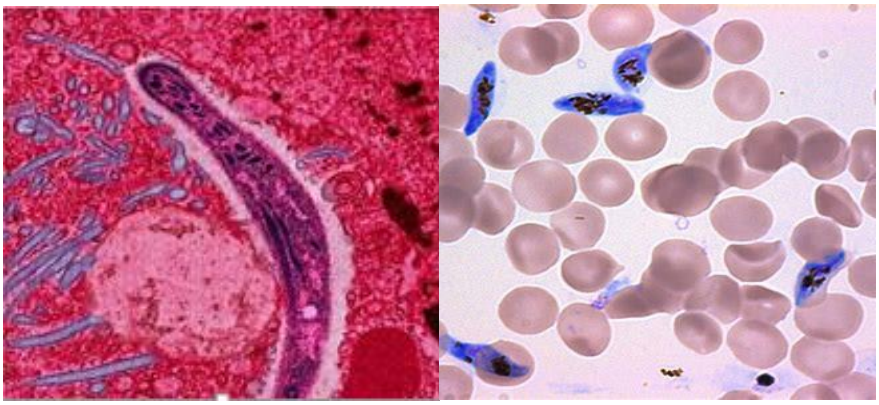
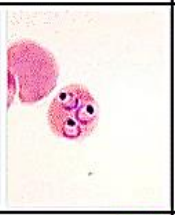
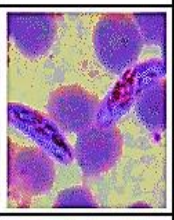
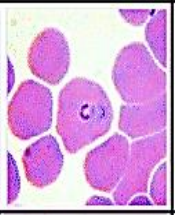
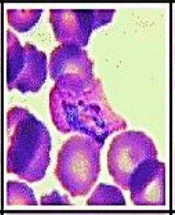
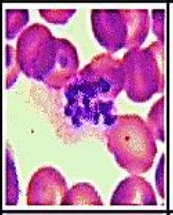
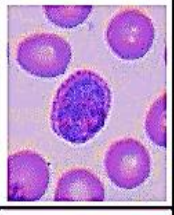
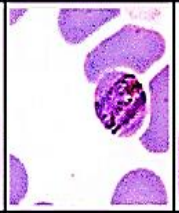
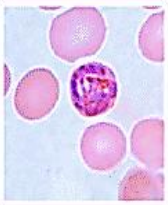
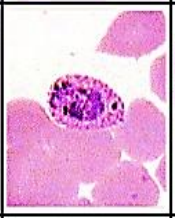
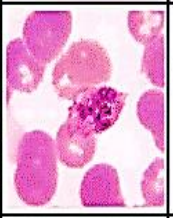
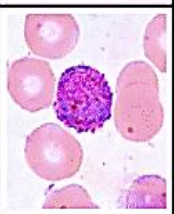


Figure 19. *Plasmodium falciparum observé au* microscope **(Garrido-Cardenas et al., 2019)**

Tableau 1. Principales espèces de plasmodium responsables de la malaria **(Poostchi et al., 2018)**

Human Malaria					
Stages / Species	Ring	Trophozoite	Schizont	Gametocyte	
P. falciparum					• *Parasitised red cells (pRBCs) not enlarged.* • *RBCs containing mature trophozoites sequestered in deep vessels.* • *Total parasite biomass = circulating parasites + sequestered parasites.*
P. vivax					• *Parasites prefer young red cells* • *pRBCs enlarged.* • *Trophozoites are amoeboid in shape.* • *All stages present in peripheral blood.*
P. malariae					• *Parasites prefer old red cells.* • *pRBCs not enlarged.* • *Trophozoites tend to have a band shape.* • *All stages present in peripheral blood*
P. ovale					• *pRBCs slightly enlarged and have an oval shape, with tufted ends.* • *All stages present in peripheral blood.*

1.1. Cycle de vie du plasmodium

Lorsque le parasite est injecté par le moustique *Anopheles*, il est au stade de sporozoïte. Ensuite, il migre rapidement vers le foie par le biais de la circulation sanguine. Il pénètre dans les cellules hépatiques et amorce sa division (**Figure 20**). En quelques jours, des dizaines de milliers de nouveaux parasites apparaissent. Ce sont des mérozoïtes, qui sont finalement libérés dans le sang. Ils colonisent les globules rouges et poursuivent leur multiplication jusqu'à l'éclatement des globules (**Figure 21**) **(Scholte et al., 2006).**

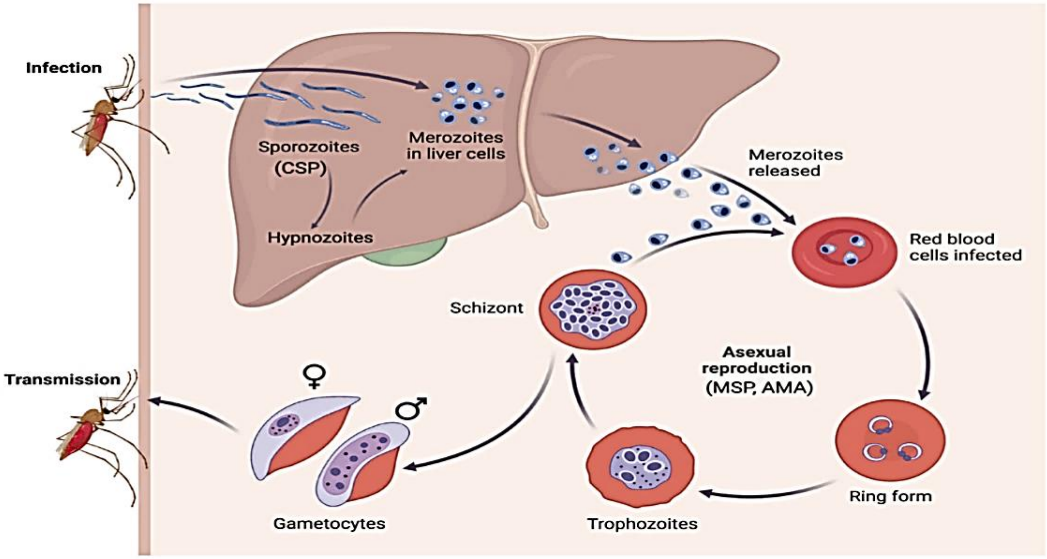


Figure 20. Cycle évolutif du plasmodium **(Scholte et al., 2006)**

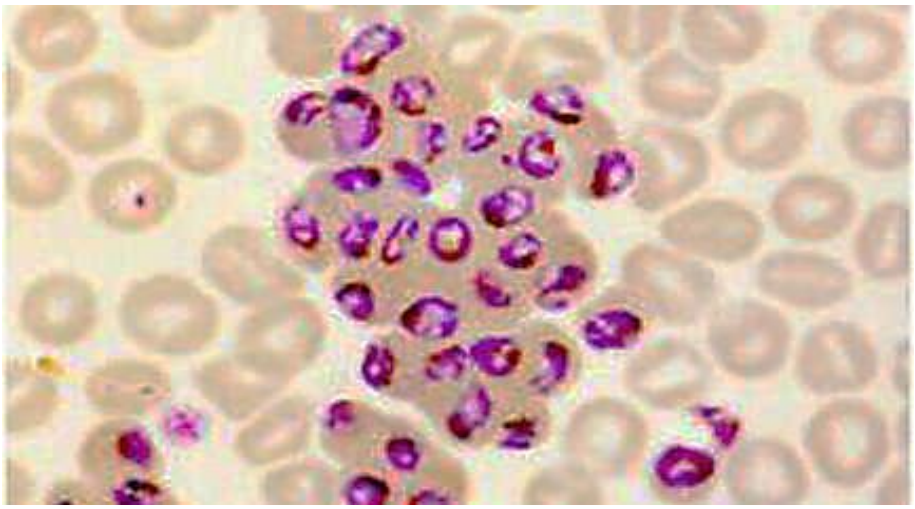


Figure 21. Globules rouges infectés **(Scholte et al., 2006)**

2. Moustiques vecteurs du plasmodium

La transmission du paludisme est assurée uniquement par les moustiques du genre Anopheles (**Figure 22**). Parmi la diversité des espèces d'anophèles, seules une cinquantaine sont actuellement impliquées dans la transmission du paludisme, dont 20 sont responsables de la majorité des cas de transmission à l'échelle mondiale. La variabilité des comportements, observée entre les différentes espèces d'anophèles, ainsi qu'au sein d'une même espèce, est influencée par des facteurs tels que les conditions climatiques, géographiques et l'impact de l'activité humaine sur l'environnement. Ces éléments déterminent le degré de contact entre l'homme et le vecteur, ainsi que les divers profils épidémiologiques du paludisme. Les anophèles sont principalement des moustiques présents en milieu rural et sont théoriquement moins fréquents en milieu urbain. En pratique, la capacité d'adaptation de certaines espèces à l'environnement urbain et la culture maraîchère en milieu urbain, ou en périphérie des grandes agglomérations, contribuent à maintenir des populations d'anophèles en milieu urbain. On observe une variation significative du risque à l'intérieur d'un pays, d'une région ou même à une

distance de quelques kilomètres seulement. La transmission du virus connaît des variations saisonnières et interannuelles en fonction du niveau des événements climatiques, comme indiqué par **Scholte et al. (2008)**.

Figure 22. Anophèle femelle vectrice du plasmodium **Scholte et al. (2008)**

2.1. Cycle de vie de l'anophèle

Le cycle de vie du moustique comprend deux phases distinctes : une phase aquatique d'environ dix jours, comprenant les stades juvéniles, l'œuf, la larve et la nymphe, suivie d'une phase aérienne où le moustique atteint l'âge adulte. La durée de vie moyenne de l'anophèle femelle adulte est d'environ un mois **(Pagès et al., 2007)**.

Le cycle du paludisme met en jeu deux hôtes principaux : le moustique anophèle femelle et l'homme. Les différentes phases du cycle sont présentées dans la **Figure 23** :

1. Le moustique femelle se nourrit du sang d'une personne infectée par le paludisme, ingérant ainsi les gamétocytes du parasite.
2. Les parasites se reproduisent et se déplacent vers les glandes salivaires du moustique.
3. Avant de se nourrir de sang, le moustique introduit des sporozoïtes dans l'organisme humain d'un autre hôte.
4. Les sporozoïtes se dirigent vers le foie où ils subissent une division cellulaire.

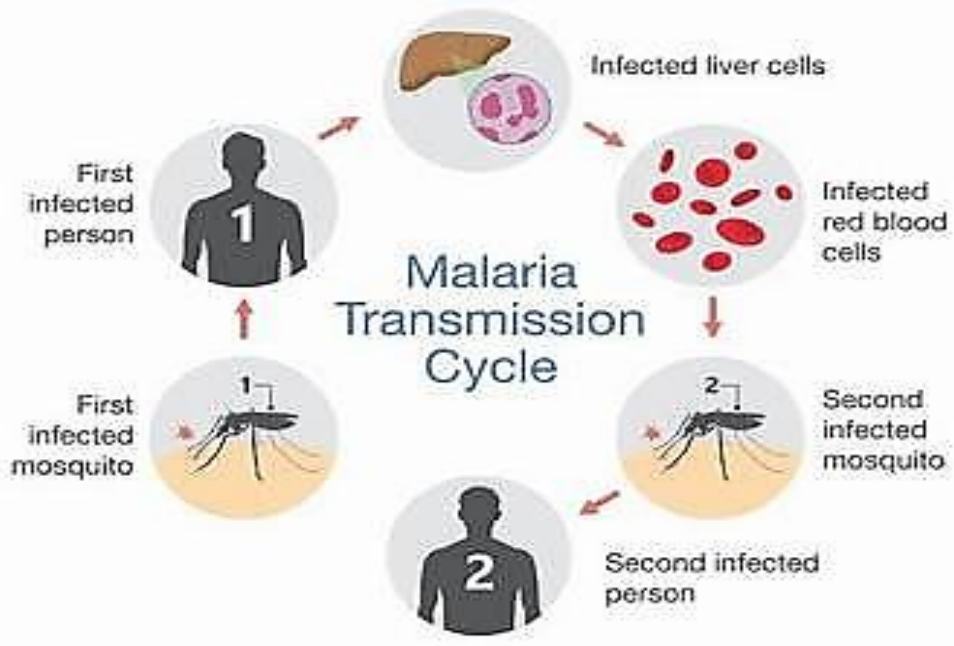


Figure 23. Cycle de transmission du paludisme **(Pagès et al., 2007)**

3. Symptômes du paludisme et complications

Les parasites responsables du paludisme suivent un cycle spécifique qui détermine l'apparition des symptômes. Dans les situations les plus sévères et chez les personnes vulnérables, la malaria peut provoquer des complications graves, voire le décès **(Coetzee et Fontenille, 2004).** Les symptômes du paludisme présentent une grande diversité clinique et peuvent être assimilés à ceux de la grippe. Entre huit et trente jours après l'infection, la fièvre apparaît comme le premier symptôme de la maladie. Cette dernière peut s'accompagner d'autres symptômes qui ne sont pas systématiques tels que des céphalées, des douleurs musculaires, une asthénie, des nausées et des vomissements, des diarrhées, une toux... Les cycles se caractérisent par des périodes de fièvre suivies de phases de tremblements, de sueurs froides, de transpiration abondante et de somnolence. Il est question d'accès palustre. La fréquence de ces cycles est déterminée par le parasite impliqué et correspond à sa reproduction ainsi qu'à la rupture des globules rouges. Ce phénomène entraîne un risque de développement d'anémie et d'obstruction des vaisseaux sanguins qui alimentent le cerveau, connu sous le nom de neuropaludisme **(Scholte et al., 2008).**

En général, les décès liés au paludisme surviennent principalement en raison de complications. En présence de neuropaludisme, il est possible de constater une altération cérébrale se manifestant par des symptômes tels que le délire, la perte de conscience, le coma, voire le décès du sujet infecté. Parmi d'autres complications éventuelles, on peut mentionner :

- les problèmes respiratoires et l'œdème pulmonaire ;
- l'insuffisance rénale et hépatique ;
- la rupture de la rate ;
- l'anémie sévère ;

- la chute du taux de sucre dans le sang (hypoglycémie).

Les catégories de personnes les plus susceptibles de développer des complications comprennent les femmes enceintes (risque de fausse couche et de retard de croissance intra-utérin), les jeunes enfants, les personnes âgées, les individus souffrant de maladies chroniques et les patients immunodéprimés (notamment ceux ayant subi une greffe ou atteints du VIH). En règle générale, les premiers signes du paludisme se manifestent entre 10 et 15 jours après l'inoculation du parasite par la piqûre d'un moustique infecté. La fièvre, les céphalées et les frissons sont habituellement observés, bien que ces manifestations puissent être légères et peu spécifiques pour le diagnostic du paludisme. Dans les régions endémiques, les individus qui ont acquis une immunité partielle peuvent être infectés sans manifester de symptômes (infections asymptomatiques).

Il est recommandé par l'OMS de procéder à un diagnostic rapide des cas suspects de malaria. En l'absence de traitement du paludisme à *Plasmodium falciparum* dans les 24 heures suivant l'infection, il existe un risque d'évolution vers une forme sévère, potentiellement mortelle. Le paludisme sévère peut entraîner une défaillance de plusieurs organes chez les adultes, tandis que chez les enfants, il se caractérise fréquemment par une anémie sévère, une insuffisance respiratoire ou du neuropaludisme. Le paludisme chez l'homme, provoqué par des espèces de Plasmodium différentes, peut conduire à une forme sévère et potentiellement mortelle de la maladie **(Carnevale et al., 1993)**.

4. Diagnostic du paludisme

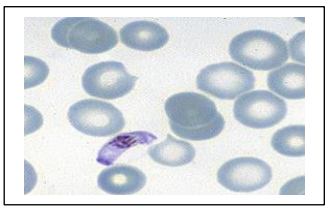
Il est recommandé par l'Organisation mondiale de la Santé d'effectuer un diagnostic rapide du paludisme, que ce soit par microscopie optique ou par des tests de diagnostic rapide (TDR), pour tous les cas suspects de cette parasitose, avant d'initier un traitement. Il est impératif d'établir un diagnostic précoce et précis afin d'assurer une gestion efficace de la maladie et une surveillance rigoureuse de la malaria **(Girod et al., 2006)**.

Le diagnostic du paludisme repose sur diverses techniques telles que la microscopie, les tests immunochromatographiques et la biologie moléculaire (**Figure 24**). Un résultat négatif d'un frottis ne permet pas d'exclure le diagnostic de paludisme. Afin d'obtenir un résultat négatif pour le paludisme lors d'une recherche, il est essentiel de procéder à la réalisation d'une goutte épaisse, d'un frottis ou d'une technique de biologie moléculaire rapide. En cas de non-réalisation d'une de ces techniques au sein du laboratoire, le prélèvement doit être envoyé à un centre spécialisé. Les tests diagnostiques qui se fondent sur la détection de parasites contribuent de manière significative à la réduction de la morbidité et de la mortalité en permettant aux professionnels de santé de différencier rapidement les fièvres d'origine palustre des autres, et de sélectionner le traitement le plus adéquat. Ils contribuent à améliorer la gestion globale des patients souffrant d'une maladie fébrile et peuvent également aider à limiter l'émergence et la propagation de la résistance aux médicaments **(Pilly, 2018).**

➢ **Différents outils diagnostiques**

- ***Microscopie***

- ***Immunochromatographie***



- **PCR**

4.1. Diagnostic biologique

Le diagnostic biologique du paludisme repose sur la mise en évidence du parasite dans le sang (**Figure 24**). Il s'agit d'un diagnostic d'urgence qui doit être fait dans les deux heures qui suivent le prélèvement sanguin **(Pilly, 2018)**.

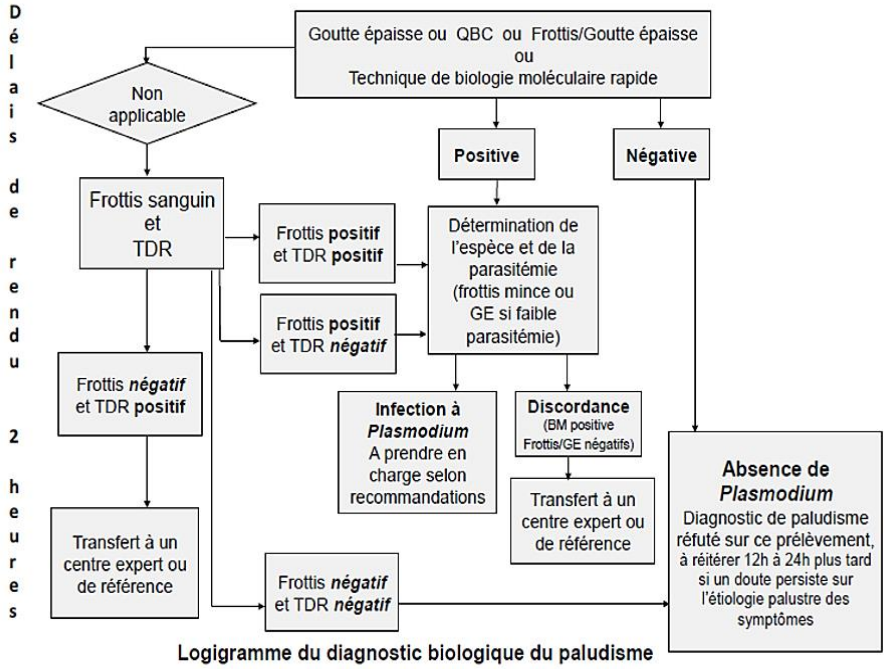


Figure 24. Logigramme du diagnostic du paludisme **(Pilly, 2018)**

1.4.1. Examen sanguin au microscope (frottis-goutte épaisse)

Les examens microscopiques sont utilisés pour identifier divers parasites responsables du paludisme (*P. falciparum*, *P. vivax*, *P. malariae* et *P. ovale*), ainsi que pour observer les différents stades parasitaires, y compris les gamétocytes et pour mesurer la densité parasitaire, afin de surveiller la réponse au traitement. Les analyses microscopiques constituent la méthode privilégiée pour investiguer les causes des échecs thérapeutiques. Le Giemsa représente la technique de coloration traditionnelle, utilisée dans les analyses microscopiques, pour détecter la malaria. Le processus diagnostique implique l'examen à la fois d'un échantillon mince (**Figure 25**) et d'un échantillon épais (**Figure 26**) provenant du même patient.

a) **Frottis** :

- Sensibilité : **100-300 parasites / µL**
- Minimum 20 minutes sur 100/200 champs
- Identification d'espèce(s)

Figure 25. Frottis sanguin mince coloré au Giemsa **(Pilly, 2018)**

- Parasitémie (critère de gravité+++)

b) Goutte épaisse :

- Sensibilité : **10-20 parasites /µL**
- Difficulté de lecture
- Microscopiste expérimenté

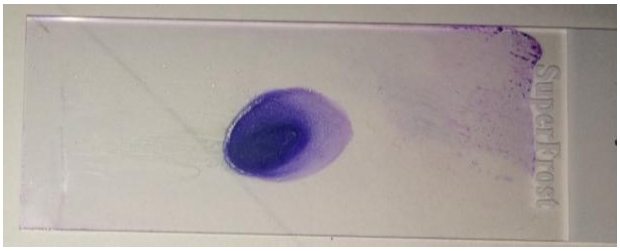


Figure 26. Goutte de sang épaisse colorée au Giemsa **(Pilly, 2018)**

e) Types d'examens microscopiques

La méthode de diagnostic la moins chère, la plus fiable et la plus répandue est l'examen au microscope optique d'un frottis sanguin et d'une goutte épaisse de sang (**Figures 27 et 28**).

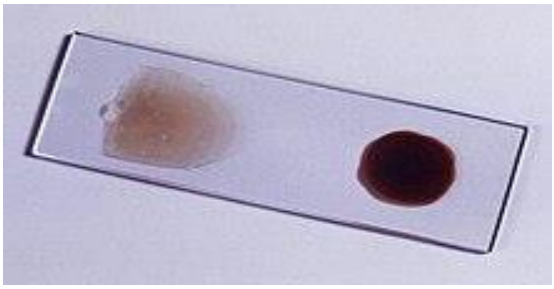


Figure 27. Plaquette de verre, avec une goutte fine (frottis) et une goutte épaisse de sang, prête à être examinée au microscope **(Pilly, 2018)**

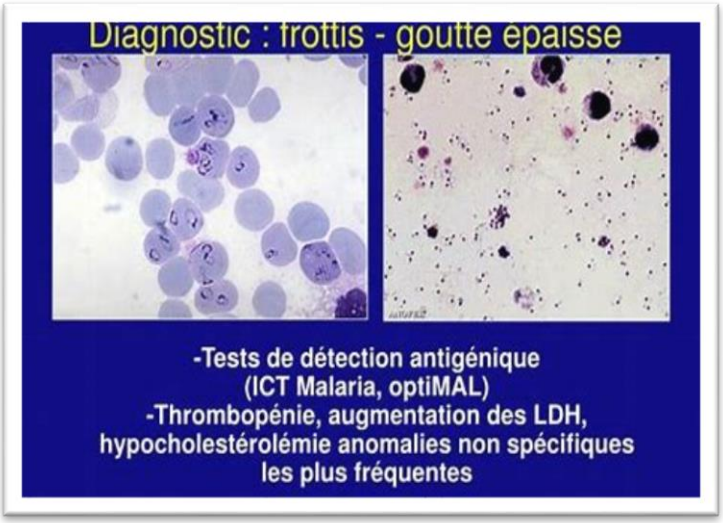


Figure 28. Observation des plasmodiums par microscopie sur frottis et goutte épaisse **(OMS, 2023)**

Le frottis est un outil permettant de mettre en évidence les caractéristiques distinctives de chacune des quatre espèces du parasite (**Figure 29**), car celles-ci sont mieux préservées lors de ce prélèvement. La goutte de sang épais facilite l'exploration d'un volume sanguin plus important afin d'assurer un diagnostic précis et de ne pas manquer la présence de Plasmodium (**Figure 30**). Ces examens doivent être effectués par un biologiste qualifié et expérimenté, comme indiqué par **Pilly (2018)**.

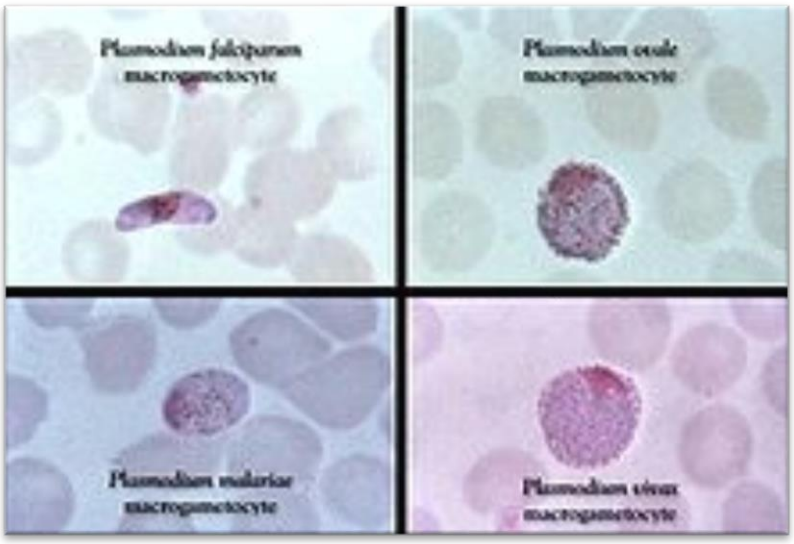


Figure 29. Macrogamétocytes (gamétocytes femelles) identifiés par goutte fine (**Pilly, 2018)**

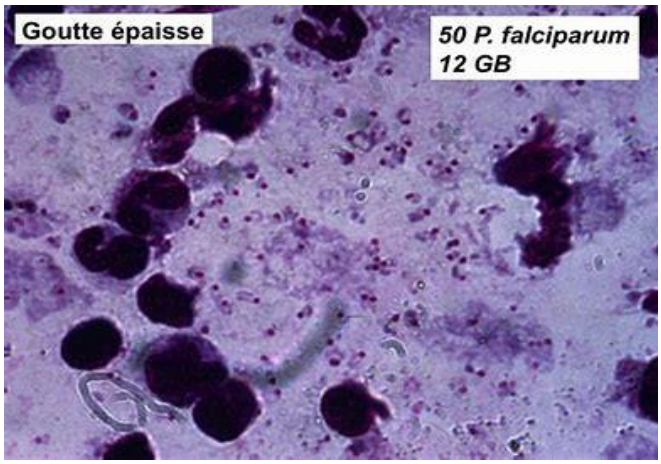


Figure 30. Plasmodium observés par une goutte épaisse (**Pilly, 2018)**

Dans la **Figure** 31, nous pouvons observer plusieurs globules rouges comprennent des anneaux. Vers le centre, une schizonte est visible, et un trophozoïte à gauche (**Pilly, 2018).** De même la **Figure 32**, nous montre des gamétocytes de *P. falciparum* dans les globules rouges observés par microscope (**Pilly, 2018)**.

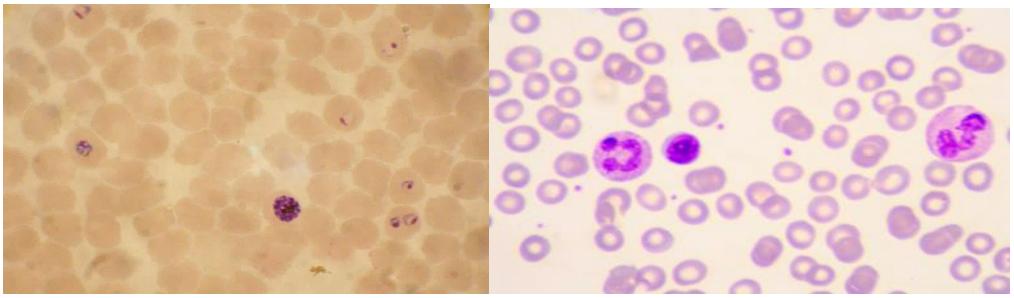


Figure 31: Frottis sanguin d'une culture de *P. falciparum* (**Pilly, 2018)**

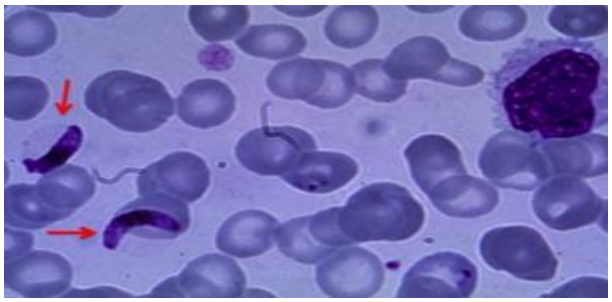


Figure 32. Gamétocytes de *P. falciparum* dans les globules rouges observés au microscope (**Pilly, 2018)**

- Trois éléments sont essentiels dans le diagnostic du paludisme **(Pilly, 2018)** :
    - Confirmation de la présence du *Plasmodium* ;
    - Identification de l'espèce parasitaire, (Ex. *P. falciparum),* pour évaluer la gravité potentielle de l'infection ;
    - Mesure de la parasitémie, exprimée en pourcentage d'hématies parasitées ou en nombre de parasites par microlitre de sang, afin de suivre l'efficacité du traitement.
- Contrôle de qualité d'un diagnostic microscopique :

La prédiction négative des deux méthodes combinées (frottis et goutte épaisse) n'atteint pas 100 %. Un résultat négatif d'un examen doit être répété dans un délai de 12 à 24 heures en cas de persistance de la suspicion clinique. La précision du diagnostic du paludisme peut être compromise en raison de la qualité souvent insatisfaisante de l'examen microscopique, ce qui affecte la sensibilité et la spécificité des résultats.

1.4.2. Tests de diagnostic rapide (TDR)

En l'absence d'un microscope, il est envisageable de recourir à des tests de détection rapide d'antigènes, qui requièrent uniquement une petite quantité de sang et ne demandent pas de compétences spécialisées.

Les tests de diagnostic rapide **(TDR)** du paludisme identifient des antigènes particuliers, c'est-à-dire des protéines, qui sont générés par les plasmodies se trouvant dans la circulation sanguine des individus infectés. Certains tests de diagnostic rapide sont capables de détecter des infections causées par une seule espèce de parasite (soit *P. falciparum*, soit *P. vivax*), tandis que d'autres sont capables de détecter des infections mixtes impliquant plusieurs espèces telles que *P. falciparum, P. vivax, P. malariae* et *P. ovale*. En outre, certains tests permettent de différencier les infections causées par *P. falciparum* des infections causées par d'autres espèces de parasites. Le prélèvement sanguin est généralement effectué par une piqure digitale et les résultats peuvent être obtenus en 15 à 30 minutes. Bien que les quelques 200 produits de tests de diagnostic rapide (TDR) disponibles sur le marché puissent présenter des variations, les principes des tests demeurent similaires **(Pilly, 2018)**.

Les tests immunochromatographiques, connus également sous le nom de tests de diagnostic rapide du paludisme (TDR), peuvent adopter la configuration d'une bandelette réactive ou d'un "dipstick" (**Figure 33**).

La sensibilité des tests employés sur 200 produits commerciaux est de plus de 95 % lorsque la parasitémie dépasse 100 par µL de sang, et de 70 % pour des parasitémies plus faibles, et encore moindre pour *P. ovale*. Leur spécificité se situe entre 90 et 95 %. Le **Tableau 2** résume les antigènes détectés par les TDR.

Dans les pays développés, ces tests rapides sont généralement utilisés en conjonction avec la technique du frottis-goutte épaisse.

Tableau 2. Les différents antigènes détectés par les TDR

Antigène(s) détecté(s)	Sous types	Espèce(s) cibles identifiées(s)
HRP-2	-	P. falciparum
LDH	Pf pLDH	P. falciparum
	Pv pLDH	P. vivax
	Pan pLDH	toutes
Aldolase	Pv aldolase	P. vivax
	Pan aldolase	toutes

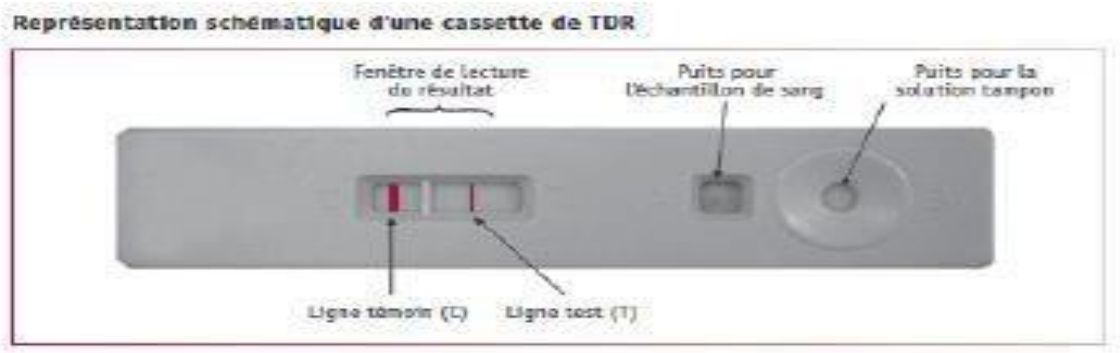


Figure 33. Test de Diagnostic Rapide

a) ***Inconvénients des TDR :***

L'apparition de souches de *P. falciparum* avec une délétion des gènes hrp2 et/ou hrp3 entraîne l'absence de production de la protéine HRPII par le parasite, qui est une des cibles majeures des tests de diagnostic rapide, les plus couramment utilisés pour détecter le paludisme. Par conséquent, les patients atteints par une souche de Plasmodium falciparum qui ne produit pas de HRP auront un résultat négatif au test de diagnostic rapide HRPII. En conséquence, ils pourraient être catégorisés comme non atteints de paludisme et pourraient ne pas bénéficier d'un traitement antipaludique adéquat, augmentant ainsi le risque de complications graves, de décès et favorisant la dissémination du parasite.

4.2 Biologie moléculaire

Cette approche repose sur l'amplification des acides nucléiques des parasites à l'aide de la réaction de polymérisation en chaîne PCR, une méthode plus sensible et spécifique que l'observation microscopique. Elle facilite la détection de parasitémies très faibles et l'identification précise des espèces de Plasmodium **(Pilly, 2018)**.

Dans les nations les plus développées, la PCR est en voie de devenir la technique privilégiée pour le diagnostic de recours, notamment dans les cas de difficultés diagnostiques. Des techniques d'analyse plus rapides telles que la PCR en temps réel, permettant d'obtenir des résultats en moins d'une heure, sont accessibles dans les laboratoires de pointe et pourraient être adaptées à des fins de diagnostic d'urgence ou de routine.

PCR :

- **Excellente spécificité et sensibilité** (0,005 – 1 parasite/µL)
- Délai de réalisation long / coût important
- Technique de référence pour le diagnostic d'espèce **(coinfection +++)**
- Diagnostic rétrospectif chez un patient prémuni ou traité

LAMP :

- Temps réduit (50mn) / **simplicité**
- Résultat uniquement **qualitatif,** diagnostic de genre
- **Excellente VPN** +++ Sensibilité (2 parasites /µl pour *P. falciparum* à 0,12/µl pour *P. vivax*)

Inconvénients des PCR et LAMP :

- Pas de diagnostic d'espèce ;
- Pas de distinction gamétocyte/schizonte/trophozoïte (gravité de l'infection?)
- **Pas de quantification** (parasitémie indispensable)
- Si positif : infection ancienne/aiguë ?
- Coût

Une PCR négative = Diagnostic d'exclusion

Une PCR positive & microscopie négative = Rechercher les autres causes possibles de fièvre

4.3 Bilan de gravité

L'examen biologique standard permet d'évaluer le degré de gravité d'une pathologie et de contribuer au diagnostic différentiel, notamment à travers l'analyse de paramètres tels que l'hémogramme, la protéine C-réactive, l'ionogramme, ainsi que les marqueurs biologiques d'hémolyse, de dysfonction hépatique ou rénale. L'analyse sanguine peut être employée afin de détecter d'autres infections, notamment d'origine virale, en tenant compte du contexte et en se conformant aux directives établies **(Vignier, 2019)**.

Une co-infection avec le paludisme reste une éventualité, cependant, la détection du paludisme est prioritaire en raison de sa prévalence plus élevée et de sa gravité immédiate potentielle **(Vignier, 2019)**.

IV. Antipaludéens : mécanisme d'action et indications

Diverses molécules antipaludiques peuvent être employées à des fins préventives à court terme, notamment lors de voyages en zones endémiques, ou être administrées sous forme de médicaments par voie orale ou intraveineuse (les formes injectables étant réservées à l'administration hospitalière) **(Vignier, 2019)**.

Ci-dessous sont énumérés les médicaments les plus fréquemment utilisés pour traiter le paludisme :

- ✓ Les thérapies combinées à base d'artémisinine pour le traitement du paludisme à *Plasmodium falciparum.*
- ✓ La chloroquine est recommandée pour le traitement de l'infection par *Plasmodium vivax* uniquement dans les régions où ce parasite reste sensible à ce médicament (**Figure 34**).
- ✓ En complément du traitement principal, l'administration de la primaquine peut contribuer à la prévention des rechutes de l'infection causée par *Plasmodium vivax* et *Plasmodium ovale*.

Depuis un certain nombre d'années, les parasites ont acquis des résistances aux agents antipaludiques.

Actuellement, le seul vaccin contre le paludisme disponible est le vaccin nommé « RTS,S ». Cependant, son efficacité est modérée et se concentre exclusivement sur le parasite *P. falciparum*. Il est conseillé de l'utiliser en complément d'autres mesures pour prévenir les formes sévères de la maladie.

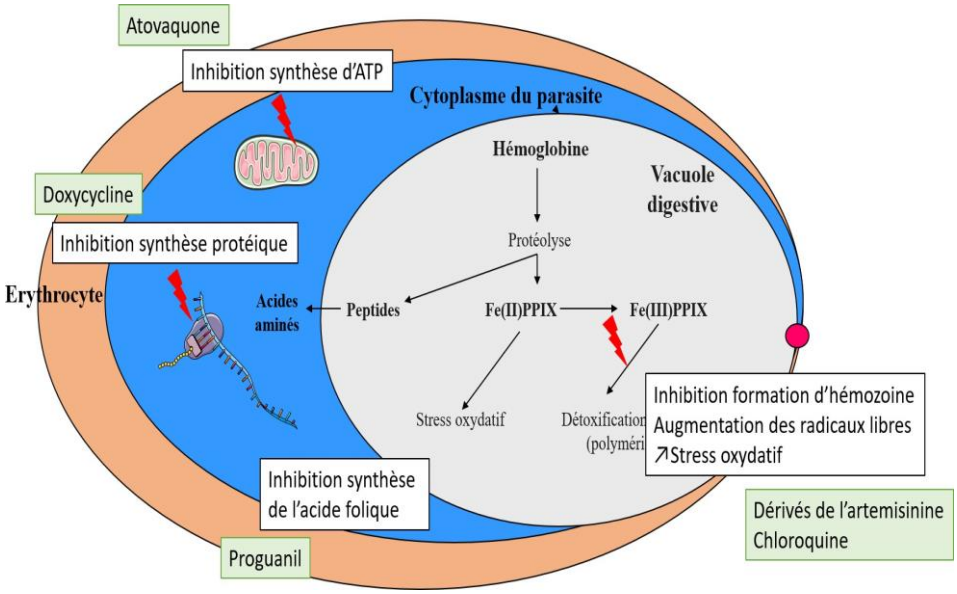


Figure 34. Mode d'action : synthèse **(diapositive du Dr Marjorie Cornu)**

V. Conclusion

Pour conclure, le paludisme demeure l'une des maladies infectieuses les plus meurtrières de notre époque, affectant de manière disproportionnée les populations les plus vulnérables. Chaque avancée dans la prévention et le traitement est une victoire, mais elle reste menacée par l'émergence rapide des résistances. La bataille contre le paludisme est donc une véritable course contre la montre qui exige un engagement politique, scientifique et financier sans faille si nous voulons transformer les progrès d'aujourd'hui en une victoire durable pour demain.

CHAPITRE III

TRAVAUX CONNEXES

CHAPITRE III

Travaux Connexes

Approches de Classification et explicabilité dans la détection du paludisme

Sommaire

I. Introduction

Le paludisme reste l'une des maladies infectieuses les plus meurtrières au monde, causée principalement par le parasite *Plasmodium falciparum*, transmis par la piqûre de moustiques infectés du genre *Anopheles*. Malgré les efforts de prévention et de traitement, la maladie continue de représenter un défi majeur pour la santé publique, en particulier dans les régions tropicales et subtropicales. Un diagnostic précoce et précis est crucial pour lutter contre la propagation de cette maladie, et c'est dans ce domaine que les progrès récents en intelligence artificielle (IA) apportent des solutions innovantes.

Ce chapitre présente un état de l'art sur les avancées récentes dans la classification du paludisme, en mettant l'accent sur les méthodes de diagnostic basées sur l'IA et l'intégration des techniques d'explicabilité de ces modèles, également connues sous le nom de XAI.

La première partie de ce chapitre se concentre sur les méthodes de diagnostic utilisées pour détecter la présence du parasite *Plasmodium* dans les frottis sanguins, ainsi que sur l'application des modèles d'IA pour améliorer la précision, la rapidité ainsi que l'accessibilité au diagnostic. En particulier, l'utilisation des réseaux de neurones convolutifs (CNN), des architectures comme YOLO **(K. Anand et al., 2022)**, Mask R-CNN **(Kassim et al., 2021a)** et d'autres modèles avancés que nous allons présenter.

La seconde partie explore les travaux récents sur l'explicabilité des modèles d'IA dans le domaine du diagnostic du paludisme. Bien que les modèles IA puissent surpasser les méthodes

traditionnelles en termes de précision, leur manque de transparence soulève des préoccupations, notamment dans des contextes cliniques où la confiance des praticiens est essentielle. Des techniques comme LIME, Grad-CAM ou encore SHAP ((SHapley Additive exPlanations) sont désormais intégrées aux modèles IA afin de rendre leurs prédictions plus compréhensibles et de favoriser leur adoption dans les environnements médicaux. Ce chapitre examine comment ces techniques peuvent non seulement améliorer la confiance des utilisateurs, mais aussi permettre une meilleure interprétation des décisions des modèles d'IA à travers les articles de recherche récemment publiés.

L'objectif de ce chapitre est de présenter un aperçu complet des progrès réalisés avec l'utilisation de l'IA pour le diagnostic du paludisme. Il s'agit d'examiner les dernières innovations dans la classification du paludisme à travers des modèles IA, tout en mettant l'accent sur les méthodes visant à améliorer l'explicabilité de ces modèles. En explorant à la fois l'impact des techniques avancées de classification et l'importance croissante de la transparence des modèles, ce chapitre vise à fournir une compréhension approfondie des avancées actuelles et des défis restants dans l'application de l'IA pour le diagnostic du paludisme.

II. Importance du Diagnostic Précis des Maladies Malaria

1. Impact sanitaire et économique

Le diagnostic rapide, précoce et précis du paludisme est essentiel pour réduire le taux de morbidité et de mortalité associé à la maladie. Un diagnostic précoce permet un traitement rapide, réduisant ainsi les risques de complications graves comme l'anémie, les troubles neurologiques, et l'insuffisance rénale aiguë. Selon l'OMS, le paludisme représente également un lourd fardeau économique pour les pays en voie de développement, due aux coûts de traitement élevés, aux hospitalisations prolongées et à une perte de productivité causée par les absences des malades.

Les pays à faible revenu, en particulier en Afrique, sont confrontés à des coûts de santé élevés et à un manque de ressources humaines et matérielles pour diagnostiquer et traiter efficacement le paludisme. Par conséquent, un diagnostic rapide et fiable est indispensable pour optimiser les traitements et limiter l'impact économique.

2. Approches et techniques de diagnostic traditionnelles

Comme nous l'avons vu au chapitre II, la microscopie, qui consiste à examiner des frottis sanguins sous un microscope, reste la méthode de référence pour diagnostiquer la malaria. Bien qu'elle soit fiable, elle dépend fortement des compétences des techniciens et nécessite un équipement spécialisé, deux conditions, difficiles à réunir dans les zones considérées comme « désert médical ». De plus, le processus peut être long et coûteux.

Les tests rapides de diagnostic (TDR) ont été développés pour fournir des résultats rapides et portables. Ces tests détectent des antigènes spécifiques des parasites du paludisme, permettant un diagnostic immédiat, mais leur précision varie selon les conditions du terrain et le stade de la maladie. De plus, comme nous l'avons déjà mentionné au chapitre II, l'émergence

de nouvelles souches présentant une délétion du gène hrp2 et/ou du gène hrp3 rendent ce type de test obsolète **(Pilly, 2018)**.

III. Histoire et Évolution des Méthodes Traditionnelles en Malaria

1. Les premières méthodes de diagnostic (avant 1900)

Avant le 20e siècle, le diagnostic de la malaria reposait principalement sur l'observation clinique des symptômes. Ce n'est qu'en 1880 que le médecin français Alphonse Laveran a découvert le parasite responsable du paludisme dans le sang des patients. Cette découverte a permis une meilleure compréhension de la maladie, mais des méthodes de diagnostic fiables restaient inexistantes **(Bastianelli et al., 1898 ; Celli et al., 1933)**.

2. Évolution de la microscopie et des tests de laboratoire

Au début du 20e siècle, la microscopie est devenue la méthode principale pour détecter les parasites. L'introduction de colorants spécifiques a amélioré la visualisation des parasites. Depuis les années 2000, les tests de diagnostic rapide (TDR) ont été développés pour offrir une alternative plus rapide et portable, mais leur fiabilité reste un problème dans des environnements de terrain difficiles.

IV. L'Intelligence Artificielle et son Rôle dans le Diagnostic de la Malaria

L'intelligence artificielle, et plus particulièrement les méthodes du ML et DL, ont révolutionné le domaine du diagnostic des maladies infectieuses, dont la malaria. Le diagnostic traditionnel du paludisme repose sur l'examen microscopique de frottis sanguins, une méthode lente et exigeante en expertise. Cependant, l'IA a permis d'automatiser ce processus, augmentant ainsi la vitesse et la précision du diagnostic **(Yang et al., 2020)**.

1. Introduction à l'IA dans le domaine médical

L'IA connaît une croissance rapide dans le domaine médical, avec des systèmes capables de traiter d'énormes volumes de données pour détecter des anomalies et effectuer des diagnostics plus rapidement. Dans des domaines comme la radiologie, la pathologie et la microbiologie, l'IA permet d'analyser les images médicales et les données cliniques plus efficacement et plus précisément que les méthodes humaines traditionnelles (H. Yang et al., 2020). Grâce à des algorithmes d'apprentissage automatique, l'IA peut également identifier des schémas invisibles à l'œil humain, améliorant ainsi les résultats du diagnostic **(Kassim et al., 2021a).** Des applications de l'IA dans le domaine de la médecine, telles que le diagnostic du cancer et des maladies cardiaques, ont déjà démontré des résultats impressionnants, servant de modèle pour l'application dans le diagnostic du paludisme **(Amin et al., 2024)**.

Les recherches récentes sur la détection du paludisme ont exploré diverses méthodes, allant de la classification binaire de la présence du paludisme dans les frottis sanguins à des tâches plus spécifiques telles que la distinction entre les espèces de Plasmodium.

2. L'IA pour l'analyse des images de frottis sanguins

L'un des modèles de l'IA, des plus performants dans le diagnostic du paludisme repose sur l'utilisation des réseaux neuronaux convolutifs (CNN). Ces réseaux sont capables d'analyser des images de frottis sanguins et d'identifier des signes d'infection par *Plasmodium* avec une précision supérieure à celle des méthodes traditionnelles. Les CNN excèlent dans le traitement des données visuelles complexes, comme celles des frottis sanguins, où la localisation et l'identification des parasites sont cruciales. Ces modèles d'apprentissage profond peuvent détecter des caractéristiques microscopiques subtiles des parasites, souvent invisibles à l'œil humain, et classifier les images en fonction de la présence ou de l'absence de l'infection **(Goni et al., 2023)**. De plus, les CNN sont capables d'identifier des espèces spécifiques de *Plasmodium*, comme *P. falciparum* et *P. vivax*, grâce à des réseaux spécialement entraînés pour cette tâche **(Goni et al., 2023)**. Ces progrès dans la classification des espèces permettent une prise en charge plus ciblée et plus rapide des patients **(Ufuktepe et al., 2021 ; Qadri et al., 2023a)**.

L'algorithme YOLO, par exemple, a été utilisé pour détecter *Plasmodium vivax* dans les images de frottis sanguins avec une grande efficacité, surpassant ainsi les performances des méthodes conventionnelles **(Yang et al., 2020)**. D'autres approches, comme l'utilisation du modèle Mask R-CNN et ResNet50, ont également montré des résultats impressionnants dans la détection des parasites **(Kassim et al., 2021a)**. Des méthodes comme le Channel Feature Pyramid Network (CFPNet-M) sont également prometteuses, permettant de mieux traiter les variations de taille des parasites dans les images **(Amin et al., 2024)**. Ces modèles permettent non seulement de détecter la présence de l'infection, mais aussi de classifier précisément les parasites en fonction des espèces, offrant ainsi une solution complète pour le diagnostic.

3. Avantages de l'IA dans les pays en développement

Dans les pays en développement, où les ressources humaines et les infrastructures médicales peuvent être limitées, l'IA offre des solutions peu coûteuses et accessibles pour diagnostiquer le paludisme. L'IA peut être intégrée dans des dispositifs mobiles, permettant ainsi des diagnostics rapides et précis même dans des zones reculées. Ces systèmes mobiles, équipés de modèles d'IA pour l'analyse des images de frottis sanguins, permettent de réduire le besoin d'expertise spécialisée et de former des techniciens de laboratoire locaux à effectuer des analyses efficaces **(Islam et al., 2022)**. Cette accessibilité et rapidité sont essentielles pour gérer les épidémies de paludisme et améliorer l'efficacité des campagnes de diagnostic et de traitement **(Khan et al., 2020)**.

En outre, des modèles tels que les réseaux de neurones Bi-LSTM, qui sont utilisés pour classifier les images médicales, ont montré leur efficacité dans le diagnostic du paludisme, ce qui permet d'augmenter l'efficacité des systèmes de diagnostic dans les pays à faible revenu **(Rajab et al., 2023)**. Ces innovations permettent de pallier le manque de spécialistes en utilisant des technologies d'IA accessibles via des plateformes mobiles.

4. Intégration de l'IA dans des systèmes de diagnostic assisté par ordinateur (CAD)

Les modèles d'IA peuvent également être intégrés dans des systèmes de diagnostic assisté par ordinateur (CAD), qui soutiennent les cliniciens et techniciens de laboratoire dans la prise de décision. Ces systèmes fournissent des résultats d'analyse en temps réel, permettant aux professionnels de santé de se concentrer sur l'interprétation des données plutôt que sur l'analyse brute des images. L'IA réduit ainsi les erreurs humaines, améliore la qualité du diagnostic et aide à standardiser les pratiques dans les environnements médicaux **(Alanazi and Alaerjan, 2023)**. L'intégration des systèmes CAD avec des outils d'IA assure également une plus grande efficacité dans le traitement des grandes quantités de données provenant de tests de paludisme, ce qui est crucial dans les régions à forte incidence de la maladie **(Qadri et al. 2023b)**.

L'IA ne se limite pas à une simple amélioration des outils existants, elle permet de redéfinir les processus de diagnostic et de soins dans des contextes à ressources limitées, contribuant ainsi à un contrôle plus efficace de la maladie **(Lee et al., 2023)**.

V. Méthodes traditionnelles versus IA dans l'analyse des images histopathologiques

L’analyse des images histopathologiques, utilisées pour détecter les parasites du *Plasmodium* dans les frottis sanguins, repose sur des techniques avancées d’imagerie. Dans ce domaine, une distinction claire émerge entre les méthodes traditionnelles et celles basées sur l’intelligence artificielle (IA). Les méthodes traditionnelles, souvent fondées sur des algorithmes comme KNN (k-nearest neighbors) ou SVM (Support Vector Machines), ont montré des performances satisfaisantes dans des scénarios simples, mais leur efficacité diminue lorsque les données sont volumineuses, complexes ou de faible qualité. Ces techniques requièrent une intervention humaine importante, notamment pour l'extraction manuelle des caractéristiques pertinentes des images, ce qui ralentit considérablement le processus de diagnostic et augmente le risque d'erreur humaine. De plus, elles peinent à gérer la variabilité des images histopathologiques, surtout lorsqu'il s'agit de diverses espèces de *Plasmodium* et de parasites dans des stades évolutifs différents **(Yang et al., 2020)**.

En revanche, les modèles d'intelligence artificielle, notamment les CNN et les RNN dont fait partie les LSTM, Stacked-LSTM et BI-LSTM, surpassent les méthodes traditionnelles de manière significative. Ces techniques d’IA sont capables d'analyser automatiquement des images en haute résolution sans nécessiter un prétraitement complexe. Grâce à l’apprentissage profond, ces modèles peuvent non seulement extraire des caractéristiques subtiles des images de manière autonome, mais aussi reconnaître des patterns complexes qui échappent souvent à l’œil humain. Cela permet de détecter les parasites à des stades précoces de manière plus précise et rapide, et ce, même avec des images de qualité médiocre ou avec une forte variabilité des données ou de luminosité. De plus, ces modèles d'IA peuvent être entraînés sur des bases de données volumineuses, ce qui améliore leur capacité à généraliser et à s'adapter à différents types de tissus et de conditions d'imagerie **(Yang et al., 2020)**.

Les méthodes basées sur l’IA ne se contentent pas d'améliorer la précision du diagnostic, mais elles offrent également des possibilités d'intégration dans des systèmes d’assistance

médicale à grande échelle, qui peuvent être déployés dans des zones à ressources limitées. Par exemple, des dispositifs mobiles équipés de modèles d'IA peuvent être utilisés sur le terrain pour diagnostiquer rapidement le paludisme dans des environnements ruraux et reculés, réduisant ainsi le besoin de spécialistes et augmentant l'accessibilité au diagnostic **(Kassim et al., 2021a)**. En outre, l'utilisation de l'IA permet de rationaliser le processus de diagnostic en offrant des résultats en temps réel, ce qui est crucial pour une prise en charge rapide et efficace des patients **(Jothi et al., 2016)**.

En comparaison avec les méthodes traditionnelles, les systèmes d'IA offrent donc une meilleure évolutivité et une plus grande fiabilité, en particulier pour traiter des volumes de données importants et des images complexes. Cependant, l'application généralisée de l'IA dans ce domaine nécessite encore des améliorations en termes de standardisation des protocoles d'entraînement et de validation des modèles, ainsi que de la création de bases de données diversifiées et de haute qualité pour l'entraînement des systèmes **(Jothi et al., 2016)**.

L'analyse des images histopathologiques utilisées pour détecter les parasites du *Plasmodium* dans les frottis sanguins a considérablement évolué grâce à l'introduction de l'IA. Cette évolution a permis une amélioration significative de la précision et de l'efficacité du diagnostic par rapport aux méthodes traditionnelles.

1. Méthodes d'IA traditionnelles : KNN, SVM, et prétraitement manuel

Les méthodes traditionnelles, telles que KNN et SVM, reposent généralement sur l'extraction manuelle de caractéristiques à partir des images. Ces techniques sont efficaces dans des situations simples, mais présentent plusieurs limites :

- **KNN** : Cet algorithme de classification fonctionne en identifiant les K échantillons les plus proches de l'échantillon à classer. Par exemple, dans une étude de **Jothi et al. (2016)**, l'algorithme KNN a été utilisé pour classer des images histopathologiques de tissus thyroïdiens en catégories (normal vs. carcinome thyroïdien). L'algorithme nécessite un prétraitement des images, comme le seuillage d'Otsu, pour améliorer la qualité des données avant application de l'algorithme. Bien que KNN puisse atteindre une précision de 100 % sur des ensembles de données limités, il devient inefficace et sujet à des erreurs lorsque les données sont complexes ou de mauvaise qualité **(Jothi et al., 2016).**

  Algorithme de KNN :

  (1) **Prétraitement** des images pour améliorer la qualité (ex. seuillage d'Otsu).

  (2) **Mesure de la distance** entre les points de données (euclidienne, Manhattan, etc.).

  (3) **Sélection des K voisins les plus proches** pour effectuer la classification.

  (4) **Assignation de la classe** en fonction de la majorité des voisins.

- **SVM** : Cet algorithme fonctionne en trouvant un hyperplan qui sépare les différentes classes de manière optimale dans un espace à plusieurs dimensions. Pour les images histopathologiques, un SVM peut classifier des cellules ou des tissus comme normaux ou

anormaux. Toutefois, comme KNN, SVM nécessite un prétraitement approfondi pour extraire les caractéristiques pertinentes, ce qui peut rendre l'approche plus lente et dépendante de l'intervention humaine.

Algorithme de SVM :

(1) **Extraire les caractéristiques** pertinentes des images (par exemple, les textures ou les formes cellulaires).

(2) **Transformation des données** dans un espace à haute dimension pour mieux séparer les classes.

(3) **Trouver l'hyperplan optimal** qui sépare les différentes classes avec une marge maximale.

(4) **Classification** des nouvelles données selon l'hyperplan.

2. L'intelligence artificielle : Réseaux neuronaux convolutifs (CNN) et autres algorithmes

Bien que les méthodes traditionnelles comme KNN et SVM aient été des étapes importantes dans le développement du diagnostic assisté par ordinateur, elles montrent leurs limites face aux exigences croissantes en termes de volume de données, de qualité des images et de diversité des parasites. Les techniques d'IA, notamment les CNN, Mask R-CNN, les LSTM et les réseaux Transformer, non seulement surpassent ces limitations en termes de précision et d'efficacité, mais elles offrent également de nouvelles perspectives pour l'automatisation du diagnostic du paludisme à grande échelle. En particulier, elles ouvrent la voie à des solutions de diagnostic rapides et accessibles, adaptées aux ressources limitées des pays en développement.

2.1. Modèles basés sur les CNN

A. Mask R-CNN et ResNet50 : Identification des espèces

Une autre étude pertinente, menée par **Kassim et al. (2021a)**, a utilisé un modèle plus avancé, à savoir Mask R-CNN, pour détecter les parasites dans des images de frottis sanguins. Mask R-CNN combine la détection d'objets (localisation des parasites) avec une segmentation plus fine, permettant de localiser chaque parasite avec un masque précis pour chaque instance. Combiné avec **ResNet50**, qui est un réseau de neurones profond capable d'extraire des caractéristiques plus détaillées des images, ce modèle a permis d'identifier des espèces de *Plasmodium* avec une précision de 98 %. Ce type de réseau neuronal est particulièrement adapté à l'analyse des images où les parasites peuvent apparaître dans des tailles et formes variées **(Kassim et al., 2021a)**.

Algorithme de Mask R-CNN :

(1) **Détection d'objets** : Le réseau détecte les objets (parasites) dans l'image.

(2) **Segmentation** : Chaque objet détecté est accompagné d'un masque, délimitant précisément sa forme.

(3) **Réseau de segmentation** : Le modèle applique une couche de segmentation pour chaque objet, permettant une segmentation fine des parasites dans les frottis sanguins.

L'architecture de Mask R-CNN repose sur deux réseaux principaux : un pour la détection d'objets et un autre pour générer des masques de segmentation pour chaque instance d'objet (Dans ce cas, chaque parasite).Cela permet non seulement de localiser le parasite mais aussi de mieux délimiter sa forme dans l'image, améliorant la précision du diagnostic ce qui améliore la localisation et l'identification des parasites dans les images.

Dans ces travaux, les auteurs conçoivent quatre pipelines en partant d'un modèle simple basé sur la détection Mask R-CNN, et ajoutons trois autres classifieurs pour discriminer entre P. vivax, P. falciparum et les patients non infectés tel qu'il est illustré dans la **Figure 35**. Leur Framework nommé PlasmodiumVF-Net, est basé sur plusieurs indicateurs, compteurs et scores pour calculer les décisions au niveau de l'image et du patient. La **Figure 36** illustre le diagramme de flux de l'ensemble du Framework.

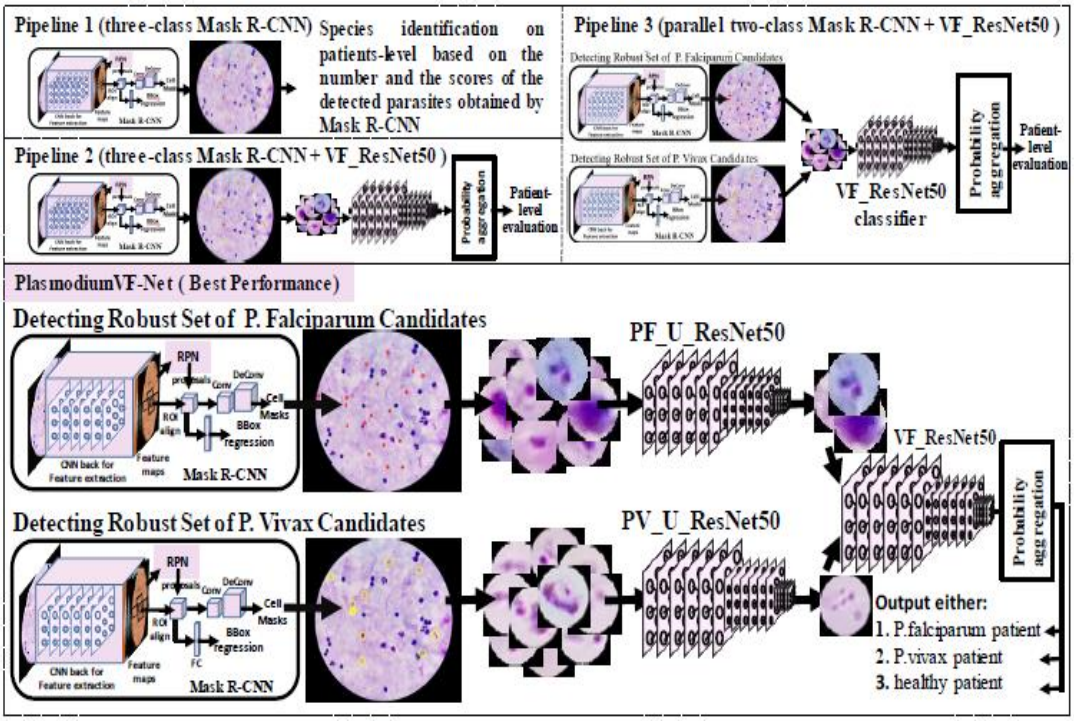


Figure 35. Les différents Pipelines conçus utilisant Mask R-CNN et ResNet50 **(Kassim, et al., 2021a)**

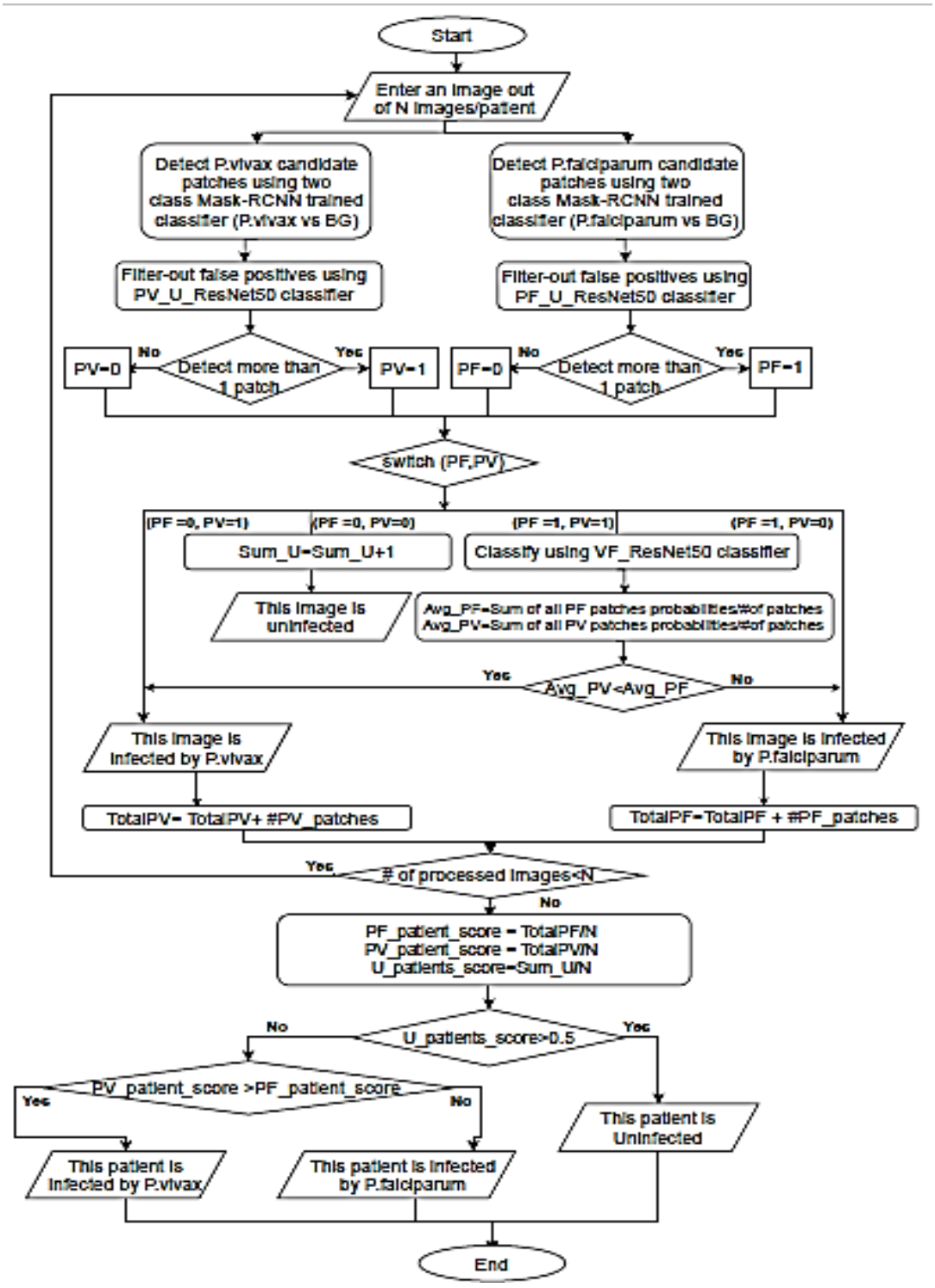


Figure 36. Organigramme pour PlasmodiumVF-Net **(Kassim, et al., 2021a)**

B. YOLO (You Only Look Once):

L'algorithme YOLO de **Yang et al. (2021)** est un réseau neuronal conçu pour effectuer des détections d'objets en temps réel. Contrairement aux approches traditionnelles où l'image est traitée par plusieurs étapes, YOLO procède à la détection en une seule passe (une seule passe du réseau pour localiser les objets et prédire leurs classes). Cela permet d'identifier les parasites

du *Plasmodium vivax* dans les images de frottis sanguins avec une précision de 95 % et un temps de traitement réduit de plusieurs heures à quelques minutes.

Algorithme de YOLO :

(1) **Convolution** : Extraction des caractéristiques essentielles de l'image avec plusieurs couches convolutives.

(2) **Détection d'objets** : L'image est divisée en une grille, et chaque cellule prédit la position, la taille et la classe des objets (ici, les parasites).

(3) **Classification et localisations** : Prédiction des boîtes de délimitation et de la classe pour chaque objet détecté.

La **Figure 37** présente L'architecture de YOLO. Elle repose sur des **couches convolutives** qui extraient des informations spatio-contextuelles et des couches entièrement connectées qui génèrent des prédictions en termes de classe et de position des parasites.

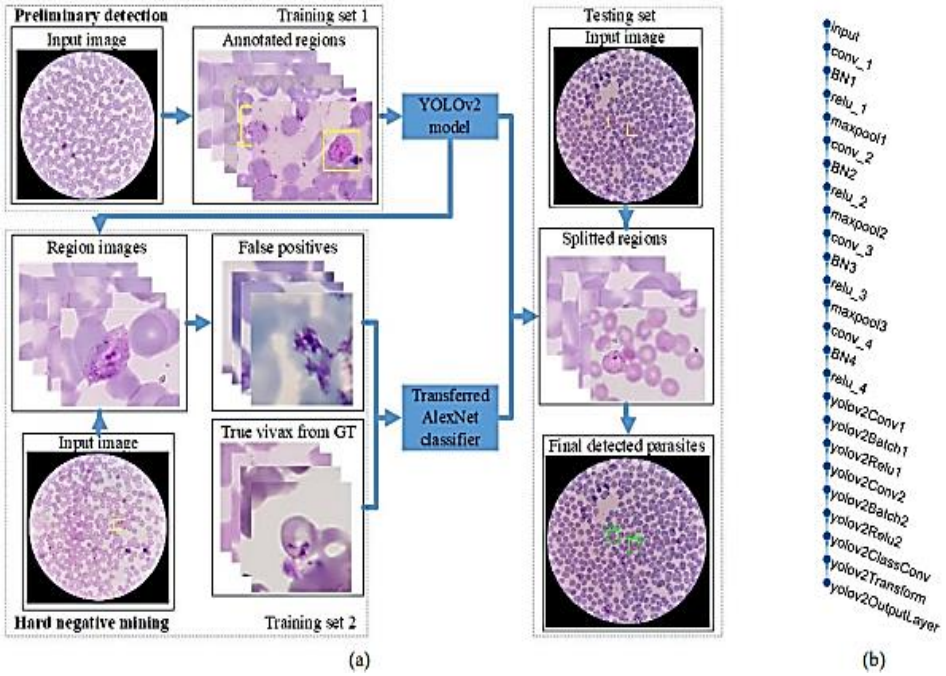


Figure 37: (a) Organigramme du modèle YOLO en cascade proposé. GT indique la vérité de terrain. **(b)** Structure du réseau du modèle YOLOv2 dans (a). Notez que conv indique une couche de convolution, BN indique une couche de normalisation par lot, et relu représente une unité linéaire rectifiée (ReLU) (**Yang, et al., 2021)**

3. Modèles basés sur les réseaux LSTM pour l'Analyse Temporelle et Spatiale

Outre les CNN, d'autres modèles basés sur les RNN et les LSTM en particulier, ont donné d'excellents résultats tels que l'approche d'augmentation des caractéristiques basée sur le classement des manifolds **(Pereira-Ferrero et al, 2023)**, dans laquelle les auteurs ont proposé une approche d'apprentissage de représentation, basée sur l'augmentation des caractéristiques pour améliorer les performances des LSTM, montrant jusqu'à 20 % d'amélioration de la précision de classification des images. Dans ces travaux de recherche, les auteurs ont proposé une méthode qui vise à exploiter les informations de similarité contextuelle disponibles grâce à

un apprentissage de variété (manifold) basé sur le rang, utilisé pour définir et attribuer des poids aux échantillons utilisés dans l'augmentation. L'approche est validée en utilisant des caractéristiques basées sur CNN et des modèles LSTM, pour obtenir des résultats de précision encore plus élevés sur les tâches de classification d'images.

L'architecture proposée est présentée dans la **Figure 38**, où on peut remarquer que les auteurs ont procédé à une hybridation non seulement entre les modèles d'apprentissage CNN-LSTM, mais aussi une hybridation entre les méthodes d'apprentissage profonds (CNN et LSTM) et les méthodes d'apprentissage par manifold (pour le classement de l'importance des variétés).

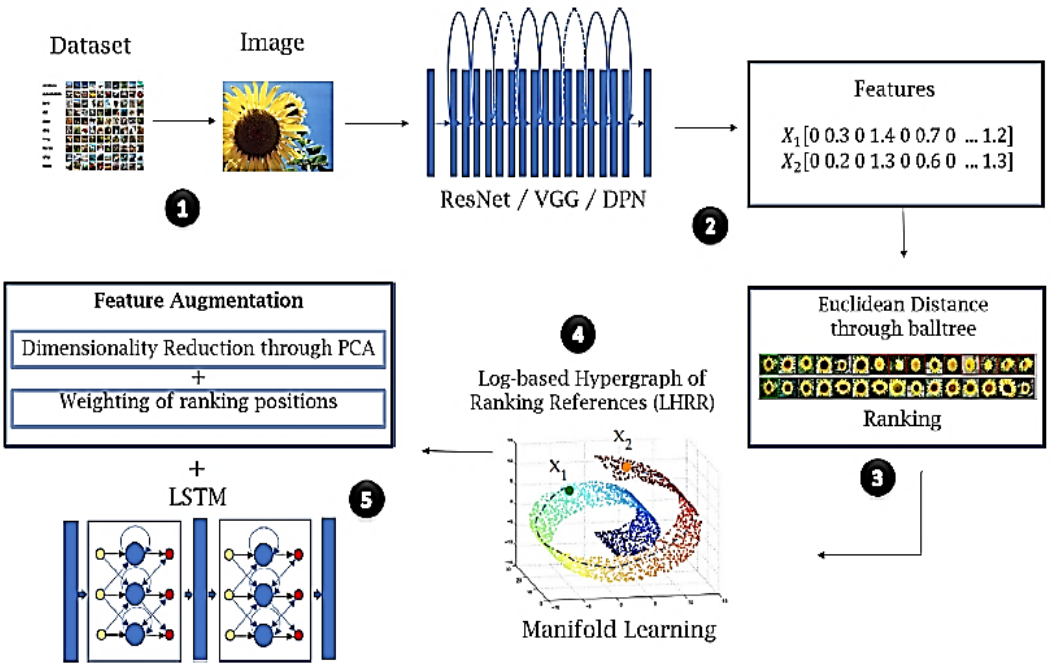


Figure 38. Architecture approche proposée pour le classement et l'augmentation des caractéristiques dans les tâches de classification d'images **(Pereira-Ferrero, et al, 2023)**

Une étude intéressante menée par **Alanazi et al. (2023)** présente une architecture Bi-LSTM (Bidirectional LSTM), qui a montré des résultats très prometteurs en termes de précision, de justesse et de robustesse. L'approche proposée repose sur la méthode Speed Up-Robust Feature (SURF) pour l'extraction des caractéristiques à partir du jeu de données. Les caractéristiques requises sont sélectionnées après leur extraction, en se basant sur l'optimisation par recherche de corbeaux (CSO). Le mécanisme Bi-LSTM complète la procédure de classification à la dernière étape, comme on peut clairement le distinguer dans la **Figure 39**. Les métriques de qualité telles que la précision, le rappel, la précision et le score F1 sont comparées à d'autres classificateurs de pointe. Les résultats montrent l'efficacité du paradigme proposé pour l'identification et la classification du COVID-19, montrant une haute spécificité, une haute sensibilité et une faible complexité de calcul.

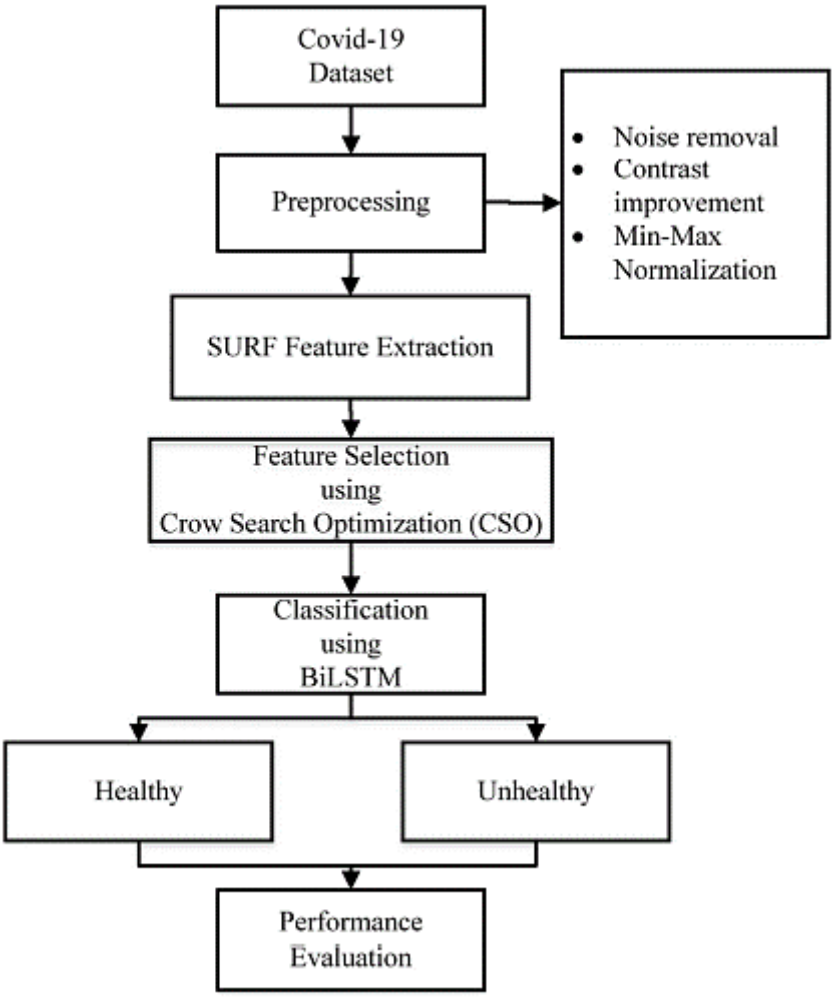


Figure 39. Architecture du modèle Bi-LSTM **(Alanazi, et al., 2023)**

Dans une autre étude, on a combiné un LSTM avec l'apprentissage par Transfert **(Lanjewar et al., 2024)** —MobileNetV2, ResNet50 et VGG16— pour l'extraction de caractéristiques à partir d'images échographiques dans le diagnostic du cancer du sein. Le modèle basé sur VGG16 a atteint une Aire Sous la Courbe (AUC) de 1,0, un score F1 de 99%, un Coefficient de Corrélation de Matthews (MCC) de 98,9 %, et un coefficient Kappa de 98,9%, avec une validation croisée K-fold donnant un score F1 moyen de 96 %.

Dans le même contexte, **Kim et al. (2023)**, ont développé leur modèle LSTM-TL. Une technique pour prévoir la consommation d'énergie des bâtiments en utilisant des LSTM et l'apprentissage par transfert (TL). Pour appliquer cette méthode, l'étude a d'abord analysé les données brutes pour créer des ensembles de données ; un type d'hôpital a été choisi pour une étude de cas au début. Le modèle de prototype d'hôpital, créé par le Département de l'énergie des États-Unis, a servi à créer des données pour s'entraîner et tester avant de commencer le transfert d'apprentissage. Pour transférer les connaissances vers un nouveau domaine, une analyse par simulation a été faite pour créer des ensembles de données, en tenant compte de périodes de données courtes et de différentes conditions météo. Cette étude montre que le modèle LSTM-TL est plus efficace que le modèle de prédiction qui utilise seulement LSTM, même avec différents types de temps. De plus, les résultats peuvent changer selon les méthodes

de transfert de connaissances utilisées, avec des couches fixes ou ajustées, et aussi selon les lieux.

La **Figure 40**, montre l'architecture générale de cette approche et met en avant les différentes méthodes de transfert de connaissances expérimentées.

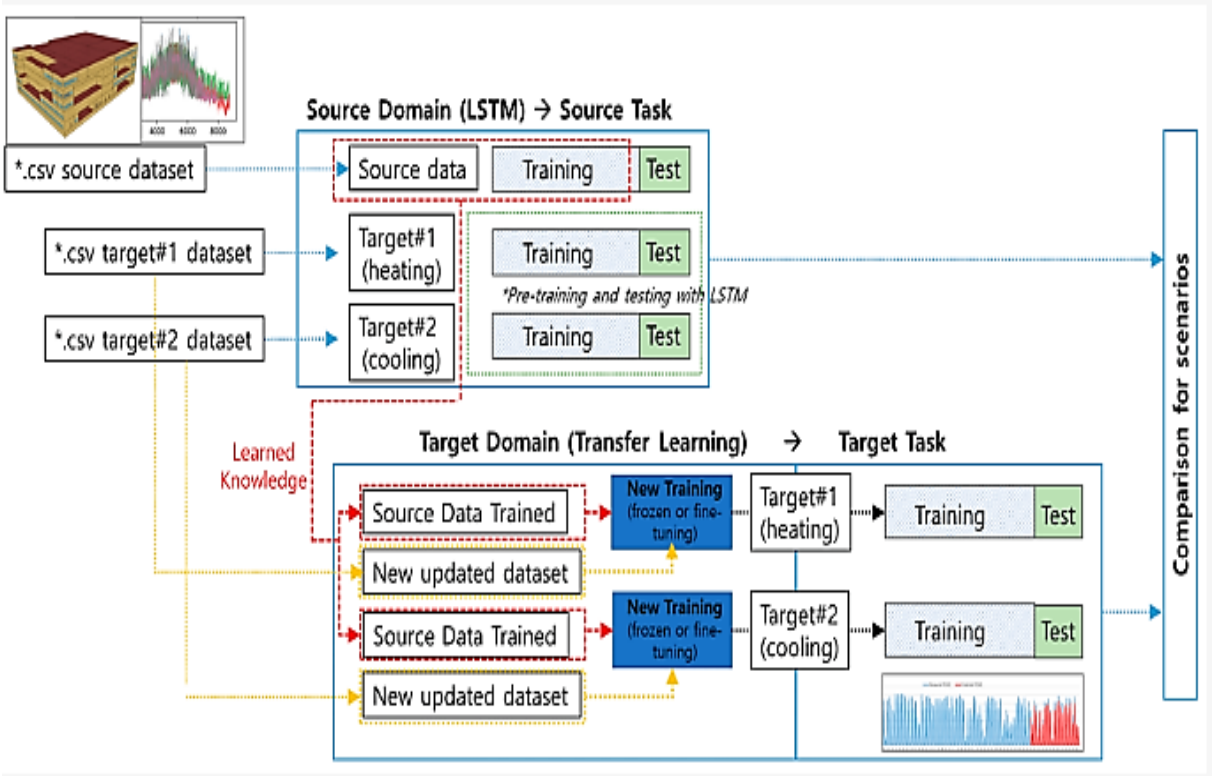


Figure 40. Cadre de travail du LSTM avec le transfert de connaissances **(Kim, et al., 2023)**

Une autre étude originale basée sur les LSTM et plus précisément sur les Bi-LSTM, est celle de **Fei et al. (2019)**. Ce document a pour objectif d'examiner l'analyse des sentiments des textes chinois sur les réseaux sociaux en intégrant des réseaux Bidirectionnels de Long-Short Term Memory (Bi-LSTM) avec un mécanisme d'Attention Multi-Head (MHAT) afin de remédier aux insuffisances de l'analyse des sentiments effectuée par les méthodes d'apprentissage automatique conventionnelles. Les réseaux BiLSTM non seulement résolvent le problème de dépendance à long terme, mais ils saisissent également le contexte réel du texte. Étant donné que le mécanisme MHAT peut acquérir des informations pertinentes à partir d'un sous-espace de représentation distinct en utilisant des calculs distribués multiples, l'objectif est d'incorporer des poids d'influence à la séquence de texte élaborée. Les résultats des simulations indiquent que le modèle proposé surpasse les méthodes établies existantes en termes de performance, atteignant un F1 score de 95.33% et une précision de 92.11%.

Cette approche est résumée dans la **Figure 41**, ci-après.

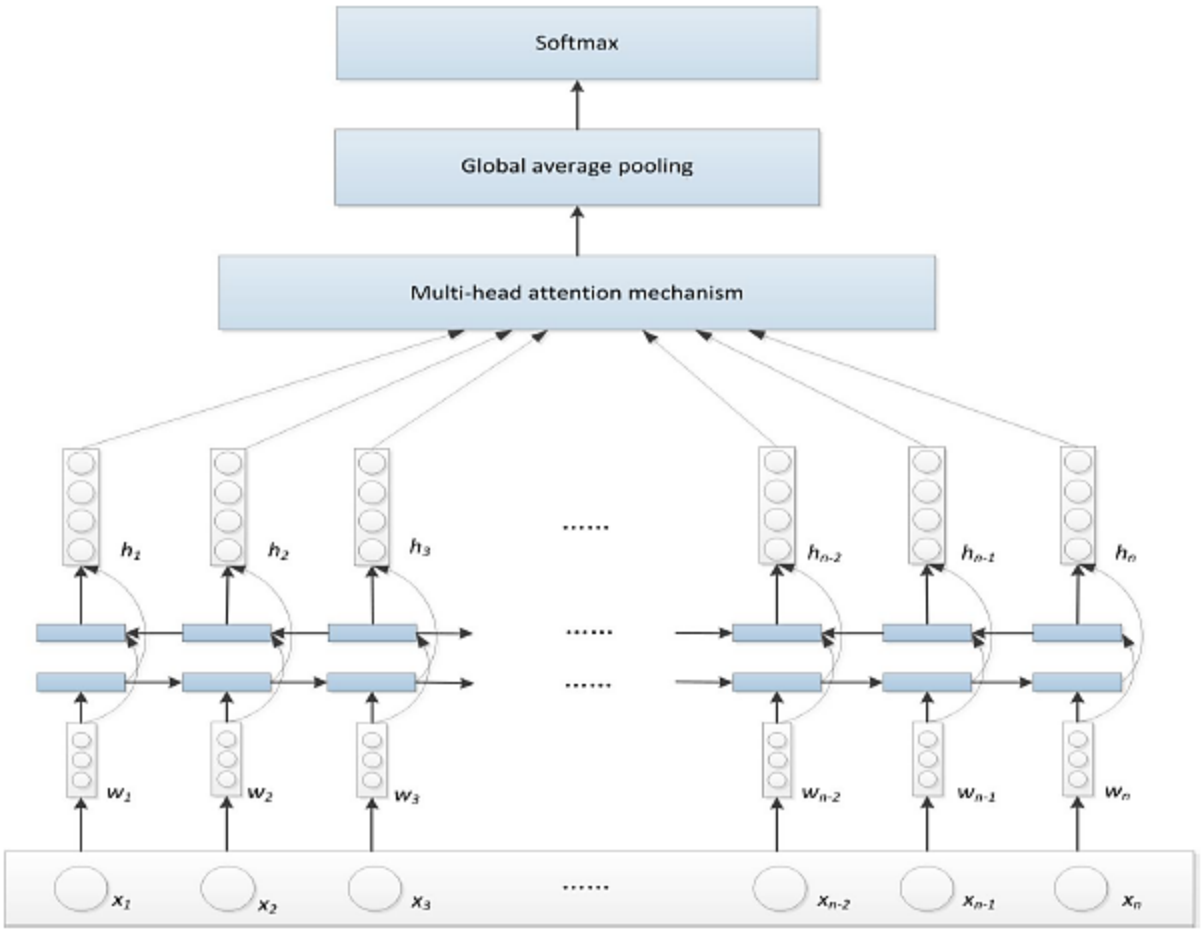


Figure 41. Structure d'un Bi-LSTM muni d'un mécanisme d'attention Multi-Head **(Fei, et al., 2019)**

4. Approches récentes : Transformer Networks

En raison de leurs limitations conceptuelles et de leur incapacité à capturer les longues dépendances, les CNN ont intégré les mécanismes d'attention, qui leur permettent d'ajuster dynamiquement les poids en fonction des caractéristiques d'entrée, afin d'améliorer leur capacité de modélisation non locale **(Vaswani et al., 2017 ; Bello et al., 2019 ; Ramachandran et al., 2019)**. Inspirés par cette ligne de recherche, de nombreux chercheurs ont déployé des efforts significatifs pour proposer des modèles avec des variantes d'attention dans le domaine de l'imagerie médicale **(Azad et al., 2020 ; Bozorgpour et al., 2021 ; Sang et al., 2021 ; Yao et al., 2021 ; Al-Shabi et al., 2022)**. Bien que ces mécanismes d'attention permettent de modéliser l'information contextuelle complète de l'image, la complexité computationnelle de ces approches croît généralement de manière quadratique par rapport à la taille spatiale, ce qui implique une charge computationnelle intensive, les rendant inefficaces dans le cas des images médicales à haute résolution pixel par pixel **(Gonc et al., 2022)**. De plus, malgré le fait que la combinaison du mécanisme d'attention avec l'opération de convolution entraîne des gains de performance systématiques, ces modèles souffrent inévitablement de contraintes dans l'apprentissage des interactions à longue portée. Le Transformer originel présenté dans la **Figure 42 (Vaswani et al., 2017)** a d'abord été appliqué à la tâche de la traduction automatique, en tant que nouveau bloc de construction basé sur l'attention, il a démontré des performances impressionnantes dans une large gamme de tâches, y compris le traitement du langage naturel

(NLP), la traduction automatique, la classification de texte et la réponse aux questions. Le succès des Transformers a conduit à l'application généralisée de cette technique dans les modèles modernes de vision par ordinateur (CV), donnant naissance aux Vision-Transformers (ViT) **(Dosovitskiy et al., 2020)**. Les ViTs se sont rapidement imposés comme des alternatives viables aux CNN dans diverses tâches telles que la reconnaissance d'images **(Dosovitskiy et al., 2020)**, la détection d'objets **(Zhu et al., 2020)**, la segmentation d'images **(Chen et al., 2021)**, la compréhension vidéo **(Arnab et al., 2021)** et la super-résolution d'images **(Chen et al., 2023)**. En tant que pièce centrale du Transformer, le mécanisme d'auto-attention a la capacité de modéliser les relations entre les éléments d'une séquence, apprenant ainsi les interactions à longue portée.

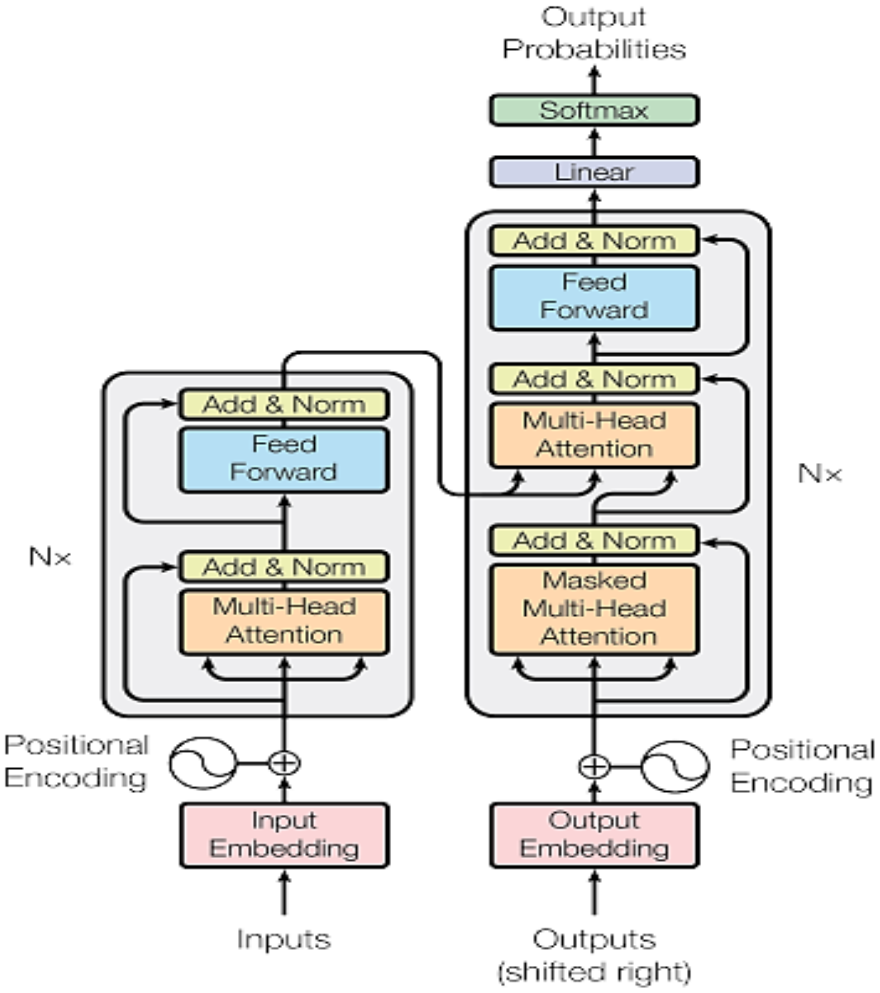


Figure 42. Architecture originelle du premier Transformer **(Vaswani et al., 2017)**

Le Vision Transformer (ViT) est une architecture neuronale innovante qui a révolutionné la vision par ordinateur, en utilisant des techniques d'auto-attention issues du traitement du langage naturel et en les adaptant aux données visuelles. ViT segmente les images en patches de taille fixe, les intégrant de manière linéaire avant de les passer à un encodeur de transformateur **(Qian et al., 2022 ; Ma et al., 2023)**. Cette méthode innovante facilite l'apprentissage complet des propriétés des images. En permettant des échanges entre les patches dans les deux sens, le ViT capture des dépendances étendues, ce qui améliore sa capacité à

représenter le contexte plus large des images. Le succès exceptionnel du ViT a non seulement repoussé les limites de la classification d'images, mais a également stimulé des avancées dans diverses tâches de vision par ordinateur, dévoilant de nouvelles avenues de recherche en intelligence artificielle et des applications pratiques **(Al-Hammuri et al., 2023 ; Han et al., 2023)**. Les auteurs dans **(Aladhadh et al., 2022)** ont proposé une architecture à deux niveaux pour une classification efficace du cancer de la peau. Le premier niveau implique des techniques d'augmentation des données pour augmenter le nombre d'échantillons présents dans le jeu de données HAM10000. Dans la deuxième couche, les auteurs ont exploité l'efficacité des transformateurs de vision médicale (MVT) utilisés dans le traitement d'images médicales pour concevoir un modèle basé sur MVT pour la classification du cancer de la peau. Une grande version du modèle Vision Transformer a été utilisée, combinée avec une tête MLP à sa sortie.

Nous avons enregistré une précision de 96%, une sensibilité de 96%, un score F1 de 97%, et une précision de 0,96. Dans **(Yang et al., 2023)**, un nouveau modèle ViT pour la classification du cancer de la peau a été présenté. La méthode était basée sur l'apprentissage par transfert en utilisant un modèle ViT pré-entraîné et en le peaufinant avec le jeu de données HAM10000. Le processus de fine-tuning a été réalisé en intégrant un segment de classification dans le segment final du transformateur encodeur, composé d'une couche aplatie et de deux normalisations par lot, séparées par une couche dense activée par GeLU. L'expérience a atteint une précision de 94%, surpassant toutes les techniques comparées.

Tout comme dans **(Xin et al., 2022)**, un ViT pré-entraîné affiné avec un MLP sur le HAM10000 a été utilisé. Cependant, une méthode basée sur l'apprentissage contrastif a été mise en œuvre. L'apprentissage contrastif repose sur une fonction de perte spécifique pour diminuer la similarité entre les échantillons de la même classe tout en augmentant la similarité entre les échantillons de classes distinctes. Ce modèle a atteint un taux de précision de 94%.

5. Recherches Basées sur les Méthodes Traditionnelles

5.1. État de l'art des recherches en microscopie

Les recherches récentes sur la microscopie se concentrent sur l'automatisation et l'amélioration de la sensibilité et de la précision. Des microscopes automatisés et des colorants spécifiques sont en développement pour rendre le diagnostic plus rapide et plus fiable.

5.2. Utilisation des tests rapides de diagnostic (TDR)

Les innovations dans les TDR visent à améliorer leur sensibilité, en particulier dans les environnements de terrain difficiles. Les nouveaux réactifs et technologies sont conçus pour augmenter la précision des tests, même dans les conditions les plus contraignantes.

VI. Recherches basées sur l'apprentissage automatique pour la malaria

Les recherches récentes dans le domaine de l'apprentissage automatique appliqué à la détection de la malaria ont démontré des progrès notables, en particulier dans l'utilisation de

modèles d'apprentissage profond tels que les réseaux de neurones convolutifs (CNN) et les approches hybrides comme les CNN combinés avec des machines à vecteurs de support (SVM). Ces techniques ont permis d'améliorer significativement les performances du diagnostic en automatisant l'analyse des images de frottis sanguins et en détectant les parasites responsables de la malaria.

1. Avancées récentes en apprentissage automatique pour la malaria

1.1. Modèles basés sur l'apprentissage profond (DL) pour la détection de la malaria

Comme nous l'avons déjà expliqué dans le chapitre I, les CNN, font partie des modèles profonds et sont largement utilisés dans le domaine de la vision par ordinateur, et ont prouvé leur efficacité pour la détection du Plasmodium, grâce à leur capacité à extraire des caractéristiques complexes à partir d'images de frottis sanguins. Un exemple clé de ces travaux est l'étude **d'Agus Eko et al. (2024)**. Le modèle CNN conçu dans cette étude, est composé de trois couches convolutionnelles et de deux couches entièrement connectées tel qu'il est montré dans la **Figure 43**. Il est appliqué pour classifier les images microscopiques des frottis sanguins afin de détecter le paludisme. L'architecture proposée utilise Maxpooling pour le pooling, en ajoutant un taux de dropout de 0,1 pour éviter le surapprentissage et une normalisation par lot pour chaque couche. Pendant le processus d'entraînement, les images sont redimensionnées à 64 x 64 pixels. Le modèle conçu ainsi, a atteint une précision de 96 % en classifiant les parasites et a montré des résultats prometteurs pour une détection rapide en milieux à ressources limitées.

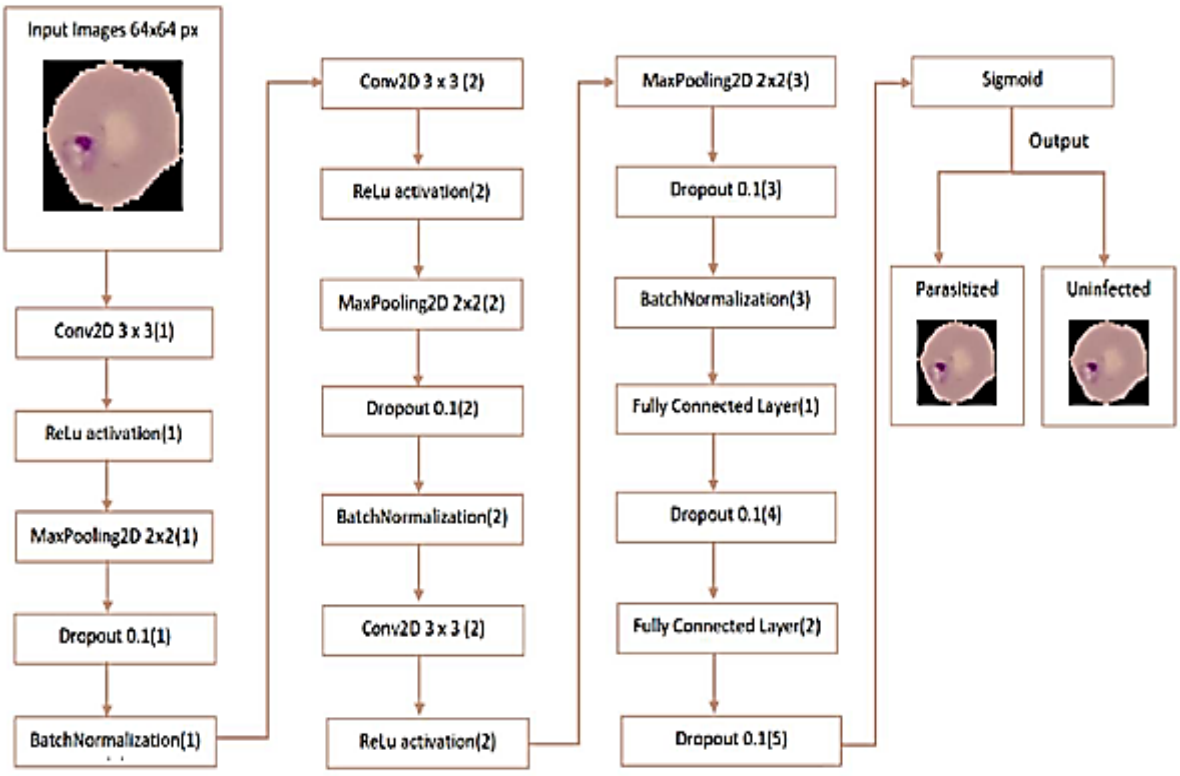


Figure 43 : Architecture du CNN pour la détection de la malaria (Minarno et al., 2024)

À partir de cette Figure nous pouvons remarquer, que L'architecture CNN de Minarno, et al. (2024) se compose de trois blocs identiques composés des couches suivantes :

1. **Entrée :**
    - Images de frottis sanguins prétraitées, redimensionnées à une taille standard de 64x64 pixels.
2. **Couche convolutionnelle 2D de dimension 3x3 pixels :**
    - Le rôle de cette couche est d'extraire les caractéristiques visuelles importantes à partie d'images de frottis sanguin.
    - Cette couche est suivie d'une fonction d'activation, **ReLU.**
3. **Couche de pooling 2D de dimension 2x2 pixels :**
    - Une couche de **max pooling** pour la réduction de la dimensionnalité des données tout en conservant les caractéristiques essentielles. Elle agit en regroupant les valeurs dans des blocs de 2×2 pixels dans une image ou une carte de caractéristiques, elle est munie d'une fonction Dropout de 0.1, ce qui signifie que **10 % des neurones** de cette couche seront désactivés à chaque itération, tandis que les 90 % restants contribueront normalement à la sortie, et ce pour éviter le surapprentissage.
4. **Blocs de normalisation :**
    - Cette couche **(batch normalization en Anglais)** est utilisée pour accélérer l'entraînement et stabiliser les gradients.

Après ces 3 blocs, on trouve une première couche entièrement connectées (Fully Connected Layer **FC**), munie d'une fonction Dropout, suivie d'une deuxième couche FC identique à la première. La sortie de cette dernière est activée à l'aide d'une fonction d'activation Sigmoïde dont le rôle est d'effectuer la classification.

5. **Couches entièrement connectées (Fully Connected - FC) :**
    - Les deux couches entièrement connectées, connectent **tous les neurones** de la couche précédente à **tous les neurones** de la couche actuelle. Chaque connexion est pondérée, et le modèle ajuste ces poids pendant l'entraînement pour minimiser l'erreur entre les prédictions et les véritables étiquettes.
6. **Sortie :**
    - Une couche **Sigmoïde** pour la classification binaire :
        - Non infecté
        - Infecté par *Plasmodium*

Un autre travail original par sa technique est celui de **Rajaraman et al. (2018a)**, qui a utilisé un réseau CNN modifié pour analyser les frottis sanguins et détecter Plasmodium

falciparum avec une précision de 95,2 % comme le montre la **Figure 44**. Le CNN proposé comporte trois couches convolutionnelles et deux couches entièrement connectées. L'entrée du modèle constitue des cellules segmentées d'une résolution de 100×100×3 pixels. Les couches convolutionnelles utilisent des filtres de 3 × 3 avec des pas de 2 pixels. Les première et deuxième couches convolutionnelles ont 32 filtres et la troisième couche convolutionnelle a 64 filtres. La conception en sandwich des unités convolutionnelles/linéaires rectifiées (ReLU) et une initialisation appropriée des poids améliorent le processus d'apprentissage **(Shang et al., 2016)**. Les couches de max-pooling avec une fenêtre de pooling de 2 × 2 et des pas de 2 pixels suivent les couches convolutionnelles pour résumer les sorties des groupes neuronaux voisins dans les cartes de caractéristiques. La sortie poolée de la troisième couche convolutionnelle est transmise à la première couche entièrement connectée qui contient 64 neurones, et la deuxième couche entièrement connectée alimente le classificateur Softmax. La régularisation par abandon **(Srivastava et al., 2014)** avec un taux d'abandon (Dropout) de 0,5 est appliquée aux sorties de la première couche entièrement connectée. Le modèle est entraîné en optimisant l'objectif de régression logistique multinomiale en utilisant la descente de gradient stochastique (SGD) **(LeCun et al., 2015)** et le momentum de Nesterov **(Botev et al., 2017)**. Le modèle personnalisé est optimisé pour les hyperparamètres par une méthode de recherche aléatoire sur grille **(Bergstra and Bengio, 2012)**.

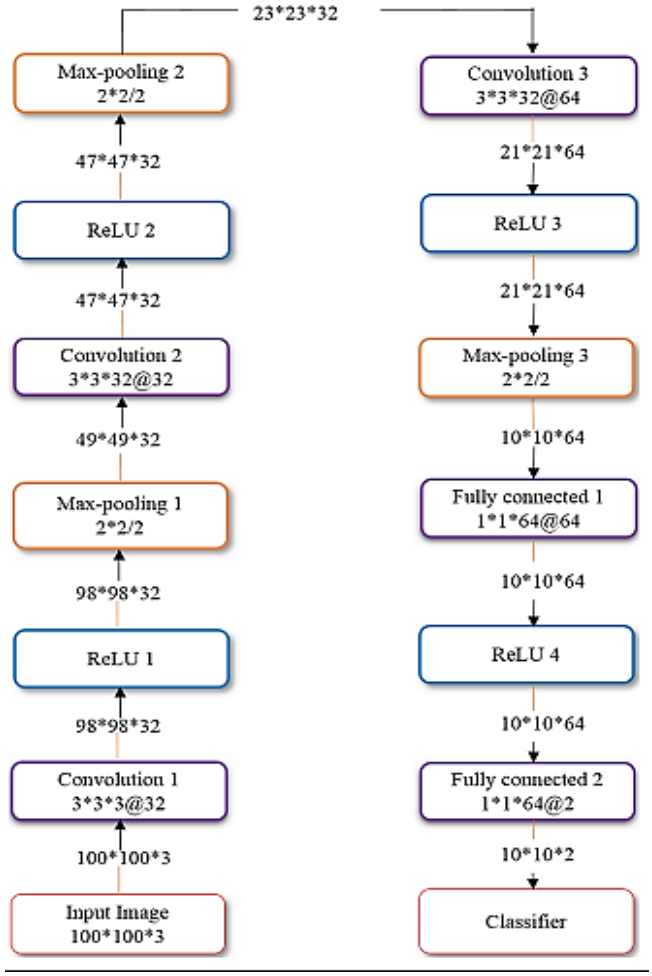


Figure 44 : Architecture CNN **(Rajaraman et al., 2018a)**

2. Modèles hybrides pour la détection de la malaria

2.1. Combinaison CNN et SVM

En plus des modèles CNN purs, plusieurs chercheurs ont combiné CNN et SVM pour améliorer encore plus la précision et la robustesse de la détection. Dans leurs travaux de recherches **Amin et al. (2024b)** ont proposé une méthode pour classer les parasites du paludisme, qui se compose de trois phases tel qu'illustré dans la **Figure 45**. Le filtre bilatéral est appliqué pour améliorer la qualité de l'image. Après cela, des caractéristiques basées sur la forme et des caractéristiques profondes sont extraites. Dans les histogrammes pyramidaux de gradients orientés (PHOG) basés sur la forme, les caractéristiques sont dérivées avec une dimension de N×300. Les caractéristiques profondes sont dérivées du réseau résiduel (ResNet)-50, et ResNet-18 aux couches entièrement connectées ayant respectivement une dimension de N×1000. Les caractéristiques obtenues sont fusionnées de manière sérielle, ce qui donne une dimension de N×2 300. À partir de cet ensemble, N×498 caractéristiques sont choisies en utilisant la méthode d'optimisation de la distribution normale généralisée (GNDO), ces dernières sont envoyées à un classifieur SVM pour prendre une décision sur la classe de l'échantillon examiné. La méthode proposée a été évaluée sur un ensemble de données d'imagerie microscopique de parasites du paludisme, offrant une précision de classification de 99%.

La **Figure 45** ci-après résume cette architecture hybride :

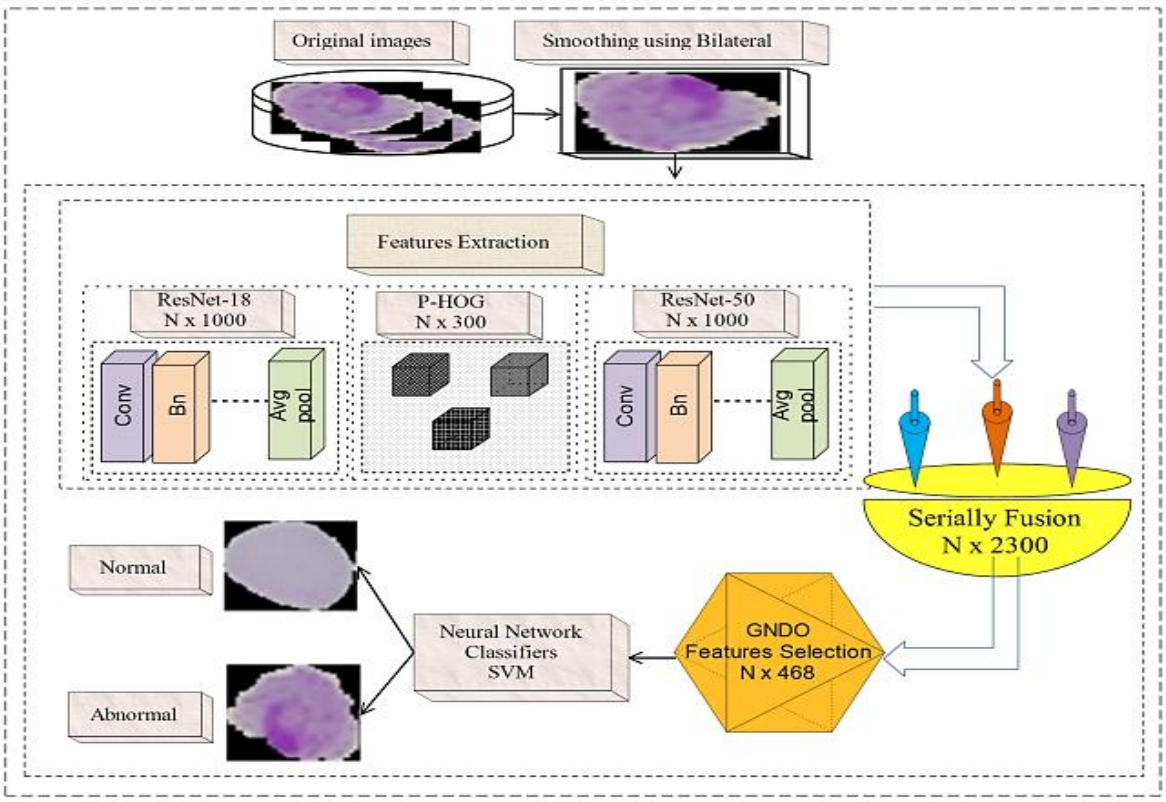


Figure 45 : Classification du paludisme en utilisant la fusion de caractéristiques artisanales et profondes par un modèle hybride CNN-SVM. **(Amin et al., 2024b)**

2.2. Modèle hybride CNN-KNN

Dans cette étude de **Wisit et al. (2019)**, les auteurs ont réalisé un modèle hybride utilisant deux modèles d'apprentissage un CNN et un KNN (voir la **Figure 46**). Comme on peut le constater à partir de l'architecture exposée dans la **Figure 44**, le CNN avec la configuration proposée, est utilisé principalement pour extraire les caractéristiques pertinentes à partir des images en entrée produisant des cartes de caractéristiques qui serviront d'entrée à un classifieur KNN. Le processus de reconnaissance des données utilisé dans l'enseignement est obtenu à partir de la couche entièrement connectée du CNN. Ce processus permet d'envoyer la valeur des principales caractéristiques ou variables des 1000 variables à l'algorithme KNN **(Zamil et al., 2019)** qui est une approche adaptée à la classification des données d'images médicales. En ce qui concerne les caractéristiques d'apprentissage de l'algorithme KNN, aucun modèle de classification n'est préétabli. Si de nouvelles données d'image doivent être classées, elles peuvent être comparées aux données existantes de la manière de classification ou de catégorisation des ensembles de données ayant des caractéristiques similaires. Le processus de traitement pour classifier les données d'image nécessite la détermination de k, qui est le nombre de données existantes proches de celles à classifier, en choisissant une valeur k appropriée, qui peut être considérée à partir des caractéristiques des classes de données. Afin d'obtenir une haute précision de classification, les auteurs, n'ont pas choisie une valeur de k élevée. En effet, le k choisit est égal à 3, ce qui permet d'avoir un taux de précision élevé car les données ont un petit nombre de caractéristiques principales et de classes, ce qui les rend plus faciles à classifier.

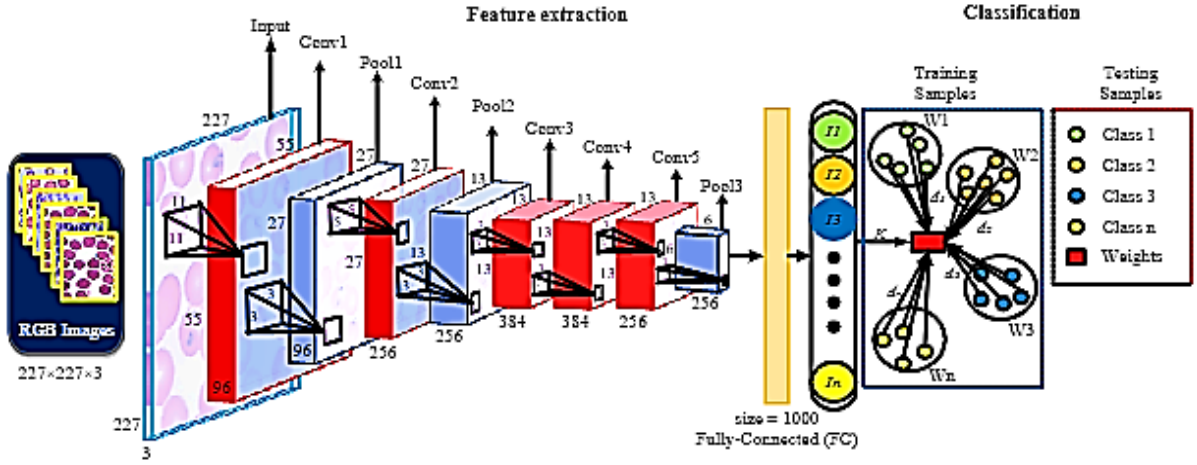


Figure 46. Architecture d'un modèle hybride CNN et KNN **(Wisit et al., 2019)**

2.3. Détection en temps réel avec YOLO

Le modèle **YOLO (You Only Look Once)** est utilisé dans des applications nécessitant une détection en temps réel. Dans leurs travaux de recherches portants sur la détection du pathogène du paludisme sur un ensemble de données d'images microscopiques à frottis épais capturées avec un téléphone portable, les auteurs ont développé deux modèles personnalisés, le premier, ayant trois couches YOLO-mp-3l et le second, quatre couches YOLO-mp-4l comme le montre les **Figures 47** et **48**. Ils ont obtenu les meilleurs scores avec une mesure de précision

moyenne (mesure-Average-Precision mAP) de 93,99 et 94,07, respectivement, surpassant le YOLOv4 standard qui a obtenu 92,56 pour la même mesure. Aussi, ils ont pu alléger leurs modèles pour une exécution rapide et optimale avec une taille de 21.8Mb pour le YOLO-mp-3l (**Figure 47**) et 25.4Mb pour le YOLO-mp-4l (**Figure 48**), ce qui est nettement meilleur que le YOLOv4 standard possédant une taille égale à 244Mb, prouvant ainsi leur capacité pour fonctionner sur des dispositifs à faibles ressources **(Anand et al. 2022)**.

Architectures YOLO-mp-3l et YOLO-mp-4l d'Anand, et al. (2022).

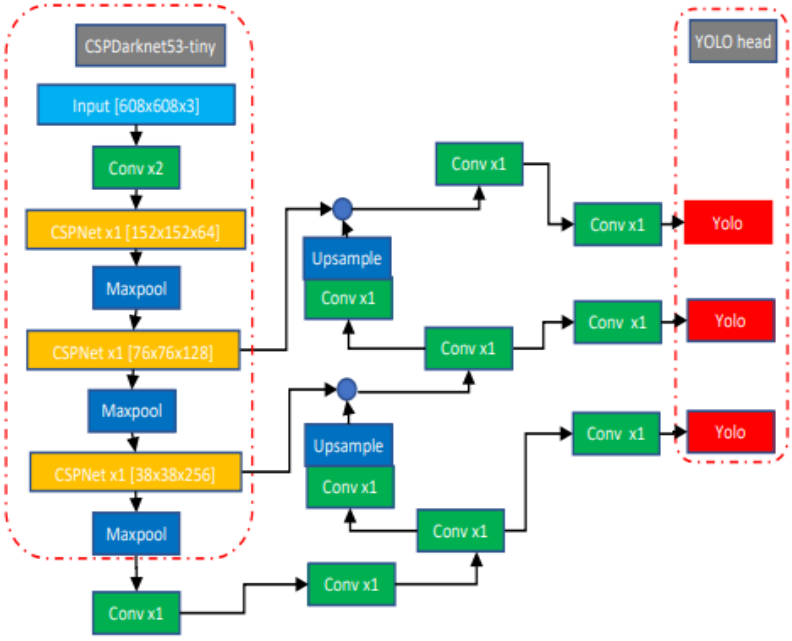


Figure 47. Architecture du YOLO-mp-3l **(Anand et al. 2022)**

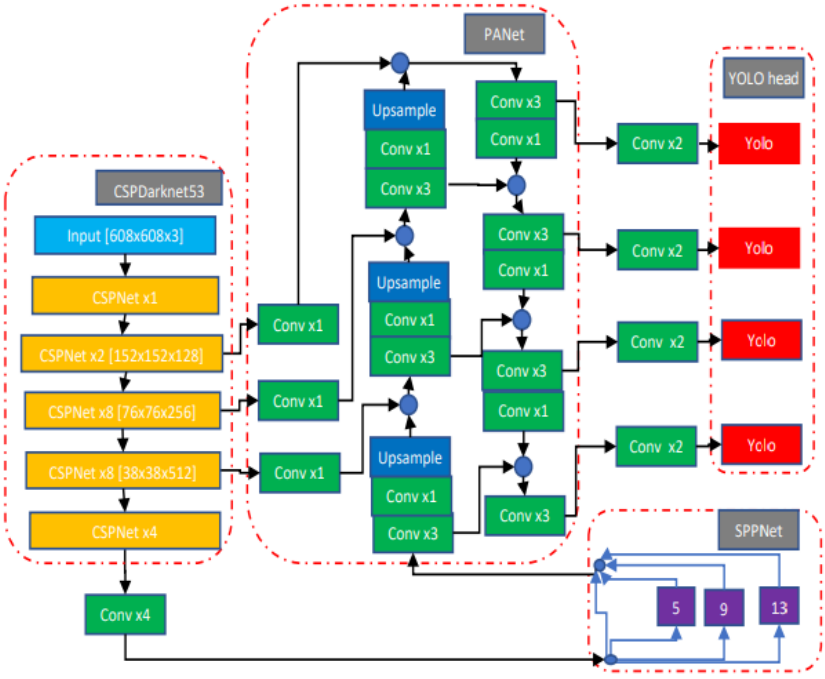


Figure 48. Architecture du YOLO-mp-4l **(Anand et al., 2022)**

Les progrès récents dans l'utilisation de l'apprentissage automatique pour la détection de la malaria, en particulier avec des architectures comme les CNN, les modèles hybrides CNN-SVM et YOLO, ont permis d'améliorer considérablement la précision et la rapidité du diagnostic. Ces technologies offrent des solutions viables pour la détection précoce de la malaria, ce qui est essentiel dans les régions à ressources limitées. Les architectures présentées ci-dessus ont montré qu'elles sont non seulement efficaces mais aussi adaptées à des environnements à faibles ressources, rendant ainsi le diagnostic plus accessible et plus rapide.

3. Techniques Avancées pour l'Explicabilité des Modèles IA

3.1. L'Importance de l'Explicabilité en Médecine : cas de la détection du paludisme

Un des défis majeurs de l'intelligence artificielle en médecine est la **compréhension des décisions** des modèles. L'IA, bien que puissante, peut manquer de transparence, ce qui peut freiner son adoption, notamment dans des domaines sensibles comme la médecine. L'explicabilité des modèles intelligents **XAI**, est cruciale pour que les médecins puissent comprendre et valider les décisions des systèmes de l'IA.

Des techniques comme **Grad-CAM**, **LIME** ou **SHAP** sont de plus en plus utilisées pour rendre ces modèles plus **transparents** et accessibles. Ces techniques permettent de **décoder** les décisions des modèles d'IA, ce qui est essentiel pour garantir leur **acceptation** clinique (**Rajab et al. 2023)**.

Des travaux de recherche notables dans le domaine du XAI dans le contexte de la détection automatique de la malaria, ont été menés. Offrant des visualisations expliquant les décisions prises par les modèles de l'IA et de l'apprentissage profond en particulier.

Selon Islam et al. (2022), un réseau basé sur la méthode Grad-CAM (**Figure 13**) a été utilisé pour expliquer les décisions prises par un modèle de transformer pour le diagnostic de la malaria. Ce modèle d'IA permet non seulement de prédire la présence de *Plasmodium* mais aussi de visualiser quelles parties de l'image ont été les plus influentes dans la décision, ce qui améliore la transparence et la confiance des cliniciens dans les résultats. Ce processus d'explication a été combiné avec des réseaux Transformer pour améliorer encore la précision du modèle, en particulier dans les cas où les parasites soient visibles sous un faible contraste **(Goni et al., 2023)**.

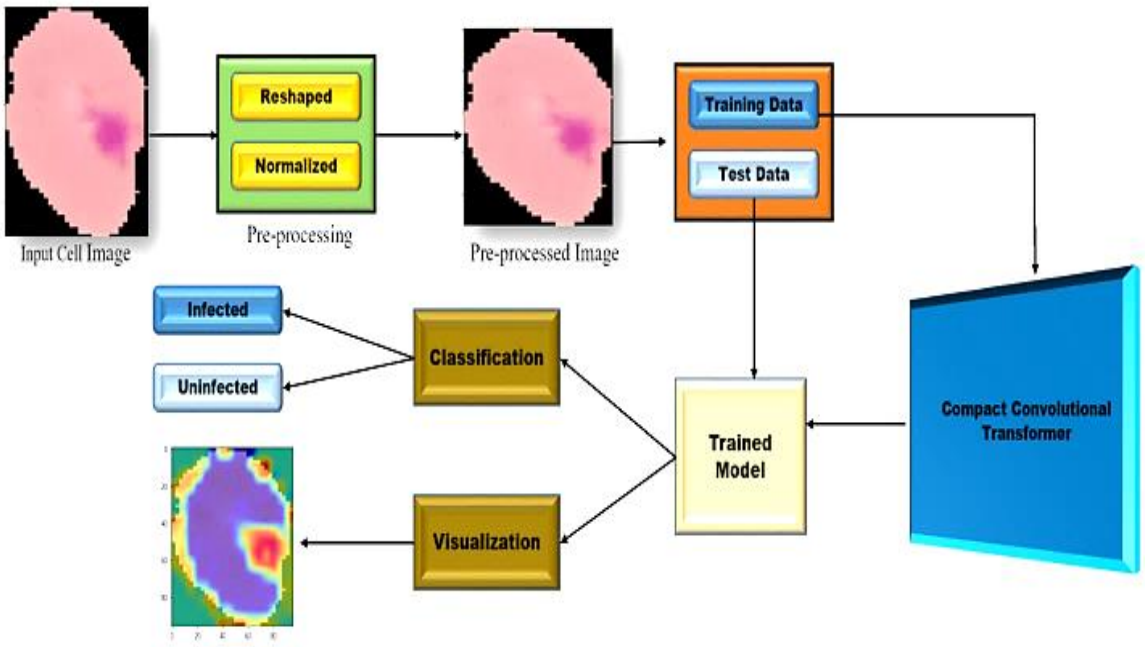


Figure 49. Explicabilité basée sur Grad-CAM dans la détection de la Malaria **(Islam et al., 2022)**

- **Grad-CAM** : L'utilisation de Grad-CAM dans le cadre de réseaux Transformer permet non seulement de prédire la présence de parasites, mais aussi de visualiser les zones de l'image les plus influentes dans la décision du modèle. Cela rend le modèle plus transparent, ce qui est essentiel pour l'interprétabilité et la confiance des cliniciens dans le diagnostic.

Algorithme de Grad-CAM :

1. **Calcul du gradient** : Le gradient de la sortie par rapport aux activations de la couche est calculé.

2. **Mappage d'activation** : Un mappage d'activation pondéré est généré pour indiquer les zones importantes de l'image influençant la décision du modèle.

3. **Superposition** : Le mappage est superposé à l'image pour visualiser les régions d'activation.

La **Figure 49** présente une illustration détaillée de l'architecture de l'algorithme Grad-CAM, représentant les trois étapes principales : le calcul du gradient, la génération du mappage d'activation et la superposition sur l'image d'origine. Cette visualisation peut nous aider à mieux comprendre le fonctionnement de Grad-CAM pour la détection de parasites dans les images.

Grad-CAM fonctionne en générant des cartes d'activation qui montrent quelles zones d'une image ont le plus influencé la prédiction du modèle, ce qui aide à comprendre le raisonnement du modèle et à assurer sa fiabilité. Cela est particulièrement utile pour l'interprétabilité du modèle d'IA et pour confirmer que le modèle se concentre sur les bonnes zones de l'image (les parasites).

Dans les travaux menés par **Rajab et al. (2023)**, les auteurs expliquent comment utiliser l'IA pour diagnostiquer le paludisme de manière claire. Ils utilisent des méthodes comme SHAP

et LIME pour donner des explications compréhensibles sur les prédictions de cas graves de paludisme faites par des modèles d'apprentissage automatique. Différents modèles comme Extreme Gradient Boosting, K-means, K-Nearest Neighbor, Support Vector Machine (SVM), Arbre de Décision, Régression Logistique (RL), Forêt aléatoire, Naive Bayes, AdaBoost et Explainable Boosting Machines (EBM) ont été utilisés pour cette tâche. L'étude a révélé que Random Forest et Explainable Boosting Machines ont obtenu la meilleure précision de 84 %. L'EBM a aussi aidé à mieux comprendre les traits qui permettent de faire des prévisions claires. La RL a obtenu une précision de 81 % en utilisant GridSearchCV pour améliorer ses prédictions. En plus, Ils ont utilisé la validation K-fold avec XGBoost pour évaluer la performance du modèle sur de nouvelles données. L'IA explicable (XAI) a aidé à mieux comprendre des éléments qui causent le paludisme sévère. Utiliser ces techniques peut améliorer beaucoup la précision des prévisions de paludisme grave et aider les médecins à faire de bons choix. Cet article explique pourquoi il est important d'utiliser des techniques XAI pour mieux diagnostiquer et traiter le paludisme grave en offrant une aide à la décision claire et compréhensible par les professionnels de la santé. Cette approche a été résumée dans la **Figure 50**.

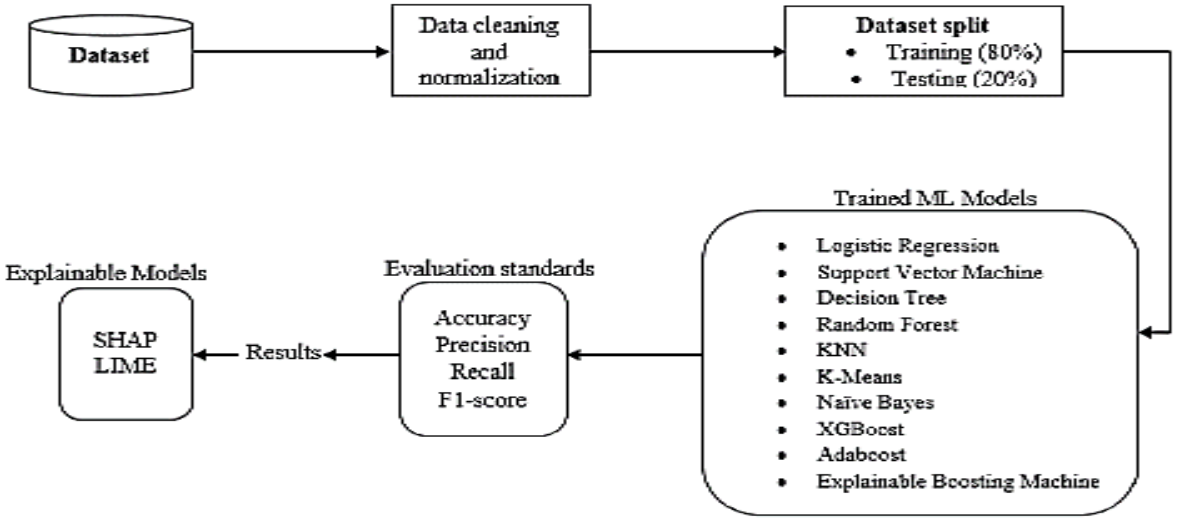


Figure 50. IA et techniques XAI LIME et SHAP pour la détection et la classification du paludisme **(Rajab et al. 2023)**

Les techniques de l'apprentissage supervisé, utilisent un très grand nombre de caractéristiques (qui peuvent ou non être pertinentes) et par conséquent, l'entraînement de tels modèles est coûteux en termes de calcul. De plus, et plus important encore, le grand espace de caractéristiques rend très difficile l'interprétation des caractéristiques qui sont réellement importantes pour les prédictions. **Khan et Al., (2020)**, abordent ces problèmes, dans leur article, et proposent une approche pour extraire un très petit nombre de caractéristiques agrégées qui sont faciles à interpréter et à calculer, et montrent empiriquement qu'ils obtiennent une grande précision de prédiction même avec un espace de caractéristiques considérablement réduit pour la détection de la malaria.

Dans cette étude, les auteurs ont remarqué que les globules rouges infectés par le paludisme ont une forme en anneau, alors que ceux qui ne sont pas infectés n'ont pas cette

forme. Leur méthode consiste à repérer l'entité en forme d'anneau. La **Figure 51**, résume cette démarche composée des étapes suivantes : D'abord, ils récoltent les informations générales des images de globules rouges. Après, ils utilisent un classificateur de type forêt aléatoire avec les caractéristiques qu'ils ont obtenues. Après un bon entraînement, ils utilisent ce classificateur pour identifier les globules rouges infectés par le paludisme et expliquer leurs choix. Pour être clairs et fiables, les auteurs avantagent les caractéristiques faciles à comprendre plutôt que celles cachées dans plusieurs niveaux d'un modèle complexe rendant leur modèle explicable et interprétable. Les métriques exposées pour cette étude sont résumées dans le **Tableau 3**.

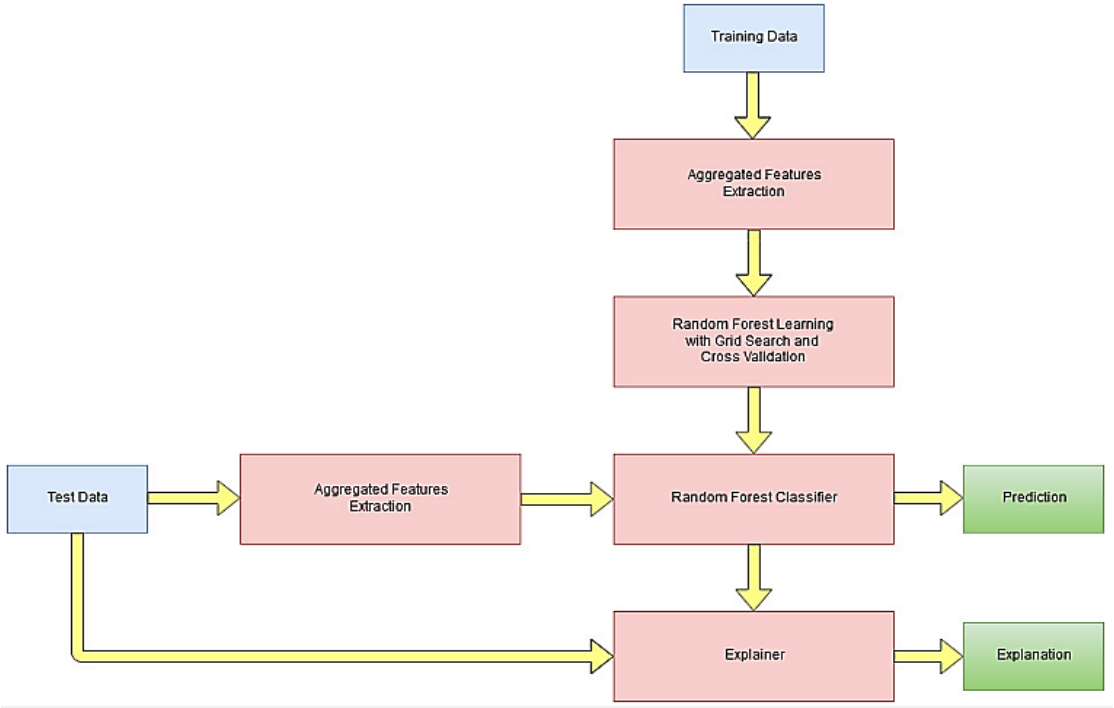


Figure 51. Diagramme de l'approche proposée par **Khan, et al., (2020)**

Tableau 3. Comparatifs des performances obtenues pour chaque algorithme **(Khan, et al., 2020)**

Learning Algorithm	Precision	Recall	F1 Score
Logistic Regression	0.84	0.75	0.79
Decision Tree	0.78	0.75	0.76
Random Forest	**0.82**	**0.86**	**0.84**

Dans le même contexte de l'explicabilité des modèles profonds pour la détection du paludisme, les travaux menés par **Mridha, et al. (2023)**, présentent l'automatisation du processus de diagnostic par le biais des technologies d'apprentissage profond, notamment les CNN, fondés sur les caractéristiques d'intensité des parasites Plasmodium et des érythrocytes.

L'approche implique l'injection d'images dans des modèles CNN tels que ResNet50, CNN et MobileNet, le modèle MobileNet affichant les performances globales les plus élevées. La principale innovation de cet article réside dans la mise à jour des modèles pré-entraînés, ce qui engendre des résultats supérieurs. Le système proposé combine l'apprentissage profond et l'intelligence artificielle explicable (XAI), offrant ainsi des explications claires et interprétables

pour les processus décisionnels, facilitant le développement d'outils de diagnostiques plus efficaces par l'utilisation de Grad-CAM et Grad-CAM++ pour identifier les zones impactées sur les images de frottis sanguin. Une étude de performance approfondie démontre que l'automatisation du processus permet de détecter avec précision et efficacité le parasite du paludisme dans les échantillons de sang, affichant une sensibilité supérieure à 95 % avec une complexité inférieure.

3.2. Explicabilité Basée sur Les Grands Modèles Linguistiques (Large Language Model, LLM) comme GPT :

L'une des approches les plus originales pour la détection du paludisme et de la typhoïde, est celle proposée par **Attai et al. (2024)**. L'originalité de cette étude demeure non seulement dans l'utilisation des modèles d'IA explicable (XAI), comme LIME, mais aussi et surtout, l'utilisation des LLM tels que GPT (Generative Pretrained Transformer), afin d'élucider les résultats du diagnostic pour les professionnels de la santé (**Figure 52**). Les résultats ont indiqué que le modèle Random Forest surpassait les performances des autres modèles évalués, atteignant un F1-Score = 71.45% ; de plus, des caractéristiques significatives ont été trouvées en utilisant des graphiques LIME, tandis que ChatGPT 3.5 a montré un avantage comparatif par rapport aux autres grands modèles de langage. La recherche combine RF, LIME et GPT pour développer une application mobile visant à améliorer l'interprétabilité et la transparence dans le diagnostic du paludisme et de la typhoïde. Malgré ses résultats encourageants, l'efficacité du système est limitée par la qualité du jeu de données. De plus, bien que LIME et GPT améliorent la transparence, ils peuvent compliquer la mise en œuvre en temps réel en raison de leurs exigences informatiques et de la nécessité d'une connexion Internet pour garantir la pertinence et la précision.

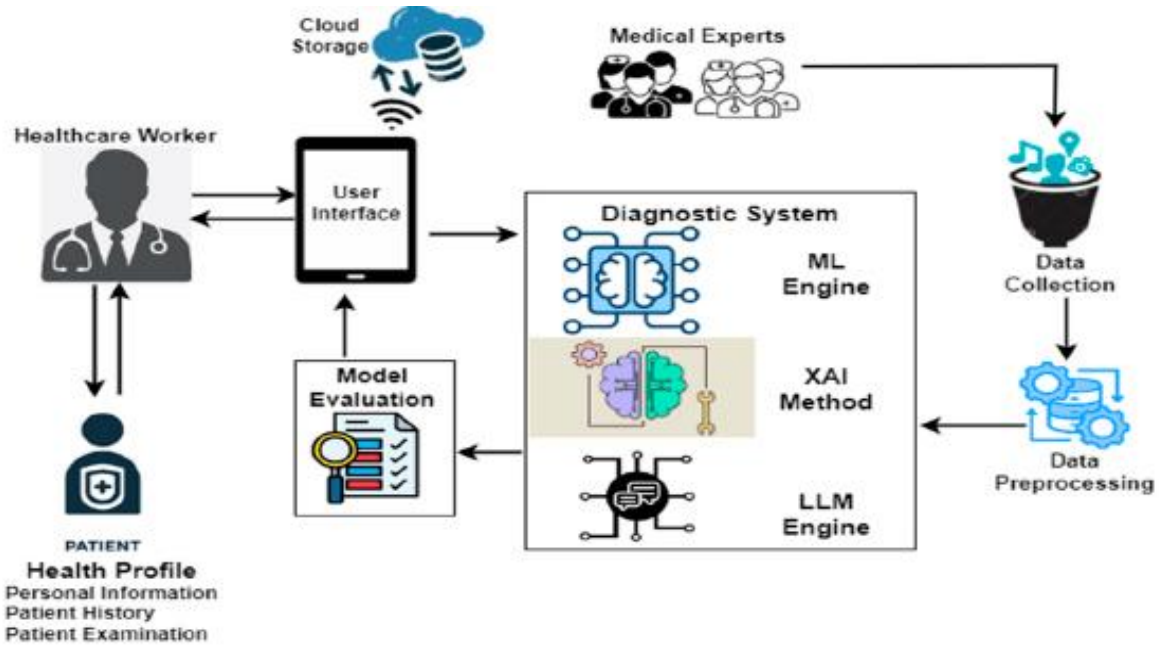


Figure 52. Cadre de diagnostic du paludisme et de la fièvre typhoïde **(Attai et al., 2024)**

3.3. Amélioration de l'explicabilité par les ViT (Vision Transformer)

L'une des dernières avancées dans le domaine du diagnostic du paludisme par l'**IA** implique l'utilisation de réseaux de type **Transformer.** Initialement utilisés en traitement du langage naturel, ils ont été récemment appliqués à l'analyse des images. Ces réseaux utilisent des mécanismes d'attention pour traiter efficacement des séquences d'images et extraire des informations pertinentes même dans des contextes complexes **(Islam et al., 2022)**. La **Figure 53**, montre l'architecture du Transformateur convolutionnel compact (CCT) tel que présenté dans cette étude. Ce dernier a été légèrement modifié du transformateur de vision auquel les auteurs ont ajouté une visualisation Grad-CAM pour l'explicabilité du modèle.

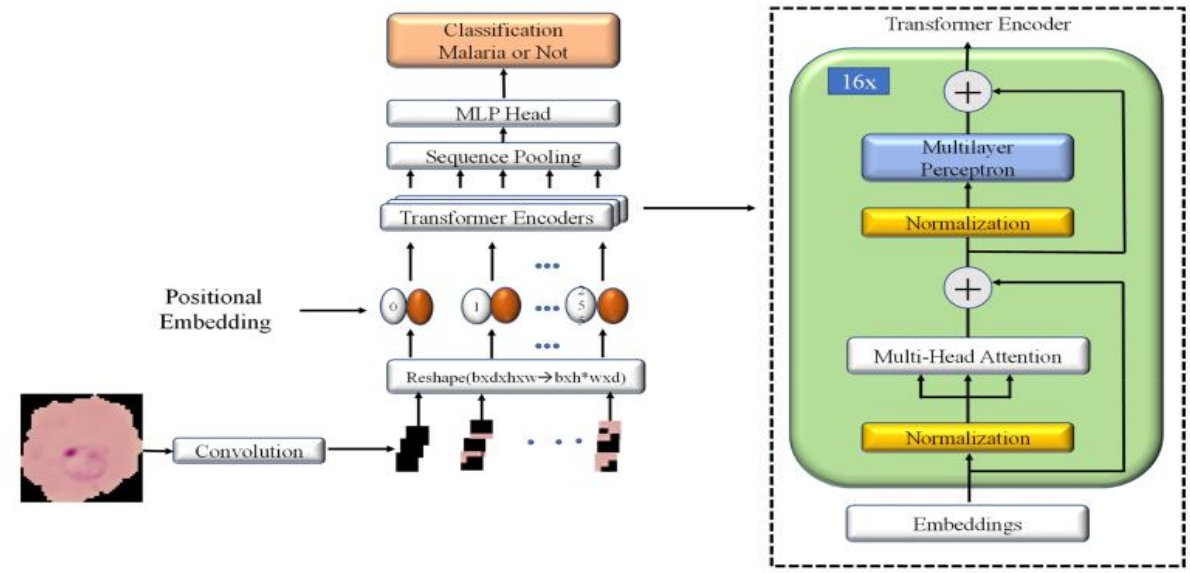


Figure 53. Architecture du modèle de transformateur convolutionnel compact (CCT) **(Islam et al., 2022)**

Dans le **Tableau 4** ci-après, nous résumons les travaux cités dans la détection automatisée et les techniques avancées pour l'explicabilité des modèles IA dans la détection du paludisme :

Tableau 4. Récapitulatif des travaux connexes

Référence	Année	Méthode	Technique	Précision obtenue	Procédé	Avantages
Khan et al. (2020)	2020	LIME (Local Interpretable Model-Agnostic Explanations)	Explicabilité locale des décisions	Varia	Interprétation locale des prédictions pour comprendre la logique du modèle	Facilite la transparence et la confiance des utilisateurs dans les décisions automatisées.
Yang et al. (2020)	2020	YOLO (You Only Look Once)	Détection d'objets en temps réel	79 %	Détection binaire de la présence de Plasmodium vivax dans les frottis sanguins	Rapide, adapté aux applications en temps réel.

Kassim et al. (2021**a**)	2021	PlasmodiumVF-Net	Mask R-CNN + ResNet50	90 %	Mask R-CNN pour la détection initiale des parasites et ResNet50 pour la classification	Amélioration de la classification des espèces de Plasmodium, segmentation précise.
Kassim et al. (2021**b**)	2021	RBCNet	U-Net + Faster R-CNN	97 %	U-Net pour la segmentation, Faster R-CNN pour la détection et la classification	Segmentation et localisation précises des cellules infectées.
Ufuktepe et al. (2021)	2021	Classificateurs SVM	Machines à vecteurs de support	99 %	Identification des cellules infectées après prétraitement et sélection de caractéristiques	Haute précision, mais moins flexible pour les images complexes ou de grande taille.
Islam et al. (2022)	2022	XAI (Explained AI)	Grad-CAM	Varia (84 % pour LIME)	Visualisation des zones influentes de l'image sur la décision du modèle	Amélioration de l'interprétabilité, aide à comprendre les décisions du modèle.
Goni et al. (2023)	2023	CNN (Convolutional Neural Networks)	Réseaux neuronaux convolutifs	99,66 %	Extraction des caractéristiques des frottis sanguins et classification des cellules infectées	Gestion des images bruitées, performances élevées sur des données de qualité variable.
Qadri et al. (2023)	2023	NASNet + Random Forest	Apprentissage par transfert + Forêts aléatoires	99 %	NASNet pour l'extraction de caractéristiques, Random Forest pour la classification	Apprentissage par transfert pour des caractéristiques robustes, amélioration de la précision.
Amin et al. (2024)	2024	CFPNet-M	Réseau de pyramides de fonctionnalités par canal	92,2 %	Extraction de caractéristiques multiresolutions pour la détection et la classification	Meilleure robustesse face aux variations de taille et forme des cellules infectées.

4. Impact de l'Intelligence Artificielle dans la Lutte Contre la Malaria

4.1. Réduction du fardeau sanitaire grâce à l'IA

L'intégration de l'IA dans le diagnostic du paludisme permet de réduire le fardeau sanitaire en améliorant la rapidité et la précision des diagnostics. Les systèmes d'IA peuvent analyser les images de manière quasi instantanée, permettant ainsi un traitement rapide et une gestion efficace des cas. Ces systèmes, notamment les réseaux de neurones convolutifs (CNN), ont montré des résultats prometteurs, non seulement pour diagnostiquer les infections à *Plasmodium falciparum*, mais aussi pour détecter les autres espèces du parasite avec une grande précision.

Les diagnostics rapides améliorent les taux de guérison et contribuent à la réduction des complications graves. En permettant des interventions précoces, l'IA aide à réduire les hospitalisations prolongées et les traitements coûteux. Cela est particulièrement crucial dans les zones à ressources limitées, où l'accès aux soins est souvent restreint.

4.2. Accélération des processus diagnostiques

L'IA permet de traiter un plus grand nombre de cas en peu de temps. Contrairement aux méthodes manuelles qui nécessitent du temps pour préparer les échantillons, les tests et l'analyse des résultats, les systèmes d'IA peuvent analyser des images et des échantillons de manière autonome, libérant ainsi les ressources humaines et accélérant le processus de diagnostic. Ce gain de temps est particulièrement important dans les situations d'urgence, où chaque minute compte pour sauver une vie.

Les modèles d'IA utilisés pour l'analyse des frottis sanguins peuvent, dans certains cas, analyser des milliers d'échantillons par jour, ce qui était inimaginable avec des méthodes manuelles. Cela permet de traiter un plus grand nombre de patients, réduisant ainsi la charge sur les laboratoires et les médecins, tout en augmentant la capacité de diagnostic.

5. Défis et Limitations de l'IA dans le Diagnostic de la Malaria

5.1. Limites de l'IA en contexte réel

Bien que l'IA offre de nombreuses possibilités pour améliorer le diagnostic du paludisme, son utilisation n'est pas sans défis. L'un des principaux obstacles à l'adoption de l'IA dans les pays en développement est l'accès limité aux technologies nécessaires pour implémenter ces solutions. Les systèmes d'IA, notamment ceux utilisés pour l'analyse d'images médicales, nécessitent un matériel informatique sophistiqué (comme des processeurs graphiques puissants) et une connectivité Internet stable, ce qui peut être problématique dans des régions rurales ou isolées.

De plus, bien que les modèles d'IA aient montré d'excellents résultats en laboratoire, leur performance en situation réelle peut parfois être affectée par des facteurs externes, tels que la qualité des images ou la variabilité des échantillons. Les systèmes peuvent également rencontrer des difficultés dans l'identification des stades précoces ou des formes atypiques de l'infection, ce qui représente un défi supplémentaire dans les régions où la malaria est particulièrement endémique.

5.2. Problèmes liés à l'interprétabilité des modèles

L'un des défis majeurs de l'IA dans le domaine médical est l'interprétabilité des modèles. Alors que les modèles d'apprentissage automatique peuvent fournir des prédictions très précises, il peut être difficile pour les médecins de comprendre comment ces prédictions ont été formulées. Cette absence de transparence peut réduire la confiance des praticiens dans les résultats fournis par l'IA, ce qui pourrait freiner son adoption.

Des efforts sont en cours pour améliorer l'explicabilité des modèles IA à travers des techniques comme les cartes de chaleur (Heatmaps) ou les méthodes d'explication de l'importance des caractéristiques (Feature Importance). Cependant, il reste encore beaucoup de travail pour garantir que ces systèmes d'IA sont non seulement efficaces, mais aussi compréhensibles et fiables pour les professionnels de la santé.

5.3. Besoin de données de qualité et de formation

La performance des modèles d'IA dépend fortement de la qualité et de la diversité des données utilisées pour les entraîner. Dans le cas de la malaria, cela signifie qu'il faut disposer d'un grand nombre d'images de frottis sanguins provenant de diverses régions géographiques, types de patients et stades de la maladie. Les systèmes d'IA doivent être formés pour reconnaître les parasites dans des conditions variées (qualité d'image, éclairage, type de matériel utilisé, etc.), ce qui nécessite un ensemble de données robuste et bien annoté.

Le manque de données annotées et de collaboration internationale pour la collecte de ces données demeure un défi majeur. De plus, les professionnels de la santé doivent être formés à l'utilisation des systèmes d'IA, et les programmes de formation doivent être intégrés aux cursus de formation médicale pour garantir une adoption réussie.

6. Applications Pratiques de l'IA dans la Lutte Contre la Malaria

6.1. Développement d'applications mobiles pour le diagnostic

L'une des applications les plus prometteuses de l'IA dans le diagnostic du paludisme est le développement d'applications mobiles accessibles. Ces applications permettent aux professionnels de santé, même dans des zones reculées, de diagnostiquer rapidement la malaria en utilisant des smartphones et des dispositifs portables comme des microscopes ou des caméras de faible coût.

Des start-ups et des entreprises technologiques ont développé des applications qui intègrent des algorithmes d'IA pour analyser les images des frottis sanguins, permettant ainsi un diagnostic rapide et fiable. Ces technologies ont le potentiel de révolutionner le diagnostic de la malaria dans les régions les plus touchées, en fournissant une solution pratique et évolutive.

6.2. Automatisation des processus de laboratoire

L'automatisation des laboratoires est une autre application de l'IA dans la lutte contre le paludisme. Les systèmes automatisés peuvent non seulement analyser les échantillons de manière plus rapide et précise, mais aussi réduire les risques d'erreurs humaines. Cela permet

d’augmenter l'efficacité du traitement des échantillons et de réduire les coûts associés à l'utilisation de technologies avancées.

L'IA pourrait également être utilisée pour analyser les tendances épidémiologiques, en étudiant les facteurs environnementaux, climatiques et sociaux qui influencent la propagation de la maladie. Cela pourrait fournir des informations cruciales pour la gestion des épidémies de malaria et aider les autorités sanitaires à prendre des décisions éclairées sur les stratégies de lutte.

VII. Conclusion

L'intelligence artificielle (IA) représente un progrès majeur dans le diagnostic du paludisme, apportant des solutions plus rapides, plus accessibles et plus fiables pour détecter cette maladie endémique. Bien que des défis demeurent, notamment en matière d'infrastructure, de disponibilité des données et de l'interprétabilité des modèles, les technologies avancées comme les réseaux neuronaux et l'apprentissage automatique ouvrent des perspectives prometteuses. Ces outils, combinés à des solutions mobiles, peuvent transformer la manière dont le paludisme est diagnostiqué, notamment dans les régions rurales et mal desservies.

L'intégration de méthodes d'IA explicables, telles que **LIME** et l'utilisation de **réseaux LSTM**, offre des avancées notables pour non seulement améliorer la précision des diagnostics, mais aussi garantir une meilleure compréhension des décisions prises par les systèmes. L’interprétabilité des modèles est essentielle pour renforcer la confiance des professionnels de santé et faciliter l’adoption des technologies dans des contextes cliniques réels.

En combinant les techniques traditionnelles de diagnostic et les solutions innovantes offertes par l'IA, l'avenir de la lutte contre la malaria pourrait se dessiner comme une alliance entre l'humain et la machine, visant à rendre les soins de santé plus rapides, plus fiables et plus accessibles à une population mondiale en constante croissance. Ce chapitre met en lumière l'énorme potentiel de l'IA dans ce domaine, et souligne l'importance de continuer à investir dans la recherche, l'amélioration des infrastructures et l'expansion de l'accès à ces technologies pour une lutte mondiale durable contre le paludisme.

Enfin, nous dirons que cette revue de la littérature (bien qu’elle n’en soit pas une) met en lumière l'évolution du paysage du diagnostic du paludisme, où les avancées en IA et en explicabilité entraînent des améliorations significatives de la précision et de l'accessibilité des diagnostics.

CHAPITRE IV

MALARIASCOPE : DÉTECTION AUTOMATIQUE DU PALUDISME & EXPLICABILITÉ DES RÉSULTATS

CHAPITRE IV

MalariaScope : Détection Automatique du Paludisme et Explicabilité des Résultats

Sommaire

I. Introduction

Au cours de cette dernière décennie, l'apprentissage profond, communément nommé Deep Learning (**DL**), a connu un essor des plus impressionnants dû à ses performances extraordinaires, notamment en apprentissage automatique (Machine Learning) (**ML**) **(Kumar et al., 2020a)**, ou, comme nous l'avons vu plus récemment, l'apprentissage automatique quantique (QML), qui offre des opportunités pour de futures avancées dans le traitement des données de santé de haute dimension et l'amélioration des résultats cliniques **(Chow, 2025)**.

De même, la vision par ordinateur est devenue incontournable, surtout dans le domaine de la reconnaissance et la classification des images, qui constituent la base d'autres tâches aussi importantes dans le domaine de la vision par ordinateur telles que la localisation, la détection et la segmentation **(Rawat and Wang, 2017)**.

Grâce aux nouvelles technologies, nous avons vécu, ces dernières années, dans un monde hautement visuel. Le hasard fait bien les choses, plus d'un tiers de l'ensemble du cerveau humain est sollicité dans le traitement et la compréhension optique et visuel. En effet, des études ont montré que le système visuel humain est hautement performant pour capter des sémantiques de haut niveau à partir des scènes réelles désordonnées, tels que des objets, des classes de scènes, des activités et des histoires à partir de l'observation des images **(Fei-Fei and Li, 2010)**.

Dans le même contexte, il faut savoir que l'une des premières étapes de classification d'images, relève de la capacité d'un programme informatique à découvrir et à encoder les concepts sémantiques des images en formats numériques. Pour ce faire, l'algorithme informatique, doit d'abord être capable de comprendre la sémantique qui se cache derrière les images analysées, afin d'extraire des paramètres visuels efficaces et performants, lui permettant de construire des modèles intelligents de classification d'images. Ce processus est communément nommé **« identification et extraction des caractéristiques »**. Notons tout de

même à quel point le type et la pertinence des ressources extraites jouent un rôle prépondérant dans la multitude des tâches de traitement multimédia **(Gkelios et al., 2021)**.

Entre autres objectifs, la vision par ordinateur a cherché des moyens d'identifier, de représenter et de classifier les informations visuelles existantes, présentes en grande quantité pour différents usages, allant de la détection et l'identification des objets et des personnes à des diagnostics médicaux complexes et difficiles. Dans le même contexte, et en s'inspirant de la biologie humaine et animale des réseaux de neurones complexes, ont émergé parmi eux, les Réseaux de Neurones Convolutifs (CNN) qui sont sans doute les plus connus et les plus anciens, dont les origines remontent aux années 1960 ; présentant une architecture de connectivité entre les neurones, inspirée par l'organisation du cortex visuel de l'animal.

Dans des travaux de recherches scientifiques menés par **Hubel et Wiesel en 1959**, ils ont prouvé que ce n'est que dans des régions restreintes du champ visuel, connues sous le nom de champs récepteurs, que les neurones corticaux individuels répondent aux stimuli. Ainsi, l'ensemble du champ de vision est couvert par le chevauchement partiel des champs récepteurs des différents neurones. Ils ont aussi observé et proposé pour la première fois le concept de « champ réceptif » et le mécanisme de traitement hiérarchique de l'information dans les voies corticales visuelles. En effet, ils ont révélé que les cellules simples détectent les informations de localisation et que les cellules complexes intègrent les informations stimulées par les cellules simples **(Wang et al., 2019)**.

Les CNN ont pour principal avantage leur capacité à apprendre des représentations, c'est-à-dire à extraire des caractéristiques de données visuelles. De cette manière, ces modèles tendent à exiger un niveau minimal de prétraitement par rapport à d'autres algorithmes de classification d'images **(Pereira-Ferrero et al., 2023)**. Les CNN ont été très utilisés dans la reconnaissance d'images et le traitement vidéo, bien qu'ils aient déjà été appliqués avec succès dans des expériences impliquant le traitement de la voix et le langage naturel **(Matsugu et al., 2003)**.

Bien que les modèles profonds DL aient connus des avancées extraordinaires, soutenues principalement par des caractéristiques profondes basées sur les CNN, un obstacle majeur demeure lié à l'absence d'informations contextuelles dans de telles représentations. En effet, ces dernières se trouvent souvent sur des variétés, dans un espace de haute dimension **(Iscen et al., 2018)**, où la formulation par paire de la mesure de similarité est insuffisante pour révéler la relation intrinsèque entre les images.

Dans un tel état des lieux, notre travail se concentre sur la manière de trouver une représentation des caractéristiques plus efficace en tenant compte des relations de similarité contextuelles, définies à partir des caractéristiques extraites. Pour atteindre ce but, nous avons utilisé des Réseaux de Neurones Récurrents (RNN), plus spécialement les réseaux Long Short-Term Memory empilés (Stacked-LSTM networks) que nous verrons en détails un peu plus loin dans cette partie.

Ce chapitre est consacré à la présentation de l'ensemble de méthodes et de techniques expérimentées dans le cadre des travaux menés dans cette thèse de doctorat, pour la détection automatique et précoce du paludisme, en utilisant les techniques de traitement d'images,

d'apprentissage automatique supervisé et d'apprentissage profond, tout en mettant l'accent sur l'interprétabilité des résultats par le biais des techniques d'explicabilité XAI.

II. Matériels et Méthodes

L'approche proposée dans cette étude comprend un pipeline constitué de trois étapes principales :

La première étape consiste à prétraiter des données, afin de préparer les images de frottis sanguins pour l'analyse.

La deuxième étape se concentre sur la classification. Cela nous permet de tester et de comparer les résultats de six modèles différents, à savoir VGG-16, VGG-19, MobileNetV2, ViT, Stacked-LSTM, sans mécanisme d'attention et Stacked-LSTM avec mécanisme d'attention. Comme nous le verrons plus loin dans ce chapitre, ces modèles ont été personnalisés pour détecter les cas de paludisme avec une grande précision.

Enfin, la troisième étape compare les résultats en utilisant deux techniques d'explicabilité, Grad-CAM et LIME, pour interpréter et comprendre les décisions prises. Cette méthodologie a été proposée pour combiner à la fois une performance de classification solide et une transparence accrue, offrant ainsi un cadre robuste et interprétable pour la détection du paludisme.

1. Collecte des données et description du dataset (NLM - Malaria Data)

A. Source des données du Dataset

Le dataset nommé NLM - Malaria Data, utilisé dans le cadre de nos travaux sur la détection automatique de la malaria, est disponible publiquement sur le site web de la bibliothèque nationale de médecine[3].

B. Caractéristiques des données

Le dataset est équilibré, avec un nombre égal d'images pour les cellules infectées et non infectées, à divers endroits, tels que l'hôpital de la Faculté de Médecine de Chittagong au Bangladesh et Bangkok, en Thaïlande, comme on peut le constater dans la section suivante.

Ce jeu de données ainsi diversifié et équilibré, garantit un apprentissage efficace des modèles intelligents que nous allons entraîner sur les deux classes ; ce qui va nous permettre de fournir une base solide pour la détection du paludisme.

[3] https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html

C. Processus de collecte des données

Selon les propriétaires du dataset, la construction de ce dernier s'est faite par étapes, comme indiqué ci-dessous :

- Étape 1 : Échantillons de sang de 150 individus infectés par *Plasmodium falciparum* à l'Hôpital de Médecine de Chittagong, Bangladesh **(Yu et al., 2020 ; Yang et al., 2020)**.
- Étape 2 : échantillons de 150 individus infectés par *Plasmodium vivax* **(Yang, et al., 2019 ; Kassim et al., 2021a)**.
- Étape 3 : échantillons de 50 individus non infectés **(Yang et al., 2019 ; Kassim et al., 2021a).**
- Étape 4 : échantillons de 148 individus infectés par *Plasmodium falciparum* et 45 individus non infectés **(Rajaraman et al., 2018a ; Rajaraman et al., 2018b ; Rajaraman et al., 2019 ; Kassim et al., 2021b).**
- Étape 5 : échantillons de 171 individus avec *Plasmodium vivax** à Bangkok, Thaïlande **(Yang et al., 2020).**

D. Composition du dataset

Le dataset final se compose de 27 558 images, représentant équitablement des cellules parasitées et non parasitées.

L'ensemble de données est divisé en deux parties distinctes : un ensemble de polygones et un ensemble de points. La différence entre ces deux ensembles réside dans la méthode d'annotation. Dans l'ensemble des polygones, tous les globules rouges (GR) et les globules blancs (GB) ont été délimités manuellement avec des polygones en utilisant l'outil d'annotation Firefly, tandis que dans l'ensemble des points, les cellules ont été marquées en plaçant un point sur chaque cellule.

E. Acquisition des images

Les auteurs des travaux cités dans la section **C** ont permis de collecter les données, tel que nous l'avons expliqué dans la section **C**. Ils ont photographié des frottis sanguins fins, colorés au Giemsa, de 714 patients (suivant la répartition exposée dans la section **C**) en utilisant un appareil photo de smartphone qu'ils ont fixé à l'oculaire d'un microscope optique (ce type de microscope peut aller jusqu'à 1000X pour agrandir les zones observées).

Cet échantillonnage, a permis de capturer 27558 images, avec un agrandissement de 100x dans l'espace colorimétrique RGB, avec une très haute résolution de 3024×4032 pixels et sont fournies au format JPG. Par la suite, un filtre a été appliqué pour éliminer les bruits d'acquisition. Les images ainsi obtenues sont de très haute qualité comme on peut le voir dans les **Figures 54, 55, 56, 57 et 58**.

Des experts en frottis sanguins ont annoté manuellement chacune des images ainsi capturées. Ensuite, les chercheurs ont désidentifié (ils les ont rendues anonymes) toutes les

images et leurs annotations, et les ont archivées dans la Bibliothèque Nationale de Médecine (NLM - Malaria Data).

Outre les images, le dataset fournit aussi des documents explicatifs ainsi que des statistiques définissants le nombre de GR sains et infectés, ainsi que le nombre de GB.

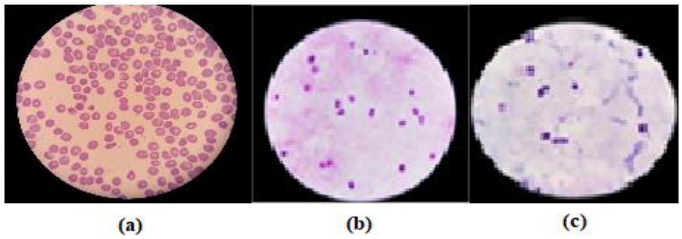


Figure 54. Exemples d'images du dataset ***NLM - Malaria Data***. (a) représente un frottis sanguin, (b) représente un globule rouge sain, (c) représente un globule rouge infecté.

Comme nous l'avons vu au chapitre I, il existe 4 types de parasites responsables du paludisme. Les **Figure 55** et **56** présentent un exemple de ces derniers par images fournies par le dataset ***NLM - Malaria Data.***

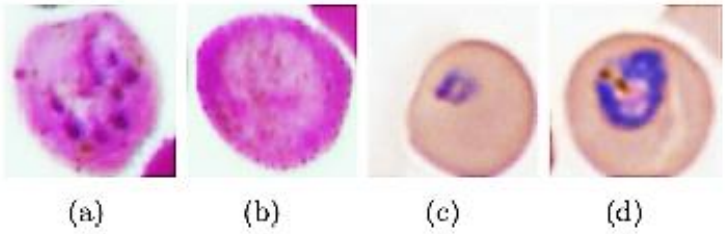


Figure 55. Exemples d'images décrivant les différents types du parasite Plasmodium. (a) *P. falciparum*, (b) *P. vivax*, (c) *P. ovale* et (d) *P. malariae*.

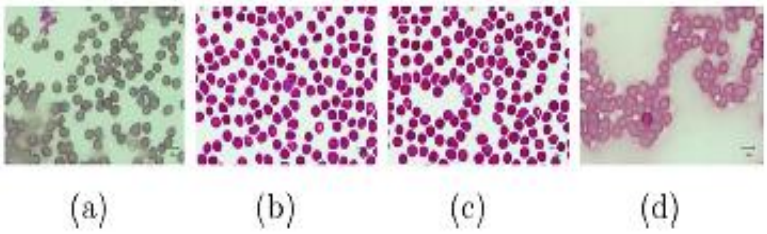


Figure 56. Exemples d'images du dataset : (a) anneau, (b) schizonte, (c) trophozoïte, et (d) trophozoïte de vivax

Figure 57. Gamétocytes de *P. falciparum* dans les globules rouges observés au microscope.

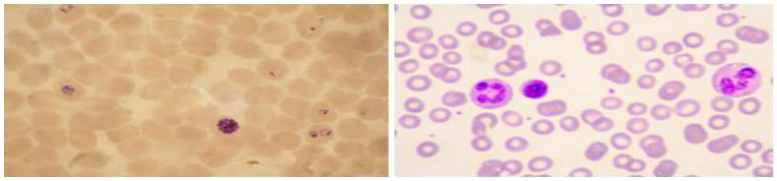


Figure 58. Frottis sanguin d'une culture de *P. falciparum*. Plusieurs globules rouges comprennent des anneaux. Vers le centre, une schizonte est visible, et un trophozoïte à gauche.

Au cours des 7 à 10 jours suivant l'infection, ils prolifèrent dans les cellules hépatiques sans déclencher de symptômes notables. 8 à 30 jours après l'infection, une fièvre se développe. Cela peut être accompagné de faiblesse, de maux de tête, de douleurs musculaires, de vomissements, de diarrhée et/ou de toux. La fièvre accompagnée de tremblements, de sueurs froides et de transpiration intense peut survenir de manière cyclique en raison des différentes phases du cycle du parasite. Des symptômes plus graves peuvent survenir, tels que des difficultés respiratoires, des saignements, de la jaunisse, une fatigue extrême et des convulsions. Dans certains cas, les globules rouges infectés peuvent obstruer les vaisseaux sanguins qui alimentent le cerveau, ce qui peut être fatal.

Pour confirmer ou exclure un diagnostic de paludisme, un échantillon de sang doit être analysé à l'aide de tests parasitologiques. Le test standard consiste en un examen microscopique des frottis sanguins (frottis minces et épais). Une formation et une expérience approfondies sont requises et essentielles pour analyser correctement les frottis sanguins, en particulier pour interpréter les frottis épais, identifier les espèces de parasites et quantifier la parasitémie.

L'absence de personnel expérimenté peut limiter la précision du diagnostic du paludisme, et, dans la majorité des cas, le paludisme ne peut pas être diagnostiqué dans les premières heures suivant l'infection du patient avec la méthode conventionnelle. C'est pourquoi nous voulons dépister et détecter le parasite tôt (dans les premières heures de contamination), en analysant les images de frottis sanguins avec nos modèles intelligents, avant qu'il ne se développe et ne devienne contagieux. D'une part, pour prévenir sa propagation au sein de la population, et d'autre part, pour s'assurer que l'individu commence le traitement avant que sa santé ne se détériore ou que son pronostic ne devienne menaçant pour sa vie **(Baer et al., 2007)**.

2. Prétraitements des données

Dans le cadre de nos travaux, le Dataset a été divisé en trois sous-ensembles différents, pour évaluer rigoureusement les performances des différents modèles implémentés : 50 % pour l'entraînement, 20 % pour la validation et 30 % pour les tests. Cette répartition garantit une distribution équilibrée pour un entraînement robuste et une évaluation fiable.

Une série d'étapes de prétraitement a été appliquée aux images des cellules infectées et non infectées, pour préparer les données aux modèles d'apprentissage automatique.

2.1. Redimensionnement des images

Tout d'abord, les images ont été chargées à partir du Dataset et redimensionnées à 224×224 pixels afin de les alléger et rendre leurs traitements plus rapides, tout en gardant les informations essentielles pour la suite des prétraitements.

2.2. Augmentation des données

Les techniques d'augmentation des données, y compris le recadrage aléatoire, les rotations de 45° et 75°, et le flou gaussien avec un noyau de 10 × 10, ont été appliquées pour améliorer la diversité du jeu de données et prévenir le surapprentissage. Nous avons assigné à chaque image ainsi obtenue une étiquette : 1 pour les cellules infectées et 0 pour les cellules non infectées.

2.3. Conversion en niveaux de gris

Après l'augmentation des données, les images ont subi une conversion en niveaux de gris pour réduire la complexité computationnelle en éliminant les informations colorimétriques. Cette conversion a été réalisée à l'aide du modèle de luminance perceptuelle :

$$\text{Niveau de gris} = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \quad \textbf{(10)}$$

Avec R, G et B les trois canaux de l'image couleur. Notez que ces coefficients sont basés sur le modèle de luminance perceptuelle et donnent plus de poids au vert, car l'œil humain est plus sensible à cette couleur.

2.4. Segmentation et identification des régions d'intérêts (Region Of Interest (ROI))

2.4.1. Segmentation

Après la conversion en niveaux de gris, une segmentation par seuillage a été appliquée en utilisant la méthode d'Otsu, qui nous a permis de binariser les images (noir et blanc), tout en mettant en évidence les zones d'intérêt potentielles et identifier efficacement les objets (cellules) de l'arrière-plan, comme le montre la **Figure 59**.

L'extraction des caractéristiques discriminantes, facilite l'analyse, l'interprétation et l'extraction d'informations pertinentes, parmi lesquelles, nous pouvons citer :

- **Caractéristiques géométriques :**
  - Taille (aire des cellules ou des parasites) ;
  - Forme (circularité, aspect ratio, convexité).
- **Caractéristiques d'intensité et de texture :**
  - Moyenne et variance de l'intensité (histogramme de niveaux de gris).
  - Caractéristiques d'Haralick pour la texture (contraste, homogénéité, entropie).
- **Caractéristiques biologiques :**

Présence d'anomalies spécifiques (anneaux de Plasmodium, pigments d'hémozine).

Dans la **Figure 59**, nous illustrons un exemple du résultat obtenu par une telle segmentation.

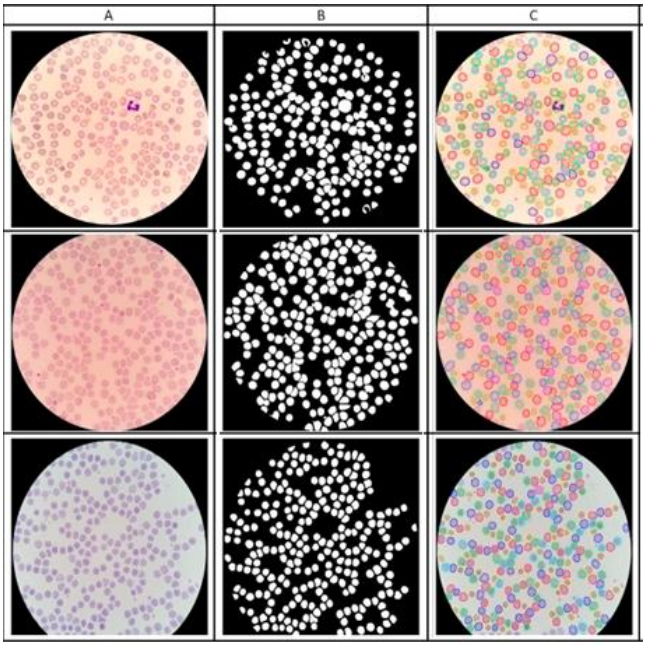


Figure 59. Détection et segmentation par seuillage d'Otsu des GR.**(A)** Image d'entrée. **(B)** Masque de la segmentation finale des GR. **(C)** Résultats de la segmentation, superposés sur l'image originale.

2.4.2. Détection des régions d'intérêts ROI

Par la suite, les ROI ont été identifiés sur la base de l'analyse des contours et des formes, permettant l'identification des parasites Plasmodium à divers stades de développement. Cette approche sélective réduit le bruit en se concentrant uniquement sur les zones pertinentes, afin d'améliorer l'efficacité de la classification :

- **Identification des régions pertinentes** : Une méthode comme l'utilisation de contours est appliquée sur les images binaires résultantes, pour identifier les zones potentiellement intéressantes.
- **Extraction des ROIs** : Les contours détectés sont utilisés pour extraire les sous-régions pertinentes de l'image. Si plusieurs régions sont détectées, des heuristiques basées sur des caractéristiques telles que la taille ou la forme sont appliquées pour choisir la région la plus significative.
- **Normalisation des ROIs** : Les ROIs extraites ont ensuite été redimensionnées uniformément à 50×50 pixels.

Dans la Figure 60, nous montrons l'effet de la combinaison segmentation par seuillage et masque de ROI, sur nos images.

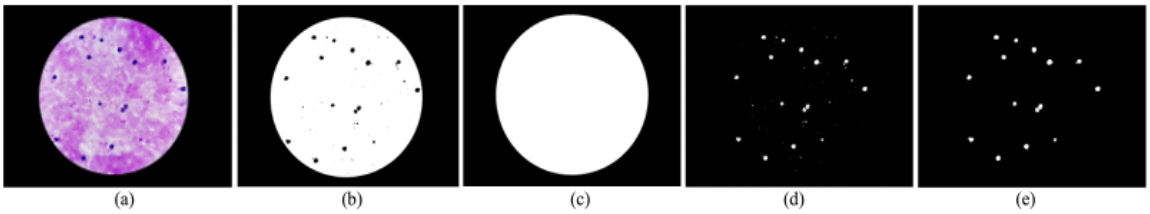


Figure 60. Exemple de détection de parasites. **(a)** Une image d'échantillon d'un frottis sanguin acquise avec un smartphone. **(b)** parasites détectés après utilisation du seuillage d'Otsu. **(c)** Masque ROI du champ de vision détecté. **(d)** parasites détectés, y compris de petites zones de bruit. **(e)** Détection des parasites après filtrage du bruit dans **(d) (Gaouar et al., 2025)**

2.4.3. Transformation des images en patchs

Cette étape de preprocessing des données ne concerne que les deux modèles Stacked-LSTM (avec et sans mécanisme d'Attention).

La transformation d'une image en **patchs** est une étape fondamentale dans les architectures de type Vision Transformer (ViT). Son rôle principal est d'adapter le traitement des images aux mécanismes d'attention conçus initialement pour des données séquentielles (comme le texte) **(Arnab et al., 2021).**

Dans notre cas, nous avons opté pour cette transformation dans le but d'atteindre un double objectif. D'une part, cela nous a permis de réduire la complexité computationnelle car dans le cas des LSTM nous avons besoins de traiter chaque pixel de la ROI comme un Token afin de construire une séquence d'entrée pour le LSTM, ce qui aurait pu nous mener vers une complexité quadratique insoutenable. En effet, traiter une image de 224×224 pixels en Tokens, nous donne 50176 tokens → **$(50176)^2$ = 2.5 milliards de paires d'attention !** La transformation des images en patchs de 16x16 pixels nous a permis de réduire cette complexité de **256 fois**, comme on peut le constater à travers ces calculs :

$\left(\frac{224}{16}\right)^2 = \mathbf{196}\ \boldsymbol{Tokens}$ Alors :

La réduction de la complexité computationnelle $=\left(\frac{50176}{196}\right) = \mathbf{256}\ \boldsymbol{X}$**.** D'autre part, cette transformation nous a permis de capturer des informations locales (car chaque patch encode des caractéristiques locales comme les textures, les bords et les motifs élémentaires) et des informations globales permettant de relier tous les patchs d'une même image entre eux.

Ainsi, chaque image ROI a été ensuite divisée en plus petits patchs de 16x16 pixels pour construire une séquence qui sert d'entrée pour les couches LSTM. Le choix d'un patch ayant une taille fixe de 16x16 pixels a été motivé par le fait que cette taille nous permet de capturer de très petits objets comme c'est le cas dans notre étude, où on fait face à des cellules sanguines et aux parasites qui s'y trouvent dedans (de très petites tailles).

Cette approche convertit les données spatiales en représentations séquentielles, facilitant la capture des dépendances spatiales et améliorant l'interprétabilité du processus de prise de décision du modèle.

3. Modèles Expérimentés

3.1. Explicabilité des Modèles

Les techniques XAI ont émergé pour relever le défi de la transparence dans l'apprentissage automatique, en particulier pour les applications de santé à enjeux élevés où l'interprétabilité est essentielle pour l'acceptation clinique, la responsabilité éthique et la conformité réglementaire.

Ces méthodes facilitent une compréhension plus approfondie des décisions des modèles, cruciale pour des tâches telles que la détection automatisée du paludisme. Les termes interprétabilité et explicabilité sont souvent utilisés de manière interchangeable dans la littérature sur l'IA, mais ils se réfèrent à des concepts distincts. L'interprétabilité désigne la mesure dans laquelle le processus de prise de décision d'un modèle est intrinsèquement compréhensible sans nécessiter d'outils ou de modifications externes **(Arrieta, et al., 2020)**. Les modèles transparents, tels que la régression linéaire et les arbres de décision, présentent cette propriété par conception.

L'explicabilité en revanche, fait référence à l'utilisation de techniques supplémentaires pour élucider le fonctionnement interne des modèles complexes, en particulier ceux manquant de transparence intrinsèque, tels que les réseaux de neurones profonds (DNN) (**Gleicher 2016 ; Doshi-Velez and Kim, 2017 ; Fernandez et al., 2019)**. Les méthodes d'explicabilité, y compris l'attribution de caractéristiques, les cartes de saillance et les modèles de substitution, permettent aux cliniciens et aux chercheurs d'analyser les résultats générés par l'IA, facilitant ainsi la validation des modèles et renforçant la confiance dans les systèmes de l'IA médicale. Les méthodes XAI sont largement classées en techniques spécifiques au modèle et techniques indépendantes du modèle (voir la **Figure 13** et la section V.3 du Chapitre I). Les méthodes spécifiques au modèle exploitent les structures internes du modèle, comme on le voit avec les approches basées sur le gradient telles que Grad-CAM **(Selvaraju et al., 2017)**, qui visualise les régions importantes de l'image, influençant les prédictions des CNN. Les méthodes indépendantes du modèle, telles que LIME, fonctionnent indépendamment de l'architecture du modèle en approximant le comportement du modèle, localement à l'aide de modèles de substitution interprétables **(Ribeiro et al., 2016a)**.

Les méthodes XAI peuvent également être classées en fonction de leur granularité en approches globales ou locales. Les méthodes globales, comme SHAP, fournissent des informations sur le comportement général d'un modèle à travers les ensembles de données **(Lundberg and Lee, 2017)**, tandis que les méthodes locales, telles que Grad-CAM et LIME, expliquent les prédictions individuelles en identifiant des caractéristiques d'entrée spécifiques et influentes.

Dans cette thèse, Grad-CAM et LIME ont été utilisés pour améliorer l'interprétabilité du modèle proposé Stacked-LSTM avec mécanisme d'attention pour la détection du paludisme. Grad-CAM est utilisé pour visualiser les régions d'une image qui ont le plus influencé la décision du modèle. Comme montré dans la **Figure 61 (Gaouar et al,. 2025)**, Grad-CAM

calcule les gradients de la classe cible par rapport aux cartes de caractéristiques de la couche convolutionnelle finale, puis utilise ces gradients pour générer une carte de chaleur. Cette dernière met en évidence les régions de l'image, qui ont le plus grand impact sur la décision du modèle, offrant une visualisation globale et aidant à interpréter quelles parties de l'image ont contribué à un résultat de classification spécifique **(Selvaraju et al., 2017)**.

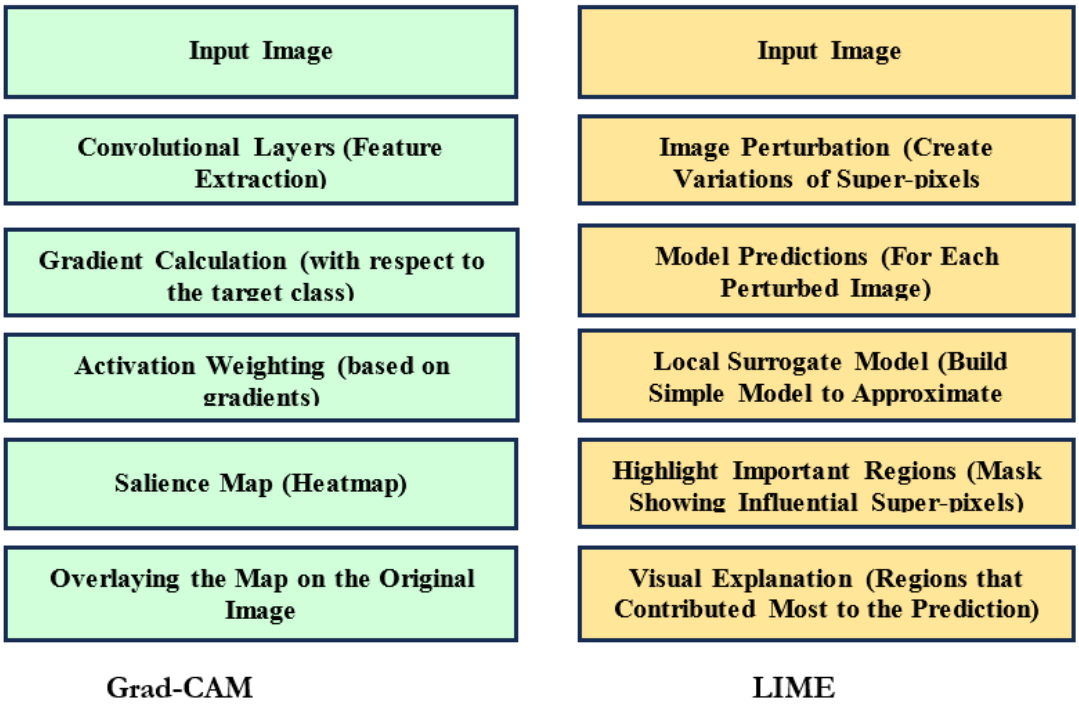


Figure 61. Diagrammes des modèles Grad-CAM et LIME **(Gaouar et al., 2025)**

En revanche, LIME explique les prédictions individuelles en approximant le comportement du modèle, avec un modèle de substitution plus simple et interprétable. Cette technique génère des versions perturbées des données d'entrée pour déterminer quelles caractéristiques des données ont affecté une prédiction spécifique, rendant ainsi le processus de décision plus transparent **(Ribeiro et al., 2016a)**. Ici, nous utilisons LIME pour améliorer l'interprétabilité de notre modèle d'apprentissage profond pour la classification des cellules de paludisme. LIME génère des superpixels, en perturbant l'image d'entrée et en examinant l'effet de ces perturbations sur la sortie du modèle. L'explication visuelle met en évidence les parties importantes de l'image (par exemple, la forme, la texture ou la couleur de la cellule) responsables de la décision du modèle. Une telle transparence est cruciale pour valider les prédictions et confirmer que le modèle se concentre sur des caractéristiques biologiquement pertinentes, en particulier pour des applications médicales telles que le paludisme.

L'utilisation de ces techniques complémentaires a permis une analyse complète du comportement du modèle. Grad-CAM a fourni une compréhension globale des régions de l'image, considérées comme significatives par le modèle, tandis que LIME a offert des interprétations détaillées et localisées, identifiant des caractéristiques précises à l'origine de chaque décision. L'intégration de ces méthodes a démontré la robustesse et la fiabilité du

modèle proposé et a confirmé son accent sur les caractéristiques d'image biologiquement pertinentes, essentielles pour un diagnostic précis du paludisme.

3.2. Modèles de Classification

3.2.1. Modèles basés sur les CNN

Les réseaux de neurones profonds sont utilisés pour la classification d'images en raison de leur meilleure performance par rapport à d'autres algorithmes **(Chowdary et al., 2020).** Mais entraîner un réseau de neurones profond est coûteux et long, car cela nécessite une grande quantité de puissance de calcul et d'autres ressources **(Marketting, 2019).**

Le principal avantage des CNN est leur capacité à apprendre des représentations et à extraire des caractéristiques à partir de données visuelles. De cette manière, ces modèles nécessitent un niveau minimal de prétraitements par rapport à d'autres algorithmes de classification d'images **(Pereira-Ferrero et al., 2022)**. Dans ce travail, nous avons expérimenté six modèles de Deep Learning bien établis.

L'apprentissage par transfert, basé sur l'apprentissage profond, est en cours de développement afin de rendre l'entraînement du réseau plus rapide et plus rentable. L'apprentissage par transfert permet de transférer les informations apprises par le réseau de neurones en termes de poids paramétriques vers le nouveau modèle (**Sangha, 2020**). Même lorsqu'il est entraîné sur un petit ensemble de données, le transfert d'apprentissage améliore l'efficacité du nouveau modèle. Plusieurs modèles pré-entraînés, tels que InceptionV3, Xception, MobileNet, MobileNetV2, VGG-16, VGG-19, ResNet50, et autres, ont été entraînés en utilisant 14 millions d'images du dataset ImageNet **(Chowdary et al., 2020)**.

Dans le cadre des travaux menés dans cette thèse, nous avons centré nos expérimentations sur trois modèles faisant partie de la famille des CNN, à savoir VGG-16, VGG-19 et MobileNetV2.

Pour adapter ces modèles à notre tâche de classification d'images cellulaires, nous avons remplacé les couches finales entièrement connectées, par des couches personnalisées adaptées à la classification binaire. Nous avons eu recours à l'apprentissage par transfert en utilisant des poids pré-entraînés provenant de l'ensemble de données ImageNet et nous les avons adaptés à notre tâche de détection du , à partir d'images de frottis sanguin. Cela nous a permis d'exploiter les capacités d'extraction de caractéristiques pré-entraînées de VGG-16, VGG-19 et MobileNetV2, améliorant ainsi l'efficacité de l'entraînement et les performances de chaque modèle.

A. VGG-16 et VGG-19

Initialement, nous avons testé les modèles VGG-16 et VGG-19 pour la classification des images de cellules infectées et non infectées. Les deux modèles, VGG-16 et VGG-19, sont des CNN largement adoptés pour les tâches de classification d'images en raison de leur architecture efficace et simple, et se distinguent par leur utilisation de petits filtres convolutionnels (3×3), qui sont empilés dans une architecture profonde pour capturer des motifs visuels complexes **(Simonyan et Zisserman, 2014)**.

L'architecture VGG-16 comprend 16 couches (**Figure 62**), dont 13 couches convolutives et 3 couches entièrement connectées. Les couches convolutives sont regroupées en cinq blocs, chaque bloc contenant 2 ou 3 couches convolutives suivies d'une couche de mise en commun maximale. Les filtres convolutifs 3x3 permettent au modèle d'apprendre des caractéristiques visuelles de bas niveau telles que les bords et les textures, tandis que les couches de mise en commun maximale réduisent les dimensions spatiales, améliorant ainsi l'efficacité des calculs. Les couches entièrement connectées effectuent la classification finale. Les deux premières couches denses, composées respectivement de 1024 et 256 unités, utilisent des fonctions d'activation ReLU, suivies d'une couche d'abandon (taux 0,5) pour atténuer l'adaptation excessive. La dernière couche de sortie comporte une fonction d'activation sigmoïde, qui fournit une classification binaire, indiquant si une cellule est infectée ou non.

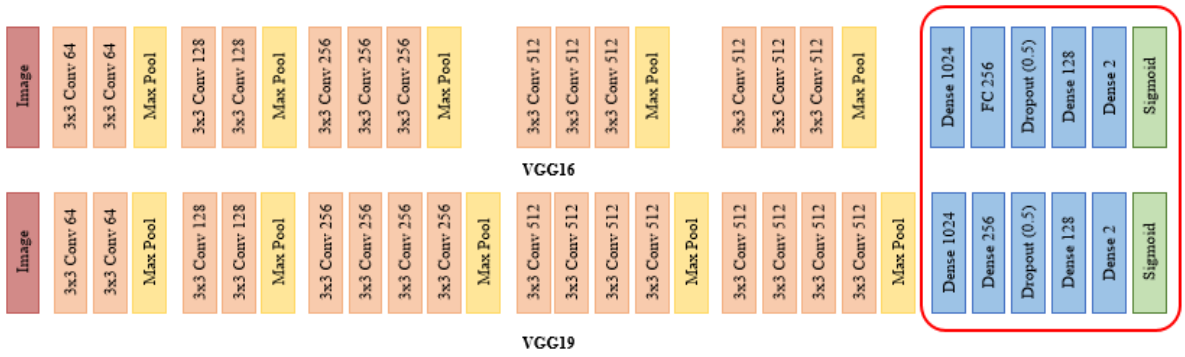


Figure 62. Architecture des modèles VGG-16 et VGG-19 [4]

Le VGG-19 est une variante plus profonde du VGG-16, avec trois couches convolutives supplémentaires (**Figure 62**), ce qui en fait un modèle à 19 couches. Comme le VGG-16, le VGG-19, il se compose également de cinq blocs convolutifs, mais avec quatre couches convolutives dans les troisième, quatrième et cinquième blocs. L'architecture plus profonde du VGG-19 lui permet de capturer des caractéristiques plus complexes et d'offrir potentiellement de meilleures performances de classification. Malgré les couches supplémentaires, la structure générale du VGG-19 reflète celle du VGG-16, les couches convolutives utilisant des filtres 3x3 et les couches entièrement connectées ayant la même configuration, ce qui conduit à une classification binaire grâce à l'activation sigmoïde.

B. Modèle MobileNetV2

Bien que les deux premiers modèles nous aient donné des performances notables, nous avons pensé qu'il serait très intéressant de tester un modèle léger et profond qui peut être utilisé sur des appareils à ressources limitées (smartphones, microscopes connectés, Raspberry Pi, etc.)

[4] https://medium.com/coinmonks/paper-review-of-vggnet-1st-runner-up-of-ilsvlc-2014-image-classification-d02355543a11

et qui présente également un bon compromis entre précision et taille du modèle, ce qui est crucial pour les applications médicales où la latence et la puissance de calcul peuvent être limitées. De plus, la détection du paludisme est souvent nécessaire dans des régions avec un accès limité à l'électricité ou à des équipements coûteux.

Ces motivations nous ont conduits à concevoir une architecture de classification légère et efficace basée sur MobileNetV2, un réseau de neurones convolutifs profonds optimisé pour les environnements à ressources limitées **(Sandler et al., 2018 ; Zou et al., 2025)**. MobileNetV2 est conçu pour être léger grâce à son architecture basée sur des convolutions depthwise et des blocs résiduels inversés (**Figure 63**), permettant une réduction significative du nombre de paramètres tout en maintenant de bonnes performances de généralisation. Cela permet une inférence rapide même sur des appareils à ressources limitées.

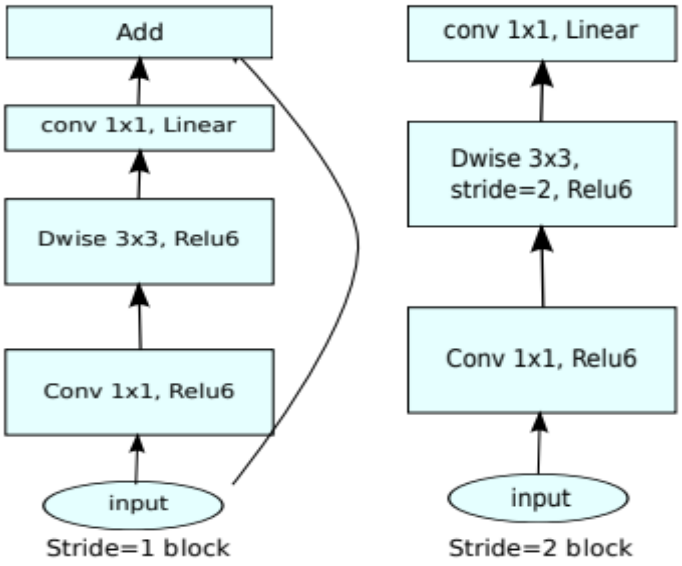


Figure 63. Architecture de MobileNetV2 **(Sandler et al., 2018).**

Bien que les modèles CNN aient fait d'excellents progrès, principalement soutenus par leurs caractéristiques profondes, un obstacle majeur demeure lié au manque d'informations contextuelles dans de telles représentations. En effet, ces représentations se trouvent souvent sur des variétés dans un espace de haute dimension (**Iscen et al., 2018)**, où la formulation par paires de la mesure de similarité est insuffisante pour révéler la relation intrinsèque entre les images.

3.2.2. Modèle Vision Transformer

Afin d'atteindre une précision de classification et d'interprétabilité des résultats importante, nos travaux se sont centrés sur la recherche d'une représentation plus efficace des caractéristiques, en prenant en compte les relations de similarité contextuelles définies par les caractéristiques extraites.

Pour atteindre cet objectif, nous avons testé les ViT (VisionTransformer), qui révolutionnent progressivement l'imagerie médicale, en offrant des solutions innovantes là où les CNN traditionnels montrent des limitations, telles que leur incapacité à capturer des motifs

subtils (micro-calcifications, anomalies cellulaires, présence de formes lunaires dans les cellules…etc.) grâce à l'attention multi-têtes ; ou leur capacité à préserver les relations spatiales à longue distance, cruciales pour les structures anatomiques complexes ou les grandes images, contrairement aux CNN, qui ne peuvent pas le faire **(Arnab et al., 2021 ; Aladhadh et al., 2022 ; Al-Hammuri et al., 2023 ; Yang et al., 2023).**

Les ViT constituent une avancée majeure dans le domaine de la vision par ordinateur. Introduits par **Dosovitskiy et al. (2020)**, ces modèles s'inspirent de l'architecture des Transformers initialement conçue pour le traitement du langage naturel **(Vaswani et al., 2017).** Contrairement aux réseaux convolutifs classiques (CNN), les ViT abandonnent les convolutions au profit d'un mécanisme d'attention globale qui traite des patchs d'image comme des tokens, à l'instar des mots dans une phrase comme on peut clairement le distinguer dans la **Figure 64**.

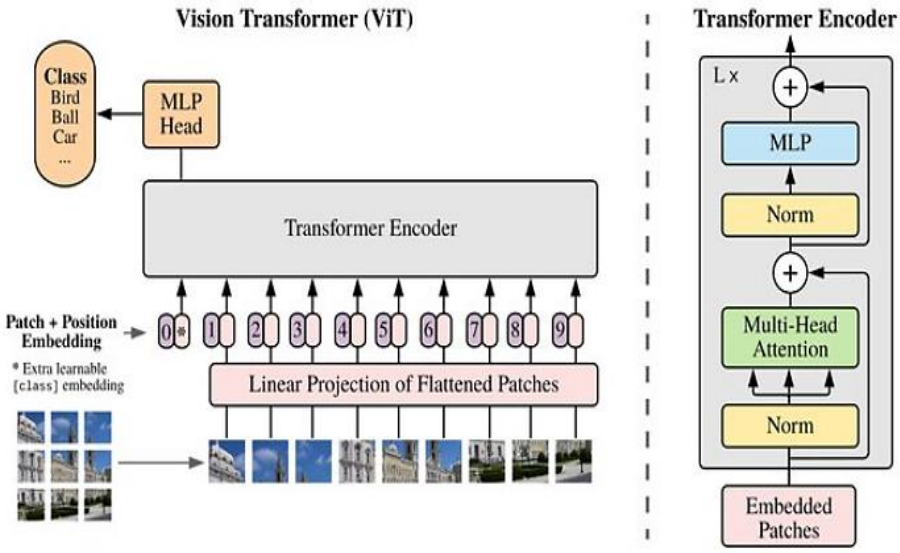


Figure 64. Architecture Originelle du ViT **(Dosovitskiy et al., 2020)**

Concrètement, une image est d'abord découpée en petits patchs fixes (souvent 16×16 pixels). Chaque patch est ensuite transformé en un vecteur (embedding) auquel on ajoute un encodage positionnel pour conserver l'information spatiale. Ces vecteurs sont ensuite traités par une série de blocs Transformers, composés de couches de self-attention multi-têtes et de perceptrons multicouches. Un token est utilisé pour agréger l'information globale de l'image et sert à la classification finale.

Malgré leur efficacité, les ViT exigent une quantité massive de données pour l'entraînement depuis zéro. Pour pallier cette contrainte, leur utilisation en transfert learning s'est largement démocratisée. Ainsi, un modèle ViT préentraîné sur un vaste corpus d'images (comme ImageNet-21k) peut être réutilisé et ajusté (fine-tuned) sur une tâche spécifique à moindre coût en données et en temps **(Touvron et al., 2021)**. Cette approche a démontré des performances remarquables, en particulier dans des contextes médicaux où les données annotées sont rares **(Chen et al., 2021 ; Aladhadh et al., 2022)**.

Les ViT offrent une meilleure modélisation des relations globales dans une image et se sont avérés compétitifs, voire supérieurs, aux CNN dans de nombreuses tâches de classification, de segmentation, et de détection.

En somme, les ViT présentent des avantages théoriques notables, mais leur application à des contextes médicaux spécifiques comme la détection du paludisme nécessite des ajustements importants ou l'intégration d'approches hybrides combinant CNN et ViT les rendant trop lourds à déployer surtout dans des endroits à ressources matériels limitées. En effet, les ViT sont très gourmands en ressources computationnelles, en particulier à cause de la complexité quadratique du mécanisme de self-attention. Dans le cas d'images microscopiques à haute résolution, comme celles des frottis sanguins, cela engendre des coûts de calcul élevés, limitant leur déploiement dans des environnements à ressources réduites **(Zhou et al., 2022)**.

De plus, les ViT ne disposent pas naturellement des inductifs locaux utiles pour détecter des motifs morphologiques spécifiques, comme la présence de parasites intracellulaires dans les globules rouges. Ce manque de perception locale peut nuire à la capacité de détection de détails subtils **(Khan et al., 2022)**.

Bien que le transfert learning à partir de jeux de données larges tels qu'ImageNet, présente de sérieux avantages, les images naturelles de ce type de base diffèrent radicalement des images médicales en microscopie, limitant la transférabilité des représentations apprises **(Touvron et al., 2021)**.

Enfin, même si les ViT offrent une forme d'explicabilité via les cartes d'attention, celles-ci sont souvent difficiles à interpréter dans un contexte clinique, ce qui peut poser un obstacle majeur à leur acceptation et leur adoption par les professionnels de santé **(Chefer et al., 2021)**.

Pour des raisons évidentes liées aux limitations et inconvénients des modèles intelligents basés sur les CNN (VGG-16, VGG-19 et MobileNetV2) et les ViT, nous avons expérimenté les RNN, en particulier les réseaux Stacked-LSTM, qui offrent à la fois la possibilité de capturer les motifs subtils et de préserver les relations spatiales à longue distance, sans être gourmands en termes de données et de ressources.

Étant donné que notre modèle phare repose sur les RNN et plus précisément les Stacked-LSTM, nous nous proposons de les présenter dans la section suivante.

3.2.3. Réseaux de Neurones Récurrents (Recurrent Neural Network RNN) :

L'inception des RNN remonte aux années 70, avec les travaux importants et fondamentaux de Werbos **(Werbos et al., 1990)**, qui a introduit pour la première fois le concept de rétro-propagation à travers le temps (RPTT) et qui a posé les bases de l'entraînement des réseaux de neurones récurrents.

Les RNN sont conçus pour traiter des données séquentielles en maintenant un état caché qui capture des informations sur les entrées précédentes **(Tsantekidis et al., 2022)**. L'architecture de base se compose d'une couche d'entrée, d'une couche cachée et d'une couche de sortie. Contrairement aux réseaux de neurones à propagation avant, les RNN ont des

connexions récurrentes, comme le montre la **Figure 65**, permettant à l'information de circuler au sein des réseaux.

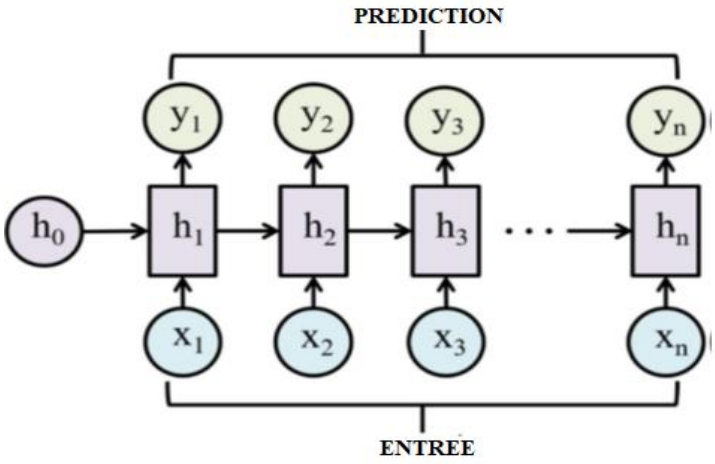


Figure 65. Architecture de base d'un RNN **(Tsantekidis et al., 2022)**

À chaque étape de temps t, le RNN prend un vecteur d'entrée $\mathbf{X}_t$, et met à jour son état caché $\mathbf{h}_t$, (voir Figure 66) en utilisant l'équation suivante :

$$\mathbf{h}_t = \boldsymbol{\sigma}_h (\mathbf{W}_{xh} \mathbf{X}_t + \mathbf{W}_{hh} \mathbf{h}_{t-1} + \mathbf{b}_h) \quad \textbf{(11)}$$

Où :

- $\mathbf{W}_{xh}$ est la matrice de poids entre l'entrée et la couche cachée,
- $\mathbf{W}_{hh}$ est la matrice de poids pour la connexion récurrente,
- $\mathbf{b}_h$ est le vecteur de biais,
- $\boldsymbol{\sigma}_h$ est la fonction d'activation, typiquement la fonction tangente hyperbolique (**tanh**) ou la fonction de l'unité linéaire rectifiée (Rectified Linear Unit **ReLU**) **(Mienye, I.D. et al., 2024 ; Mienye, I.D. et al., 2023)**. La sortie à chaque étape temporelle, t, est donnée par ce qui suit :

$$\mathbf{y}_t = \boldsymbol{\sigma}_y (\mathbf{W}_{hy} \mathbf{h}_t + \mathbf{b}_y) \quad \textbf{(12)}$$

Où:

- $\mathbf{W}_{hy}$ est la matrice de poids entre les couches cachée et de sortie,
- $\mathbf{b}_y$ est le vecteur de biais,
- et $\boldsymbol{\sigma}_y$ est la fonction d'activation pour la couche de sortie.

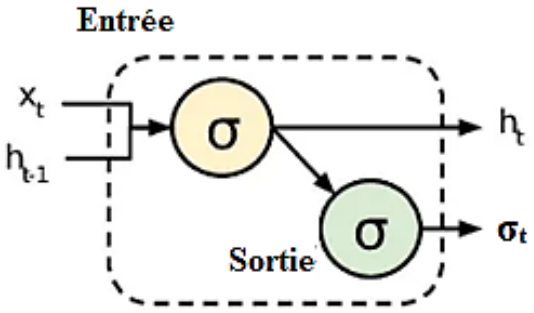


Figure 66. Architecture d'une cellule RNN **(Tsantekidis et al., 2022)**

3.2.4. Réseaux de mémoire à long et à court terme (Long Short-Term Memory LSTM)

Les RNN ont vite rencontré des difficultés dans les applications pratiques réelles, dues au problème du gradient qui disparaissait, où les gradients augmentaient ou diminuaient de façon exponentielle durant le processus de rétropropagation **(Lalapura et al., 2021)**. En même temps, l'arrivée des réseaux LSTM proposés par Hochreiter et Schmidhuber **(Hochreiter and Schmidhuber, 1997)** a été un moment clé pour les RNN, car cela a permis l'apprentissage des dépendances sur des périodes de temps beaucoup plus longues, pour résoudre le problème de la disparition du gradient inhérent aux RNN de base.

La nouveauté principale des LSTM est qu'ils utilisent des portes pour gérer le flux d'informations dans le réseau. Cela aide les réseaux LSTM à maintenir et de mettre à jour leur état pendant longtemps, ce qui les rend efficaces pour des tâches qui nécessitent la modélisation des dépendances à long terme. Chaque cellule LSTM possède trois portes : **la porte d'entrée, la porte d'oubli** et **la porte de sortie** (**Figure 67**). Ces portes régulent l'état de la cellule $\mathbf{c_t}$ et l'état caché $\mathbf{h_t}$ **(Mienye et al., 2023)**. Ces portes décident combien d'informations garder de l'entrée, combien d'informations de l'état précédent oublier et combien d'informations de la cellule construire.

Les équations pour mettre à jour un LSTM sont les suivantes **(Mienye et al., 2023)**:

$$\mathbf{i_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + b_i)} \quad \textbf{(13)}$$

$$\mathbf{f_t = \sigma\,(W_{xf}X_t + W_{hf}h_{t-1} + b_f)} \quad \textbf{(14)}$$

$$\mathbf{o_t = \sigma\,(W_{xo}X_t + W_{ho}h_{t-1} + b_o)} \quad \textbf{(15)}$$

$$\mathbf{\check{c}_t = tanh\,(W_{xg}X_t + W_{hg}h_{t-1} + b_g)} \quad \textbf{(16)}$$

$$\mathbf{c_t = f_t \odot c_{t-1} + i_t \odot \check{c}_t} \quad \textbf{(17)}$$

$$\mathbf{h_t = o_t \odot tanh(c_t)} \quad \textbf{(18)}$$

Où :

- $\mathbf{i_t}$ est la porte d'entrée,
- $\mathbf{f_t}$ est la porte d'oubli,
- $\mathbf{o_t}$ est la porte de sortie,
- $\mathbf{\check{c}_t}$ est l'entrée de la cellule,
- $\mathbf{c_t}$ est l'état de la cellule,
- $\mathbf{h_t}$ est l'état caché, $\boldsymbol{\sigma}$ représente la fonction sigmoïde,
- **tanh** est la fonction tangente hyperbolique,
- $\odot$ désigne la multiplication élément par élément.

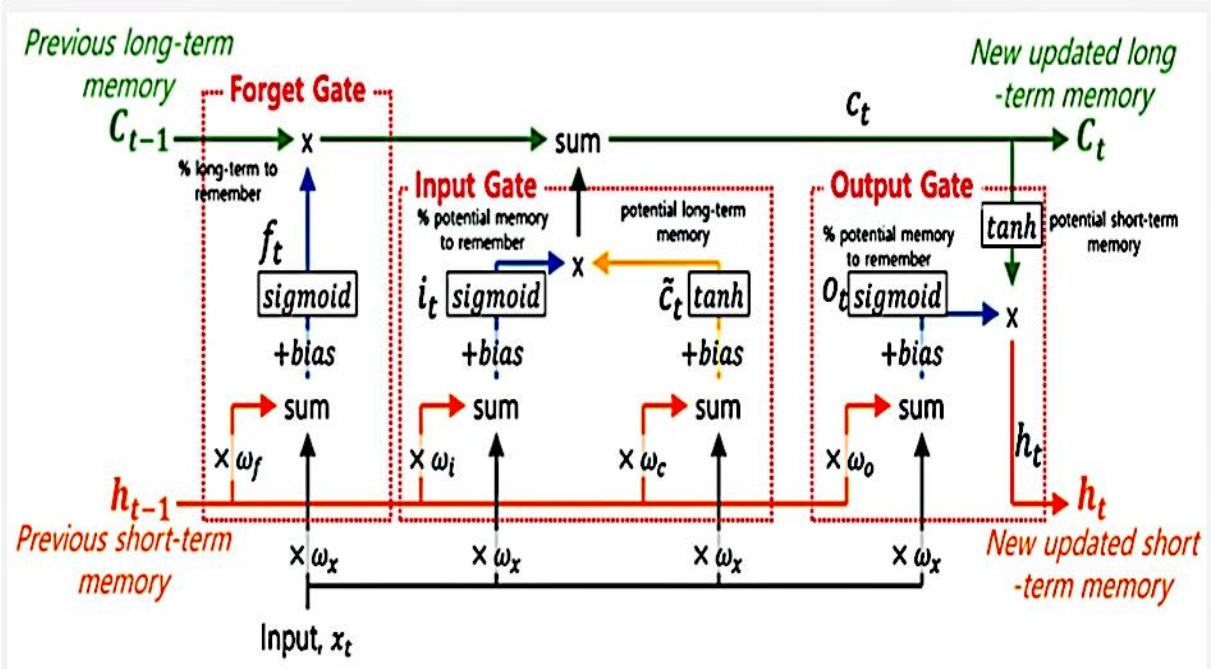


Figure 67. Architecture d'une cellule LSTM **(Mienye et al., 2023)**

La **Figure 67** illustre l'architecture interne d'une cellule LSTM, qui gère efficacement les dépendances à long terme dans les données de séquence, en utilisant trois mécanismes de porte cruciaux. Chacune de ces portes joue un rôle particulier dans le contrôle du flux d'information par le biais de la cellule :

- **Porte d'entrée $\mathbf{i_t}$** : Des informations utiles supplémentaires à l'état de la cellule sont ajoutées par la porte d'entrée. Cette porte contrôle combien d'informations de la nouvelle entrée $\mathbf{X_t}$ est écrite dans l'état de la cellule $\mathbf{C_t}$
- **Porte d'oubli $\mathbf{f_t}$** : Elle supprime les informations qui ne sont plus utiles dans l'état de la cellule et décide de la quantité de l'état de cellule précédent $\mathbf{C_{t-1}}$ qui doit être conservée.

- **Porte de sortie O_t**: Des informations utiles supplémentaires à l'état de la cellule sont ajoutées par la porte de sortie qui détermine combien de l'état de cellule C_t est utilisé pour calculer l'état caché h_t.

- L'entrée de cellule $\check{C}_t$ est une valeur candidate qui est ajoutée à l'état de la cellule après avoir été modulée via la porte d'entrée

L'intégration de ces mécanismes de porte confère aux réseaux LSTM la possibilité de se souvenir ou d'oublier sélectivement des informations, leur permettant de gérer les dépendances à long terme plus efficacement que les RNN traditionnels.

L'état de la cellule C_t gère la récurrence interne au sein de la cellule LSTM et agit comme un convoyeur, transportant les informations pertinentes à travers les différentes étapes temporelles.

Ainsi, le mécanisme de récurrence permet au LSTM de maintenir et de mettre à jour sa mémoire sur de longues séquences, capturant ainsi efficacement les dépendances à long terme. De plus, les opérations de multiplication élément par élément entre les portes et leurs entrées respectives garantissent que les interactions entre les différents composants du LSTM sont fluides et efficaces. Cela permet au LSTM d'effectuer des transformations complexes sur les données d'entrée tout en maintenant la stabilité du processus d'apprentissage. Durant ce temps, les réseaux LSTM utilisent une récurrence interne au sein de chaque cellule pour gérer les dépendances à long terme, la récurrence se produisant à travers l'état de la cellule alors que l'information est transmise d'un pas de temps à l'autre **(Yu et al., 2019)**. D'autres variantes des réseaux LSTM existent, telles que les Bi-LSTM (Bidirectional LSTM) et les Stacked-LSTM **(Mienye et al., 2024)**.

A. LSTM Bidirectionnel (Bi-LSTM) :

Le LSTM bidirectionnel (Bi-LSTM), décrit dans la **Figure 68**, augmente le modèle LSTM classique en analysant la séquence dans les deux sens, passe avant et passe arrière. Cette méthode aide le réseau à saisir le contexte du passé et du futur, ce qui améliore sa capacité à comprendre les liens dans la séquence de manière plus claire.

Dans les Bi-LSTM, on a deux états cachés pour chaque phase temporelle : un pour l'avance ($h_t \rightarrow$) et un pour le recul ($\leftarrow h_t$). Ces états invisibles sont calculés comme expliqué dans les Équations (19) et (20). Les Bi-LSTM traitent les données en allant dans les deux sens, vers l'avant et vers l'arrière, et gardent des états cachés, séparés pour chaque direction. L'inconvénient majeur de cette variante est le risque élevé du surapprentissage (Overfitting) **(Chorowski et al., 2015 ; Dong et al., 2019)**.

$$\overrightarrow{h_t} = \sigma_h(W_{xh}X_t + W_{hh}\overrightarrow{h_{t-1}} + b_h) \quad \textbf{(19)}$$

$$\overleftarrow{h_t} = \sigma_h(W_{xh}X_t + W_{hh}\overleftarrow{h_{t+1}} + b_h) \quad \textbf{(20)}$$

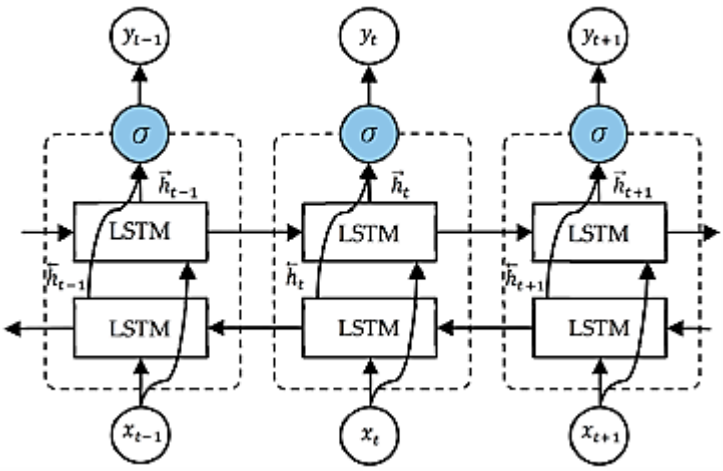


Figure 68. Architecture d'un Bi-LSTM **(Mienye et al., 2024)**

B. LSTM empilés (Stacked-LSTM):

Le Stacked-LSTM consiste à empiler plusieurs couches LSTM. La sortie d'une couche devient l'entrée de la couche suivante, comme illustré dans la **Figure 69**. Cette architecture profonde confère au réseau la capacité de comprendre des patterns plus complexes dans les données, en apprenant des représentations à différents niveaux. Pour un LSTM avec L couches, les états cachés de la couche l au moment t sont mis à jour selon les Équations (21) et (22). Le Stacked-LSTM utilise plusieurs couches LSTM. Chaque couche envoie sa sortie à la couche suivante. Ce qui permet au réseau à mieux saisir des motifs dans le temps de façon plus complexe.

$$\boldsymbol{h_t^{(l)} = \sigma_h(W_{xh}^{(l)} h_t^{(l-1)} + W_{hh}^{(l)} h_t^{(l-1)} + W_{hh}^{(l)} h_{t-1}^{(l)} + b_h^{(l)})} \qquad \textbf{(21)}$$

Où $\boldsymbol{h_t^{(0)} = X_t}$ représente l'entrée à la première couche. La sortie à la couche la plus haute est ensuite calculée en utilisant la même procédure que dans les RNN de base :

$$\boldsymbol{y_t = \sigma_y(W_{hy} h_t^{(l)} + b_y)} \qquad \textbf{(22)}$$

Figure 69. Architecture d'un Réseau LSTM empilé **(Mienye et al., 2024)**

L'empilement de couches LSTM aide le réseau à apprendre des caractéristiques et des représentations de plus en plus complexes. Les couches inférieures détectent des motifs à court terme, alors que les couches supérieures capturent des caractéristiques plus abstraites et des dépendances à long terme **(Yu et al., 2019)**. Cet apprentissage par couche est utile pour des tâches comme la modélisation du langage, où il faut comprendre différentes informations sur la grammaire et le sens, ou pour l'analyse de vidéos, où il est important de saisir les dépendances

temporelles à différentes échelles. Les LSTM empilés améliorent la modélisation, mais ils nécessitent plus de ressources et augmentent le risque de surapprentissage.

4. Approche Proposée MalariaScope

Dans cette thèse, nous proposons **pour la première fois** une approche novatrice nommée **MalariaScope** basée sur les réseaux Stacked-LSTM, équipés d'un mécanisme d'attention pour détecter et classifier le paludisme à partir d'images de frottis sanguins. Nous avons développé cette approche dans l'intention de pallier aux limitations susmentionnées relatives aux modèles précédemment expérimentés d'une part, et pour améliorer la transparence et la confiance dans les systèmes de détection du paludisme, afin d'améliorer l'explicabilité et l'interprétabilité des résultats obtenus par notre classifieur d'autre part, offrant ainsi un outil efficace pour l'aide au diagnostic médical, dans le contexte de la détection précoce de la malaria.

Pour ce faire, nous avons apporté les contributions suivantes :

- **Proposition d'un modèle de classification robuste basé sur un réseau Stacked-LSTM muni d'un mécanisme d'attention** ;
    - Une approche novatrice, exploitant les capacités des Stacked-LSTM à générer les informations contextuelles dans le processus d'extraction des caractéristiques, à partir de séquences temporelles ou spatiales extraites des images de frottis sanguin, pour détecter efficacement les globules rouges infectés, permettant une classification hautement précise.
    - Une optimisation des paramètres pour garantir une robustesse face aux variations des conditions d'acquisition des images.
    - Une démonstration de l'efficacité du mécanisme d'attention pour la classification d'images médicales et les tâches complexes de vision par ordinateur, ainsi que son impact sur l'explicabilité et l'interprétabilité des résultats de classifications, obtenus par le Stacked-LSTM.
- **Intégration d'un mécanisme d'explicabilité par les techniques XAI** :
    - Ajout d'un mécanisme d'explicabilité après notre classifieur, basé sur les réseaux Stacked-LSTM, avec Attention pour concentrer le modèle sur les régions d'images les plus pertinentes, permettant d'extraire et d'expliquer les motifs ou les patterns responsables de la décision de notre modèle intelligent.
    - Évaluation de deux méthodes XAI, la carte d'activation de classe pondérée par le gradient (Grad-CAM) et les explications locales interprétables indépendamment du modèle (LIME), pour interpréter et expliquer les prédictions faites par le classificateur Stacked-LSTM. Ces techniques d'explicabilité nous permettent d'analyser comment le modèle prend des décisions et d'identifier les régions d'intérêt influençant ses prédictions. Cette approche vise à améliorer l'interprétabilité des modèles d'IA et à garantir leur application fiable dans les contextes critiques de soins de santé.

- **Détection précoce du paludisme :**

La "détection précoce" fait référence à l'identification de l'infection par le paludisme aux premiers stades de la présentation des symptômes, généralement avant que des complications cliniques ne se présentent. Notre modèle est conçu pour analyser les échantillons de laboratoire (images de frottis sanguins) avec une grande sensibilité, permettant un diagnostic précoce, rapide et efficace par rapport à une suspicion clinique tardive ou à la microscopie traditionnelle nécessitant une grande expertise humaine.

- **Étude comparative avancée entre plusieurs modèles de DL :**

Une évaluation comparative de cinq modèles d'apprentissage profond—VGG-16, VGG-19, Vision Transformer (ViT), MobileNetV2, Stacked-LSTM **avec** Attention, Stacked-LSTM **sans** Attention —pour la détection du paludisme à l'aide d'images de frottis sanguins.

- **Développement d'une plateforme web applicative** :
    - Une interface interactive permettant d'intégrer le diagnostic automatisé avec une visualisation des explications générées par le modèle.
    - Un système léger et déployable dans des environnements à faibles ressources.

Ces contributions combinent précision, transparence et accessibilité, ouvrant ainsi la voie à une adoption clinique plus large de l'IA dans l'aide au diagnostic médical.

1.1. Protocole Expérimental de la Classification

Dans la majorité des travaux utilisant des LSTMs dans le domaine de la classification d'images, les chercheurs précèdent cette classification par une étape d'extraction de caractéristiques basée sur des CNNs **(Long et al., 2019 ; Alanazi et Alaerjan, 2023 ; Kim et al., 2023 ; Lanjewar et al., 2023 ; Pereira-Ferrero et al., 2023)**. Bien qu'une telle approche ait donné de bons résultats en termes de précision et de justesse, leurs temps d'entraînement sont relativement lents et d'un point de vue computationnel, ils sont coûteux et complexes, ajoutant une couche d'opacité aux « modèles boîte noire », les rendant non interprétables et difficilement explicables.

Comme notre objectif principal n'est pas seulement d'obtenir de bonnes performances mais aussi, et surtout, de présenter des modèles aussi interprétables et explicables que possible, avec un temps d'apprentissage court, pouvant obtenir de très bonnes performances même lorsqu'il est utilisé dans des environnements à ressources limitées, nous avons préféré éviter cette hybridation.

Étant donné que les LSTM traitent des séquences temporelles et que les images médicales télescopiques ne présentent pas de telles séquences, nous avons transformé les ROI obtenues lors de l'étape de prétraitement des images en des patches d'images de 16x16 pixels, afin de les organiser en une séquence spatiale. Le fait de transformer une image en un ensemble de patchs pose un problème majeur ; le classifieur va traiter les tokens de manière désordonnée, perdant toute notion de la position relative du token, ce qui conduit de façon inhérente à l'incapacité de capturer les dépendances entre les différents tokens, ce qui a pour résultat, la **perte quasi totale**

du sens et de la structure de l'image. En résumé, la position relative n'est pas une simple information technique, elle est le ciment qui lie les différents tokens pour former une image cohérente et compréhensible ; la perdre revient à transformer une image structurée en un simple puzzle inexploitable.

Pour pallier à ce problème, nous avons effectué un encodage positionnel **(Vaswani et al., 2017 ; Dosovitskiy et al., 2020)**, comme on le fait pour les Transformers, en utilisant les formules suivantes :

$$PE_{(x,y,2i)} = Sin\left(\frac{X}{1000^{\frac{2i}{d}}}\right) \quad (23)$$

$$PE_{(x,y,2i+1)} = Cos\left(\frac{X}{1000^{\frac{2i}{d}}}\right) \quad (24)$$

$$PE_{(x,y,2j)} = Sin\left(\frac{X}{1000^{\frac{2j}{d}}}\right) \quad (25)$$

$$PE_{(x,y,2j+1)} = Cos\left(\frac{X}{1000^{\frac{2j}{d}}}\right) \quad (26)$$

Où :

• **(x, y)** sont les coordonnées du patch dans la grille.

• **i** et **j** itèrent sur les dimensions du vecteur d'encodage.

• **d** est la dimension du vecteur d'encodage.

Cela donne à chaque patch un vecteur unique encodant sa position dans l'image. L'ensemble des patches est ensuite organisé en une séquence selon leur position. L'objectif est de fournir à chaque patch des informations sur sa position spatiale afin que la première couche LSTM comprenne les relations spatiales lors du traitement de la séquence comme s'il s'agissait d'une séquence temporelle pouvant être exploitée par la première couche LSTM de notre modèle, comme le montre la **Figure 70 (Gaouar et al., 2025)**.

La première couche LSTM se compose de 64 cellules LSTM. Elle analyse la séquence des patches, leurs dépendances et leurs relations spatiales, transformant chacun en une représentation riche et contextuelle et produisant une séquence d'états cachés $\mathbf{h_t}$. Nous obtenons une séquence de représentations $\mathbf{H = [h_1, h_2, \ldots, h_n]}$, où **n** est le nombre de patches.

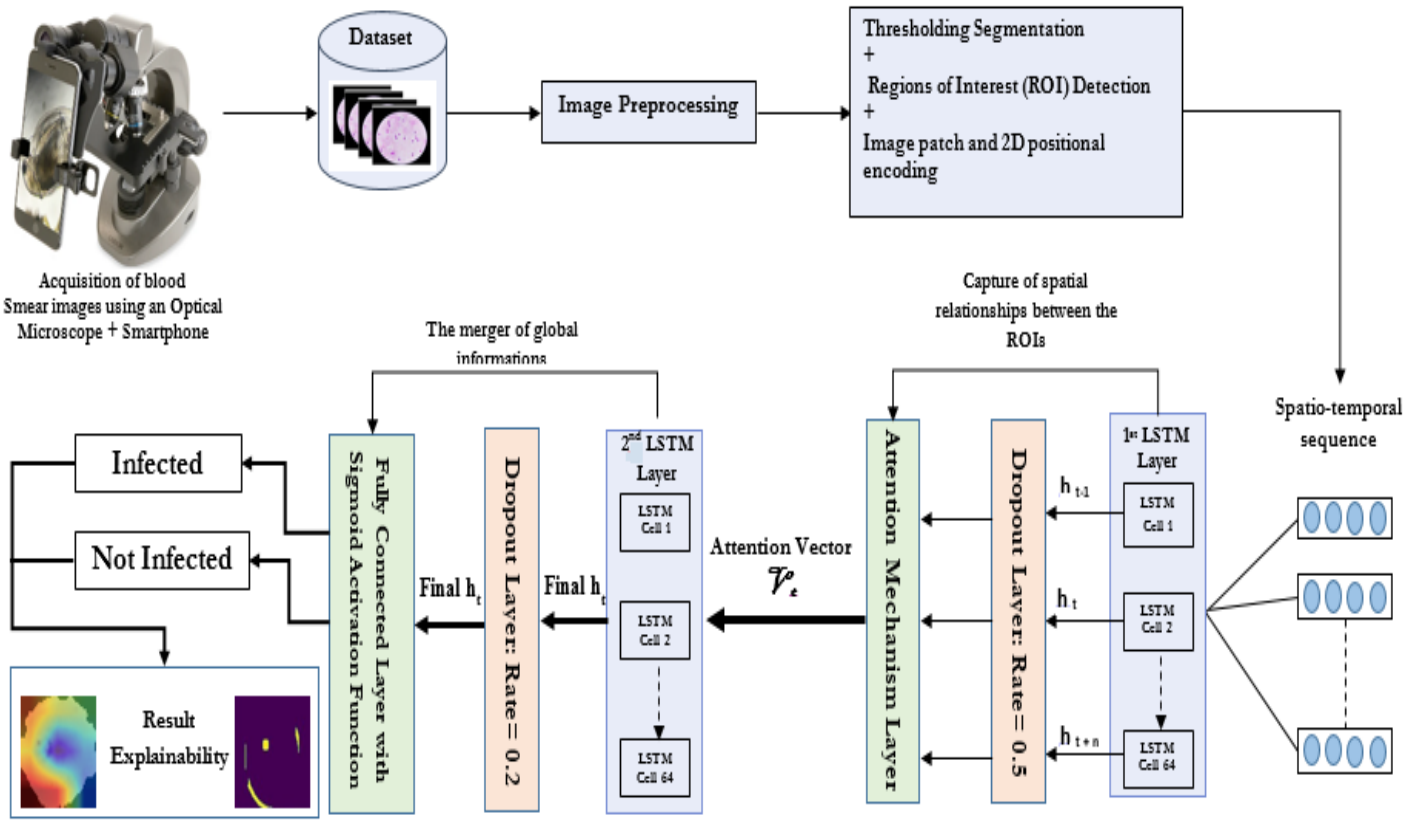


Figure 70. Architecture générale de notre système **MalariaScope (Gaouar et al., 2025)**

Nous avons appliqué un dropout de 50 % sur les sorties obtenues de la première couche LSTM, pour éviter le surapprentissage et fournir de la robustesse à notre modèle.

Les séquences restantes seront dirigées vers une couche d'attention, intégrant un mécanisme d'attention additive **(Bahdanau et al., 2014)**. Ce mécanisme d'attention joue un rôle crucial dans la détection du paludisme. Cela permet au modèle de se concentrer sur les zones les plus pertinentes des images de frottis sanguins, en particulier les régions d'intérêt (ROIs) détectées par segmentation.

Ce mécanisme commence par extraire les caractéristiques spatiales des patches et calcule un vecteur d'attention pour chacune d'elles, reflétant ainsi son importance et sa pertinence pour la classification. Les patchs les plus influents (notamment ceux qui présentent des caractéristiques du plasmodium) se verront attribuer un score d'attention plus élevé. Le score d'importance est calculé pour chaque vecteur $\mathbf{h_t}$ et est comparé à un vecteur de contexte **w** appris par le modèle. Le score d'importance $\mathbf{e_t}$ est calculé pour chaque patch **t** par la formule suivante :

$$e_t = score(h_t, w) = v^T \tanh(wh_t + b) \quad (27)$$

Où :

- **w** et **b** sont des paramètres appris.
- $\mathbf{v^T}$ est un vecteur de pondération appris.

Nous passons ensuite à la phase de normalisation des scores en utilisant la fonction Softmax, en l'appliquant à tous les scores d'importance pour obtenir les poids d'attention normalisé $\boldsymbol{\alpha_t}$:

$$\boldsymbol{\alpha_t} = \frac{\boldsymbol{exp(e_t)}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{exp(e_j)}} \quad \textbf{(28)}$$

La normalisation garantit que les poids d'attention $\boldsymbol{\alpha_t}$ sont compris entre 0 et 1 et que leur somme est égale à 1 ; cela nous permet d'interpréter les poids de manière probabiliste. Les caractéristiques extraites sont pondérées par le vecteur d'attention $\mathbf{V_t}$, donnant plus de poids aux zones importantes ; permettant au modèle de se concentrer sur les régions importantes plutôt que de traiter tous les patchs de manière similaire. Le vecteur $\mathbf{V_t}$ est obtenu par la somme pondérée de tous les états cachés $\mathbf{h_t}$:

$$\boldsymbol{V_t} = \sum_{\boldsymbol{t=1}}^{\boldsymbol{n}} \boldsymbol{\alpha_t}\, h_t \quad \textbf{(29)}$$

L'intégration du mécanisme d'attention dans notre classificateur a été motivée par des objectifs liés à la prédiction et à la classification, ainsi que par l'augmentation de la transparence du modèle pour le rendre plus interprétable et explicable, permettant ainsi aux professionnels de la santé de comprendre pleinement les résultats obtenus. Le premier objectif a été atteint principalement en concentrant le modèle sur les régions d'attention. En effet, cela nous a permis de réduire le bruit et d'augmenter la précision de la classification, ainsi que la complexité computationnelle et le temps d'entraînement et d'exécution, car nous ne traitons pas l'image entière mais uniquement des régions spécifiques d'attention. Le deuxième objectif, relatif à l'augmentation de la transparence du modèle, le rendant plus interprétable et plus facile à expliquer, a été atteint en utilisant la visualisation des régions d'attention pour comprendre quelles zones du frottis sanguin le modèle considère comme importantes. Cette visualisation nous permet également d'expliquer pourquoi le modèle a prédit la présence ou l'absence de paludisme en indiquant les régions d'intérêt pertinentes. De même, les experts peuvent vérifier les décisions du modèle en analysant ces régions, leur donnant confiance dans ses prédictions.

Contrairement à la première couche LSTM, qui reçoit une séquence de patches et produit une séquence de sorties, la deuxième couche LSTM reçoit un seul vecteur d'attention $\mathbf{V_t}$ et ne produit aucune séquence. Elle agit comme un encodeur final qui compresse cette information en une représentation d'image unique, intégrant les informations contextuelles extraites par le mécanisme d'attention et produisant une représentation globale cohérente.

Cette opération est comparable au fonctionnement d'un LSTM récurrent en un seul pas de temps, transformant le vecteur $\mathbf{V_t}$ en une représentation plus riche, produisant une sortie qui représente l'état caché final (Final $\mathbf{h_t}$) utilisé pour la classification. De la même manière que pour la sortie de la première couche LSTM, nous appliquons un dropout à la sortie de la deuxième couche LSTM, désactivant seulement 20 % des connexions.

La sortie sera ensuite envoyée à la couche entièrement connectée, équipée d'une fonction sigmoïde pour la classification binaire : infecté ou non-infecté.

La dernière étape de notre pipeline concerne l'explicabilité et la visualisation des résultats obtenus, grâce aux techniques XAI, Grad-CAM et LIME.

En résumé, notre approche Stacked-LSTM avec attention comprend les éléments clés suivants :

A. Première couche LSTM

- 64 unités LSTM configurées pour l'extraction des caractéristiques à partir des images segmentées et retournées des séquences.
- Objectif : Extraire des caractéristiques temporelles et spatiales significatives.

B. Première couche de Dropout

- Dropout de 20% pour éviter le surapprentissage et améliorer la généralisation.

C. Seconde couche LSTM

- 64 unités LSTM configurées pour ne pas retourner de séquences.
- Objectif : Résumer les caractéristiques en une représentation fixe.

D. Seconde couche de Dropout

- Dropout de 20%, similaire à la première.

E. Mécanisme d'attention : Pour renforcer l'interprétabilité et la performance du modèle, un mécanisme d'attention est ajouté, comprenant :

- **Poids attentionnels** : Un vecteur de poids est appris pour attribuer une importance variable aux différentes caractéristiques extraites par les LSTM.
- **Agrégation pondérée** : Les caractéristiques sont combinées en utilisant les poids attentionnels pour produire une représentation finale enrichie.
- **Sortie interprétable** : Le mécanisme d'attention permet de visualiser les régions ou caractéristiques auxquelles le modèle accorde le plus d'importance, améliorant ainsi l'explicabilité.

F. Couche Dense : Une couche entièrement connectée est utilisée pour la classification binaire, où le modèle produit des probabilités, indiquant la probabilité d'infection par le paludisme dans chaque image d'une cellule GR. Elle comprend une seule unité neuronale avec activation sigmoid pour produire une probabilité indiquant si un individu est infecté ou non-infecté.

G. Explicabilité : Elle est donnée par :

- **Grad-CAM** : Identifie les patchs contribuant le plus à la prédiction.
- **LIME** : Perturbe localement les patchs pour interpréter la décision.

1.2. Protocole Expérimental de l'explicabilité (XAI)

L'interprétabilité des modèles d'apprentissage profond est cruciale en diagnostic médical, où la confiance clinique repose sur la transparence décisionnelle. Ce protocole évalue l'explicabilité du modèle **Stacked-LSTM avec mécanisme d'attention** proposé par Gaouar et al. (2025) pour la détection du paludisme sur des images de frottis sanguin, en utilisant **LIME** et **Grad-CAM**. L'objectif est de quantifier l'alignement entre les régions identifiées par le modèle et les signaux parasitologiques réels (*Plasmodium*).

A. Génération des Explications

Étant donné que nous avons intégré un mécanisme d'attention dans notre approche, afin d'augmenter significativement les performances de notre classifieur Stacked-LSTM avec mécanisme d'attention, ainsi que pour améliorer l'explicabilité des résultats obtenus par ce dernier, l'ensemble des formules mathématiques liées aux calculs des métriques de l'explicabilité pour les deux techniques LIME et Grad-CAM ont été adaptées pour prendre en compte les poids d'attention générés par notre classifieur.

a. Grad-CAM adapté aux poids d'attention

Le principe de base repose sur le calcul des gradients de la sortie « parasité » par rapport aux **poids d'attention.** Pour identifier les régions critiques et exploiter pleinement les avantages et apports du mécanisme d'attention pour l'explicabilité de notre modèle, nous nous sommes basés sur la formule **(7)**, pour avoir les formules **(30)** et **(31)** suivantes, permettant de calculer les cartes de chaleurs :

$$\nabla_{\alpha} = \frac{\partial \mathcal{L}_{\text{parasite}}}{\partial \alpha_i} \qquad \textbf{(30)}$$

Où :

- $\mathcal{L}_{\text{parasite}}$ est la fonction de perte spécifique à la classe "parasite" dans le modèle de classification binaire. Elle quantifie l'erreur entre :
  - **La prédiction du modèle** : Probabilité estimée que l'image contienne des *Plasmodium*.
  - **La vérité terrain** : Étiquette réelle (1 = parasite, 0 = sain).
- α_i **:** Poids d'attention du patch i
- $\nabla_{\alpha i}$ **:** Gradient de la perte $\mathcal{L}$ par rapport à α_i (calculé via la rétropropagation) et permet de mesurer l'impact dynamique de α_i sur la décision.

$$\text{Si} \begin{cases} \nabla \alpha i < 0 : \textit{Augmanter } \alpha i, \textbf{réduit } l'\textbf{erreur} \Rightarrow \textbf{Le patch est critique} \\ \nabla \alpha i > 0 : \textit{Réduire } \alpha i, \textbf{réduit } l'\textbf{erreur} \Rightarrow \textbf{Le patch est Trompeur} \end{cases}$$

Ainsi, nous obtenons la carte de chaleur $\boldsymbol{H}$ par la formule suivante :

$$H_{Grad-CAM} = \sum_i \alpha_i * \nabla_{\alpha i} \qquad \textbf{(31)}$$

b. LIME

L'explicabilité avec LIME se déroule comme suit :

- Tout d'abord, génération de 1000 échantillons perturbés, en masquant aléatoirement 50% des patches.
- Entraînement d'un modèle linéaire local pour approximer la prédiction générée par le classifieur. Cette approximation est calculée par la fonction suivante :

$$\boldsymbol{arg\ min}_{\boldsymbol{\beta}} \parallel \boldsymbol{f}(\boldsymbol{x_{pert}}) - \boldsymbol{\beta}.\boldsymbol{x_{pert}} \parallel^2 + \boldsymbol{\lambda} \parallel \boldsymbol{\beta} \parallel_1 \quad \textbf{(32)}$$

Où :

- Extraction des top-10 patches par magnitude des poids β.

B. Métriques d'explicabilité

a. Drop in Confidence (DC)

DC, mesure la baisse de confiance du modèle après masquage des régions critiques par la formule suivante :

$$\boldsymbol{DC} = \frac{\boldsymbol{P_{orig} - P_{masked}}}{\boldsymbol{P_{orig}}} \quad \textbf{(33)}$$

Où :

- P_{orig} : Représente la confiance initiale pour la classe « Parasité ».
- P_{masked} : Représente la confiance après masquage des ***k*** patches les plus importants.

- Si **DC est élevé > 0.7** : La région est ***critique*** et indique une forte dépendance aux patches identifiés.
- Si **DC est faible < 0.7** : La région est ***peu fiable*** ou la région est ***non exclusive***.

b. Mean Average Precision (MAP)

- Le MAP évalue l'accord spatial entre les régions expliquées et les annotations "experts" :

$$\boldsymbol{MAP} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{AP_i} \quad \textbf{(34)}$$

Où : AP (Average Precision) est donnée par :

$$\boldsymbol{IoU} = \frac{\boldsymbol{Region_{explic} \cap Region_{expert}}}{\boldsymbol{Region_{explic} \cup Region_{expert}}} \quad \textbf{(35)}$$

- Un "True Positive" est comptabilisé si **IoU > 0.5**

- **Le seuil clinique** est considéré comme excellent si **MAP > 0.7**

III. Expériences, Résultats et Discussion

Nous avons réalisé nos expériences en utilisant Google Colab, une plateforme basée sur le cloud qui offre un accès gratuit aux GPU, ce qui était idéal pour gérer les exigences computationnelles de l'entraînement et du test des modèles d'apprentissage profond. Google Colab prend en charge Python et des frameworks de deep learning populaires comme TensorFlow et Keras, que nous avons utilisés pour implémenter les modèles VGG-16, VGG-19, MobileNetV2, ViT et Stacked-LSTM, ainsi que les techniques d'explicabilité (Grad-CAM, LIME). Cette plateforme nous a permis de traiter efficacement de grands ensembles de données, d'effectuer des évaluations de modèles et d'appliquer diverses méthodes d'interprétabilité, tout en bénéficiant de son intégration avec Google Drive, pour une gestion et un partage des données facilités. L'environnement basé sur le cloud a également facilité la collaboration et assuré la reproductibilité de nos expériences.

L'entraînement et l'évaluation des modèles susmentionnés ont été effectués en utilisant un ensemble cohérent d'hyperparamètres fondamentaux, pour garantir une comparaison équitable. Toutes les images ont été redimensionnées à 224 × 224 pixels avec trois canaux RGB pour standardiser les dimensions d'entrée entre les modèles. L'apprentissage par transfert a été utilisé pour VGG-16, VGG-19 et MobileNetV2 en initialisant les réseaux avec des poids préentraînés d'ImageNet, tandis que les modèles ViT et Stacked-LSTM ont traité des représentations séquentielles dérivées des images d'entrée. L'entraînement a été effectué en utilisant une taille de lot de 32 pour tous les modèles, en optimisant les paramètres avec l'optimiseur Adam à un taux d'apprentissage de 0,0001. La fonction de perte d'entropie croisée binaire a été utilisée pour traiter la tâche de classification binaire, et l'entraînement a été effectué sur 100 époques avec un arrêt anticipé (patience = 5) pour éviter le surapprentissage.

Des techniques de régularisation ont également été appliquées, avec des mécanismes de dropout intégrés dans les couches du réseau, pour améliorer la généralisation. Les fonctions d'activation ont été soigneusement sélectionnées : ReLU a été utilisé dans les couches intermédiaires des modèles VGG, tandis que l'activation sigmoïde a été appliquée aux couches de sortie de tous les modèles pour produire des scores de probabilité.

Au-delà de ces hyperparamètres communs, des configurations spécifiques aux modèles ont été introduites pour adapter les architectures à leurs paradigmes d'apprentissage respectifs. Pour les modèles convolutionnels, VGG-16 et VGG-19, les couches denses incluaient un taux de dropout de 50 % pour atténuer le surapprentissage. Ces modèles ont tiré parti de leurs capacités d'extraction de caractéristiques convolutionnelles profondes, suivies de couches entièrement connectées pour la classification.

Concernant le modèle MobileNetV2, nous l'avons utilisé comme une épine dorsale de caractéristiques et ce, en désactivant sa tête de classification originale et en la remplaçant par une tête de classification ajoutée au-dessus de MobileNetV2. C'est-à-dire que nous l'avons configuré et utilisé uniquement comme un extracteur de caractéristiques, sans les couches de classification d'origine. La couche Global Average Pooling 2D nous a permis de réduire chaque carte de caractéristiques à une valeur moyenne, tout en diminuant le nombre de paramètres ; ensuite, un Dropout de 30 % a été appliqué deux fois pour limiter le surapprentissage ; la couche Dense composée de 64 neurones, ReLU est une couche entièrement connectée, utilisée pour apprendre des représentations abstraites ; enfin, nous avons utilisé une couche de sortie représentée par une fonction Sigmoid composée d'un seul neurone, pour la

classification binaire produisant une probabilité pour les classes « Infecté » vs « Non infecté ». L'ensemble des couches du modèle a été conçu pour maintenir un bon équilibre entre légèreté, performance, robustesse et efficacité computationnelle.

Le modèle Vision Transformer (ViT), conçu dans notre étude, représente un modèle léger, adapté à la classification binaire des images de globules rouges en classes infectées et non infectées. Contrairement au modèle CNN présenté, le modèle ViT exploite le mécanisme d'auto-attention pour capturer les dépendances à long terme et les relations contextuelles entre les régions de l'image. Chaque image d'entrée est redimensionnée à 224x224 pixels et partitionnée en blocs de taille fixe, non chevauchants, de 16x16 pixels. Ces patches sont aplatis et projetés linéairement dans un espace d'incorporation de dimension inférieure (dimension 64), formant une séquence de 196 tokens. Des embeddings positionnels sont ajoutés à chaque jeton de patch, pour préserver les informations spatiales perdues lors de l'aplatissement. Cela permet au modèle de distinguer les emplacements des patches au sein de la structure 2D de l'image. Les patches encodés sont passés à travers une pile de quatre blocs Transformer. Chaque bloc se compose de Normalisation de Couche, d'Auto-Attention Multi-Têtes (MHSA) avec 4 têtes, de connexions résiduelles, et d'un réseau de neurones à propagation avant à deux couches (MLP) avec activation GeLU et dropout. Ces blocs Transformer permettent au modèle d'apprendre des interactions complexes entre des patches éloignés, améliorant ainsi sa capacité à détecter des motifs globaux, caractéristiques des cellules infectées par le paludisme. Au lieu d'utiliser le token conventionnel (CLS) appliqué dans les implémentations originales de ViT, nous agrégeons les informations de tous les patches, en utilisant une couche GlobalAveragePooling1D. Cela est suivi d'une couche entièrement connectée avec 128 neurones (activation ReLU) et d'une couche dense activée par sigmoïde finale qui produit la probabilité d'infection.

En revanche, le modèle Stacked-LSTM, conçu pour capturer les dépendances temporelles dans les représentations des caractéristiques, intégrait un traitement séquentiel. Chaque couche LSTM se composait de 64 cellules LSTM, et la séquence d'entrée était structurée en 50 étapes temporelles par instance. La régularisation a été appliquée différemment dans cette architecture, avec un taux de dropout de 50 % dans la première couche LSTM et un taux de dropout réduit de 20 % dans la deuxième couche LSTM, pour maintenir un équilibre entre la rétention des caractéristiques et la prévention du surapprentissage.

Contrairement aux modèles VGG, où les activations ReLU dominaient les couches intermédiaires, le modèle Stacked-LSTM s'appuie sur une fonction d'activation sigmoïde dans sa couche de sortie entièrement connectée, pour générer des probabilités de classification.

1. Métriques et Évaluation de Classification

Pour évaluer les performances des modèles entraînés, nous nous sommes basés sur les métriques les plus communément utilisées pour la tâche de classification et nous les avons testés sur les données de test, afin de déterminer dans quelle mesure ils peuvent se généraliser aux données non vues et faire des prédictions précises dans la tâche de détection des cellules infectées par le paludisme par rapport aux cellules non infectées.

Principales métriques d'évaluation

- **Justesse (Accuracy) :** C’est une métrique courante, utilisée pour évaluer la performance globale d'un modèle de classification. Elle est définie comme le ratio des instances correctement prédites (à la fois les vrais positifs (TP) et les vrais négatifs (TN)) par rapport

au nombre total d'instances dans l'ensemble de données. Autrement dit, c'est la proportion d'instances positives correctement prédites parmi toutes les instances positives prédites, elle est calculée par la formule :

$$\boldsymbol{Accuracy = \frac{TP+FN}{TP+FP+TN+FN}} \quad \textbf{(36)}$$

Où :

- **TP (Vrai Positif. VP) :** nombre d'instances positives correctement prédites.
- **TN (Vrai Négatif. VN) :** nombre d'instances négatives correctement prédites.
- **FP (Faux Positif. FP) :** nombre d'instances positives prédites incorrectement.
- **FN (Faux Négatif. FN) :** nombre d'instances négatives incorrectement prédites.

- **Précision :** La précision mesure l'exactitude des prédictions positives faites par le modèle, elle est calculée par la formule :

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \quad \textbf{(37)}$$

- **Rappel ou sensibilité (Recall or Sentivity) :** La sensibilité est le taux de vrais positifs ; dans la détection des maladies, un positif fait référence au patient ayant cette maladie. Donc la sensibilité (aussi appelée rappel) est la capacité du modèle à diagnostiquer correctement un patient malade, considéré comme tel. Elle représente la proportion d'instances positives, correctement prédites, par rapport à toutes les instances positives réelles, et est donnée par la formule :

$$\boldsymbol{Sensitivity = \frac{TP}{TP+FN}} \quad \textbf{(38)}$$

- **F1 Score :** La moyenne harmonique de la précision et du rappel fournit une mesure unique de la performance du modèle. Le score F1 est une métrique précieuse dans les problèmes de classification d'images car il prend en compte à la fois les faux positifs (TP) et les faux négatifs (FN), offrant une évaluation équilibrée de la performance d'un modèle, en particulier dans les situations avec des ensembles de données déséquilibrés.

  Le score F1 est calculé par la formule suivante :

$$\boldsymbol{F1\ Score = 2 * (Precision * Recall)/(Precision + Recall)} \quad \textbf{(39)}$$

- **Spécificité (Specificity) :** La spécificité mesure la capacité du modèle à identifier correctement les échantillons sains (VN). La formule nous permettant de calculer cette métrique est donnée par :

$$\boldsymbol{Spécificité = \frac{TN}{TN+FP}} \quad \textbf{(40)}$$

- **AUC (Area Under the Curve):** est une métrique clé qui mesure la capacité d'un modèle à distinguer deux classes (ex: "parasite" vs "sain"). Elle évalue les performances à tous les seuils de décision possibles, offrant une vision globale de la robustesse du modèle et est donnée soit par la formule suivante :

$$AUC = \frac{1}{n_+ * n_-} \sum_{k=1}^{n_+} rang(s_+^{(k)}) - \frac{n_+ + 1}{2} \qquad (41)$$

Où :

- $\mathbf{n_+}$ **:** est le nombre d'échantillons **positifs** (classe minoritaire) ;
- $\mathbf{n_-}$ **:** est le nombre d'échantillons **négatifs** (classe majoritaire) ;
- $(s_+^{(k)})$ **:** est le score de prédiction du $k^{\text{ième}}$ échantillon **positif** (entre 0 et 1) ;
- $rang\left(s_+^{(k)}\right)$ **:** est le rang du score $(s_+^{(k)})$ dans l'ensemble trié **tous scores confondus.**

Ou bien par cette formule :

$$AUC = \sum_{i=1}^{n} \left\{ \frac{[(FPR_i - FPR_{i-1}) \times (TPR_i + TPR_{i-1})]}{2} \right\} \qquad (42)$$

Où :

- **FPR** (Taux Faux Positifs) = FP / (FP + VN)
- **TPR** (Taux Vrais Positifs (Sensibilité)) = VP / (VP + FN)

Interprétation clinique : le **Tableau 5** suivant résume des seuils relatifs à l'AUC et leurs interprétations cliniques.

Tableau 5. Interprétation clinique de l'AUC

Valeur AUC	Performance	Risque clinique	Fiabilité du modèle
0.90 - 1.00	Excellente	Faible taux d'erreur	Très fiable
0.80 - 0.89	Bonne	Surveillance recommandée	Fiable
0.70 - 0.79	Moyenne	Vérification manuelle nécessaire	Peu fiable, nécessitant une surveillance continue
< 0.70	Médiocre	Non fiable pour usage clinique	Pas fiable avec une prédominance de choix aléatoires

2. Résultats de Classification

Dans le cadre de ce travail, nous avons commencé par comparer les performances de notre modèle proposé, Stacked-LSTM avec un mécanisme d'attention, à quatre modèles Deep Learning (VGG-16, VGG-19, ViT et MobileNetV2), confirmés et couramment utilisés pour la

tâche de classification d'images médicales. Par la suite, une étude d'ablation a été menée pour évaluer l'impact du mécanisme d'attention en comparant les performances de notre modèle **avec** et **sans** attention. Enfin, une étude statistique complète a été effectuée afin de démontrer formellement la supériorité de notre approche **MalariaScope** par rapport aux autres modèles expérimentés susmentionnés.

Tableau 6. Comparaison des performances entre les modèles testés et l'approche proposée Stacked-LSTM avec un mécanisme d'attention pour la détection précoce de la malaria

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
VGG-16	0.9802	0.9839	0.9765	0.9763	0.9801	0.9780
VGG-19	0.9601	0.9681	0.9521	0.9528	0.9604	0.9588
ViT	0.9509	0.9273	0.9746	0.9734	0.9498	0.9503
MobileNetV2	0.9303	0.9475	0.9132	0.9160	0.9315	0.9258
Stacked-LSTM avec Attention	**0.9912**	**0.9867**	**0.9956**	**0.9959**	**0.9911**	**0.9912**

Le **Tableau 6** résume les performances des modèles expérimentés sur le jeu de données de la malaria. Le Stacked-LSTM proposé avec un mécanisme d'attention a atteint la plus haute précision de **0.9912**, surpassant VGG-16, VGG-19, ViT et MobileNetV2, qui ont obtenu des scores de précision de 0.9802, 0.9601, 0.9509 et 0.9303 respectivement. Les résultats mettent en évidence l'amélioration significative apportée par l'incorporation du réseau LSTM amélioré avec le mécanisme d'attention, atteignant des valeurs supérieures en termes de précision, de sensibilité, de spécificité, de précision, de F1-score et d'AUC par rapport aux autres architectures expérimentées.

Pour illustrer davantage les performances obtenues pour chaque modèle, et surtout le nombre de cas VP, VN, FN et FP générés par chaque modèle, les **Figures** allant de **71** à **76** présentent les matrices de confusion des modèles présentés.

Le VGG-16 **(Figure 71)** a classé avec précision 4068 échantillons non infectés et 4034 échantillons infectés, avec 98 cas de faux positifs et 66 de faux négatifs, démontrant une haute précision et sensibilité. Cependant, il reste de la place pour améliorer la réduction des faux négatifs.

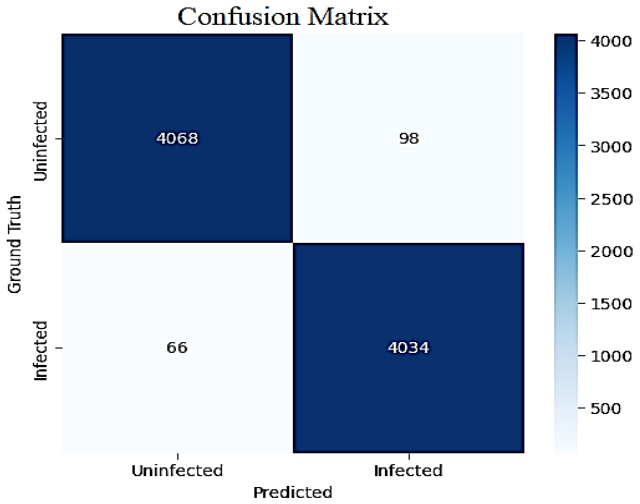


Figure 71. Matrice de Confusion du modèle VGG-16

En revanche, VGG-19 **(Figure 72)** a correctement identifié 3935 échantillons non infectés et 4001 échantillons infectés, mais a mal classé 198 échantillons non infectés comme infectés et 132 échantillons infectés comme non infectés, ce qui suggère une spécificité inférieure.

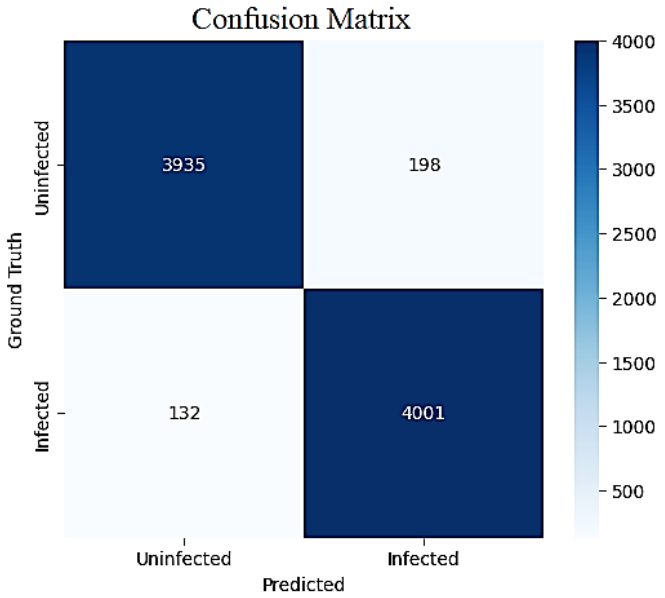


Figure 72. Matrice de Confusion du modèle VGG-19

MobileNetV2 **(Figure 73)** est un modèle léger, conçu pour les appareils mobiles et edges, qui privilégie l'efficacité et la rapidité. Cela est confirmé par sa matrice de confusion, qui montre que ce modèle a pu classer 3775 cas infectés et 3915 cas non infectés correctement. Cependant, nous notons également qu'il a le plus grand nombre de cas FN = 217, ce qui est énorme dans le contexte médical, puisque son taux de FN est de 5,25 %. Par ailleurs, il a mal classé 359 cas non infectés comme infectés, confirmant ainsi sa faible spécificité et précision.

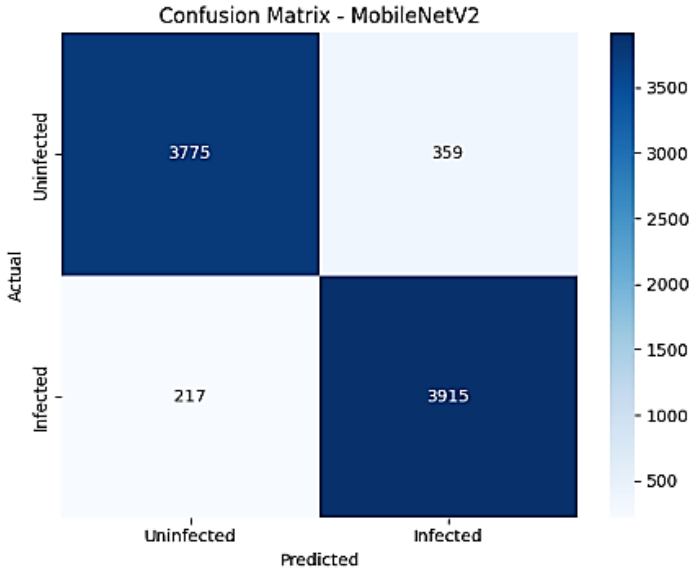


Figure 73. Matrice de Confusion du modèle MobileNetV2

Le modèle basé sur le Transformer **(Figure 74)** a correctement classé 4022 cas comme infectés et 3840 cas comme non infectés, confirmant ainsi sa haute spécificité et précision. Bien qu'il ait mal classé 105 individus en bonne santé comme infectés, cela montre que ViT est assez conservateur dans les prédictions positives. Cependant, mal classer 301 cas infectés à tort prédit comme non infectés représente un taux élevé de faux négatifs (7,27 %), ce qui traduit sa faible sensibilité.

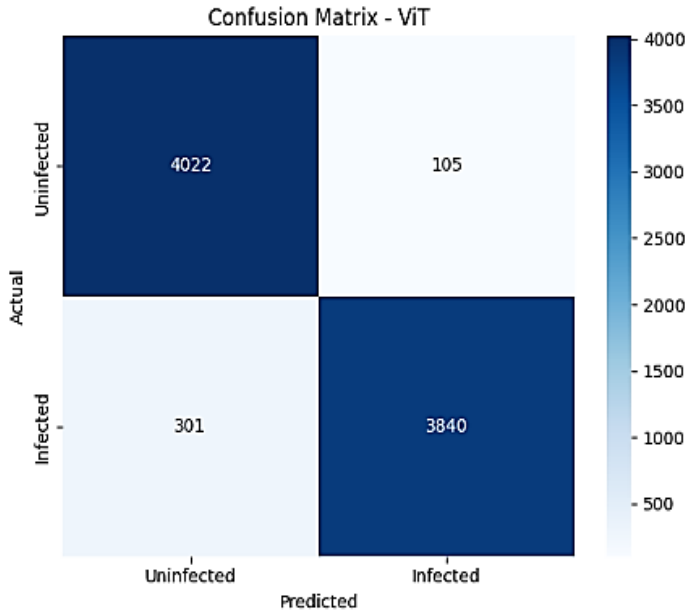


Figure 74. Matrice de Confusion du modèle ViT

Pour démontrer la contribution du mécanisme d'attention et son importance dans notre approche, nous avons mené une étude d'ablation. Le **Tableau 7** présente les résultats comparatifs pour le modèle Stacked-LSTM **avec** et **sans** mécanisme d'Attention.

Tableau 7. Tableau comparatif entre les deux modèles Stacked-LSTM avec Attention et son homologue sans attention

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
Stacked-LSTM sans Attention	0.9758	0.9877	0.9644	0.9638	0.9815	0.9761
Stacked-LSTM avec Attention	**0.9912**	**0.9867**	**0.9956**	**0.9959**	**0.9911**	**0.9912**

L'étude d'ablation confirme le rôle crucial du mécanisme d'attention dans l'amélioration des performances du modèle. Spécifiquement, la précision est passée de 0,9758 à 0,9912, tandis que la spécificité, la précision, le score F1 et l'AUC ont connu des améliorations notables. Ces résultats soulignent la capacité du mécanisme d'attention à se concentrer efficacement sur les régions critiques, améliorant ainsi la précision du modèle et réduisant les faux positifs, rendant le modèle plus robuste et efficace pour les applications médicales et pour son adoption clinique.

Les **Figures 75** et **76** illustrent davantage cette différence de performance en utilisant des matrices de confusion. Le modèle Stacked-LSTM sans attention **(Figure 75)** a classé 4000 vrais négatifs et 4067 vrais positifs, mais a mal classé 50 faux positifs et 150 faux négatifs. Ces erreurs indiquent que, bien que le modèle capture bien les dépendances temporelles, il manque encore de précision dans la focalisation sur les caractéristiques les plus pertinentes.

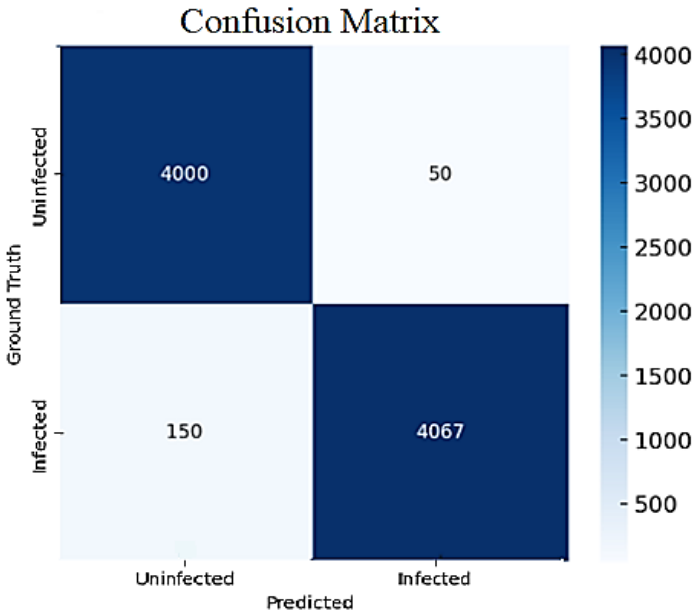


Figure 75. Matrice de confusion du modèle Stacked-LSTM sans mécanisme d'Attention

En revanche, le modèle Stacked-LSTM avec attention **(Figure 76)** a considérablement amélioré la classification, identifiant correctement 4115 échantillons non infectés et 4078 échantillons infectés tout en réduisant les faux positifs à 18 et les faux négatifs à 55. Cette réduction substantielle des erreurs de classification démontre la capacité du mécanisme d'attention à mettre en évidence les régions critiques dans la séquence d'entrée, améliorant ainsi la robustesse du modèle.

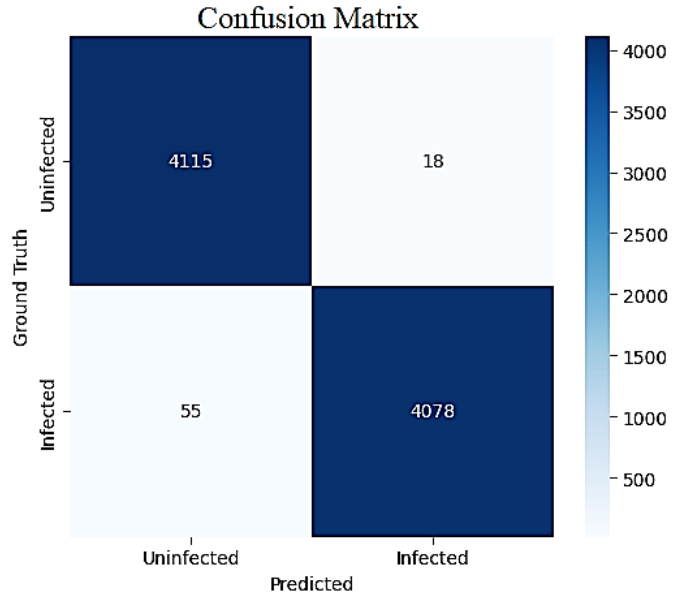


Figure 76. Matrice de confusion du modèle Stacked-LSTM avec mécanisme d'Attention

La **Figure 77** illustre typiquement les courbes de précision / perte de l'entraînement et du test, montrant les progrès d'apprentissage du modèle et sa stabilisation au cours des 100 époques.

Ces tests confirment le fait que la performance du modèle s'est améliorée de manière constante pendant l'entraînement, finissant par se stabiliser. Elle met également en évidence l'observation courante d'une perte de test plus élevée par rapport à la perte d'entraînement, ce qui indique la capacité de généralisation du modèle aux nouvelles données.

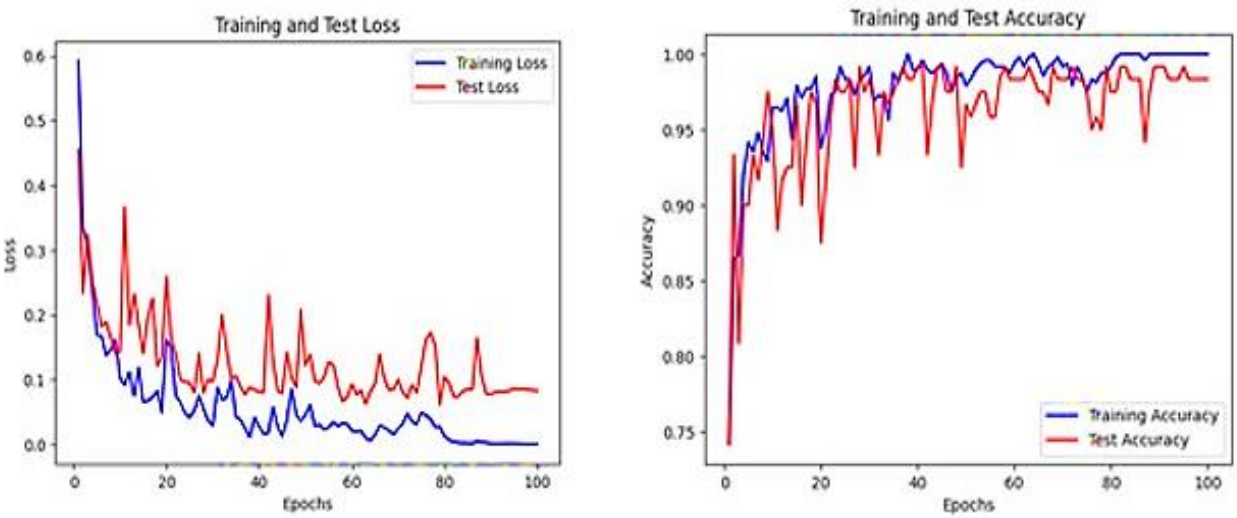


Figure 77. Résultats de tests et d'entrainement du modèle Stacked-LSTM avec attention

3. Étude Statistique

3.1. Analyse de la Courbe de Décision (Decision Curve Analysis (DCA))

La DCA est une méthode statistique innovante, développée en 2006 par Vickers et Elkin **(Vickers and Elkin, 2006)** pour évaluer l'utilité clinique des modèles prédictifs, des tests diagnostiques ou des biomarqueurs. Elle répond à une question clé : "Quand faut-il utiliser un modèle pour guider les décisions médicales, en tenant compte des bénéfices et des risques ?"

Conformément au contexte clinique, elle équilibre les résultats vrais positifs et faux positifs, en quantifiant le **bénéfice net** sur une plage de seuils de probabilités.

a. Concepts Fondamentaux

- ***Seuil de Probabilité p_t :*** C'est la probabilité minimale à partir de laquelle un clinicien décide d'un traitement, impliquant que les bénéfices du traitement surpassent ses risques. Typiquement, la plage du seuil $\mathbf{p_t}$ est définie comme suit : 0% < $\mathbf{p_t}$ < 50% **(Vickers and Elkin, 2006)**.

- ***Bénéfice Net (Net Benefit) :*** C'est une métrique centrale combinant :
  - **Vrais positifs** : patients correctement traités.
  - **Faux positifs** : traitements inutiles aux conséquences négatives.
  - **Formule :** $$\boldsymbol{Bénéfice\ Net} = \frac{VP}{n} - \frac{FP}{n} * \frac{p_t}{1-p_t} \quad \textbf{(43)}$$

Où :

n = Taille de l'échantillon.

b. Interprétation Graphique

- **Axe-X :** Seuil de probabilité $\mathbf{p_t}$ (de 0% à 100%).
- **Axes-Y :** Bénéfice net.
- **Courbes comparatives :**
  - **La courbe "Traiter tous" (Treat all)**: Ligne droite maximisant les VP mais ignorant les FP.
  - **La courbe "Ne traiter personne" (Treat None)** : Ligne à 0.
  - **Modèle évalué :** Courbe du bénéfice net calculé pour chaque $\mathbf{p_t}$.
  - **Règle d'or :** Le modèle est cliniquement utile si sa courbe est au-dessus des stratégies ***"traiter tous"*** et ***"ne traiter personne"*** pour une plage de $\mathbf{p_t}$ pertinente.

Cette section évalue tous les modèles d'apprentissage profond qui ont été présentés et utilisés pour classifier le paludisme à partir des images de frottis sanguins dans notre étude. Les paramètres d'évaluation les plus couramment utilisés ont été examinés de manière à être facilement compréhensibles par les médecins. Ils ont également été calculés et utilisés de manière pratique pour interpréter et évaluer les résultats du modèle **(Abbasian Ardakani et al., 2024)**. Comparer le potentiel de soutien à la décision clinique de six modèles—VGG-16, VGG-19, Stacked-LSTM avec Attention, Stacked-LSTM sans Attention, Vision Transformer (ViT), et MobileNetV2— est l'objectif de l'analyse de la courbe DCA affichée dans la **Figure 78**. Une plage de probabilité seuil de 0,01 à 0,99 est utilisée pour analyser le bénéfice net.

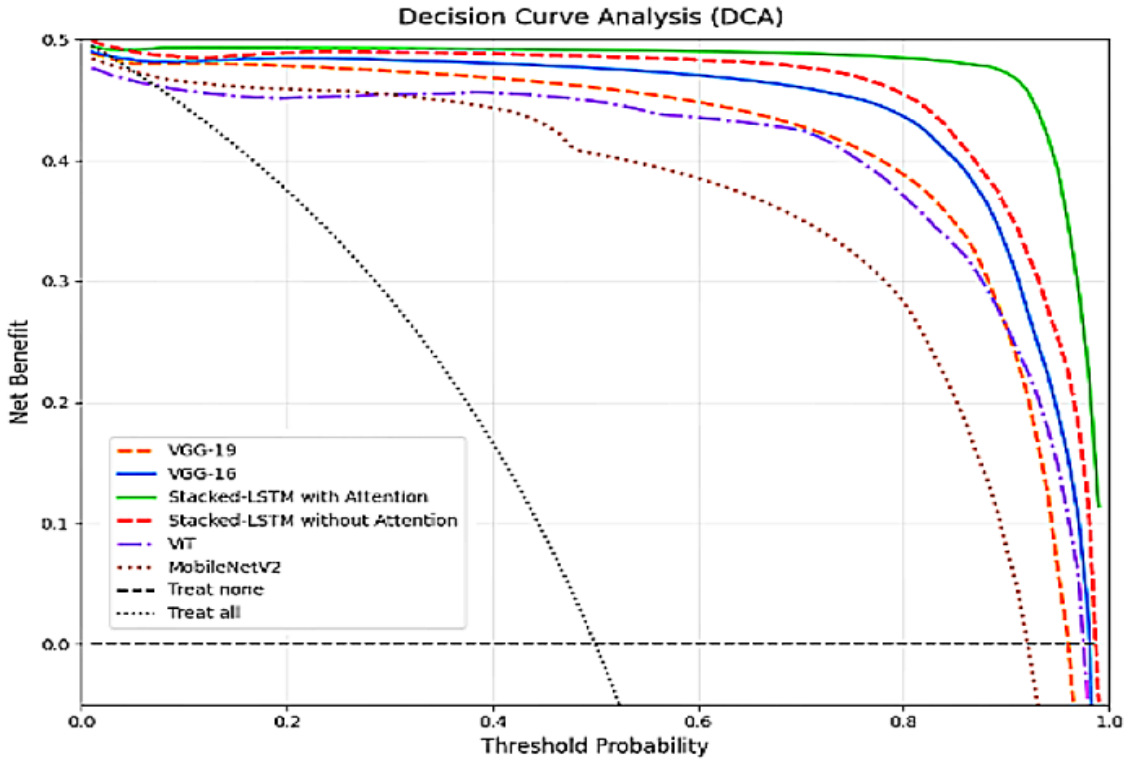


Figure 78. Courbes DCA combinées **(Gaouar et al., 2025)**

Le Stacked-LSTM avec Attention (Courbe Verte) présente une robustesse et une grande utilité clinique, maintenant une performance exceptionnelle à des seuils bas, moyens et élevés,

tout en affichant le plus grand bénéfice net à toutes les probabilités de seuil. La modélisation séquentielle et les mécanismes d'attention permettent à cette architecture de capturer des motifs spatiaux et des dépendances complexes, la rendant parfaite pour des applications cliniques à enjeux élevés où la réduction des faux négatifs est cruciale.

Le VGG-19 (courbe en pointillés oranges) performe de manière cohérente à travers les seuils et offre un soutien clinique solide dans les scénarios à seuil intermédiaire, il présente un bénéfice net légèrement inférieur à celui du modèle Stacked-LSTM avec attention. Il offre également un compromis utile entre sensibilité et précision. Selon les métriques standards, VGG-16 (courbe bleue solide) performe légèrement mieux que VGG-19 (précision = 0,9802, AUC = 0,9765). Malgré sa haute précision, sa courbe DCA indique un bénéfice net légèrement inférieur, ce qui indique que son utilité dans la prise de décision clinique est plus sensible au seuil.

Bien qu'il fonctionne bien, un Stacked-LSTM sans Attention (courbe en pointillés rouges) n'est pas aussi bon que son homologue amélioré par attention. Sa courbe DCA illustre ses limitations dans les cas difficiles à catégoriser, en affichant des bénéfices décroissants à des seuils plus élevés.

Avec une AUC de 0,9503, le modèle ViT (courbe en pointillés violets) présente un bénéfice net modéré à travers les seuils. Il obtient de bons résultats aux seuils précoces (détection sensible), mais son bénéfice diminue rapidement aux seuils supérieurs à 0,6, suggérant une sensibilité mais une résilience réduite, probablement en raison de la petite taille de l'ensemble de données.

Enfin, le bénéfice net le plus faible est montré par MobileNetV2 (courbe pointillée brune), qui est clairement au-dessus du seuil de 0,5. Sa courbe DCA indique un impact plus élevé des faux positifs, malgré une précision respectable de 0,9303 et un score F1 de 0,9315.

Les modèles au-dessus des lignes de référence « Traiter tous » et « Ne traiter personne » ont une utilité clinique positive, selon l'impact sur la décision clinique. L'option la plus prometteuse pour l'incorporation dans des pipelines de diagnostic automatisé du paludisme et la plus adaptée aux applications critiques, telles que le diagnostic précoce dans les zones endémiques, est le modèle Stacked-LSTM avec attention, qui a surpassé tous les autres modèles. ViT et MobileNetV2 doivent être utilisés avec prudence en raison de leur instabilité à des seuils moyens à élevés, tandis que les modèles basés sur VGG restent fiables.

3.2. Courbes Précision-Rappel (PRC : Precision-Recall Curves)

La **PRC** est un outil fondamental pour évaluer les performances des modèles de classification, **surtout quand les données sont déséquilibrées**. Contrairement à la courbe ROC, elle se concentre sur la qualité des prédictions positives **(Davis and Goadrich, 2006)**.

a. Concepts Fondamentaux

- La métrique "**Précision**" : permet de répondre à la question *"Parmi les prédictions positives, combien sont correctes ?"* ;

- La métrique "**Rappel**" répond à la question *"Parmi les vrais positifs, combien ont été détectés ?"* ;
- **Formules :** Les deux métriques sont données par les formules **(37)** et **(38)** respectivement.

b. Interprétation Graphique

- **Axe-X :** Rappel (de 0 à 1).
- **Axe-Y :** Précision (de 0 à 1).
- **Courbe :** Trace la précision en fonction du rappel pour différents seuils de décision.
- **Point idéal : (1, 1) =** Précision parfaite et rappel parfait.
- **Métrique clé : Précision Moyenne (Average Precision) (AP) =** Aire sous la courbe PRC (valeur entre 0 et 1).
    - **AP = 1 :** Modèle parfait.
    - **AP = 0.5 :** Performances aléatoires dans un cas déséquilibré.
    - **AP < 0.5 :** Pire que l'aléatoire.

Ce rapport d'analyse examine la performance Précision-Rappel de nos six modèles d'apprentissage profond pour la détection du paludisme à partir d'images de frottis sanguins. Comme nous pouvons le voir dans la **Figure 79**, les modèles sont comparés en utilisant les scores de Précision Moyenne (AP) et leur capacité à maintenir une haute précision à tous les niveaux de rappel (0,0 à 1,0).

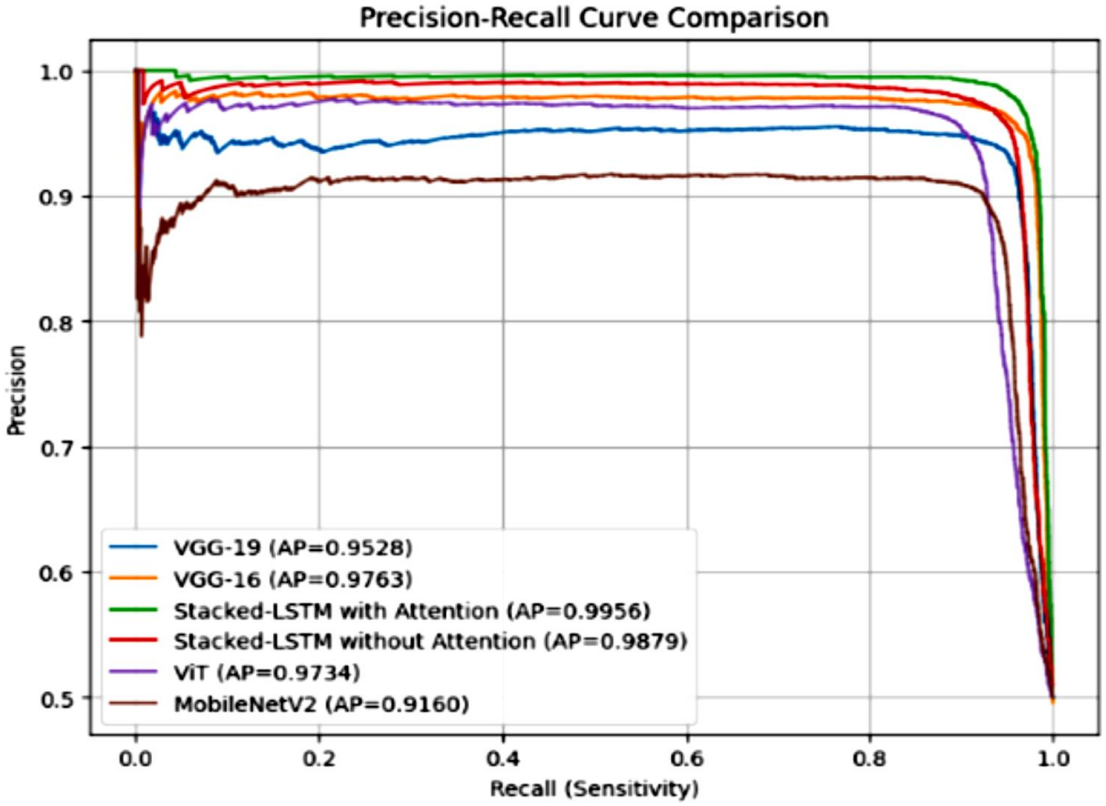


Figure 79. Courbes Précision-Rappel combinées **(Gaouar et al., 2025)**

Les principales conclusions de cette analyse mettent en évidence le modèle Stacked-LSTM avec Attention, qui obtient un score AP de 0,9956, montrant ainsi que le mécanisme

d'attention se concentre dynamiquement sur les régions des cellules infectées ; minimisant ainsi fortement les faux négatifs et le rendant cliniquement idéal pour les contextes nécessitant à la fois une haute sensibilité (rappel) et une fiabilité (précision) ; maintenant une précision presque parfaite même à un rappel élevé, ce qui en fait le meilleur dans l'ensemble. D'autre part, le modèle Stacked-LSTM sans attention, avec un AP de 0,9879, montre un bon compromis. Cependant, nous notons que l'absence d'attention fine entraîne de légères diminutions de la précision pour des valeurs de rappel supérieures à 0,8. En effet, l'absence d'un mécanisme d'attention peut entraîner une légère diminution de la capacité du modèle à se concentrer sur les caractéristiques clés, ce qui conduit à une baisse de la précision à des taux de rappel plus élevés. Les modèles basés sur CNN, VGG-16 et VGG-19, sont des modèles robustes, légèrement surpassés par les modèles basés sur LSTM, mais toujours fiables. Paradoxalement, VGG-16 surpasse VGG-19 (AP : 0.9763, 0.9528, respectivement), ce qui suggère que la profondeur seule ne garantit pas de meilleures performances pour cette tâche. Le modèle VGG-19 (AP = 0.9528) montre une baisse notable de la précision à des rappels plus élevés.

En revanche, VGG-16 (AP=0.9763) démontre une précision élevée et stable sur la majeure partie de la plage de rappel. Cliniquement, le classificateur VGG-16 a moins de FN (FN = 98) que le modèle VGG-19 (FN=198), ce qui le rend plus adapté à une utilisation clinique.

Le modèle Transformer (AP = 0.9734) donne de bons résultats, mais il est plus lourd en calculs que les LSTMs et n'offre aucun avantage apparent. Il est également légèrement moins stable que le modèle VGG-16. La baisse de précision à des valeurs de rappel élevées suggère qu'un pré-entraînement plus adapté ou plus de données sont nécessaires pour bien généraliser. MobileNetV2 (AP = 0.9160) est considéré comme le modèle le moins performant ; bien que léger et adapté aux appareils embarqués, il a du mal à maintenir une haute précision à mesure que le rappel augmente. Comme nous pouvons le voir dans la **Figure 79**, la précision de ce modèle s'effondre à 0,85 lorsque le rappel est de 0,6, le rendant inadapté à une utilisation clinique. Nous considérons que ce modèle est inadapté aux tâches de diagnostic à enjeux élevés.

Enfin, nous pouvons affirmer que le modèle Stacked-LSTM avec mécanisme d'attention émerge comme le modèle le plus fiable pour les tâches de détection du paludisme, combinant une haute précision (AP > 0,99) et une interprétabilité clinique. L'écart de performance significatif entre les meilleurs et les moins bons modèles (différence de 8 % en AP) souligne l'importance de la sélection du modèle pour cette application médicale. Les CNN et ViT sont des alternatives viables, mais nécessitent un réglage minutieux pour égaler leurs performances.

3.3. Comparaison Statistique des Modèles basée sur le test de McNemar

Le **Tableau 8** présente des comparaisons par paires entre nos modèles d'apprentissage profond (VGG-16, VGG-19, Stacked-LSTM avec et sans mécanisme d'attention, ViT et MobileNetV2) entraînés pour classifier le paludisme à partir d'images de frottis sanguins. L'évaluation se concentre principalement sur les faux négatifs (FN)—le type d'erreur le plus critique dans les diagnostics cliniques—car ils représentent des infections manquées.

Tableau 8. Résumé du test de McNemar comparant les modèles par paires

Model 1	Model 2	p-value	OR (FN2/FN1)	IC 95%	Taux de Réduction des FN
VGG-19	**VGG-16**	0.000000	0.50	[0.38, 0.68]	49.6%
MobileNetV2	**VGG-16**	0.000017	0.31	[0.23, 0.40]	69.3%
ViT	**VGG-16**	0.000000	0.22	[0.17, 0.29]	77.8%
Stacked-LSTM sans Attention	**VGG-16**	0.000000	0.44	[0.33, 0.59]	55.6%
MobileNetV2	**VGG-19**	0.000006	0.61	[0.49, 0.76]	39.0%
ViT	**VGG-19**	0.000000	0.44	[0.36, 0.54]	56.0%
Stacked-LSTM sans Attention	**VGG-19**	0.000000	0.88	[0.70, 1.11]	11.9%
VGG-16	**Stacked-LSTM avec Attention**	0.000000	0.84	[0.59, 1.19]	16.4%
VGG-19	**Stacked-LSTM avec Attention**	0.000000	0.42	[0.31, 0.58]	57.9%
MobileNetV2	**Stacked-LSTM avec Attention**	0.000258	0.26	[0.19, 0.34]	74.3%
ViT	**Stacked-LSTM avec Attention**	0.000553	0.19	[0.14, 0.25]	81.5%
Stacked-LSTM sans Attention	**Stacked-LSTM avec Attention**	0.000000	0.37	[0.27, 0.50]	62.9%

Les métriques dérivées du test de McNemar, des rapports de côtes (OR) et de l'intervalle de confiance à 95 % (IC) indiquées dans le **Tableau 8,** sont appliquées aux matrices de confusion et visent à évaluer si un modèle surpasse significativement un autre dans la réduction des cas de faux négatifs (FN).

Ces tests statistiques sont complémentaires. En fait, le test de McNemar répond à la question : Y'a-t-il une différence statistiquement significative ? Cependant, le rapport de côtes et l'intervalle de confiance répondent à la question : Quelle est l'ampleur et la fiabilité de la différence ? Rapporter les trois (P-value, OR et IC) donne une image complète, c'est-à-dire la signification, la direction et l'ampleur de la comparaison des modèles. Les colonnes Modèle 1 et Modèle 2 représentent les deux modèles comparés. Une colonne P-value montre le résultat du test de McNemar. Il évalue s'il existe une différence significative dans les erreurs de classification (en particulier les paires discordantes) entre les deux modèles ; si la valeur p est < 0,05, cela indique **une différence statistiquement significative**.

La colonne du Rapport de Côtes OR (FN2/FN1) compare les FN du Modèle 2 au Modèle 1. Si OR <1, le Modèle 2 a moins de FN (mieux), mais si OR >1, le Modèle 2 a plus de FN (moins bien). La colonne IC 95% montre l'intervalle de confiance à 95% du rapport de côtes. Si l'intervalle n'inclut pas 1, la différence est statistiquement significative. Enfin, la colonne Taux de Réduction des FN indique le pourcentage de réduction des FN lors de l'utilisation du Modèle 2 au lieu du Modèle 1. Si nous obtenons une valeur positive, cela implique une réduction des FN (amélioration) ; en revanche, si cette valeur est négative, cela suggère une augmentation des FN (détérioration). Les résultats révèlent des différences significatives dans les taux de faux négatifs (FN), avec des implications critiques pour le diagnostic du paludisme.

Les principales conclusions de cette analyse démontrent la supériorité du modèle Stacked-LSTM avec Attention, réduisant les faux négatifs (FN) de 57,9 % par rapport à VGG-19 (OR=0,42, IC à 95 % [0,31-0,58]), de 16,4 % par rapport à VGG-16 (OR=0,84), de 81,5 % par rapport à ViT (OR=0,19), et de 74,3 % par rapport à MobileNetV2 (OR=0,26). Pour mettre

en évidence l'impact des mécanismes d'attention, nous avons comparé le Stacked-LSTM avec attention à sa version sans attention. Cette comparaison a montré une réduction importante des FN = 62,9% et un OR=0,37.

À ce stade, nous pouvons confirmer la performance exceptionnelle d'un Stacked-LSTM avec attention, surtout que toutes les comparaisons montrent un OR < 1 et une valeur P proche de 0, confirmant sa supériorité statistique. D'un point de vue clinique, notre modèle stacked-LSTM avec attention présente significativement moins de cas de paludisme manqués, ce qui est essentiel pour prévenir des complications graves.

La comparaison des modèles CNN montre que le VGG-16 est nettement meilleur que le VGG-19 et réduit les FN de 49,6 % (OR=0,50, 95 %, IC[0,38-0,68]). Le VGG-16 et le VGG-19 réduisent les FN de 77,8 % (OR= 0,22) et 56 % (OR= 0,44) par rapport au ViT, ce qui implique que ces modèles CNN sont cliniquement plus efficaces et fiables.

Ces statistiques considèrent MobileNetV2 comme le pire modèle car il augmente les cas de FN de 39% à plus de 74% par rapport aux autres modèles.

4. Discussion

Les résultats de performance obtenus au cours des travaux menés dans cette thèse, ont été comparés à plusieurs approches de pointe pour la détection du paludisme à partir d'images de frottis sanguins microscopiques, comme résumé dans le **Tableau 9**. Cette analyse comparative démontre que le Stacked-LSTM avec mécanisme d'attention proposé a atteint une performance supérieure par rapport aux autres modèles évalués sur le même ensemble de données NIHNLM.

Plus précisément, les premières approches basées sur les CNN de **Rajaraman et al. (2018)**, telles que les architectures VGG-16 et ResNet-50 pré-entraînés, ont atteint des taux de précision compétitifs d'environ 0,9510. Par contre les méthodes d'apprentissage par ensemble ultérieures ont encore amélioré cette performance à une précision d'environ 0,9650 **(Rajaraman et al., 2019)**. Les méthodes CNN personnalisées par **Yang F et al. (2020a)** ont initialement atteint une précision de 0,8180, qui a été considérablement améliorée en incorporant des techniques de filtrage itératif du minimum global (IGMS) à 0,9346 **(Yang F. et al., 2020b)**, mettant en évidence le potentiel des méthodes de sélection de caractéristiques ciblées. Cependant, une autre approche basée sur CNN utilisant un YOLO en cascade basé sur YOLOv2 par **Yang H. et al. (2020)** a montré une précision relativement inférieure de 0,7922, démontrant les défis des modèles de détection en une seule passe pour la détection de petits parasites.

Tableau 9. Tableau comparatif des performances entre nos modèles et les modèles fondamentaux utilisant le même ensemble de données

Auteur(s)	Modèle(s) Utilisés	Accuracy
Kassim et al. (2021a)	PlasmodiumVF-Net Mask R-CNN and ResNet50	0.9000
Kassim et al. (2021b)	RBCNet pipeline, U-Net and Faster R-CNN	0.9776
Koirala et al. (2022)	YOLO-mp-3l	0.9020
	YOLO-mp-4l	0.9632
Rajaraman et al. (2018)	VGG-16, ResNet-50, Xception, Inception-V3, DenseNet-121, Simple CNN	0.9510
Rajaraman et al. (2019)	Ensemble deep neural networks	0.9650
Yang F et al. (2020a)	Android application and customized CNN	0.8180
Yang F et al. (2020b)	Iterative Global Minimum Screening (IGMS) and Customized CNN	0.9346
Yang H et al. (2020)	Cascading YOLO based-on YOLOv2	0.7922
Yu et al. (2020)	MalariaScreener: a smartphone-based system and customized CNN	0.9870
Travaux Actuels	VGG-16	0.9802
(Gaouar et al., 2025)	VGG-19	0.9601
	ViT	0.9509
	MobileNet-V2	0.9303
	Stacked-LSTM with attention mechanism	**0.9912**

Les approches hybrides qui combinent les tâches de segmentation et de classification, telles que les variantes de Mask R-CNN (PlasmodiumVF-Net et RBCNet), ont également montré des résultats prometteurs. **Kassim et al. (2021a)** ont atteint une précision de 0,90 avec PlasmodiumVF-Net, tandis que leur pipeline RBCNet utilisant U-Net et Faster R-CNN a amélioré la précision à 0,9776 **Kassim et al. (2021b)**. Ces résultats soulignent l'efficacité des approches multi-étapes en isolant et en analysant explicitement les régions d'intérêt.

Les solutions intégrées et basées sur des dispositifs mobiles ont également démontré un potentiel notable. Le système MalariaScreener a atteint une précision impressionnante de 0,9870 **Yu et al. (2020),** validant la faisabilité de déployer des systèmes de détection du paludisme sur des plateformes mobiles.

De même, les modèles YOLO légers de **Koirala et al. (2022)** ont atteint des précisions de 0,9020 (YOLO-mp-3l) et 0,9632 (YOLO-mp-4l), illustrant les compromis entre la complexité du modèle et la performance dans les applications en temps réel.

En comparaison, notre approche en deux étapes, impliquant une segmentation initiale des régions d'intérêt (ROI) suivie d'une classification utilisant un Stacked-LSTM avec mécanisme d'attention, a atteint la plus haute précision de 0,9912, surpassant la performance des méthodes précédentes, évaluées sur le même ensemble de données. Cette performance supérieure est attribuée à la capacité du modèle à capturer et à mettre en évidence efficacement les caractéristiques spatiales et contextuelles pertinentes dans les images de frottis sanguins. Le mécanisme d'attention a contribué de manière significative en mettant en évidence sélectivement les régions critiques, réduisant ainsi efficacement les faux positifs et les faux négatifs, améliorant ainsi la fiabilité du diagnostic de notre système MalariaScope et encourageant son adoption clinique.

De plus, nos résultats montrent des améliorations considérables par rapport aux architectures CNN standard telles que VGG-16 (0,9802) et VGG-19 (0,9601), confirmant ainsi les avantages de l'intégration des réseaux de neurones récurrents avec les mécanismes d'attention pour des tâches visuelles complexes.

5. Résultats de l'explicabilité et de la visualisation

Dans cette section, nous avons évalué l'explicabilité du classificateur Stacked-LSTM avec un mécanisme d'attention en utilisant deux méthodes, Grad-CAM et LIME, appliquées à l'ensemble du test.

La **Figure 80** ci-dessous illustre un exemple représentatif des résultats obtenus pour l'interprétabilité d'une image d'une cellule infectée et non infectée. LIME a identifié une région spécifique pour la cellule infectée, en mettant en évidence un petit cercle jaune correspondant au parasite. Cela démontre la capacité de LIME à isoler les caractéristiques critiques influençant la prédiction du modèle. Grad-CAM a également mis en évidence cette région-clé, la marquant avec des teintes rouges et jaunes intenses pour signifier son importance élevée dans la décision du modèle. Cependant, Grad-CAM a également révélé des activations dans des zones adjacentes, qui apparaissent en bleu et violet, indiquant une moindre pertinence. D'autres parties de la cellule, en particulier à gauche, étaient marquées en rouge et jaune, reflétant des activations significatives, bien que moins précises que celles détectées par LIME. Pour la cellule non infectée, LIME a principalement mis en évidence le contour de la cellule en jaune, suggérant que le modèle s'est concentré sur les bords pour prendre sa décision. D'autre part, Grad-CAM a affiché une distribution plus large des activations : les zones en haut et en bas de la cellule étaient marquées en rouge et en jaune, tandis que les régions centrales, à gauche et à droite, apparaissaient en bleu et en violet. Ce schéma d'activation plus large indique que Grad-CAM capture des zones d'intérêt plus générales, tandis que LIME fournit une interprétation plus ciblée et spécifique. Ces résultats mettent en évidence les similitudes et les différences entre les deux méthodes d'explicabilité.

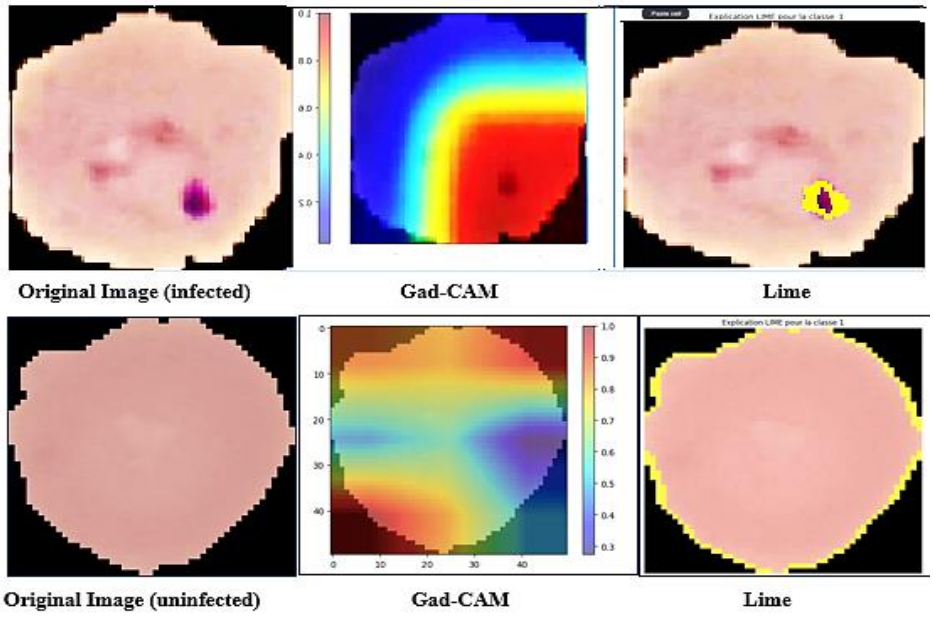


Figure 80. Exemple de résultat en Grad-Cam et LIME pour des images de cellule infectée et non-infectée **(Gaouar et al., 2025)**

Les deux techniques identifient des régions influentes pour les prédictions, mais leurs approches diffèrent : LIME fournit une interprétation localisée et précise, ce qui le rend bien adapté pour analyser des caractéristiques spécifiques, tandis que Grad-CAM offre une perspective globale en visualisant les activations sur l'ensemble de l'image. Les deux méthodes convergent vers la région parasitaire pour la cellule infectée, bien que Grad-CAM mette également en évidence des activations supplémentaires dans les zones environnantes. Pour la cellule non infectée, Grad-CAM révèle des activations plus diffuses, tandis que LIME se concentre explicitement sur les contours de la cellule. En conclusion, l'utilisation combinée de Grad-CAM et LIME fournit une interprétation complémentaire des décisions prises par le modèle Stacked-LSTM avec mécanisme d'attention. Grad-CAM fournit une vue d'ensemble en visualisant les activations à travers l'image, tandis que LIME ajoute une granularité locale en identifiant les caractéristiques les plus influentes. Cette complémentarité améliore notre compréhension des mécanismes de prise de décision du modèle et valide sa robustesse et sa fiabilité dans le contexte de la détection du paludisme.

Pour évaluer en profondeur l'explicabilité et l'interprétabilité de notre modèle, nous nous sommes concentrés sur l'explication des prédictions correctes et incorrectes.

5.1. Prédictions Correctes

(1) Nous avons commencé par sélectionner aléatoirement un échantillon de nos données de test.

(2) L'étiquette réelle de cet échantillon a été comparée avec l'étiquette prédite générée par notre modèle pour confirmer la justesse.

(3) Pour visualiser le processus de prise de décision, nous avons utilisé une instance explicative générée par le modèle, qui a produit une image et un masque pour la visualisation.

(4) Cette visualisation a mis en évidence les pixels spécifiques qui ont contribué positivement ou négativement à la prédiction du modèle, nous permettant de comprendre pourquoi le modèle a classé l'image de cette manière (**Figure 81**).

Cette approche a fourni des informations précieuses sur la manière dont le modèle traite les images d'entrée et les caractéristiques qu'il juge significatives pour faire des prédictions précises, renforçant ainsi la confiance dans sa fiabilité et son efficacité.

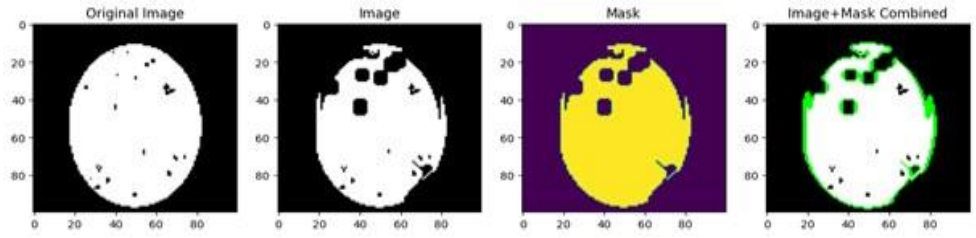


Figure 81. Les pixels contribuant positivement à la prédiction sont mis en évidence **(Gaouar et al., 2025)**

En plus d'analyser les prédictions correctes, nous avons également créé des visualisations pour illustrer les pixels qui ont contribué négativement à la catégorie de prédiction (**Figure 82**).

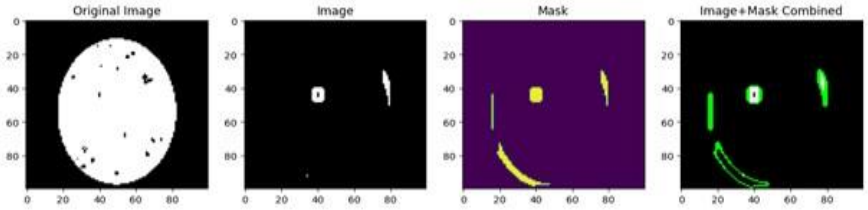


Figure 82. Les pixels qui contribuent négativement à la prédiction de la catégorie **(Gaouar et al., 2025)**

Ces visualisations ont révélé que les pixels des zones périphériques et centrales de l'image ont contribué positivement à la prédiction. Cependant, certains pixels de ces mêmes régions ont contribué négativement, indiquant une influence potentielle sur une autre catégorie.

5.2. Analyse des fausses prédictions

Pour obtenir des informations plus approfondies sur le comportement de notre modèle, nous avons examiné les cas où le modèle a fait des prédictions incorrectes. Cette analyse était cruciale pour comprendre quels pixels ont contribué à ces erreurs.

(1) ***Identification des échantillons mal prédits :*** Nous avons commencé par identifier les indices des échantillons que le modèle a mal classés.

(2) ***Sélection aléatoire pour l'analyse :*** Un échantillon a été sélectionné au hasard dans ce sous-ensemble d'instances mal classées, et les résultats ont été visualisés dans la **Figure 83** et la **Figure 84**.

(3) ***Figure 83 :*** Montre les pixels qui ont contribué positivement à la prédiction incorrecte.

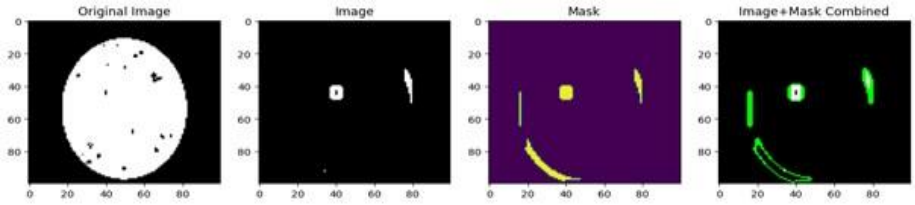


Figure 83. Les pixels ayants contribués positivement à la prédiction de faux positifs **(Gaouar et al., 2025)**

(4) ***Figure 84 :*** Montre les pixels qui ont contribué négativement à la catégorie de prédiction correcte.

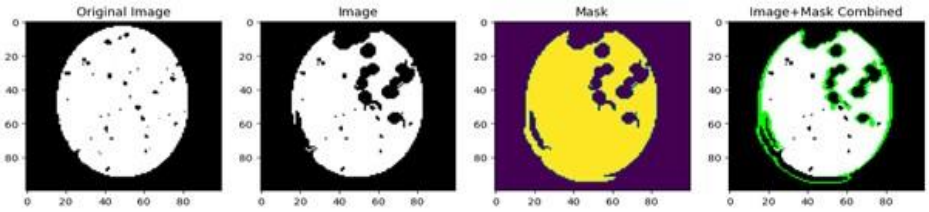


Figure 84. Pixels contribuant négativement à la prédiction de faux négatifs **(Gaouar et al., 2025)**

5.2.1. Observations Clés

A. Influence des pixels de bord

(1) ***Influence positive :*** Les pixels de bord, en particulier ceux formant des motifs continus, ont joué un rôle significatif dans les prédictions incorrectes. Ces pixels ont peut-être créé des caractéristiques trompeuses que le modèle a interprété comme appartenant à la mauvaise catégorie.

(2) ***Influence Négative :*** Lors de la visualisation des pixels qui ont contribué négativement à la catégorie correcte, nous avons observé que les pixels du milieu, qui auraient idéalement dû soutenir une prédiction précise, étaient éclipsés par l'influence des pixels de bord discontinu. Ces motifs discontinus, surtout lorsqu'ils sont disposés en forme arrondie, ont conduit à une contribution négative à la prédiction.

(3) ***Pixels centraux :*** Les pixels centraux semblaient moins influents dans les prédictions incorrectes, ce qui suggère que le modèle pourrait s'appuyer davantage sur les motifs des bords lorsqu'il commet des erreurs. Cependant, lorsque ces pixels centraux étaient censés indiquer positivement une cellule infectée, leur influence était diminuée par des signaux contradictoires provenant des pixels de bord.

B. Implications

Ces résultats sont très prometteurs, offrant une clarté précieuse sur le processus de prise de décision du modèle. Les informations obtenues grâce à cette analyse mettent en évidence des zones spécifiques de l'image, telles que les pixels le long des contours et au milieu de la forme, qui influencent de manière significative les prédictions—positivement et négativement.

Exploiter ces informations a le potentiel d'améliorer encore la précision de nos résultats de classification. Comprendre l'influence de régions spécifiques de pixels peut conduire à des améliorations dans le prétraitement, l'architecture du modèle, ou même les étapes de post-traitement, augmentant la fiabilité et l'utilité du modèle pour les praticiens dans le domaine du diagnostic du paludisme.

IV. Implémentation de la plateforme Web

En réponse aux statistiques alarmantes sur les décès liés au paludisme et aux coûts considérables associés aux méthodes de diagnostic existantes, nous avons développé une plateforme web accessible et efficace : "MalariaScope."

MalariaScope est une plateforme en ligne conçue pour simplifier le processus de diagnostic du paludisme, en tirant parti de notre modèle avancé d'apprentissage automatique. La plateforme offre une interface conviviale qui permet aux utilisateurs, y compris les professionnels de la santé et les techniciens de laboratoire, de télécharger des images de frottis sanguins directement sur le site. Lors du téléchargement d'une image, la plateforme la traite en utilisant notre modèle basé sur LSTM et renvoie rapidement un résultat diagnostique indiquant si le patient est susceptible d'avoir le paludisme.

V. Caractéristiques principales de MalariaScope

1. **Interface conviviale :** Le design intuitif de la plateforme permet aux utilisateurs de télécharger des images et de recevoir rapidement des résultats (**Figure 85**).
2. **Diagnostic Automatisé :** Une fois qu'une image est téléchargée, la plateforme applique automatiquement notre pipeline de prétraitement et notre modèle LSTM pour classer l'image comme infectée ou non infectée (**Figure 86**).
3. **Délai d'exécution rapide :** Le processus de diagnostic est rapide, fournissant des résultats presque instantanément, ce qui est crucial pour une prise de décision rapide dans les contextes médicaux.
4. **Hébergement local :** Actuellement, la plateforme est hébergée localement, permettant un déploiement et des tests contrôlés. Cette configuration garantit également la confidentialité et la sécurité des données pendant la phase de développement.

En offrant cette solution web, MalariaScope répond à des défis critiques dans le diagnostic du paludisme, en particulier dans les régions avec un accès limité aux outils de diagnostic traditionnels. Notre plateforme vise à réduire le fardeau des maladies liées au paludisme en fournissant une méthode de diagnostic rentable, précise et accessible, pouvant être utilisée dans divers établissements de santé.

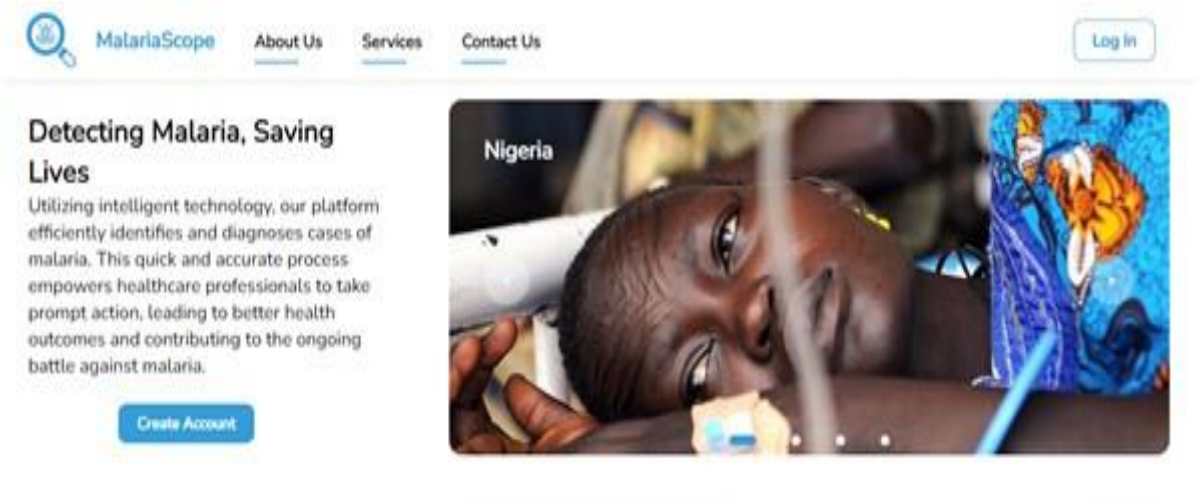


Figure 85. Interface du système MalariaScope **(Gaouar et al., 2025)**

L'application web proposée, MalariaScope, se compose de trois éléments-clés qui, ensemble, créent un outil de diagnostic complet et convivial :

1. Interface Graphique

A. **Conception et Utilisabilité :** L'interface, utilisant HTML, CSS et JavaScript, est conçue pour être intuitive et universellement compréhensible. Développé dans l'IDE Visual Studio Code, il est conçu de manière ergonomique pour faciliter son utilisation.

2. Fonctionnalité

Les utilisateurs peuvent créer des comptes et télécharger des images de frottis sanguins directement depuis leurs appareils mobiles. Après le téléchargement de l'image, l'interface traite les images rapidement et fournit des résultats diagnostiques détaillés, indiquant la présence ou l'absence de paludisme (**Figure 86**).

3. Intelligence Artificielle

A. Intégration : Le modèle LSTM pour la classification du paludisme est parfaitement intégré dans l'application, garantissant des capacités de diagnostic précises et efficaces. Ce composant est crucial pour analyser les images téléchargées et fournir des résultats fiables.

B. Opérations de base de données et de back-end

i. **Gestion de la base de données :** Ce composant gère le stockage et la récupération des données liées aux utilisateurs ainsi que l'intégration du modèle d'apprentissage. Bien que l'accent principal soit mis sur la classification binaire, la base de données suit également les interactions des utilisateurs pour des améliorations futures.

ii. **Pile technologique :** La fonctionnalité back-end est implémentée en utilisant SQL, Python et le framework Flask, fournissant une infrastructure robuste pour soutenir les opérations de l'application et améliorer ses performances.

4. Interaction utilisateur

A. Création de compte et authentification : Les utilisateurs créent des comptes personnels et s'authentifient sur la plateforme. Ce processus garantit un accès sécurisé et des expériences personnalisées.

B. Questionnaire et téléchargement d'images : Dans leurs profils, les utilisateurs remplissent un questionnaire concernant les symptômes du paludisme et des informations pertinentes, telles que vivre dans une région endémique du paludisme. Ils peuvent ensuite télécharger des images de frottis sanguins pour un traitement diagnostique utilisant la technologie IA (**Figure 86**).

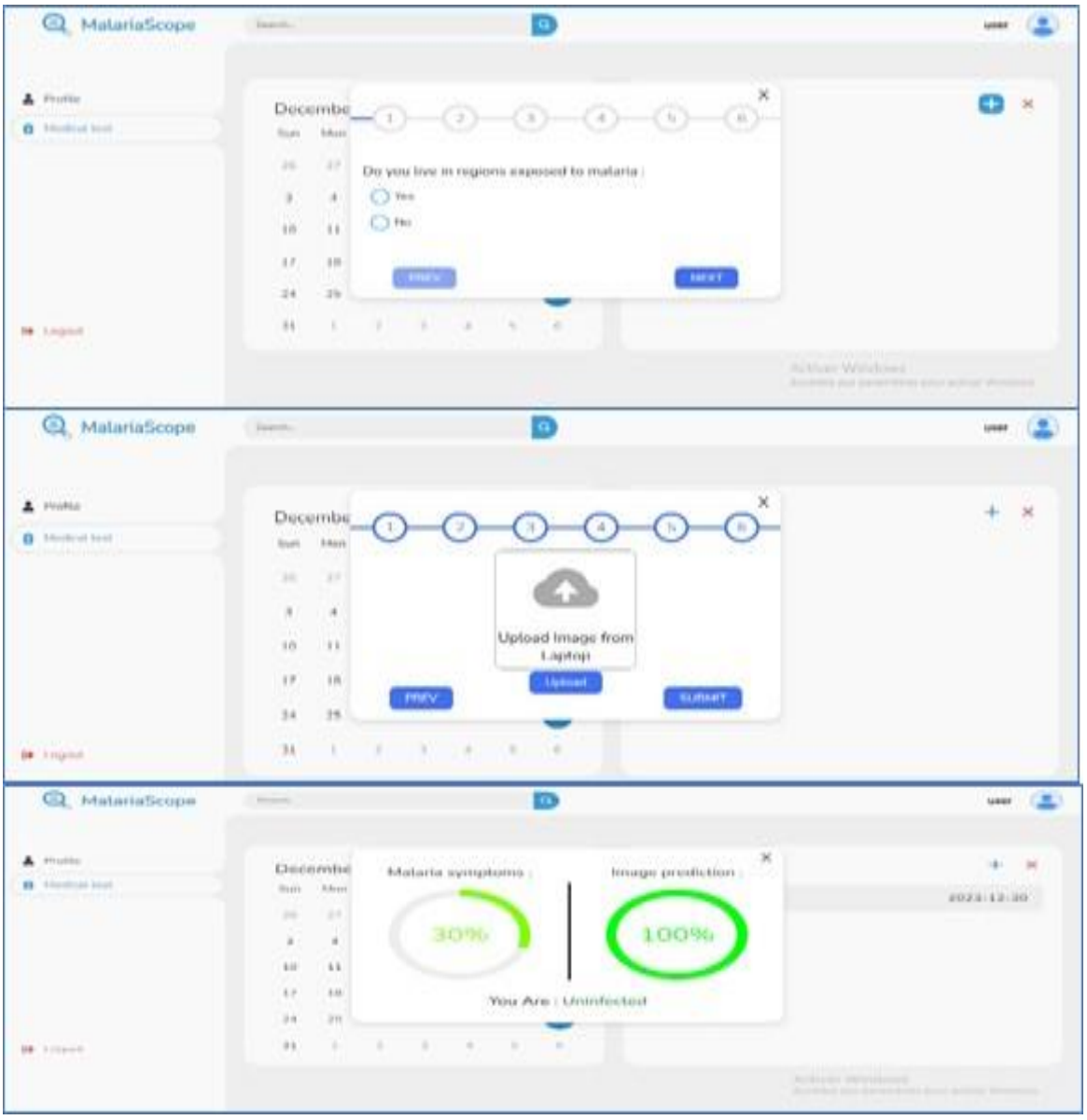


Figure 86. Diagnostic automatique avec MalariaScope **(Gaouar et al., 2025)**

VI. Conclusion

Cette recherche visait à répondre à la question : "Est-ce-que les réseaux Stacked-LSTM combinés avec des techniques XAI comme LIME et Grad-CAM peuvent-ils améliorer efficacement la précision et l'explicabilité du diagnostic du paludisme ?"

Nos expériences ont démontré que les réseaux Stacked-LSTM, lorsqu'ils sont combinés avec des techniques XAI, non seulement atteignent cet objectif mais dépassent également les attentes avec un taux de précision de plus de **0.9959**, un F1-Score de **0.9911** et une Accuracy de **0.9912**. Ce succès valide entièrement et sans équivoque notre approche **MalariaScope**, une approche innovante du diagnostic du paludisme, qui intègre des techniques avancées d'apprentissage profond avec des techniques XAI.

CONCLUSION GÉNÉRALE

CONCLUSION GENERALE

Le paludisme demeure une menace majeure pour la santé publique mondiale, en particulier dans les régions en développement où les ressources médicales sont limitées. La détection précoce et précise de cette maladie est cruciale pour réduire son impact, mais les méthodes traditionnelles de diagnostic, comme l'analyse microscopique des frottis sanguins, présentent des limites en termes de rapidité, d'accessibilité, de fiabilité et surtout elle est fortement conditionnée par le niveau de compétence et d'expertise des cytologistes. Ces défis soulignent la nécessité de solutions innovantes, robustes, explicables et interprétables par les médecins.

Dans cette thèse, nous avons évalué des architectures CNN standards, obtenant une précision notable avec VGG-16 (0,9802) et VGG-19 (0,9601) ainsi que d'autres modèles profonds plus récents tels que MobileNetV2 et ViT pour la détection du paludisme à partir d'images de frottis sanguins microscopiques. Plus significativement, nous avons introduit une approche novatrice utilisant un modèle Stacked-LSTM amélioré avec un mécanisme d'attention, qui a surpassé tous les autres modèles, atteignant une précision de 0,9912. Le mécanisme d'attention a efficacement augmenté les performances, en permettant au modèle de prioriser les régions spatiales critiques au sein des séquences d'images.

Pour améliorer la transparence de notre modèle proposé et faciliter son acceptation clinique, nous avons utilisé des méthodes d'IA explicable, en particulier Grad-CAM et LIME, pour visualiser et interpréter les décisions du modèle. Grad-CAM a fourni des visualisations globales des régions influençant les prédictions du modèle, tandis que LIME a généré des explications détaillées et localisées au niveau des pixels. Cette utilisation complémentaire des techniques XAI a considérablement amélioré l'interprétabilité de notre modèle, permettant aux professionnels de la santé de faire confiance et de valider les diagnostics générés par l'IA avec une plus grande confiance.

Dans l'introduction générale, nous avons soulevé un certain nombre de questions de recherches relatives à notre problématique. En effet, cette dernière portait globalement sur la nécessité de développer une solution capable de répondre à trois défis d'importance majeure :

1. **Confiance dans les modèles de l'IA** : En intégrant LIME et Grad-CAM, nous avons offert une transparence aux prédictions du modèle, permettant aux professionnels de comprendre et de valider chaque décision.
2. **Précision et rapidité** : Le modèle Stacked-LSTM avec mécanisme d'attention a démontré des performances élevées, avec une précision remarquable pour détecter le paludisme, tout en garantissant un diagnostic rapide et efficace.
3. **Accessibilité dans les zones démunies** : La plateforme MalariaScope a été conçue pour fonctionner avec des outils simples, comme un smartphone et un microscope, offrant ainsi une solution abordable pour les zones médicalement isolées.

Nous avons répondu à ces questions de recherches tout au long de cette thèse dont le déroulement peut être résumé comme suit :

Dans le Chapitre 1, nous avons posé les bases théoriques en présentant les concepts clés de l'intelligence artificielle (IA) et de l'intelligence artificielle explicable (XAI). Cette partie a détaillé les enjeux liés aux modèles de type « boîte noire », en insistant sur le besoin de transparence et de compréhension pour favoriser leur adoption dans le domaine médical.

Le Chapitre 2 nous a permis d'explorer le contexte médical et les défis spécifiques du diagnostic du paludisme. En effet, nous avons analysé les méthodes actuelles, comme l'analyse manuelle des frottis sanguins, en mettant en lumière leurs limites : lenteur, coût élevé et forte dépendance à l'expertise humaine. Ce constat nous a motivés pour introduire l'IA dans le processus de la détection automatisée du paludisme, à partir d'images de frottis sanguin.

Dans le Chapitre 3, nous avons réalisé une revue des approches classiques et modernes utilisées en détection d'objets et en classification. Les techniques d'apprentissage profond, comme les réseaux convolutifs (CNN) et les Stacked-LSTM, se sont révélées particulièrement adaptées à l'analyse d'images médicales grâce à leur capacité à extraire des motifs complexes.

Le Chapitre 4 nous a permis de présenter l'approche proposée, basée sur un modèle Stacked-LSTM avec mécanisme d'attention couplé à des techniques XAI. L'intégration de LIME et Grad-CAM a permis de fournir des explications compréhensibles pour chaque prédiction. Le système final, nommé MalariaScope, a été développé comme une plateforme web accessible en ligne, combinant des performances remarquables (F1-score supérieur à 0.9912 et une précision de 0.9959) avec une transparence essentielle à son adoption par les professionnels.

Bien que notre modèle ait démontré d'excellentes performances sur le jeu de données de référence, il y a plusieurs limitations à prendre en compte. Tout d'abord, le jeu de données ne provient pas d'un environnement clinique multicentrique, ce qui limite la généralisabilité. Deuxièmement, bien que les mécanismes d'attention améliorent l'interprétabilité des modèles, leur intégration dans les flux de travail cliniques en temps réel peut nécessiter des alternatives légères et la validation de leur robustesse sur divers ensembles de données cliniques, en tenant compte des scénarios de déploiement en temps réel, en particulier sur des appareils mobiles ou edge.

Les travaux futurs que nous ménerons incluront la validation dans divers contextes cliniques et le déploiement d'un moteur d'inférence compatible avec les appareils mobiles. De plus, les avancées récentes dans les modèles basés sur ViT et LLM représentent des directions prometteuses pour améliorer davantage les performances et l'interprétabilité dans les tâches d'analyse d'images médicales. Malgré ces limitations, notre modèle reste un outil prometteur pour le dépistage précoce et interprétable du paludisme à partir de frottis sanguins.

Ce travail représente une avancée significative dans l'application de l'IA explicable pour le diagnostic médical, en particulier dans le contexte du paludisme. En combinant précision, explicabilité et accessibilité, nous avons conçu un outil pratique, répondant aux besoins des professionnels de santé et des patients dans les zones médicalement isolées. Cependant, ce

projet constitue une base pour des recherches futures, avec des opportunités d'amélioration et d'extension vers de nouvelles pathologies et technologies. Les efforts continus dans cette direction pourraient transformer le paysage du diagnostic médical, rendant l'IA à la fois performante et éthique dans sa pratique.

Enfin, nous pensons que cette thèse contribue à faire progresser l'état de l'art dans la détection du paludisme, non seulement par ses performances techniques mais aussi par son engagement en faveur d'une IA responsable et explicable.

RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES BIBLIOGRAPHIQUES

- Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. 2021 Sep;298:103502. doi:10.1016/j.artint.2021.103502
- Abbasian Ardakani A, Airom O, Khorshidi H, Bureau NJ, Salvi M, Molinari F. Interpretation of artificial intelligence models in healthcare: A pictorial guide for clinicians. J Ultrasound Med. 2024 Oct;43(10):1789-818. doi:10.1002/jum.16524
- Abdelouahab K. Reconfigurable hardware acceleration of CNNs on FPGA-based smart cameras.Electronics. Université Clermont Auvergne, 2018. English. NNT : 2018CLFAC042. *(PDF)* Available from: https://www.researchgate.net/publication/331800773_Reconfigurable_hardware_acceleration_of_CNNs_on_FPGA-based_smart_cameras/figures?lo=1&utm_source=google&utm_medium=organic
- Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-60. doi:10.1109/ACCESS.2018.2870052
- Agarwal C, Ley D, Krishna S, Saxena E, Pawelczyk M, Johnson N, et al. OpenXAI: Towards a transparent evaluation of model explanations. arXiv. 2022. arXiv:2206.11104
- Agus M, Tsabita I, Munarko Y, Basuki S. Classification of Malaria Using Convolutional Neural Network Method on Microscopic Image of Blood Smear. JOIV. 2024;8(3):1469. doi:10.62527/joiv.8.3.2154
- Alanazi AF, Alaerjan MA. Bi-LSTM network for classification of medical images: Application in malaria diagnosis. Comput Mater Contin. 2023;73(1):1019-35. doi:10.32604/cmc.2023.022462
- Aladhadh S, Alsanea M, Aloraini M, Khan T, Habib S, Islam M. An effective skin cancer classification mechanism via medical vision transformer. Sensors. 2022;22(11):4008. doi:10.3390/s22114008
- Alexandrina EC, Ortigossa ES, Lui ES, Gonçalves JAS, Correa NA, Nonato LG, et al. Analysis and visualization of multidimensional time series: Particulate matter (PM10) from São Carlos-SP (Brazil). Atmos Pollut Res. 2019 Jul;10(4):1299-311. doi:10.1016/j.apr.2019.03.006
- Al-Hammuri K, Gebali F, Kanan A, Chelvan IT. Vision transformer architecture and applications in digital health: a tutorial and survey. Vis Comput Ind Biomed Art. 2023;6:14. doi:10.1186/s42492-023-00140-9
- Al-Shabi M, Shak K, Tan M. Procan: Progressive growing channel attentive non-local network for lung nodule classification. Pattern Recognit. 2022;122:108309. doi:10.1016/j.patcog.2021.108309
- Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain—Artificial intelligence explainability in clinical decision support systems. PLOS Digit Health. 2022 Feb;1(2):e0000016. doi:10.1371/journal.pdig.0000016
- Amin J, Anjum MA, Ahmad A, Sharif MI, Kadry S, Kim J. Microscopic parasite malaria classification using best feature selection based on generalized normal distribution optimization. PeerJ Comput Sci. 2024;10:e1744. doi:10.7717/peerj-cs.1744
- Amin MA, et al. CFPNet-M: Channel Feature Pyramid Network for Malaria Parasite Detection. Int J Comput Vis. 2024b;132(5):1256-75. doi:10.1007/s11263-024-01557-x

- Amparore E, Perotti A, Bajardi P. To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. PeerJ Comput Sci. 2021 Apr;7:e479. doi:10.7717/peerj-cs.479
- Angelov PP, Soares EA, Jiang RM, Arnold NI, Atkinson PM. Explainable artificial intelligence: An analytical review. Wiley Interdiscip Rev Data Min Knowl Discov. 2021;11(5):e1424. doi:10.1002/widm.1424
- Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A video vision transformer. In: Proc IEEE/CVF Int Conf Comput Vis. 2021. p. 6836-46. doi:10.1109/ICCV48922.2021.00676
- Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020 Jun;58:82-115. doi:10.1016/j.inffus.2019.12.012
- Attai K, Ekpenyong M, Amannah C, Asuquo D, Ajuga P, Obot O, et al. Enhancing the interpretability of malaria and typhoid diagnosis with explainable AI and large language models. Trop Med Infect Dis. 2024;9(9):216. doi:10.3390/tropicalmed9090216
- Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Attention DeepLabv3+: Multi-level context attention mechanism for skin lesion segmentation. In: Eur Conf Comput Vis. 2020. p. 251-66. doi:10.1007/978-3-030-66415-2_16
- Ba J, Caruana R. Do deep nets really need to be deep? In: Proc Adv Neural Inf Process Syst. 2014. p. 1-9.
- Baer K, Klotz C, Kappe SHI, Schnieder T, Frevert U. Release of hepatic plasmodium yoelii merozoites into the pulmonary microvasculature. PLoS Pathog. 2007;3(11):e171. doi:10.1371/journal.ppat.0030171
- Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. 2014. arXiv:1409.0473.
- Banachewicz K, Massaron L, Goldbloom A. The Kaggle Book: Data Analysis and Machine Learning for Competitive Data Science. Birmingham, UK: Packt Publishing; 2022.
- Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. Ann Stat. 2015 Oct;43(5):2055-85. doi:10.1214/15-AOS1337
- Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data. 2021;4:688969. doi:10.3389/fdata.2021.688969
- Bello I, Zoph B, Vaswani A, Shlens J, Le QV. Attention augmented convolutional networks. In: Proc IEEE/CVF Int Conf Comput Vis. 2019. p. 3286-95. doi:10.1109/ICCV.2019.00338
- Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281-305.
- Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, et al. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. Electronics. 2021;10(20):2470. doi:10.3390/electronics10202470
- Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, et al. Explainable machine learning in deployment. In: Proc Conf Fairness, Accountability, Transparency. 2020. p. 648-57. doi:10.1145/3351095.3375624

- Biderman S, Sai Prashanth U, Sutawika L, Schoelkopf H, Anthony Q, Purohit S, et al. Emergent and predictable memorization in large language models. arXiv. 2023a. arXiv:2304.11158
- Biderman S, Schoelkopf H, Anthony QG, Bradley H, O'Brien K, Hallahan E, et al. Pythia: A suite for analyzing large language models across training and scaling. In: Proc Int Conf Mach Learn. 2023b. p. 2397-430.
- Botev A, Lever G, Barber D. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In: Proc Int Joint Conf Neural Netw. 2017. p. 1899-903. doi:10.1109/IJCNN.2017.7966082
- Bozorgpour A, Azad R, Showkatian E, Sulaiman A. Multi-scale regional attention DeepLab3+: Multiple myeloma plasma cells segmentation in microscopic images. In: MICCAI Workshop Comput Pathol. 2021. p. 47-56. doi:10.48550/arXiv.2105.06238
- Breiman L. Random forests. Mach Learn. 2001 Oct;45:5-32. doi:10.1023/A:1010933404324
- Bruce-Chwatt LJ, de Zulueta J. The rise and fall of malaria in Europe. Oxford: Oxford University Press; 1980.
- Cantareira E, Etemad, Paulovich FV. Exploring neural network hidden layer activity using vector fields. Information. 2020 Aug;11(9):426. doi:10.3390/info11090426
- Carnevale P, Robert V, Le Coff G, Fondjo E, Manga L, Akogbeto M, et al. Données entomologiques sur le paludisme en Afrique tropicale. Cahiers santé. 1993;3:239-45.
- Carnevale P, Robert V, Molez D, Baudon. Faciès épidémiologique des paludismes en Afrique subsaharienne. Études médicales. 1984;3:123-33.
- Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. In: Proc 21st ACM SIGKDD Int Conf Knowl Discovery Data Mining. 2015. p. 1721-30. doi:10.1145/2783258.2788613
- Casalicchio G, Molnar C, Bischl B. Visualizing the feature importance for black box models. In: Proc Joint Eur Conf Mach Learn Knowl Discov Databases. 2018. p. 655-70. doi:10.1007/978-3-030-10925-7_40
- Castelvecchi D. Can we open the black box of AI? Nature. 2016 Oct;538(7623):20-3. doi:10.1038/538020a
- Celli A. A History of Malaria in the Italian Campagna from Ancient Times. London: John Bale, Sons & Danielsson; 1933.
- Chander A, Srinivasan R, Chelian S, Wang J, Uchino K. Working with beliefs: AI transparency in the enterprise. In: Workshops of the ACM Conference on Intelligent User Interfaces. 2018.
- Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proc IEEE Winter Conf Appl Comput Vis. 2018. p. 839-47. doi:10.1109/WACV.2018.00097
- Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2021. p. 782-91. doi:10.1109/CVPR46437.2021.00083
- Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv. 2021. arXiv:2102.04306.

- Chen X, Wang X, Zhou J, Qiao Y, Dong C. Activating more pixels in image super-resolution transformer. In: 2023 IEEE/CVF Conf Comput Vis Pattern Recognit. 2023. p. 22367-77. doi:10.1109/CVPR52729.2023.02142
- Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. Adv Neural Inf Process Syst. 2015;28.
- Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153-63. doi:10.1089/big.2016.0047
- Chow JCL. Quantum computing and machine learning in medical decision-making: A comprehensive review. Algorithms. 2025;18:156. doi:10.3390/a18030156
- Chow JCL. Nanomaterial-based molecular imaging in cancer: Advances in simulation and ai integration. Biomolecules. 2025;15:444. doi:10.3390/biom15030444
- Chow JCL, Li K. Ethical considerations in human-centered ai: Advancing oncology chatbots through large language models. JMIR Bioinform Biotechnol. 2024;5:e64406. doi:10.2196/64406
- Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: Current trends, challenges, and future directions in large language model-enabled medical chatbots. BioMedInformatics. 2024;4:837–52. doi:10.3390/biomedinformatics4010047
- Chowdary GJ, Punn NS, Sonbhadra SK, Agarwal S. Face mask detection using transfer learning of inceptionv3. In: Big Data Analytics: 8th International Conference, BDA 2020, Sonepat, India, December 15–18, 2020, Proceedings 8. Springer International Publishing; 2020. p. 81-90.
- Coetzee M, Fontenille D. Advances in the study of Anopheles funestus, a major vector of malaria in Africa. Insect Biochem Mol Biol. 2004;34(7):599-605. doi:10.1016/j.ibmb.2004.03.012
- Daho MEH, Li Y, Zeghlache R, Le Boit´e H, Deman P, Borderie L, et al. Discover: 2-d multiview summarization of optical coherence tomography angiography for automatic diabetic retinopathy diagnosis. Artif Intell Med. 2024;149:102803. doi:10.1016/j.artmed.2024.102803
- Das S, Javid AM, Gohain PB, Eldar YC, Chatterjee S. Neural greedy pursuit for feature selection. arXiv. 2022. arXiv:2207.09390
- Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE Symp Secur Priv. 2016. p. 598-617. doi:10.1109/SP.2016.42
- Davis J and Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25-29; Pittsburgh, PA. New York: ACM; 2006:233-240. doi:10.1145/1143844.1143874
- De Leon AR, Carrière KC. A generalized Mahalanobis distance for mixed data. J Multivariate Anal. 2005 Jan;92(1):174-85. doi:10.1016/j.jmva.2004.06.004
- Dong X, Chowdhury S, Qian L, Li X, Guan Y, Yang J, et al. Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN. PLoS One. 2019;14(9):e0222590. doi:10.1371/journal.pone.0222590
- Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. arXiv. 2017. arXiv:1710.00794

- Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017. arXiv:1702.08608
- Dosilovic FK, Brcic M, Hlupic N. Explainable artificial intelligence: A survey. In: Proc 41st Int Conv Inf Commun Technol Electron Microelectron. 2018. p. 210-5. doi:10.23919/MIPRO.2018.8400040
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. 2020. arXiv:2010.11929
- Eisemann N, Bunk S, Mukama T, Baltus H, Elsner SA, Gomille T, et al. Nationwide real-world implementation of ai for cancer detection in population-based mammography screening. Nat Med. 2025. doi:10.1038/s41591-024-03408-6
- EU Regulation. 2016/679 of the European Parliament and of the council of 27 April 2016 on the General Data Protection Regulation. 2016. Available from: http://data.europa.eu/eli/reg/2016/679/oj. Accessed: Apr 2023.
- Evandro O, Gonçalves T, Nonato L. EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. IEEE Access. 2024;12:80799-846. doi:10.1109/ACCESS.2024.3409843
- FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback. 2024. Available from: https://www.regulations.gov/document?D=FDA-2019-N-1185-0001. Accessed: Feb 2024.
- Fei L, Z K, Ou W. Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention. IEEE. 2019 Oct.
- Fei-Fei L, Li L-J. What, where and who? Telling the story of an image by activity classification, scene recognition and object categorization. In: Computer vision. Springer; 2010. p. 157-71.
- Fernandez A, Herrera F, Cordon O, Jose del Jesus M, Marcelloni F. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEEE Comput Intell Mag. 2019;14(1):69-81. doi:10.1109/MCI.2018.2881645
- Foote A, Nanda N, Kran E, Konstas I, Cohen S, Barez F. Neuron to graph: Interpreting language model neurons at scale. arXiv. 2023. arXiv:2305.19911
- Gaouar A, Hamza Cherif S, Rahmoun A, El Habib Daho M, Explainable AI for early malaria detection using stacked-LSTM and attention mechanisms, Informatics in Medicine Unlocked, Volume 57, 2025, 101667, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2025.101667.
- Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. arXiv. 2017. arXiv:1704.06857
- Garde A, Kran E, Barez F. DeepDecipher: Accessing and investigating neuron activation in large language models. arXiv. 2023b. arXiv:2310.01870
- Garrido-Cardenas JA, González-Cerón L, Manzano-Agugliaro F, Mesa-Valle C. Plasmodium genomics: an approach for learning about and ending human malaria. Parasitol Res. 2019;118:535–48. doi:10.1007/s00436-018-6127-9
- Girod R, Orlandi-Pradines E, Rogier C, Pagès F. Malaria Transmission and Insecticide Resistance of Anopheles gambiae (Diptera: Culicidae) in the French Military Camp of Port-Bouët, Abidjan

(Côte d'Ivoire): Implications for Vector Control. J Med Entomol. 2006;43(6):1082-7. doi:10.1603/0022-2585(2006)43[1082:MTAIRO]2.0.CO;2

- Gkelios S, Sophokleous A, Plakias S, Boutalis Y, Chatzichristofis SA. Deep convolutional features for image retrieval. Expert Syst Appl. 2021;177:114940. doi:10.1016/j.eswa.2021.114940
- Gleicher M. A framework for considering comprehensibility in modeling. Big Data. 2016;4(2):75-88. doi:10.1089/big.2016.0007
- Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44-65. doi:10.1080/10618600.2014.907095
- Gonçalves T, Rio-Torto I, Teixeira LF, Cardoso JS. A survey on attention mechanisms for medical applications: Are we moving towards better algorithms? IEEE Access. 2022;10:98909-35. doi:10.1109/ACCESS.2022.3206449
- Goni F, O M, Mondal MD, Nazrul I, Islam S, Riazul M, et al. Diagnosis of Malaria Using Double Hidden Layer Extreme Learning Machine Algorithm with CNN Feature Extraction and Parasite Inflator. IEEE Access. 2023;11:4117-30. doi:10.1109/ACCESS.2023.3234279
- Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a right to explanation. AI Mag. 2017;38(3):50-7. doi:10.1609/aimag.v38i3.2741
- Grignaffini F, Simeoni P, Alisi A, Frezza F. Computer-aided diagnosis systems for automatic malaria parasite detection and classification: A systematic review. Electronics. 2024;13(16):3174. doi:10.3390/electronics13163174
- Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. arXiv. 2018. arXiv:1805.10820
- Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2019;51(5):1-42. doi:10.1145/3236009
- Gunning D. Explainable artificial intelligence (xAI). Tech rep. Defense Advanced Research Projects Agency (DARPA); 2017.
- Hamilton M, Lundberg S, Zhang L, Fu S, Freeman WT. Axiomatic explanations for visual search, retrieval, and similarity learning. arXiv. 2022. arXiv:2103.00370
- Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):87-110. doi:10.1109/TPAMI.2022.3152247
- Han T, Srinivas S, Lakkaraju H. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. In: Proc Adv Neural Inf Process Syst. 2022;35:5256-68.
- Harbach R. Review of the internal classification of the genus Anopheles (Diptera: Culicidae): the foundation for comparative systematics and phylogenetic research. Bull Entomol Res. 1994;84:331-42. doi:10.1017/S0007485300032534
- Harbers M, Van den Bosch K, Meyer J-J. Design and evaluation of explainable BDI agents. In: IEEE/WIC/ACM Int Conf Web Intell Intell Agent Technol. 2010;2:125-32. doi:10.1109/WI-IAT.2010.134

- Hilali H. Application de la classification textuelle pour l'extraction des règles d'association maximales. Université du Québec à Trois-Rivières; 2009.
- Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv. 2015. arXiv:1503.02531
- Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80. doi:10.1162/neco.1997.9.8.1735
- Hohman F, Kahng M, Pienta R, Chau DH. Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Trans Vis Comput Graph. 2019;25(8):2674-93. doi:10.1109/TVCG.2018.2843369
- Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv. 2017. arXiv:1712.09923
- Hooker G, Mentch L, Zhou S. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. Stat Comput. 2021;31(6):82. doi:10.1007/s11222-021-10057-z
- Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov. 1998;2(3):283-304. doi:10.1023/A:1009769707641
- Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis Support Syst. 2011;51(1):141-54. doi:10.1016/j.dss.2010.12.003
- Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Int Conf Mach Learn. 2015. p. 448-56.
- Iscen A, Avrithis Y, Tolias G, Furon T, Chum O. Fast spectral ranking for similarity search. In: Proc IEEE Comput Soc Conf Comput Vis. 2018. p. 7632-41. doi:10.1109/CVPR.2018.00796
- Islam MR, Nahiduzzaman MD, Goni MOF, Sayeed A, Anower MS, Ahsan M, et al. Explainable Transformer-Based Deep Learning Model for the Detection of Malaria Parasites from Blood Cell Images. Sensors. 2022;22(12):4358. doi:10.3390/s22124358
- Islam O, Assaduzzaman M, Zahid Hasan M. An explainable AI-based blood cell classification using optimized convolutional neural network. J Pathol Inform. 2024 Dec;15:100389. doi:10.1016/j.jpi.2024.100389
- Jacovi A, Swayamdipta S, Ravfogel S, Elazar Y, Choi Y, Goldberg Y. Contrastive explanations for model interpretability. arXiv. 2021. arXiv:2103.01378
- Jain S, Wallace BC. Attention is not explanation. arXiv. 2019. arXiv:1902.10186
- James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. Springer; 2013.
- Jia M, Gabrys B, Musial K. A network science perspective of graph convolutional networks: A survey. IEEE Access. 2023;11:39083-122. doi:10.1109/ACCESS.2023.3268073
- Jiang P-T, Zhang C-B, Hou Q, Cheng M-M, Wei Y. LayerCAM: Exploring hierarchical class activation maps for localization. IEEE Trans Image Process. 2021;30:5875-88. doi:10.1109/TIP.2021.3089943

- John AM, Jia Y, Porter Z, Habli I. Artificial Intelligence Explainability: The Technical and Ethical Dimensions. Philos Trans R Soc A. 2021;379(2207):20200363. doi:10.1098/rsta.2020.0363
- Jothi A, et al. Automated Diagnosis of Thyroid Cancer Using KNN and Otsu Thresholding for Histopathological Images. Biomed Signal Process Control. 2016;22:1-9. doi:10.1016/j.bspc.2016.03.002
- Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling. arXiv. 2016. arXiv:1602.02410
- Karimi H, Derr T, Tang J. Characterizing the decision boundary of deep neural networks. arXiv. 2019. arXiv:1912.11460
- Kassim YM, Yang F, Yu H, Maude RJ, Jaeger S. Diagnosing Malaria Patients with Plasmodium falciparum and vivax Using Deep Learning for Thick Smear Images. Diagnostics. 2021**a**;11(11):1994. doi:10.3390/diagnostics11111994
- Kassim Y M, Palaniappan K, Yang F, Poostchi M, Palaniappan N, Maude RJ, et al. Clustering-Based Dual Deep Learning Architecture for Detecting Red Blood Cells in Malaria Diagnostic Smears. IEEE J Biomed Health Inform. 2021**b** May;25(5):1735-46. doi:10.1109/JBHI.2020.3034863
- Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In: Proc CHI Conf Hum Factors Comput Syst. 2020. p. 1-14. doi:10.1145/3313831.3376219
- Kawaguchi K. Deep learning without poor local minima. In: Adv Neural Inf Process Syst. 2016. p. 586-94.
- Khan A, Gupta KD, Venugopal D, Kumar N. CIDMP: Completely Interpretable Detection of Malaria Parasite in Red Blood Cells using Lower-dimensional Feature Space. In: Int Joint Conf Neural Netw. 2020. p. 1-8. doi:10.1109/IJCNN48605.2020.9284834
- Kim D, Lee Y, Chin K, Mago PJ, Cho H, Zhang J. Implementation of a Long Short-Term Memory Transfer Learning (LSTM-TL)-Based Data-Driven Model for Building Energy Demand Forecasting. Sustainability. 2023;15:2340. doi:10.3390/su15032340
- Kim Y, Denton C, Hoang L, Rush AM. Structured attention networks. arXiv. 2017. arXiv:1702.00887
- Kiperwasser E, Goldberg Y. Simple and accurate dependency parsing using bidirectional lstm feature representations. Trans Assoc Comput Linguist. 2016;4:313–27. doi:10.1162/tacl_a_00101
- Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: Proc Int Joint Conf Neural Netw. 2020. p. 1-7. doi:10.1109/IJCNN48605.2020.9207045
- Koirala A, Jha M, Bodapati S, Mishra A, Chetty G, Sahu PK. Deep Learning for Real-time Malaria Parasite Detection and Counting Using YOLO-mp. IEEE Access. 2022;10:98104-16. doi:10.1109/ACCESS.2022.3208270
- Kong A, Wilson SA, Yong AH, Nace D, Rogier E, Aidoo M. Hrp2 and hrp3 cross-reactivity and implications for hrp2-based rdt use in regions with plasmodium falciparum hrp2 gene deletions. Malaria Journal. 2021;20:207. doi:10.1186/s12936-021-03739-6
- Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, et al. The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv. 2022. arXiv:2202.01602

- Kumar H, Chandran J. Is Shapley explanation for a model unique? arXiv. 2021. arXiv:2111.11946
- Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: Proc Int Conf Mach Learn. 2020b. p. 5491-500.
- Kumar N, Kaur N, Gupta D. Major Convolutional Neural Networks in image classification: A survey. In: Proc Int Conf IoT Inclusive Life. 2020a. p. 243-58. doi:10.1007/978-981-15-3020-3_23
- Lacroix R, Mukabana WR, Gouagna LC, Koella JC. Malaria infection increases attractiveness of humans to mosquitoes. PLoS Biol. 2005;3(9):e298. doi:10.1371/journal.pbio.0030298
- Lakkaraju H, Arsov N, Bastani O. Robust and stable black box explanations. In: Proc Int Conf Mach Learn. 2020. p. 5628-38.
- Lalapura VS, Amudha J, Satheesh HS. Recurrent neural networks for edge intelligence: A survey. ACM Comput Surv. 2021;54(8):1-38. doi:10.1145/3448974
- Langley P, Meadows B, Sridharan M, Choi D. Explainable agency for intelligent autonomous systems. In: AAAI Conf Artif Intell. 2017. p. 4762-3.
- Lanjewar MG, Panchbhai KG, Patle LB. Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images. Comput Biol Med. 2024 Feb;169:107914. doi:10.1016/j.compbiomed.2023.107914
- Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking clever Hans predictors and assessing what machines really learn. Nat Commun. 2019 Mar;10(1):1096. doi:10.1038/s41467-019-08987-4
- LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44. doi:10.1038/nature14539
- Lee K, et al. AI-driven Decision Support Systems for Malaria Diagnosis in Sub-Saharan Africa. J Artif Intell Healthc. 2023;19(7):885-96.
- Leevy J, Khoshgoftaar TA. A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data. J Big Data. 2020;7(1):1-19. doi:10.1186/s40537-020-00382-x
- Li Y, Daho MEH, Conze PH, Zeghlache R, Le Boit´e H, Tadayoni R, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Comput Biol Med. 2024;177:108635. doi:10.1016/j.compbiomed.2024.108635
- Li Y, Ding L, Gao X. On the decision boundary of deep neural networks. arXiv. 2018. arXiv:1808.05385
- Liao QV, Gruen D, Miller S. Questioning the AI: Informing design practices for explainable AI user experiences. In: Proc CHI Conf Hum Factors Comput Syst. 2020. p. 1-15. doi:10.1145/3313831.3376590
- Lilda SD, Jayaparvathy R. Enhancing cardiovascular disease classification in ecg spectrograms by using multi-branch cnn. Comput Biol Med. 2025;186:109737. doi:10.1016/j.compbiomed.2025.109737
- Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy. 2020;23(1):18. doi:10.3390/e23010018
- Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Appl Stoch Models Bus Ind. 2001;17(4):319-30. doi:10.1002/asmb.446

- Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31-57. doi:10.1145/3236386.3241340
- Long F, Zhou K, Ou W. Sentiment analysis of text based on bidirectional lstm with multi-head attention. IEEE Access. 2019;7:141960–9. doi:10.1109/ACCESS.2019.2942614
- Lou Y, Caruana R, Gehrke J. Intelligible models for classification and regression. In: Proc 18th ACM SIGKDD Int Conf Knowl Discovery Data Mining. 2012. p. 150-8. doi:10.1145/2339530.2339556
- Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proc Int Conf Neural Inf Process Syst. 2017. p. 4768-77.
- Luo D, Cheng W, Xu D, Yu W, Zong B, Chen H, et al. Parameterized explainer for graph neural network. In: Proc Adv Neural Inf Process Syst. 2020;33:19620-31.
- Ma J, Bai Y, Zhong B, Zhang W, Yao T, Mei T. Visualizing and understanding patch interactions in vision transformer. IEEE Trans Neural Netw Learn Syst. 2023;PP:1-10. doi:10.1109/TNNLS.2023.3270479
- MacDonald G. The epidemiology and control of malaria. London: Oxford University Press; 1957.
- Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, et al. Surgical data science: From concepts toward clinical translation. Med Image Anal. 2022 Feb;76:102306. doi:10.1016/j.media.2021.102306
- Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a Convolutional Neural Networks. Neural Netw. 2003;16(5-6):555-9. doi:10.1016/S0893-6080(03)00115-1
- McClelland C. The Difference between artificial intelligence, machine learning, and deep learning, 18, 2017.
- McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv. 2018. arXiv:1802.03426
- Mienye ID, Jere N. Deep Learning for Credit Card Fraud Detection: A Review of Algorithms, Challenges, and Solutions. IEEE Access. 2024;12:96893-910. doi:10.1109/ACCESS.2024.3391906
- Mienye ID, Sun Y. A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access. 2023a;11:30628-38. doi:10.1109/ACCESS.2023.3262020
- Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019 Feb;267:1-38. doi:10.1016/j.artint.2018.07.007
- Miller T, Howe P, Sonenberg L. Explainable AI: Beware of inmates running the asylum. In: IJCAI Workshop on Explainable AI. 2017.
- Minarno AE, Izzah TN, Munarko Y, Basuki S. Classification of malaria using convolutional neural network method on microscopic image of blood smear. JOIV Int J Inform Visual. 2024;8:1469. doi:10.62527/joiv.8.3.2154
- Mishra S, Dutta S, Long J, Magazzeni D. A survey on the robustness of feature importance and counterfactual explanations. arXiv. 2021. arXiv:2111.00358
- Molnar C. Interpretable Machine Learning. Durham, NC: Lulu Press; 2019.

- Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1-15. doi:10.1016/j.dsp.2017.10.011
- Murillo DC, Astarza VR, Fagardo OP. Biology of Anopheles (Kerteszia) neivai H, D, K, 1913 (diptera: culicidae) on the pacific coast of Colombia III; Light intensity measurements and biting behaviour. Rev Saude Publica. 1988;22:102-12.
- Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.
- Nicolas V. Prise en charge de l'accès palustre simple, une urgence thérapeutique. La Revue du Praticien. 2019 Feb;69:152-8.
- Nonato LG, Aupetit M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. IEEE Trans Vis Comput Graph. 2019;25(8):2650-73. doi:10.1109/TVCG.2018.2846735
- Organisation Mondiale de la Santé. Malaria. 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/malaria. Accessed: Feb 11, 2024.
- Organisation Mondiale de la Santé. Terminologie du paludisme et de l'éradication. Genève; 1964.
- Ortigossa ES, Dias FF, Nascimento DCD. Getting over high-dimensionality: How multidimensional projection methods can assist data science. Appl Sci. 2022 Jul;12(13):6799. doi:10.3390/app12136799
- Ortigossa ES, Gonçalves T, Nonato L. Explainable artificial intelligence (xai)—from theory to methods and applications. IEEE Access. 2024;12:80799–846. doi:10.1109/ACCESS.2024.3409843
- Pagès F, Orlandi-Pradines E, Corbel V. Vecteurs du paludisme : biologie, diversité, contrôle et protection individuelle. Med Mal Infect. 2007 Mar;37(3):153-61. doi:10.1016/j.medmal.2006.10.001
- Pawar K, Raj SJ, Tiwari V. Stock Market Price Prediction Using LSTM RNN. In: Emerging Trends in Expert Applications and Security. 2019. p. 493-503.
- Pereira-Ferrero VH, Valem LP, Pedronette DCG. Feature augmentation based on manifold ranking and LSTM for image classification. Expert Syst Appl. 2023 Mar;213:118995. doi:10.1016/j.eswa.2022.118995
- Pervez K, Sohail SI, Parwez F, Zia MA. Towards trustworthy ai-driven leukemia diagnosis: A hybrid hierarchical federated learning and explainable ai framework. Informatics Med Unlocked. 2025:101618. doi:10.1016/j.imu.2025.101618
- Pilly E. Maladies infectieuses et tropicales : tous les items d'infectiologie. Paris: Alinéa Plus; 2018. p. 512-4.
- Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Transl Res. 2018; 194:36-55. doi:10.1016/j.trsl.2017.12.004
- Preethi V, Kiruthikab U. Survey on Text Transformation using Bi-LSTM in Natural Language Processing with Text Data. Turk J Comput Math Educ. 2021 Apr;12(9):2577-85.
- Qadri A, Raza A, Eid F, Abualigah L. A novel transfer learning-based model for diagnosing malaria from parasitized and uninfected red blood cell images. Decis Anal J. 2023c;9:100352. doi:10.1016/j.dajour.2023.100352

- Qadri A, et al. Advances in AI for Malaria Diagnosis: A Review of Mobile Applications in Low-Resource Settings. Glob Health Action. 2023b;15(1):1062895. doi:10.1080/16549716.2022.2062895
- Qadri A, et al. Application of NASNet Random Forest Model for Malaria Diagnosis with High Accuracy. Comput Biol Med. 2023a;157:106896. doi:10.1016/j.compbiomed.2023.106896
- Qian S, Zhu Y, Li W, Li M, Jia J. What makes for good tokenizers in vision transformer? IEEE Trans Pattern Anal Mach Intell. 2022;45(1):748-60. doi:10.1109/TPAMI.2022.3231442
- Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987;27(3):221-34. doi:10.1016/S0020-7373(87)80053-6
- Rajab S, Nakatumba-Nabende J, Ggaliwango M. Interpretable Machine Learning Models for Predicting Malaria. In: ICSTSN. 2023. p. 1-6. doi:10.1109/ICSTSN57873.2023.10151538
- Rajab T, et al. SHAP and LIME Techniques for Malaria Parasite Detection Using Explainable Machine Learning Models. J Healthc Eng. 2023;2023:8027396. doi:10.1155/2023/8027396
- Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, et al. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018a Apr;6:e4568. doi:10.7717/peerj.4568
- Rajaraman S, et al. Deep learning for malaria detection from microscopic images of blood smears. J Med Imaging. 2020;7(2):55-62.
- Rajaraman S, Jaeger S, Antani SK. Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ. 2019;7:e6977. doi:10.7717/peerj.6977
- Rajaraman S, Silamut K, Hossain MA, Ersoy I, Maude RJ, Jaeger S, et al. Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. J Med Imaging (Bellingham). 2018b;5(4):044501. doi:10.1117/1.JMI.5.4.044501
- Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attention in vision models. Adv Neural Inf Process Syst. 2019;32.
- Rauber P.E, Fadel S.G, Falcao A.X, Telea A.C. Visualizing the hidden activity of artificial neural networks. IEEE Trans. Vis. Comput. Graphics 2017, 23, 101–110
- Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform. 2016;21(1):4-21. doi:10.1109/JBHI.2016.2636665
- Rawat W, Wang Z. Deep Convolutional Neural Networks for image classification: A comprehensive review. Neural Comput. 2017;29(9):2352-449. doi:10.1162/neco_a_00990
- Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?' Explaining the predictions of any classifier. In: Proc 22nd ACM SIGKDD Int Conf Knowl Discovery Data Mining. 2016a. p. 1135-44. doi:10.1145/2939672.2939778
- Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proc AAAI Conf Artif Intell. 2018;32(1). doi:10.1609/aaai.v32i1.11491
- Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv. 2016b. arXiv:1606.05386

- Ribeiro MT, Singh S, Guestrin C. Nothing else matters: Model-agnostic explanations by identifying prediction invariance. arXiv. 2016c. arXiv:1611.05817
- Rosnelly R, Riza B, Suparni S. Comparative Analysis of Support Vector Machine and Convolutional Neural Network for Malaria Parasite Classification and Feature Extraction. J Wirel Mob Netw Ubiquitous Comput Dependable Appl. 2023;14:194-217. doi:10.58346/JOWUA.2023.I3.015
- Samek W, Müller K-R. Towards Explainable Artificial Intelligence. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer; 2019. doi:10.1007/978-3-030-28954-6_1
- Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018 June 18-22; Salt Lake City, UT. Piscataway: IEEE; 2018. p. 4510-4520. doi:10.1109/CVPR.2018.00474.
- Sang DV, Chung TQ, Lan PN, Hang DV, Long D, Thuy NT. Ag-curesnest: A novel method for colon polyp segmentation. arXiv. 2021. arXiv:2105.00402
- Sangha S. Transfer Learning: Ashortcut for training deep learningmodels. Artificial Intelligence. November, 2020, https://medium.com
- Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decis Anal J. 2023;7:100230. doi:10.1016/j.dajour.2023.100230
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradientbased localization. In: Proc IEEE Int Conf Comput Vis. 2017. p. 618–26. doi:10.1109/ICCV.2017.74
- Shang W, Sohn K, Almeida D, Lee H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In: Proc 33rd Int Conf Mach Learn. 2016;48:2217-25.
- Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv. 2013. arXiv:1312.6034
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. 2014. arXiv:1409.1556
- Situ X, Zukerman I, Paris C, Maruf S, Haffari G. Learning to explain: Generating stable explanations fast. In: Proc 59th Annu Meet Assoc Comput Linguist. 2021. p. 5340-55. doi:10.18653/v1/2021.acl-long.415
- Smagulova K, James AP. A survey on LSTM memristive neural network architectures and applications. Eur Phys J Spec Top. 2019;228:2313-24. doi:10.1140/epjst/e2019-900046-x
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929-58.
- Stepin I, Alonso JM, Catala A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access. 2021;9:11974-2001. doi:10.1109/ACCESS.2021.3051315
- Strumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory. J Mach Learn Res. 2010 Mar;11:1-18.

- Strumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647-65. doi:10.1007/s10115-013-0679-x
- Sundararajan M, Najmi A. The many Shapley values for model explanation. In: Proc Int Conf Mach Learn. 2020. p. 9269-78.
- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proc Int Conf Mach Learn. 2017. p. 3319-28.
- Tan S, Hooker G, Koch P, Gordo A, Caruana R. Considerations when learning additive explanations for black-box models. Mach Learn. 2023;112(9):3333-59. doi:10.1007/s10994-023-06342-9
- Thiagarajan JJ, Kailkhura B, Sattigeri P, Ramamurthy KN. TreeView: Peeking into deep neural networks via feature-space partitioning. arXiv. 2016. arXiv:1611.07429
- Timo S. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In: 2022 ACM Conf Fairness Account Transp. 2022. p. 1-12. doi:10.1145/3531146.3533219
- Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the International Conference on Machine Learning. PMLR; 2021. doi:10.48550/arXiv.2012.12877
- Tsantekidis A, Passalis N, Tefas A. Recurrent Neural Networks. In: Deep Learning for Robot Perception and Cognition. Elsevier; 2022. p. 101-15
- Turner R. A model explanation system. In: Proc IEEE 26th Int Workshop Mach Learn Signal Process. 2016. p. 1-6. doi:10.1109/MLSP.2016.7738825
- Ufuktepe DK, Yang F, Kassim YM, Yu H, Maude RJ, Palaniappan K, et al. Deep Learning-Based Cell Detection and Extraction in Thin Blood Smears for Malaria Diagnosis. IEEE Appl Imagery Pattern Recognit Workshop. 2021 Apr:9762109. doi:10.1109/AIPR52630.2021.9762109
- Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-605.
- Van F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J, et al. Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. IEEE J Biomed Health Inform. 2020 May;24(5):1427-38. doi:10.1109/JBHI.2019.2939121
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proc Adv Neural Inf Process Syst. 2017;30.
- Verma S, Boonsanong V, Hoang M, Hines KE, Dickerson JP, Shah C. Counterfactual explanations and algorithmic recourses for machine learning: A review. arXiv. 2020. arXiv:2010.10596
- Vickers, A. J., and Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. Medical decision making: an international journal of the Society for Medical Decision Making, 2006, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361
- Vig J. BertViz: A tool for visualizing multihead self-attention in the BERT model. In: ICLR Workshop: Debugging Mach Learn Models. 2019;23:1-6.
- Vu M, Thai MT. PGM-Explainer: Probabilistic graphical model explanations for graph neural networks. In: Proc Adv Neural Inf Process Syst. 2020;33:12225-35.
- Wang W, Yang Y, Wang X, Wang W, Li J. Development of Convolutional Neural Networks and its application in image classification: A survey. Opt Eng. 2019;58(4):040901. doi:10.1117/1.OE.58.4.040901

- Werbos P. Backpropagation through time: What it does and how to do it. Proc IEEE. 1990;78(10):1550-60. doi:10.1109/5.58337
- Wisit L, Sako LU. Image classification of malaria using hybrid algorithms: convolutional neural network and method to find appropriate k for k-nearest neighbor. Indones J Electr Eng Comput Sci. 2019;16:382. doi:10.11591/ijeecs.v16.i1.pp382-388
- Wojtas M, Chen K. Feature importance ranking for deep learning. In: Proc Adv Neural Inf Process Syst. 2020;33:5105-14.
- Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X. An improved transformer network for skin cancer classification. Comput Biol Med. 2022;149:105939. doi:10.1016/j.compbiomed.2022.105939
- Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, and Yu J. Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE J Biomed Health Inform, 5 2020**a**. doi:10.1109/JBHI.2019.2939121
- Yang F, Quizon N, Yu H, Silamut K, Maude RJ, Jaeger S. Cascading yolo: Automated malaria parasite detection for plasmodium vivax in thin blood smears. In: Proc SPIE 11314, Med Imaging 2020: Comput-Aided Diagn. 2020**b** ;11314:113141Q. doi:10.1117/12.2550143
- Yang G, Luo S, Greer P. A novel vision transformer model for skin cancer classification. Neural Process Lett. 2023. doi:10.1007/s11063-023-11204-5
- Yang H, Zhang Y, Zhang Y. Automated Malaria Detection Using YOLO for P. vivax Parasite Identification in Thin Blood Smear Images. IEEE Access. 2020;8:195104-14. doi:10.1109/ACCESS.2020.3034866
- Yao C, Tang J, Hu M, Wu Y, Guo W, Li Q, et al. Claw U-Net: A U-Net variant network with deep feature concatenation for scleral blood vessel segmentation. In: CAAI Int Conf Artif Intell. 2021. p. 67-78. doi:10.1007/978-3-030-93049-3_6
- Yeuk-Yin Chan G, Bertini E, Nonato LG, Barr B, Silva CT. Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. arXiv. 2020. arXiv:2007.10614
- Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. GNNexplainer: Generating explanations for graph neural networks. In: Proc Adv Neural Inf Process Syst. 2019;32.
- Yousefzadeh R, O'Leary DP. Investigating decision boundaries of trained neural networks. arXiv. 2019. arXiv:1908.02802
- Yu H, Yang F, Rajaraman S, Ersoy I, Moallem G, Poostchi M, et al. Malaria Screener: a smartphone application for automated malaria screening. BMC Infect Dis. 2020 Nov;20:825. doi:10.1186/s12879-020-05453-1
- Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019;31:1235-70. doi:10.1162/neco_a_01199
- Zamil YK, Ali SA, Naser MA. Spam image email filtering using K-NN and SVM. Int J Electr Comput Eng. 2019;9(1):245-54. doi:10.11591/ijece.v9i1.pp245-254
- Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Comput Vis ECCV. 2014. p. 818-33. doi:10.1007/978-3-319-10590-1_53

- Zhao Q, Hastie T. Causal interpretations of black-box models. J Bus Econ Stat. 2021;39(1):272-81. doi:10.1080/07350015.2019.1624294
- Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc IEEE Conf Comput Vis Pattern Recognit. 2016. p. 2921-9. doi:10.1109/CVPR.2016.319
- Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv. 2020. arXiv:2010.04159
- Zou Y, Wu L, Zuo C, Chen L, Zhou B, Zhang H. White blood cell classification network using mobilenetv2 with multiscale feature extraction module and attention mechanism. Biomed Signal Process Control. 2025;99:106820. doi:10.1016/j.bspc.2024.106820

Informatics in Medicine Unlocked 57 (2025) 101667

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Explainable AI for early malaria detection using stacked-LSTM and attention mechanisms

Adil Gaouar [a,*], Souaad Hamza Cherif [a], Abdellatif Rahmoun [b], Mostafa El Habib Daho [c,d,**]

[a] University Abou Bekr Belkaid, Tlemcen, Algeria
[b] École supérieure en informatique - Sidi Bel Abbès, Algeria
[c] LaTIM UMR 1101, Inserm, Brest, France
[d] University of Western Brittany, Brest, France

ARTICLE INFO

Keywords:
Malaria detection
Stacked-LSTM
Deep learning
Explainable AI
Grad-CAM
LIME
Model interpretability

ABSTRACT

Malaria remains a global public health challenge, affecting more than 247 million people and causing 619,000 deaths worldwide in 2024 (according to WHO). Rapid diagnosis is essential for effective treatment and to improve patients' chances of survival. In this study, we propose an interpretable deep learning framework for accurate malaria diagnosis using blood smear images. Also, We evaluate and compare several baseline deep learning (DL) models (fundamentals), customized VGG-16 and VGG-19, as well as newer DL models such as Vision Transformer (ViT) and MobileNet, and, for the first time, a stacked long-short-term memory network (stacked-LSTM) with an attention mechanism for automatic detection of malaria from blood smear images. These models were trained and validated on a publicly available dataset of over 27.000 labeled blood smear images. The comparative and statistical study conducted in this research showed us that the proposed Stacked-LSTM model with attention mechanism outperformed all other approaches, achieving a classification accuracy (0.9912), sensitivity, specificity, precision, F1 score (0.9911), and area under the curve (AUC) superior to all other models. Despite their solid performance, these models are often considered ''black boxes'' due to their lack of transparency in the decision-making process, which poses significant challenges in medical applications and fields where human life is at stake. To address this, we have integrated explainable AI (XAI) techniques, namely Grad-CAM and LIME, to improve the model's interpretability. Our results demonstrate the complementary value of combining high-performance deep learning models with XAI methods to enhance trust and certainty in AI-assisted medical diagnosis, suggesting that our model can support early and interpretable malaria detection in clinical environments.

1. Introduction

Artificial intelligence (AI) has become a transformative technology across multiple domains due to its ability to solve complex problems efficiently. In particular, machine learning (ML), deep learning (DL) or, as we saw more recently, quantum machine learning (QML), which presents opportunities for future advances in processing high-dimensional healthcare data and improving clinical outcomes [1], have significantly advanced AI-driven medical diagnostics, enabling automated disease detection and classification [2]. However, despite their success, Deep Neural Networks (DNNs) are often criticized for their lack of transparency, making them difficult to interpret. This "black-box" nature raises concerns about their adoption in high-stakes applications such as healthcare, where decision-making must be explainable and trustworthy. In response to these concerns, researchers have focused on Explainable AI (XAI), a field dedicated to enhancing the interpretability of machine learning models. In the healthcare sector, explainability is not just desirable but essential, as AI-generated decisions directly impact patient diagnoses and treatment plans. Ethical and legal frameworks, such as the General Data Protection Regulation (GDPR) in the European Union (EU), reinforce this need by granting medical professionals and patients the right to understand automated medical decisions [3]. Ensuring transparency in AI-based diagnostics fosters trust among healthcare professionals, supports regulatory compliance, and facilitates collaboration between clinicians and AI systems.

Malaria, a mosquito-borne disease caused by Plasmodium parasites, remains a major global health threat despite ongoing eradication efforts. Microscopic analysis of blood smear images is the gold standard for malaria diagnosis, but it is labor-intensive and requires expert

* Corresponding author.
** Corresponding author at: University of Western Brittany, Brest, France.
E-mail addresses: adil.gaouar@univ-tlemcen.dz (A. Gaouar), mostafa.elhabibdaho@univ-brest.fr (M. El Habib Daho).

https://doi.org/10.1016/j.imu.2025.101667
Received 16 March 2025; Received in revised form 29 June 2025; Accepted 4 July 2025
Available online 15 July 2025

interpretation. AI models, particularly DNNs, have demonstrated potential in automating malaria detection. However, for these models to be effectively integrated into medical workflows, they must not only achieve high accuracy but also provide interpretable explanations for their predictions.

To enhance transparency and trust in malaria detection systems, this study integrates model performance analysis with explainability. The key contributions of this work include:

- Demonstrating the effectiveness of a Stacked-LSTM model with an attention mechanism for medical image classification and complex computer vision tasks.
- Early Malria detection.
- A comparative evaluation of five deep learning models—VGG-16, VGG-19, Stacked-LSTM, Vision Transformer (ViT), and MobileNetV2—for malaria detection using blood smear images.
- An assessment of two XAI methods, Gradient-weighted Class Activation Map (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME), to interpret and explain the predictions made by the Stacked-LSTM classifier.

These explainability techniques allow us to analyze how the model makes decisions and identify the regions of interest influencing its predictions. This approach aims to improve the interpretability of AI models and ensure their reliable application in critical healthcare settings.

The remainder of this paper is organized as follows: Section 2 presents a review of related works. In Section 3, we outline the proposed methodology, detailing the entire processing workflow. Section 4 discusses the experimental results, and Section 5 concludes the paper with insights and future perspectives.

2. Related works

Rapid diagnostic tests (RDTs) have been widely adopted for their ability to provide quick, portable results in malaria diagnosis. These tests detect specific antigens of malaria parasites, allowing immediate diagnosis in field settings. However, their accuracy can vary depending on environmental conditions, parasite density, and the disease stage. Moreover, the emergence of parasite strains with deletions in the hrp2 and/or hrp3 genes can significantly compromise RDT reliability, limiting their effectiveness in certain regions [4].

Microscopic examination of blood smears remains the gold standard for malaria diagnosis due to its high reliability and capability for detailed parasite identification. It allows visualization of parasites directly within red blood cells, enabling accurate species determination and quantification of parasite load. Despite these advantages, microscopy can be resource-intensive, requiring specialized equipment and skilled technicians, conditions challenging to meet consistently, especially in resource-limited regions. These constraints motivate the integration of AI approaches into microscopy workflows, aiming to maintain diagnostic accuracy while reducing the dependency on manual expertise and improving diagnostic speed and efficiency.

AI, particularly through advancements in machine learning and deep learning, has significantly transformed the diagnosis of infectious diseases, including malaria. By automating the detection and classification of malaria parasites from medical images, AI-driven systems have increased both the speed and accuracy of diagnosis [5], enhancing clinical decision-making and patient outcomes.

The rapid growth of AI in medicine is attributed to its capability to process large-scale datasets efficiently and to identify complex patterns often undetectable by the human eye [6,7]. AI applications have already demonstrated remarkable successes in diagnosing diseases such as cancer [8], diabetic retinopathy [9] and cardiovascular conditions [10], laying a solid foundation for similar advancements in malaria detection [11].

Recent research on malaria detection has explored various methods, ranging from the binary classification of the presence of malaria in blood smears to more specific tasks such as distinguishing between Plasmodium species.

In developing countries, where human resources and medical infrastructure may be limited, AI offers low-cost and accessible solutions for diagnosing malaria. AI can be integrated into mobile devices, making quick and accurate diagnoses even in remote areas. These mobile systems reduce the need for specialized expertise and train local laboratory technicians to perform effective analyses [12]. This accessibility and speed are essential for managing malaria outbreaks and improving the effectiveness of diagnostic and treatment campaigns [13].

AI-based computer-aided diagnostic systems (CAD-AI) support clinicians and laboratory technicians in decision-making. These systems provide real-time analysis results, allowing healthcare professionals to focus on data interpretation rather than image raw analysis. Thus, AI reduces human errors, improves diagnostic quality, and helps standardize practices in medical environments [14]. Integrating CAD-AI systems also ensures greater efficiency in processing large amounts of data from malaria tests, which is crucial in regions with a high disease incidence [15].

AI is not limited to merely improving existing tools; it allows for rediscovering diagnostic and care processes in limited resource contexts, thereby contributing to more effective disease control [16]. On the other hand, AI-driven segmentation algorithms facilitate the precise delineation of tumors and nanoparticle distributions in hybrid imaging modalities, improving the detection of malignant lesions or ROIs with increased specificity and sensitivity, as demonstrated by the work carried out by Chow et al. [17].

DNNs, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) have shown impressive performance in detecting malaria from blood smear images. A relevant study conducted by Kassim et al. [6] demonstrates an advanced Mask R-CNN model for detecting parasites in blood smear images. Mask R-CNN combines object detection (localization of parasites) with finer segmentation, allowing each parasite to be located with a precise mask for each instance. Combined with ResNet50, a DNN can extract more detailed features from images. The architecture of Mask R-CNN relies on two main networks: one for object detection and another for generating segmentation masks for each object instance (In this case, each parasite). This process allows the parasite's localization and better delineates its shape in the image, improving diagnostic accuracy and enhancing the localization and identification of parasites in the images. The authors designed four pipelines starting from a simple model based on Mask R-CNN detection and added three other classifiers, each in a separate pipeline, to discriminate between P. vivax, P. falciparum, and uninfected patients. Although the first three pipelines demonstrated good performance, the fourth pipeline, presented in their framework named PlasmodiumVF-Net, performed the best and allowed for the identification of Plasmodium species with an accuracy of 0.9776.

Minarno et al. [18] developed a CNN model designed to classify malaria-infected blood smear images. Their architecture consists of three convolutional layers coupled with two fully connected layers. To enhance the model's generalizability and prevent overfitting, they incorporated max pooling operations, batch normalization, and a dropout mechanism with a rate of 0.1. The input images were standardized to a resolution of 64 × 64 pixels during training. The resulting model demonstrated a classification accuracy of 0.96, making it a promising approach for rapid malaria diagnosis in resource-limited settings.

Another original work, due to its technique, is that of Rajaraman et al. [19], who used a modified CNN network to analyze blood smears and detect Plasmodium falciparum with an accuracy of 0.9520. The CNN architecture employed in malaria detection utilizes three convolutional layers followed by two fully connected layers. The model processes input images of dimension 100 × 100 × 3 pixels, where each convolutional layer applies 3 × 3 filters with a stride of 2 pixels.

The first two convolutional layers each consist of 32 filters, while the third layer comprises 64 filters. Rectified Linear Unit (ReLU) activation functions are used to enhance non-linearity, and an optimal weight initialization strategy further supports effective training [20]. To summarize feature maps efficiently, 2×2 max-pooling layers with a stride of 2 pixels are employed. The final pooled feature map is fed into a fully connected layer with 64 neurons, followed by another fully connected layer linked to a Softmax classifier. A dropout layer with a rate of 0.5 is incorporated to mitigate overfitting [21]. The model optimization leverages stochastic gradient descent (SGD) [22] with Nesterov momentum [23], while hyperparameter tuning is performed using a randomized grid search [24] to improve generalization.

To enhance malaria parasite classification accuracy, Amin et al. [25] proposed a multi-step pipeline that integrates convolutional neural networks with a support vector machine (SVM). The approach begins with a bilateral filtering technique to improve image quality. Features are then extracted using two strategies: a shape-based descriptor, namely the Pyramid Histogram of Oriented Gradients (PHOG), and deep feature extraction utilizing ResNet-50 and ResNet-18 networks. The extracted features, initially comprising 2300 dimensions, undergo a selection process using the Generalized Normal Distribution Optimization (GNDO) method, reducing them to 498 key features. These refined features are subsequently classified using an SVM, achieving an impressive accuracy of 0.99 when tested on a malaria microscopy dataset.

Wisit et al. [26] proposed a hybrid deep learning model that combines convolutional neural networks (CNNs) with k-nearest neighbors (KNN) for malaria detection. In this approach, the CNN component is responsible for learning high-level features from input images and generating feature maps, which are then passed to a KNN classifier for final decision-making. The feature representation extracted from the CNN's fully connected layer serves as input to the KNN, enabling an efficient and robust classification process. This hybrid model effectively leverages the strengths of CNNs for feature extraction and KNN for classification, making it suitable for processing medical image data, including malaria parasite detection [27].

The YOLO (You Only Look Once) model is used in real-time detection applications. In their research on detecting the malaria pathogen using a dataset of thick blood smear microscopic images captured with a mobile phone, the authors [28] developed two custom models, the first with three layers YOLO-mp-3l and the second with four layers YOLO-mp-4l. They achieved the best scores with an average precision measure (Average-Precision mAP) of 0.9399 and 0.9407, respectively, surpassing the standard YOLOv4, which obtained 92.56% for the same measure. They were also able to streamline their models for fast and optimal execution with a size of 21.8 Mb for YOLO-mp-3l and 25.4 Mb for YOLO-mp-4l, which is significantly better than the standard YOLOv4 with a size of 244 Mb, thus proving their capability to operate on low-resource devices.

Deep learning models incorporating recurrent architectures such as LSTM networks have also demonstrated substantial improvements in malaria image classification. Pereira-Ferrero et al. [29] introduced a feature augmentation technique leveraging rank-based manifold learning to refine LSTM model performance. This method assigns contextual similarity weights to training samples, leading to an accuracy improvement of up to 20% in image classification tasks. The approach combines convolutional feature extraction with LSTM-based sequence modeling, demonstrating its potential in medical image analysis where temporal dependencies and feature enhancement play a critical role.

Due to their conceptual limitations and inability to capture long dependencies, the CNNs have integrated the attention mechanisms, which allow them to dynamically adjust weights based on input features, in order to improve their non-local modeling capability [30–32]. Inspired by this line of research, many researchers have made significant efforts to propose models with attention variants in the field of medical imaging [33–37]. Although these attention mechanisms allow for the modeling of the complete contextual information of the image, the computational complexity of these approaches generally grows quadratically with respect to the spatial size, which implies an intensive computational load, making them inefficient in the case of medical images that are dense in pixel resolution [38]. Moreover, despite the fact that combining the attention mechanism with the convolution operation leads to systematic performance gains, these models inevitably suffer from constraints in learning long-range interactions. The original Transformer [32] was first applied to the task of machine translation as a new attention-driven building block and has demonstrated impressive performance across a wide range of tasks, including natural language processing (NLP), machine translation, text classification, and question answering. The success of Transformers has led to this technique being widely applied in modern computer vision (CV) models, giving rise to Vision-Transformers (ViTs) [39]. The ViTs quickly established themselves as viable alternatives to CNNs in various tasks such as image recognition [39], object detection [40], image segmentation [41], video understanding [42], and image super-resolution [43]. As the central piece of the Transformer, the self-attention mechanism has the ability to model the relationships between the elements of a sequence, thus learning long-range interactions.

The Vision Transformer (ViT) is an innovative neural architecture that has revolutionized computer vision by using self-attention techniques from natural language processing and adapting them to visual data. ViT segments images into fixed-size patches, integrating them linearly before passing them to a transformer encoder [44,45]. This innovative method facilitates comprehensive learning of image properties. By allowing exchanges between patches in both directions, the ViT captures extended dependencies, which enhances its ability to represent the broader context of images. The exceptional success of ViT has not only pushed the boundaries of image classification but has also spurred advancements in various computer vision tasks, unveiling new avenues for research in artificial intelligence and practical applications [46,47]. The authors in [48] for efficient skin cancer classification proposed a two-level architecture. The first level involves data augmentation techniques to increase the number of samples present in the HAM10000 dataset. In the second layer, the authors exploited the efficiency of medical vision transformers (MVT) used in medical image processing to design an MVT-based model for skin cancer classification. A large version of the Vision Transformer model was used, combined with an MLP head at its output. We recorded an accuracy of 0.96, a sensitivity of 0.96, an F1 score of 0.97, and an accuracy of 0.96. In [49], a new ViT model for skin cancer classification was presented. The method was based on transfer learning by leveraging a pre-trained ViT model and fine-tuning it with the HAM10000 dataset. The fine-tuning process was carried out by integrating a classification segment into the final segment of the encoder transformer, consisting of a flattened layer and two batch normalizations, separated by a dense layer activated by GeLU. The experiment achieved an accuracy of 0.94, surpassing all the compared techniques. Just like in [50], a pre-trained ViT fine-tuned with an MLP on the HAM10000 was used. However, a method based on contrastive learning was implemented. Contrastive learning relies on a specific loss function to decrease the similarity between samples of the same class while increasing the similarity between samples of distinct classes. This model achieved an accuracy rate of 0.94.

The explainability of intelligent models is crucial for doctors to understand and validate the decisions of AI systems. Techniques like Grad-CAM, LIME, or Shapley Additive exPlanations (SHAP) are increasingly used to make these models more transparent and accessible. These techniques allow for decoding the decisions of AI models, which is essential for ensuring their clinical acceptance [51].

According to Islam et al. [12], a network based on the Grad-CAM method was used to explain the decisions made by a transformer model for malaria diagnosis. This AI model not only predicts the presence of Plasmodium but also visualizes which parts of the image were most influential in the decision, thereby improving transparency and clinicians' trust in the results. This explanation process has been combined with

Transformer networks to improve the model's accuracy, particularly when parasites are visible under low contrast [52].

Rajab et al. [51] investigated explainable AI techniques for malaria diagnosis, employing SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to enhance transparency in machine learning predictions. They evaluated several classification algorithms, including Extreme Gradient Boosting, K-means, K-Nearest Neighbors, Support Vector Machines (SVMs), Decision Trees, Logistic Regression, Random Forest, Naïve Bayes, AdaBoost, and Explainable Boosting Machines (EBM). Among these, Random Forest and EBM achieved the highest accuracy (0.84). Logistic Regression, after fine-tuning with GridSearchCV, reached an accuracy of 0.81. The study further employed k-fold cross-validation with XGBoost to ensure robust model evaluation, highlighting the role of explainability techniques in increasing clinicians' trust in AI-based malaria diagnostic systems.

Traditional supervised learning models often require a large number of features, resulting in high computational costs and reduced interpretability. To address this issue, Khan et al. [13] introduced a feature selection strategy designed to minimize the dimensionality of malaria diagnostic datasets while maintaining classification accuracy. Their approach focuses on extracting only the most relevant features, significantly improving computational efficiency and interpretability. The study observed that malaria-infected red blood cells exhibit distinctive ring-shaped structures, while uninfected cells do not. By leveraging this morphological characteristic, the proposed method efficiently identifies infected cells, providing a streamlined yet highly accurate approach to malaria detection.

One of the most original approaches for detecting malaria and typhoid is the one proposed by Attai et al. [53]. The originality of this study lies not only in the use of XAI, such as LIME, but especially in the application of LLMs, including Generative Pretrained Transformer (GPT), to facilitate the elucidation of diagnostic results for healthcare professionals. Indeed, chatbots, based on AI-driven conversational agents that are widespread in online interactions, have found considerable use in healthcare and improving customer services [54]. The work presented by Chow et al. [55] summarizes the general characteristics that healthcare professionals expect from a medical chatbot very well. These features include accurate information retrieval, symptom evaluation, and diagnostic support to help understand and treat health problems. Additionally, the chatbot should offer treatment advice, medication information, and appointment booking assistance, thereby providing a comprehensive healthcare experience. The results indicated that the Random Forest (RF) model outperformed the performance of the other evaluated models, achieving an F1 score of 0.7145; furthermore, significant features were found using LIME graphs, while ChatGPT 3.5 showed a comparative advantage over the models of large language. The research combines RF, LIME, and GPT to develop a mobile application to improve interpretability and transparency in diagnosing malaria and typhoid. Despite its encouraging results, the dataset's quality limits the system's effectiveness. Moreover, although LIME and GPT improve transparency, they can complicate real-time implementation due to their computational requirements and the need for an Internet connection to ensure relevance and accuracy.

3. Material and mathods

The approach proposed in this study includes a pipeline consisting of three main steps. The first involves data pre-processing, which prepares blood smear images for analysis. The second step focuses on classification. It allows us to test and compare the results of five different models, namely, VGG-16, VGG-19, MobileNetV2, ViT and Stacked-LSTM, with an attention mechanism. As we will see in Section 3.2, these models have been customized to detect malaria cases with high accuracy. Finally, the third step compares the results using two explainability techniques, Grad-CAM and LIME, to interpret and understand the decisions made. This methodology was proposed to combine both solid classification performance and increased transparency, thus offering a robust and interpretable framework for malaria detection.

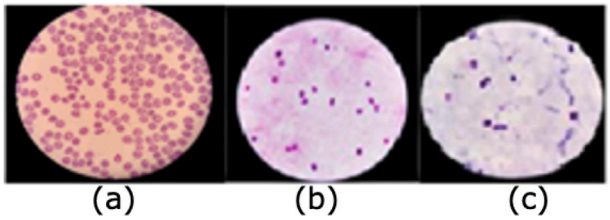


Fig. 1. Example images from the NIH-NLM - Malaria dataset. (a) represents a blood smear, (b) represents a healthy red blood cell, and (c) represents an infected red blood cell. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

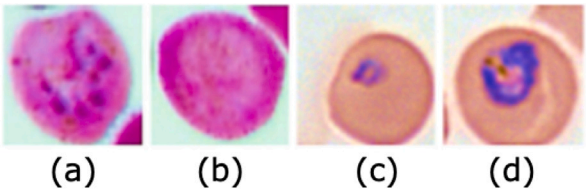


Fig. 2. Example images describing the different species of the Plasmodium parasite. (a) P. Falciparum, (b) P. Vivax, (c) P. Ovale, and (d) P. Malariae.

3.1. Dataset

The dataset used in this study, publicly available on the National Library of Medicine website, contains 27,558 images of Giemsa-stained blood smears, photographed at a 100x magnification in the RGB color space with a very high resolution of 3024 × 3024 pixels [6]. They were provided in JPG format and a filter was then applied to eliminate acquisition noise. The images thus obtained are of high quality, as can be seen inFigs. 1 and 2. The dataset is well-balanced, with an equal number of images representing infected and non-infected cells. The data collection process involved several steps, including the collection of blood samples from individuals infected with *Plasmodium falciparum* and *Plasmodium vivax*, as well as uninfected individuals, at various locations, such as the Chittagong Medical College Hospital in Bangladesh and Bangkok, Thailand. This diverse and balanced dataset ensures effective model training for both classes, providing a solid foundation for malaria detection [6,19,56,56–60].

Four species of Plasmodium are involved in the spread of malaria, as shown in Fig. 3. Plasmodium falciparum, the most pathogenic species responsible for fatal cases of malaria, is predominant in Africa. Plasmodium vivax, coexisting with P. falciparum, has been identified in temperate zones. In West Africa, the presence of P. ovale can be observed, a parasite that is not fatal but causes relapses approximately four to five years after the initial infection. Finally, on a global scale (with an uneven distribution), the presence of P. Malariae is observed, which, although not fatal, carries a risk of relapse that can occur up to twenty years after the first infection. Relapses are associated with the persistent presence of the parasite in the liver in a latent state (hypnozoite). Cases of direct transmission of this disease between individuals are rare. They can manifest, for example, when there is sharing of contaminated syringes or through transplacental transmission during pregnancy [61].

As illustrated in Fig. 3, the life cycle of the Plasmodium parasite begins when an infected female Anopheles mosquito bites a human host, introducing sporozoites into the bloodstream. These sporozoites rapidly migrate to the liver, where they invade hepatocytes and undergo asexual replication. Over the course of approximately 7 to 10 days post-infection, they proliferate within liver cells without triggering any noticeable symptoms. 8 to 30 days after infection, a fever develops. It may be accompanied by weakness, headaches, muscle pain, vomiting, diarrhea, and/or coughing. Fever accompanied by tremors,

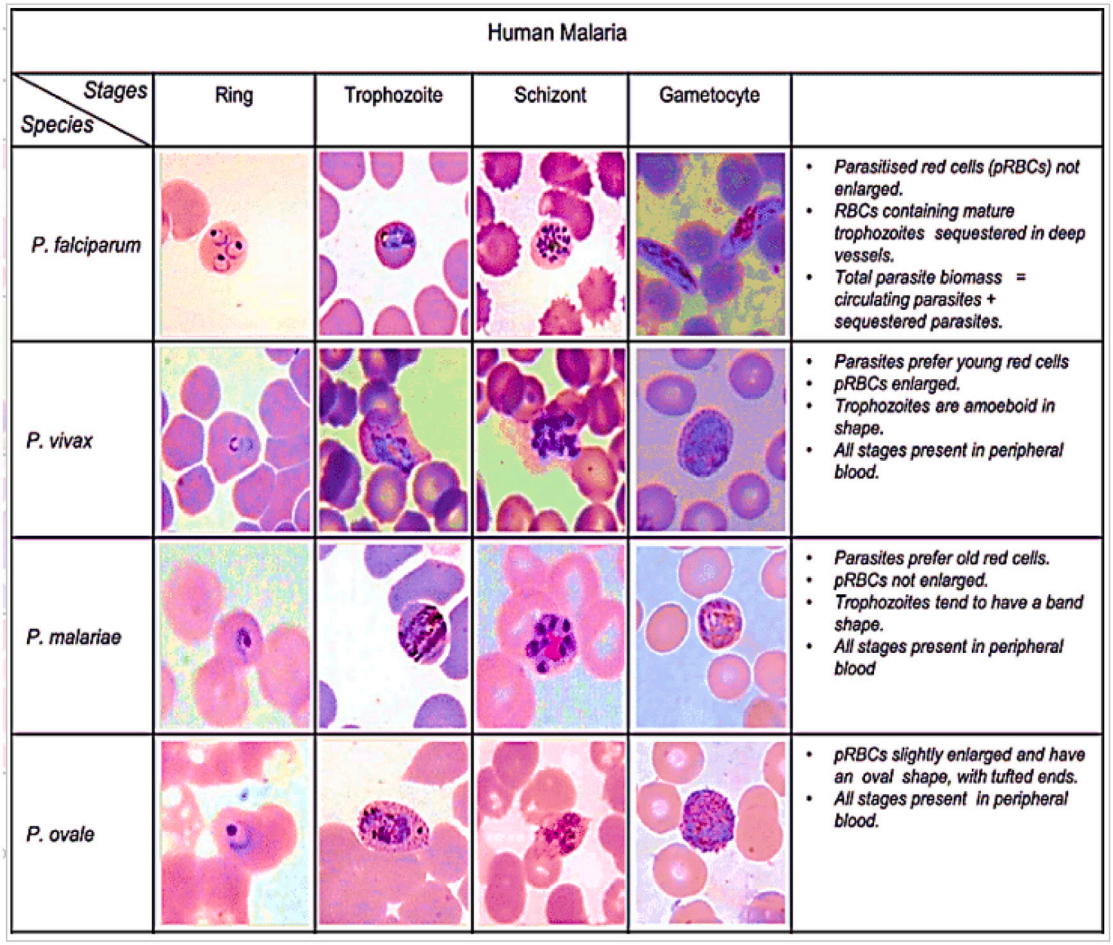


Fig. 3. Example of images that describe different species of plasmodium and their development stages.

cold sweats, and intense perspiration may occur cyclically due to the different phases of the parasite cycle. More serious symptoms may occur, such as breathing difficulties, bleeding, jaundice, extreme fatigue, and convulsions. In some cases, infected red blood cells can obstruct the blood vessels that supply the brain, which can be fatal. In the human bloodstream, merozoites invade red blood cells (erythrocytes), where they multiply until the host cells rupture. This process leads to the release of new merozoites, which then invade other erythrocytes, perpetuating the infection cycle. Each time the parasites spread and attack new blood cells, periodic fevers appear as a clinical symptom of the disease. However, a subset of the infected red blood cells diverges from this asexual replication cycle. Instead of continuing multiplication, some merozoites differentiate into sexual-stage forms known as gametocytes, which circulate within the bloodstream. When a mosquito feeds on an infected individual, it ingests these gametocytes, which then mature into sexual cells called gametes within the mosquito's gut. The fertilized female gametes develop into motile ookinetes, which penetrate the midgut lining and transform into oocysts on the external surface. Inside the oocysts, thousands of sporozoites form. Once the oocyst ruptures, these sporozoites migrate to the salivary glands of the mosquito. The transmission cycle resumes when the mosquito bites another human, injecting sporozoites into the new host and initiating another infection [40]. To confirm or exclude a diagnosis of malaria, a blood sample must be analyzed using parasitological tests. The standard test consists of microscopic examination of blood smears (thin and thick blood smears). A great deal of training and experience is required to analyze blood smears properly, particularly to interpret thick smears, identify parasite species, and quantify parasitemia. The absence of experienced personnel can limit the precision of malaria diagnosis, and, in the majority of cases, malaria cannot be diagnosed during the first hours after the patient has been infected with the conventional method. This is why we want to screen and detect the parasite early (within the first few hours of contamination) by analyzing blood smear images with our intelligent models, before it develops and becomes contagious. On the one hand, to prevent its spread within the population and, on the other hand, to ensure that the individual starts treatment before his or her health deteriorates or his prognosis becomes life-threatening [62].

3.1.1. Data preprocessing

In this study, the dataset was divided into three subsets to rigorously evaluate model performance: 50% for training, 20% for validation, and 30% for testing. This split ensures a balanced distribution for robust training and reliable evaluation.

A series of preprocessing steps were applied to infected and uninfected cell images to prepare the data for machine-learning models. First, images were loaded from the dataset and resized to 224 × 224 pixels to match the model's input requirements. Data augmentation techniques, including random cropping, rotations of 45° and 75°, and Gaussian blurring with a 10 × 10 kernel, were applied to enhance dataset diversity and prevent overfitting. Each image was assigned a label: 1 for infected cells and 0 for uninfected cells. These augmentation techniques were applied to the images used for the CNN-based models, VGG-16 and VGG-19, and the proposed Stacked-LSTM with attention mechanism ensuring balanced representation across the dataset.

For the proposed Stacked-LSTM model with attention mechanism, an extra preprocessing pipeline was adopted. The augmented images underwent grayscale conversion to reduce computational complexity

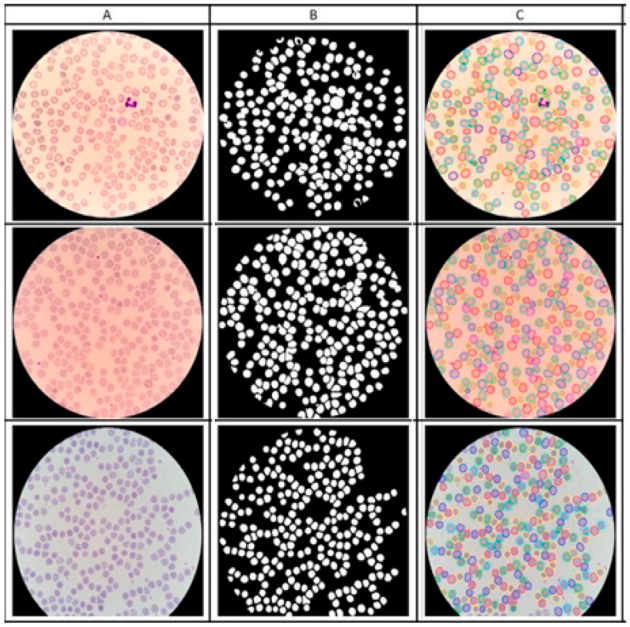


Fig. 4. Otsu thresholding segmentation and ROI detection from blood smear images. (A) Input image. (B) Final segmentation mask and ROI detection. (C) Segmentation results are superposed on the original image.

by eliminating colorimetric information, achieved using the perceptual luminance model:

$$\text{Grayscale} = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \quad (1)$$

With R, G, and B the three channels of the color image. Note that these coefficients are based on the perceptual luminance model and give more weight to green, as the human eye is more sensitive to this color.

Following grayscale conversion, threshold segmentation was applied using the Otsu method, which binarized the images and effectively separated the objects (cells) from the background (as shown in Fig. 4). Subsequently, regions of interest (ROIs) were detected based on contour and shape analysis, enabling the identification of Plasmodium parasites at various developmental stages. This selective approach reduces noise by focusing solely on relevant areas.

The extracted ROIs were then resized uniformly to 50 × 50 pixels. Each ROI image was subsequently divided into smaller patches of 5 × 5 pixels to construct a sequence that serves as the input for the LSTM layers. This approach converts spatial data into sequential representations, facilitating the capture of spatial dependencies and enhancing the interpretability of the model's decision-making process.

3.2. Proposed approach

3.2.1. Models explainability

XAI techniques have emerged to address the transparency challenge in machine learning, particularly for high-stakes healthcare applications where interpretability is essential for clinical acceptance, ethical accountability, and regulatory compliance. These methods facilitate a deeper understanding of model decisions, crucial for tasks such as automated malaria detection.

The terms interpretability and explainability are often used interchangeably in the AI literature, yet they refer to distinct concepts. Interpretability denotes the extent to which a model's decision-making process is inherently understandable without requiring external tools or modifications [63]. Transparent models, such as linear regression and decision trees, exhibit this property by design.

Explainability, by contrast, refers to the use of additional techniques to elucidate the internal workings of complex models, particularly those lacking intrinsic transparency, such as DNNs [64]. Explainability methods, including feature attribution, saliency maps, and surrogate models, enable clinicians and researchers to analyze AI-generated outputs, facilitating model validation and fostering trust in medical AI systems.

XAI methods are broadly classified into model-specific and model-agnostic techniques. Model-specific methods leverage internal model structures, as seen with gradient-based approaches like Grad-CAM, which visualizes important image regions influencing CNN predictions. Model-agnostic methods, such as LIME, operate independently of the model architecture by approximating model behavior locally using interpretable surrogate models.

XAI methods can also be categorized based on their granularity into global or local approaches. Global methods, like SHAP, provide insights into the overall behavior of a model across datasets, while local methods, such as Grad-CAM and LIME, explain individual predictions by identifying specific influential input features.

In this study, Grad-CAM and LIME were employed to enhance the interpretability of the proposed Stacked-LSTM model with an attention mechanism for malaria detection. Grad-CAM is used to visualize the regions of an image that most influenced the model's decision. As shown in Fig. 5, Grad-CAM calculates the gradients of the target class concerning the feature maps of the final convolutional layer and then uses these gradients to generate a heatmap. The heatmap highlights the regions of the input image that have the most significant impact on the model decision, offering global visualization and helping to interpret which parts of the image contributed to a specific classification result [65].

In contrast, LIME explains individual predictions by approximating the model's behavior with a simpler, interpretable surrogate model. This technique generates perturbed versions of the input data to determine which data features have affected a specific prediction, subsequently making the decision process more transparent [66]. Here, we use LIME to improve the interpretability of a deep-learning model for classifying malaria cells. LIME generates super-pixels by perturbing the input image and examining the effect of these perturbations on the model's output. The visual explanation highlights important parts in the image (e.g., the cell's shape, texture, or color) responsible for the model's decision. Such transparency is crucial for validating predictions and confirming that the model focuses on biologically relevant features, especially for medical applications such as malaria.

Using these complementary techniques allowed comprehensive analysis of model behavior. Grad-CAM provided an overall understanding of image regions considered significant by the model, whereas LIME offered detailed, localized interpretations, pinpointing precise features driving each decision. Integrating these methods demonstrated the robustness and reliability of the proposed model and confirmed its focus on biologically relevant image features essential for accurate malaria diagnosis.

3.2.2. Classification models

CNNs' main advantage is their ability to learn representations and extract features from visual data. In this way, these models tend to require a minimal level of preprocessing compared to other image classification algorithms [29]. In this study, we used five well-established deep learning models.Initially, we tested the VGG-16 and VGG-19 models for the classification of images of infected and non-infected cells. The two models, VGG-16 and VGG-19, are CNNs that have been widely adopted for image classification tasks due to their efficient and straightforward architecture and are distinguished by their use of small convolutional filters (3 × 3), which are stacked in a deep architecture to capture complex visual patterns [67].

Although the first two models gave us notable performances, we thought it would be very interesting to test a lightweight deep model

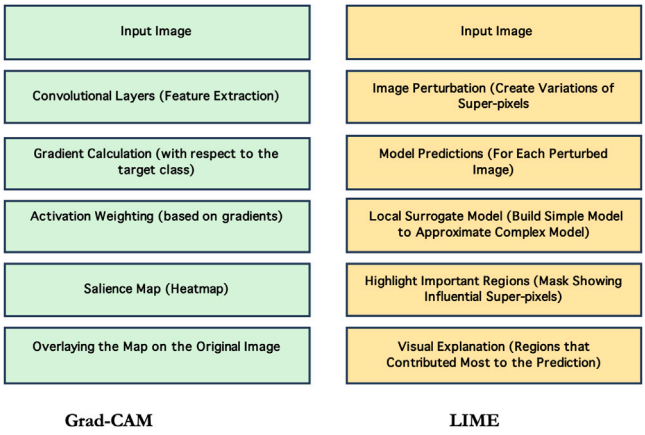


Fig. 5. Flowchart of Grad-Cam and LIME model's.

that can be used on resource-limited devices (smartphones, connected microscopes, Raspberry Pi, etc.) and that also presents a good compromise between accuracy and model size, which is crucial for medical applications where latency and computational power may be limited. Moreover, malaria detection is often needed in regions with limited access to electricity or expensive equipment. These motivations led us to design a lightweight and efficient classification architecture based on MobileNetV2, a deep convolutional network optimized for resource-constrained environments [68]. MobileNetV2 is designed to be lightweight thanks to its architecture based on depthwise convolutions and inverted residual blocks, allowing for a significant reduction in the number of parameters while maintaining good generalization performance. This leads to fast inference even on resource-limited devices.

We employed transfer learning by using pre-trained weights from the ImageNet dataset and fine-tuning them for our specific task. This allowed us to leverage the pre-trained feature extraction capabilities of VGG-16 and VGG-19, improving training efficiency and model performance. Both models were chosen for their ability to extract detailed hierarchical features, making them particularly effective for complex tasks like malaria cell detection.

Although CNN models have made excellent advances, primarily supported by their deep features, a significant obstacle remains related to the lack of contextual information in such representations. Indeed, these representations often lie on manifolds in a high-dimensional space [69], where the pairwise formulation of the similarity measure is insufficient to reveal the intrinsic relationship between the images. Our work focuses on finding a more efficient representation of features in such a context by considering the contextual similarity relationships defined by the extracted features. To achieve this goal, we have begun to test ViTs, which are gradually revolutionizing medical imaging by providing innovative solutions where traditional CNNs show limitations, such as their ability to capture subtle patterns (micro-calcifications, cellular anomalies) through multi-head attention or their ability to preserve long-distance spatial relationships, crucial for complex anatomical structures or large images, unlike CNNs, which cannot do so. After that, we used RNNs, specifically Stacked-LSTM networks. The LSTM networks proposed by Hochreiter and Schmidhuber [70] were a key moment for RNNs, as they allowed the learning of dependencies over much more extended periods of time, solving the problem of vanishing gradient inherent in basic RNNs during the backpropagation process [71].

A key innovation of Long Short-Term Memory (LSTM) networks is their ability to regulate information flow through specialized gates. This gating mechanism enables LSTMs to retain and update their internal state over extended sequences, making them particularly effective for

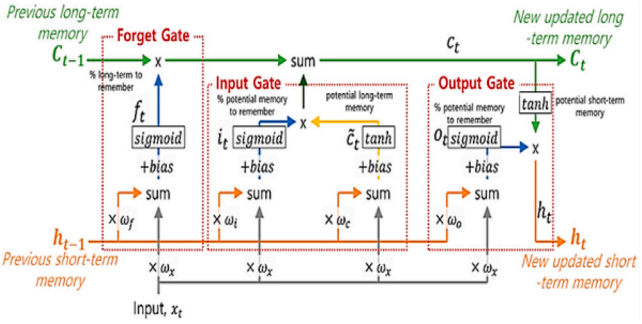


Fig. 6. Architecture of an LSTM cell.

capturing long-range dependencies in data. As illustrated in Fig. 6, an LSTM unit consists of three primary gates: the input gate, forget gate, and output gate. These components play a crucial role in controlling the cell state c_t and hidden state h_t [72,73]. Specifically, the gates determine the amount of information retained from the input, the portion of prior information to discard, and the extent of new information integrated into the cell.

The mathematical representation of the LSTM cell update process is given as follows:

$$\mathbf{i}_t = \sigma(\omega_{ix}\mathbf{x}_t + \omega_{ih}\mathbf{h}_{t-1} + b_i) \tag{2}$$

$$\mathbf{f}_t = \sigma\left(\omega_{\mathbf{x}f}\mathbf{X}_t + \omega_{\mathbf{h}f}\mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{3}$$

$$\mathbf{o}_t = \sigma\left(\omega_{\mathbf{x}o}\mathbf{X}_t + \omega_{\mathbf{h}o}\mathbf{h}_{t-1} + \mathbf{b}_o\right) \tag{4}$$

$$\tilde{c}_t = \tanh(\omega_{gx}\mathbf{x}_t + \omega_{gh}\mathbf{h}_{t-1} + \mathbf{b}_g) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

$$\mathbf{h}_t = o_t \odot \tanh(c_t) \tag{7}$$

where:

i_t is the input door.

f_t is the forget door.

o_t is the output door.

$\tilde{c}_t$ is the output door.

c_t is the state cell.

h_t is the hidden state.

σ is the sigmoid function.

tanh is the hyperbolic tangent function.

$\odot$ denotes element-wise multiplication.

There are several architectures based on LSTMs, such as bidirectional LSTMs (Bi-LSTM), which process sequences in both forward and backward directions, allowing the LSTM network to capture the context of both the past and the future [74]. Stacked-LSTM networks represent another architecture that is based on stacking LSTM layers. The stacked LSTM integrates external recurrence by connecting multiple LSTM layers in such a way that the outputs of one layer represent the input of the next LSTM layer. This deep architecture offers us the possibility to learn and capture much more complex patterns and sequences at different levels of abstraction from the input data. In fact, while the

lower layers can capture local patterns and short-term dependencies, the upper layers, on the other hand, can capture more abstract features and long-term dependencies [59].

In this study, we chose to experiment with Stacked-LSTM equipped with an attention mechanism to detect and classify malaria from blood smear images. In the majority of works using LSTMs in the field of image classification, researchers precede this classification with a feature extraction step based on CNNs [14,16,29,75,76]. Although such an approach has yielded good results in terms of accuracy and precision, their training times are relatively slow, and, from a computational point of view, they are costly and complex, adding a layer of opacity to the so-called "black box models", making them non-interpretable and difficult to explain approaches. Since our main goal is not only to achieve good performance but also, above all, to present models that are as interpretable and explainable as possible, with a very short learning time, we preferred to avoid the use of CNNs. Since LSTMs process temporal sequences and telescopic medical images do not present such sequences, we transformed the ROIs obtained in the image-preprocessing step into 5 × 5 pixel image patches in order to organize them into a spatial sequence. To do this, we performed positional coding, as done in transformers, using the following formulas:

$$\mathbf{PE}_{(\mathbf{x},\mathbf{y},\mathbf{2i})} = \sin\left(\frac{\mathbf{x}}{\mathbf{10000}^{\frac{\mathbf{2i}}{\mathbf{d}}}}\right) \tag{8}$$

$$\mathbf{PE}_{(\mathbf{x},\mathbf{y},\mathbf{2i+1})} = \cos\left(\frac{\mathbf{x}}{\mathbf{10000}^{\frac{\mathbf{2i}}{\mathbf{d}}}}\right) \tag{9}$$

$$\mathbf{PE}_{(\mathbf{x},\mathbf{y},\mathbf{2j})} = \sin\left(\frac{\mathbf{x}}{\mathbf{10000}^{\frac{\mathbf{2j}}{\mathbf{d}}}}\right) \tag{10}$$

$$\mathbf{PE}_{(\mathbf{x},\mathbf{y},\mathbf{2j+1})} = \cos\left(\frac{\mathbf{x}}{\mathbf{10000}^{\frac{\mathbf{2j}}{\mathbf{d}}}}\right) \tag{11}$$

where:

- **(x, y)** are the coordinates of the patch in the grid.
- **i and j** iterate over the dimensions of the encoding vector.
- **d** is the dimension of the encoding vector.

This gives each patch a unique vector encoding its position in the image. The set of patches is then organized into a sequence according to their position. The objective is to provide each patch with information about its spatial position so that the first LSTM layer understands the spatial relationships when processing the sequence as if it were a temporal sequence that can be exploited by the first LSTM layer of our model, as shown in Fig. 7.

The first LSTM layer consists of 64 LSTM cells. It analyses the sequence of patches, their dependencies, and their spatial relationships, transforming each one into a rich and contextual representation and outputting a sequence of hidden states h_t. We obtain a sequence of representations $H = [h_1, h_2, \ldots, h_n]$, where n is the number of patches.

We applied a 50% dropout on the outputs obtained from the first LSTM layer to avoid overfitting and provide robustness to our model. The remaining sequences will be directed towards an attention layer, integrating an additive attention mechanism [77]. This attention mechanism plays a crucial role in malaria detection. It allows the model to focus on the most relevant areas of blood smear images, particularly the regions of interest (ROIs) detected by segmentation. This mechanism begins by extracting spatiotemporal features from the patches and calculating an attention vector for each, reflecting its importance and relevance for classification. The most influential patches (notably those that exhibit characteristics of the plasmodium) will be assigned a higher attention score. The importance score is calculated for each vector ht and is compared to a context vector w learned by the model. We calculate an importance score e_t for each patch:

$$\mathbf{e_t} = \text{score}(\mathbf{h_t}, \mathbf{w}) = \mathbf{v^T} \tanh(\mathbf{wh_t} + \mathbf{b}) \tag{12}$$

where:

w and b are learned parameters.

v^t is a learned weighting vector.

We then move on to the score normalization phase using the Softmax function, applying it to all importance scores to obtain attention weights α_t:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{n} \exp(e_j)} \tag{13}$$

Normalization ensures that the attention weights α_t are between 0 and 1 and that their sum is equal to 1; this allows us to interpret the weights probabilistically. The extracted features are weighted by the attention vector V_t giving more weight to important patches. Allowing the model to focus on important regions rather than treating all patches similarly. The vector V_t is obtained by the weighted sum of all hidden states h_t:

$$v_t = \sum_{t=1}^{n} \alpha_t h_t \tag{14}$$

The integration of the attention mechanism into our classifier was motivated by objectives related to prediction and classification, as well as by the increase in model transparency to make it more interpretable and explainable, thus allowing healthcare professionals to understand the obtained results fully. The first objective was achieved primarily by focusing the model on the attention regions. Indeed, this allowed us to reduce noise and increase classification accuracy, as well as computational complexity and training time, because we do not process the entire image but only specific attention regions. The second goal, to increase the transparency of the model, making it more interpretable and easier to explain, was achieved by using the visualization of the attention regions to understand which areas of the blood smear the model considers important. This visualization also allows us to explain why the model predicted the presence or absence of malaria by pointing to the relevant regions of interest. Similarly, experts can verify the model's decisions by analyzing these regions, giving them confidence in its predictions. Unlike the first LSTM layer, which receives a sequence of patches and produces a sequence of outputs, the second LSTM layer receives a single attention vector V_t and does not produce any sequence. It acts as a final encoder that compresses this information into a unique image representation, integrating the contextual information extracted by the attention mechanism and producing a consistent global representation. This operation is comparable to the functioning of a recurrent LSTM in a single timestep, transforming the vector V_t into a richer representation, producing an output that represents the final hidden state (**Final** h_t) used for classification. Similarly to the first LSTM layer output, we apply dropout to the second LSTM layer output, deactivating only 20% of the connections. The output will then be sent to the fully connected layer, equipped with a sigmoid function for binary classification: infected or not-infected. The final step of our pipeline concerns the explainability and visualization of the obtained results, thanks to XAI methods, Grad-CAM, and LIME.

4. Experiments and results

We carried out our experiments using Google Colab, a cloud-based platform that provides free access to GPUs, which was ideal for handling the computational demands of training and testing deep learning models. Google Colab supports Python and popular deep learning frameworks like TensorFlow and Keras, which we used to implement the VGG-16, VGG-19, MobileNetV2, ViT, and Stacked-LSTM models and the explainability techniques (Grad-CAM, LIME). This platform allowed us to efficiently process large datasets, perform model evaluations, and apply various interpretability methods, all while benefiting from its integration with Google Drive for easy data management and sharing. The cloud-based environment also facilitated collaboration and ensured the reproducibility of our experiments.

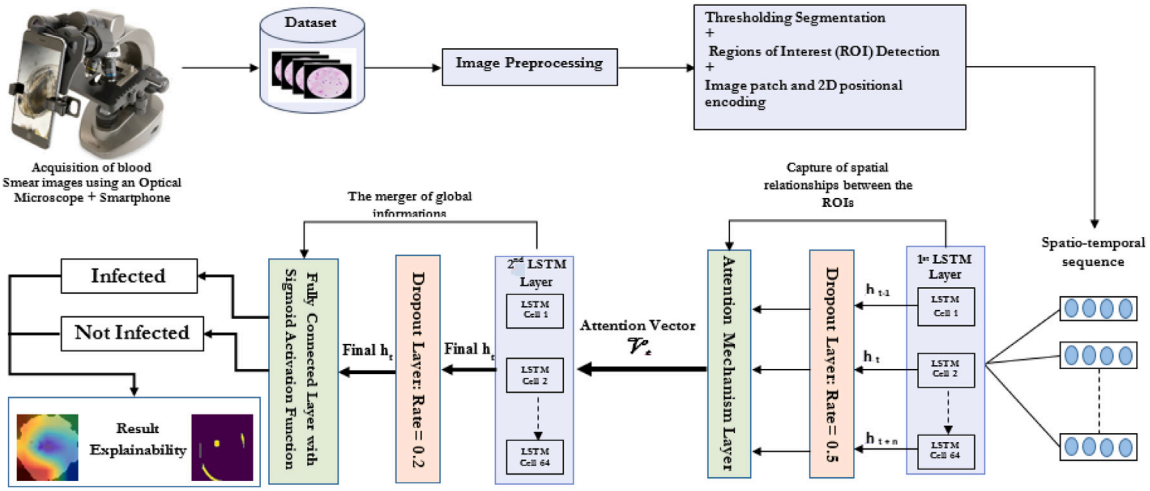


Fig. 7. Architecture of the proposed model combining Stacked-LSTM and attention mechanism.

Training and evaluation of the aforementioned models were performed using a consistent set of fundamental hyperparameters to ensure a fair comparison. All images were resized to 224 × 224 pixels with three RGB channels to standardize input dimensions across models. Transfer learning was employed for VGG-16, VGG-19, and MobileNetV2 by initializing the networks with pretrained ImageNet weights, while ViT and the Stacked-LSTM models processed sequential representations derived from the input images. Training was performed using a batch size of 32 across all models, optimizing parameters with the Adam optimizer at a learning rate of 0.0001. The binary cross-entropy loss function was used to address the binary classification task, and training was conducted over 100 epochs with early stopping (patience = 5) to prevent overfitting. Regularization techniques were also applied, with dropout mechanisms incorporated into the network layers to enhance generalization. Activation functions were carefully selected: ReLU was used in the intermediate layers of the VGG models, while sigmoid activation was applied to the output layers of all models to produce probability scores.

Beyond these common hyperparameters, model-specific configurations were introduced to tailor the architectures to their respective learning paradigms. For the convolutional models, VGG-16 and VGG-19, the dense layers included a 50% dropout rate to mitigate overfitting. These models leveraged their deep convolutional feature extraction capabilities, followed by fully connected layers for classification.

Regarding the MobileNetV2 model, we used it as a feature backbone, disabling its original classification head and replacing it with a classification head added on top of MobileNetV2. That is to say, we configured and used it solely as a feature extractor, without the original classification layers; the Global Average Pooling 2D layer allowed us to reduce each feature map to an average value, while decreasing the number of parameters; then a 30% Dropout was applied twice to limit overfitting; the Dense layer composed of 64 neurons, ReLU is a fully connected layer used for learning abstract representations; finally, we used an output layer represented by a Sigmoid function composed of a single neuron for binary classification producing a probability for the "Infected" vs. "Uninfected" classes. The entire set of layers in the model was designed to maintain a good balance between lightness, performance, robustness, and computational efficiency.

The Vision Transformer (ViT) model, designed in our study, represents a lightweight model adapted to binary classification of red blood cell images into infected and uninfected classes. Unlike the presented CNN model, the ViT model leverages the self-attention mechanism to capture long-range dependencies and contextual relationships between image regions. Each input image is resized to 224 × 224 pixels and partitioned into fixed-size, non-overlapping patches of 16 × 16 pixels. These patches are flattened and linearly projected into a lower-dimensional embedding space (dimension 64), forming a sequence of 196 tokens. Positional embeddings are added to each patch token to preserve the spatial information lost during flattening. This enables the model to distinguish patch locations within the image's 2D structure. The encoded patches are passed through a stack of four Transformer blocks. Each block consists of Layer Normalization, Multi-Head Self-Attention (MHSA) with 4 heads, Residual connections, and a two-layer feed-forward neural network (MLP) with GELU activation and dropout. These Transformer blocks enable the model to learn complex interactions between distant patches, thus improving its ability to detect global patterns characteristic of malaria-infected cells. Instead of using the conventional (CLS) token used in original ViT implementations, we aggregate the information from all patches using a GlobalAveragePooling1D layer. This is followed by a fully connected layer with 128 neurons (ReLU activation) and a final sigmoid-activated dense layer that outputs the probability of infection.

In contrast, the Stacked-LSTM model, designed to capture temporal dependencies in feature representations, incorporated sequential processing. Each LSTM layer consisted of 64 LSTM cells, and the input sequence was structured into 50 time steps per instance. Regularization was applied differently in this architecture, with a 50% dropout rate in the first LSTM layer and a reduced 20% dropout rate in the second LSTM layer to maintain a balance between feature retention and overfitting prevention. Unlike the VGG models, where ReLU activations dominated intermediate layers, the Stacked-LSTM model relied on a sigmoid activation function in its fully connected output layer to generate classification probabilities.

4.1. Classification results

In this study, we first compared the performance of our proposed model, Stacked-LSTM with an attention mechanism, against two commonly used CNN models (VGG-16 and VGG-19) for the malaria classification task. Subsequently, an ablation study was conducted to assess the impact of the attention mechanism by comparing the performance of our model with and without attention.

Table 1 summarizes the performance of these models on the malaria dataset. The proposed Stacked-LSTM with an attention mechanism achieved the highest accuracy of 0.9912, surpassing VGG-16 and VGG-19, which achieved accuracy scores of 0.9802 and 0.9601, respectively. The results highlight the significant improvement provided by incorporating the LSTM network enhanced with the attention mechanism, achieving superior values across accuracy, sensitivity, specificity, precision, F1-score, and AUC compared to traditional CNN architectures.

Table 1
Performance comparison between VGG models and the proposed Stacked-LSTM with an attention mechanism for malaria detection.

Model	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC
VGG-16	0.9802	0.9839	0.9765	0.9763	0.9801	0.9780
VGG-19	0.9601	0.9681	0.9521	0.9528	0.9604	0.9588
ViT	0.9509	0.9273	0.9746	0.9734	0.9498	0.9503
MobileNet-V2	0.9303	0.9475	0.9132	0.9160	0.9315	0.9258
Stacked-LSTM with Attention	**0.9912**	**0.9867**	**0.9956**	**0.9959**	**0.9911**	**0.9912**

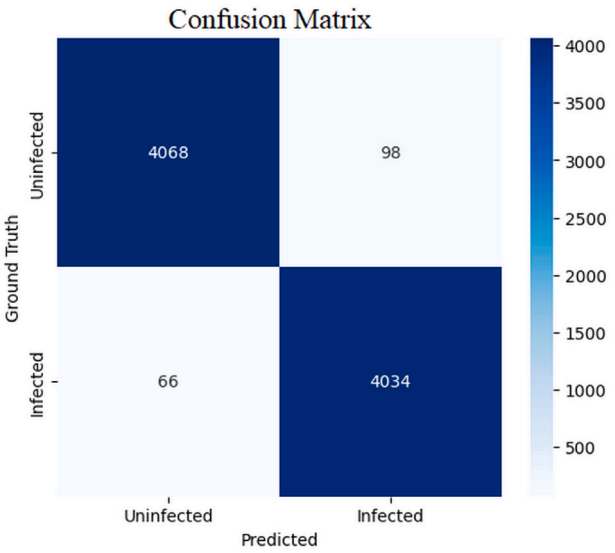


Fig. 8. Confusion matrix for VGG-16 model.

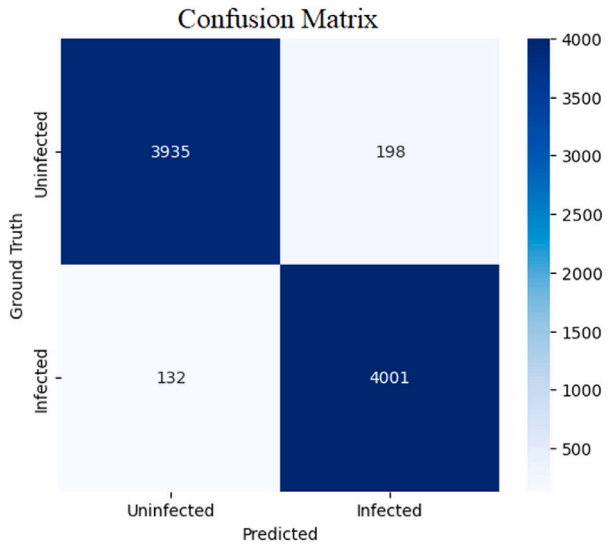


Fig. 9. Confusion matrix for VGG-19 model.

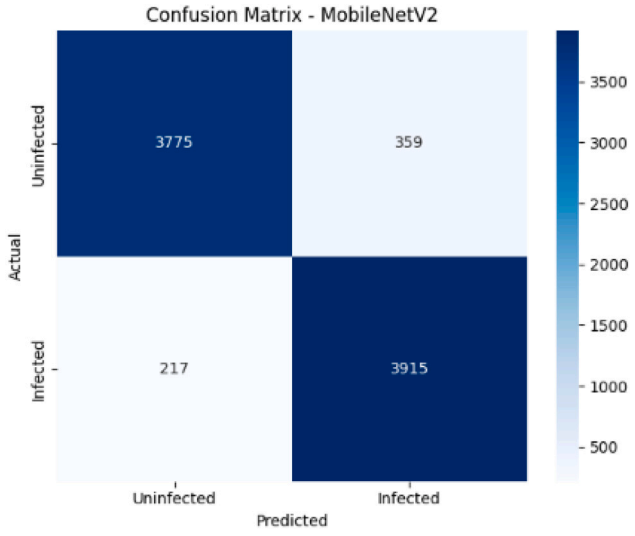


Fig. 10. Confusion matrix for MobileNet-V2 model.

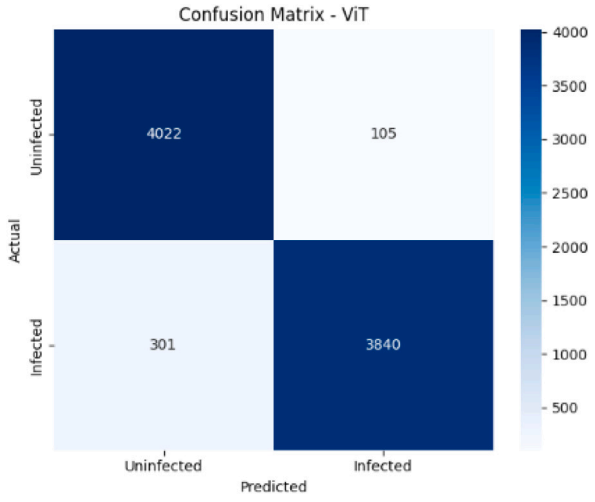


Fig. 11. Confusion matrix for ViT model.

Additionally, to further illustrate model performance, Figs. 8 and 9 present the confusion matrices for the VGG models. VGG-16 (Fig. 8) accurately classified 4068 uninfected and 4034 infected samples, with 98 false positives and 66 false negatives, demonstrating high precision and sensitivity. However, it leaves room for improvement in reducing false negatives.

In contrast, VGG-19 (Fig. 9) correctly identified 3935 uninfected and 4001 infected samples but misclassified 198 uninfected as infected and 132 infected as uninfected, suggesting lower specificity.

MobileNetV2 (Fig. 10) is designed for mobile and edge devices, and prioritizes efficiency and speed. This is confirmed by its confusion matrix, which shows that this model was able to classify 3775 infected cases and 3915 uninfected cases. However, we also note that it has the highest number of FN cases = 217, which is enormous in such a field, since it has an FN rate = 5.25%. However, it misclassified 359 uninfected cases as infected, confirming its low specificity and precision.

The Transformer-based model (Fig. 11) correctly classified 4022 cases as infected and 3840 cases as uninfected, confirming its high specificity and precision. While it misclassified 105 healthy individuals as infected, this shows that ViT is fairly conservative in positive predictions. However, misclassifying 301 infected cases wrongly predicted as uninfected represents a high FN rate (7.27%), translating his low sensitivity.

Table 2
Ablation study highlighting the performance enhancement provided by integrating an attention mechanism into the Stacked-LSTM model.

Model	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC
Stacked-LSTM (No Attention)	0.9758	**0.9877**	0.9644	0.9638	0.9815	0.9761
Stacked-LSTM with Attention	**0.9912**	0.9867	**0.9956**	**0.9959**	**0.9911**	**0.9912**

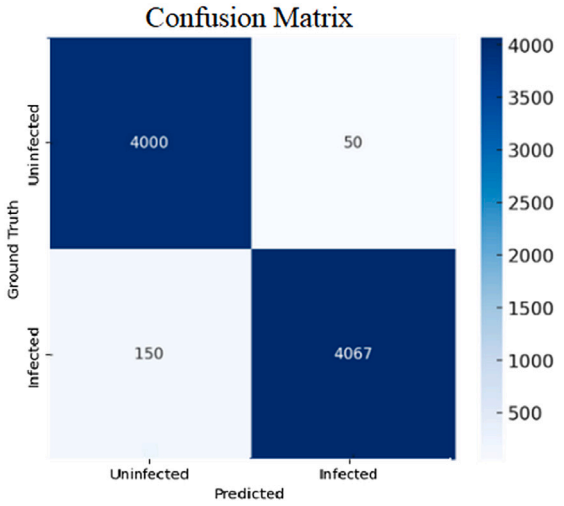


Fig. 12. Confusion matrix for Stacked-LSTM without attention mechanisms.

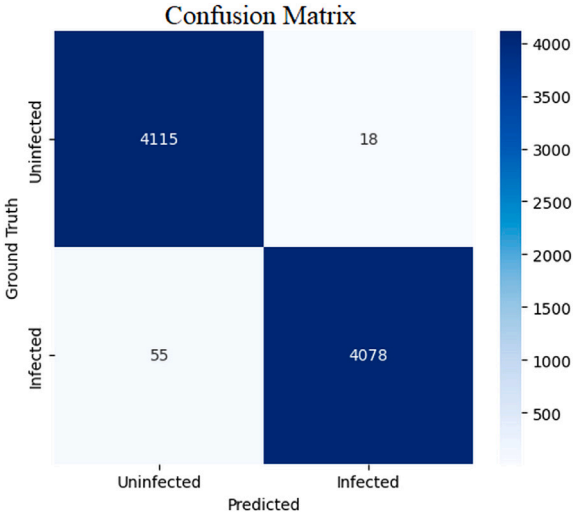


Fig. 13. Confusion matrix for Stacked-LSTM with attention mechanisms.

To demonstrate the contribution of the attention mechanism, we conducted an ablation study. Table 2 presents the comparative results for the Stacked-LSTM model with and without attention.

The ablation study confirms the critical role of the attention mechanism in enhancing the model performances. Specifically, accuracy increased from 0.9758 to 0.9912, while specificity, precision, F1 Score and AUC saw notable improvements. These results underline the attention mechanism's capability to efficiently focus on critical regions, enhancing the model's precision and reducing false positives, making the model more robust and effective for clinical applications.

Figs. 12 and 13 further illustrate this performance difference using confusion matrices. The Stacked-LSTM model without attention (Fig. 12) classified 4000 true negatives and 4067 true positives but misclassified 50 false positives and 150 false negatives. These errors indicate that while the model captures temporal dependencies well, it still lacks precise focus on the most relevant features.

In contrast, the Stacked-LSTM model with attention (Fig. 13) significantly improved classification, correctly identifying 4115 uninfected and 4078 infected samples while reducing false positives to 18 and false negatives to 55. This substantial reduction in misclassifications demonstrates the attention mechanism's ability to emphasize critical regions in the input sequence, thereby improving model robustness.

4.1.1. Statistical study

Decision Curve Analysis (DCA)

A useful technique for assessing the clinical utility of machine learning models is decision curve analysis, or DCA. In accordance with the clinical context, it balances true positive and false positive outcomes by quantifying the net benefit across a range of threshold probabilities. This section evaluates all deep learning models that have been presented and used to classify malaria from blood smear images in our study. The most commonly used evaluation parameters have been examined in a way that is easy for doctors to understand. They have also been calculated and used in a practical way to interpret and evaluate the results of the model [78] Comparing the clinical decision support potential of six models—VGG-16, VGG-19, Stacked-LSTM with Attention, Stacked-LSTM without Attention, Vision Transformer (ViT), and MobileNetV2—is the goal of the DCA-curve analysis displayed in Fig. 14. A threshold probability range of 0.01 to 0.99 is used to analyze the net benefit. The Stacked-LSTM with Attention (Green Curve) exhibits robustness and high clinical utility, maintaining outstanding performance at low, medium, and high thresholds while displaying the highest net benefit across all threshold probabilities. Sequential modeling and attention mechanisms enable this architecture to capture intricate spatial patterns and dependencies, making it perfect for high-stakes clinical applications where reducing false negatives is crucial. Although the VGG-19 (Orange Dashed Curve) performs consistently across thresholds and offers strong clinical support in mid-threshold scenarios, it has a slightly lower net benefit than the Stacked-LSTM with attention model. It also offers a useful trade-off between sensitivity and precision. According to standard metrics, VGG-16 (Blue Solid Curve) performs slightly better than VGG-19 (accuracy = 0.9802, AUC = 0.9765). Despite its high accuracy, its DCA curve indicates a somewhat lower net benefit, indicating that its usefulness in clinical decision-making is more threshold-sensitive. Although it performs well, a Stacked-LSTM without Attention (Red Dashed Curve) is not as good as its attention-enhanced counterpart. His DCA curve illustrates its limitations in cases that are difficult to categorize by displaying diminishing benefits at higher thresholds. With an AUC of 0.9503, the ViT model (Purple Dashed-Dotted Curve) exhibits a moderate net benefit across thresholds. It achieves good results in early thresholds (sensitive detection), but its benefit rapidly declines at thresholds above 0.6, suggesting sensitivity but reduced resilience, likely as a result of the small dataset size. Lastly, the lowest net benefit is shown by MobileNetV2 (Brown Dotted Curve), which is clearly above the 0.5 threshold. Its DCA curve indicates a higher false positive impact, despite having a respectable accuracy of 0.9303 and an F1 score of 0.9315. The models above the "Treat all" and "Treat none" reference lines have positive clinical utility, according to the clinical decision impact. The most promising option for incorporation into automated malaria diagnostic pipelines and the most suitable for critical applications, such as early diagnosis in endemic areas, is the Stacked-LSTM with attention model, which has been shown to outperform other models. ViT and MobileNetV2 should be used with caution due to their instability at mid-to-high thresholds, whereas VGG-based models continue to be reliable.

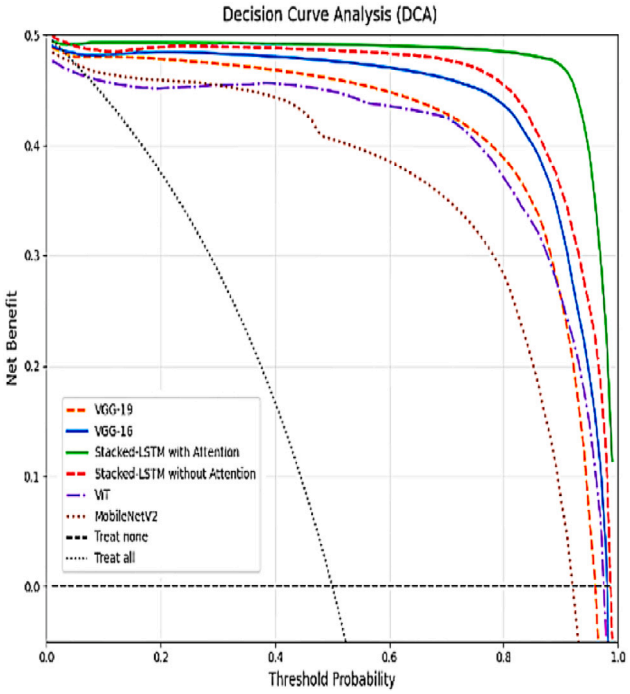


Fig. 14. Combined DCA curves. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

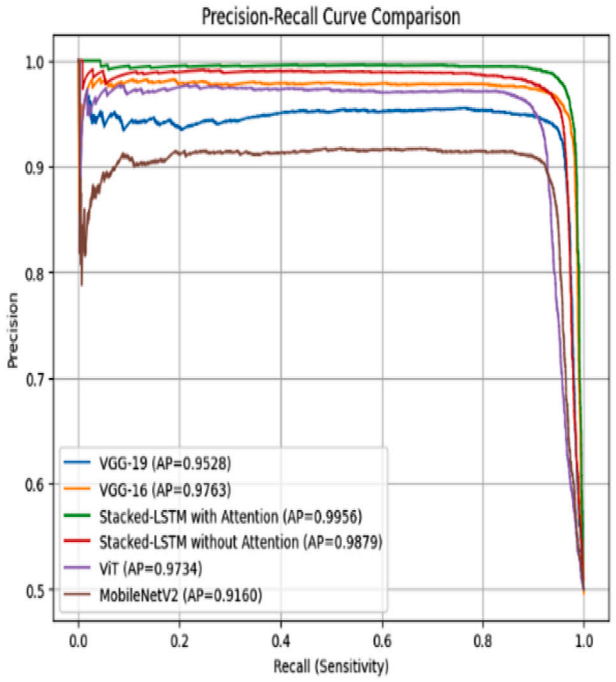


Fig. 15. Combined Precision-Recall curves.

Precision-Recall Curve

This analysis report analyses the Precision-Recall performance of our six deep learning models for the detection of malaria from blood smear images. As we can see in Fig. 15, the models are compared using Average Precision (AP) scores and their ability to maintain high precision across all recall levels (0.0 to 1.0).

The key findings of this analysis highlight the Stacked-LSTM with Attention model, which achieves an AP score of 0.9956 and shows that the attention mechanism dynamically focuses on infected cell regions, highly minimizing false negatives and making it clinically ideal for settings that require both high sensitivity (recall) and reliability (precision), maintaining near-perfect precision even at high recall, which makes it the best overall. On the other hand, the Stacked-LSTM model without attention, with AP = 0.9879, shows a good compromise. However, we note that the lack of fine attention leads to minor decreases in precision for recall values greater than 0.8. In fact, the absence of an attention mechanism may result in a marginal decrease in the model's ability to focus on key features, leading to a drop in precision at higher recall rates. CNN-based models, VGG-16 and VGG-19, are robust models, slightly outperformed by LSTM-based models, but still reliable. Paradoxically, VGG-16 outperforms VGG-19 (AP: 0.9763, 0.9528, respectively), suggesting that depth alone does not guarantee better performance for this task. The VGG-19 model (AP = 0.9528) shows a noticeable drop in precision at higher recall. In contrast, VGG-16 (AP = 0.9763) demonstrates a high and stable precision throughout most of the recall range. Clinically, the VGG-16 classifier has fewer FNs (FN = 98) than the VGG-19 model (FN = 198), which makes it more suitable for clinical utility. The Transformer model (AP = 0.9734) gives good results, but is computationally heavier than the LSTMs and does not offer any apparent advantages. It is also slightly less stable than the VGG-16 model. The drop in precision at high recall values suggests that more suitable pre-training or more data are needed to generalize well. MobileNetV2 (AP = 0.9160) is considered the least performant model. Although lightweight and suitable for embedded devices, it struggles to maintain high precision as the recall increases. As we can see in Fig. 15, the precision of this model collapses to 0.85 at recall = 0.6, making it unsuitable for clinical use. We consider this model unsuitable for high-stakes diagnostic tasks. Finally, we can affirm that the stacked-LSTM with attention mechanism emerges as the most reliable model for malaria detection tasks, combining high precision ($AP > 0.99$) and clinical interpretability. The significant performance gap between the top and bottom models (8% difference in AP) highlights the importance of model selection for this medical application. CNNs and ViT are viable alternatives, but require careful tuning to match their performance.

Statistical Comparison of Models based on McNemar's test

The Table 3 presents pairwise comparisons between our deep learning models (VGG-16, VGG-19, Stacked-LSTM with and without attention mechanism, ViT, and MobileNetV2) trained to classify malaria from blood smear images. The evaluation primarily focuses on false negatives (FN)—the most critical error type in clinical diagnostics—as they represent missed infections. The metrics derived from **McNemar's test, odds ratios (OR), and 95% Confidence Interval (CI)** are applied to confusion matrices and are intended to assess whether one model significantly outperforms another in reducing FN cases. These statistical tests are complementary. In fact, McNemar's test answers the question: **Is there a statistically significant difference?**. However, the odds ratio and confidence interval answer the question: **What is the magnitude and reliability of the difference?** Reporting all three (*p*-value, OR, and CI) gives a complete picture, i.e., significance, direction, and magnitude of the model comparison. The columns **Modèle 1** and **Modèle 2** represent the two models being compared. A **p-value** column shows the result of McNemar's test. It evaluates whether there is a significant difference in classification errors (specifically discordant pairs) between the two models; if the *p*-value is <**0.05**, that indicates a statistically significant difference. The **Odds Ratio OR (FN2/FN1)** column compares the FNs of Model 2 to Model 1. If OR <1, Model 2 has fewer FNs (better), but if OR >1, Model 2 has more FNs (worse). The **IC 95%** column demonstrates the **95% Confidence Interval** of the odds ratio. If the interval **does not include 1**, the difference is statistically significant. Finally, the **FN Reduction** column indicates the percent reduction in FN when using Model 2 instead of Model 1. If we obtain a positive value, that implies FN reduction (improvement); in contrast, if this value is negative, it suggests an increase in FN (deterioration).

Table 3
Summary of McNemar's test comparing pairs of models.

Model 1	Model 2	p-value	OR (FN2/FN1)	IC 95%	FN reduction
VGG-19	VGG-16	0.000000	0.50	[0.38, 0.68]	49.6%
MobileNetV2	VGG-16	0.000017	0.31	[0.23, 0.40]	69.3%
ViT	VGG-16	0.000000	0.22	[0.17, 0.29]	77.8%
Stacked-LSTM without Attention	VGG-16	0.000000	0.44	[0.33, 0.59]	55.6%
MobileNetV2	VGG-19	0.000006	0.61	[0.49, 0.76]	39.0%
ViT	VGG-19	0.000000	0.44	[0.36, 0.54]	56.0%
Stacked-LSTM without Attention	VGG-19	0.000000	0.88	[0.70, 1.11]	11.9%
VGG-16	Stacked-LSTM with Attention	0.000000	0.84	[0.59, 1.19]	16.4%
VGG-19	Stacked-LSTM with Attention	0.000000	0.42	[0.31, 0.58]	57.9%
MobileNetV2	Stacked-LSTM with Attention	0.000258	0.26	[0.19, 0.34]	74.3%
ViT	Stacked-LSTM with Attention	0.000553	0.19	[0.14, 0.25]	81.5%
Stacked-LSTM without Attention	Stacked-LSTM with Attention	0.000000	0.37	[0.27, 0.50]	62.9%

The results reveal significant differences in false negative (FN) rates, with critical implications for malaria diagnosis. The key findings of this analysis demonstrate the superiority of Stacked-LSTM with Attention model reducing false negatives (FN) by 57.9% compared to VGG-19 (OR = 0.42, 95% CI[0.31-0.58]), by 16.4% compared to VGG-16 (OR = 0.84), by 81.5% compared to ViT (OR = 0.19), and by 74.3% compared to MobileNetV2 (OR = 0.26). To highlight the impact of attention mechanisms, we have compared the Stacked-LSTM with attention with its version without attention. This comparison has shown an important FN reduction = 62.9% and an OR = 0.37.

At this stage, we can confirm the exceptional performance of a Stacked-LSTM with attention, especially as all comparisons show an OR < 1 and a *P*-value near 0, confirming its statistical superiority. From a clinical point of view, our stacked-LSTM with attention model presents significantly fewer cases of missed malaria, which is essential to prevent serious complications.

The CNN model comparison shows that the VGG-16 is notably better than VGG-19 and reduces FN by 49.6% (OR = 0.50, 95%, CI[0.38-0.68]). The VGG-16 and VGG-19 reduce FN by 77.8% (OR = 0.22) and 56% (OR = 0.44) in comparison with ViT, implying that these CNN models are clinically more efficient and confident.

These statistics consider MobileNetV2 the worst model because it increases the FN cases between 39% and more than 74% compared to other models.

4.2. Comparison with state-of-the-art

This study's performance results were compared with several state-of-the-art approaches for malaria detection from microscopic blood smear images, as summarized in Table 4. This comparative analysis demonstrates that the proposed Stacked-LSTM with attention mechanism achieved superior performance relative to other models evaluated on the same NIH-NLM dataset. Specifically, earlier CNN-based approaches by Rajaraman et al. [19], such as pre-trained VGG-16 and ResNet-50 architectures, achieved competitive accuracy rates of around 0.9510. Subsequent ensemble learning methods further improved this performance to an accuracy of approximately 0.9650 [56]. Customized CNN methods by Yang et al. [5] initially reached an accuracy of 0.8180, which was significantly improved by incorporating iterative global minimum screening (IGMS) techniques to 0.9346 [57], highlighting the potential of targeted feature selection methods. However, another CNN-based approach utilizing a cascading YOLO based on YOLOv2 by Yang et al. [58] showed a relatively lower accuracy of 0.7922, demonstrating challenges in single-pass detection models for small parasite detection.

Hybrid approaches that combine segmentation and classification tasks, such as Mask R-CNN variants (PlasmodiumVF-Net and RBCNet), also demonstrated promising results. Kassim et al. [6] achieved an accuracy of 0.90 with PlasmodiumVF-Net, while their RBCNet pipeline utilizing U-Net and Faster R-CNN improved accuracy to 0.9776 [60]. These results underscore the effectiveness of multi-stage approaches by explicitly isolating and analyzing regions of interest.

Table 4
Comparative performance table between our models and the fundamental models using the same dataset.

Author(s)	Model(s) used	Accuracy
Rajaraman et al. [19]	VGG-16, ResNet-50, Xception, Inception-V3, DenseNet-121, Simple CNN	0.9510
Rajaraman et al. [56]	Ensemble deep neural networks	0.9650
Yang et al. [5]	Android application and customized CNN	0.8180
Yang et al. [57]	Iterative Global Minimum Screening (IGMS) and Customized CNN	0.9346
Yang et al. [58]	Cascading YOLO based-on YOLOv2	0.7922
Yu et al. [79]	MalariaScreener: a smartphone-based system and customized CNN	0.9870
Kassim et al. [6]	PlasmodiumVF-Net Mask R-CNN and ResNet50	0.90
Kassim et al. [60]	RBCNet pipeline, U-Net and Faster R-CNN	0.9776
Koirala et al. [28]	YOLO-mp-3l	0.9020
	YOLO-mp-4l	0.9632
Current study	VGG-16	0.9802
	VGG-19	0.9601
	ViT	0.9509
	MobileNet-V2	0.9303
	Stacked-LSTM with attention mechanism	**0.9912**

Embedded and mobile-device-based solutions demonstrated notable potential as well. The MalariaScreener system achieved an impressive accuracy of 0.9870 [79], validating the feasibility of deploying malaria detection systems on mobile platforms. Similarly, lightweight YOLO models by Koirala et al. [28] achieved accuracies of 0.9020 (YOLO-mp-3l) and 0.9632 (YOLO-mp-4l), illustrating the trade-offs between model complexity and performance in real-time applications.

In comparison, our proposed two-stage approach, involving initial segmentation of regions of interest (ROIs) followed by classification using a Stacked-LSTM with an attention mechanism, achieved the highest accuracy of 0.9912, surpassing the performance of previous methods evaluated on the same dataset. This superior performance is attributed to the model's capability to effectively capture and emphasize relevant spatial and contextual features within blood smear images. The attention mechanism significantly contributed by selectively highlighting critical regions, effectively reducing false positives and negatives, and thus enhancing diagnostic reliability.

Additionally, our results show considerable improvements over standard CNN architectures like VGG-16 (0.9802) and VGG-19 (0.9601), further confirming the advantages of integrating recurrent neural networks with attention mechanisms for complex visual tasks.

4.3. Explainability and visualization results

In this section, we evaluated the explainability of the Stacked-LSTM classifier with an attention mechanism using two methods, Grad-CAM

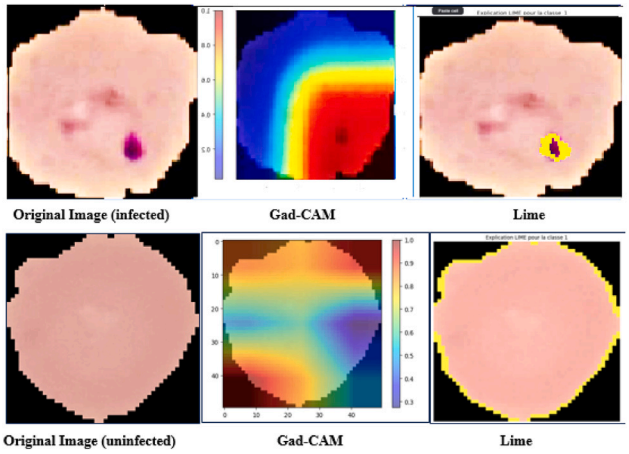


Fig. 16. Example for Grad-Cam and LIME result for infected and uninfected image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and LIME, applied to the test set. Fig. 16 below illustrates a representative example of the results obtained for the interpretability of an image of an infected and non-infected cell. LIME identified a specific region for the infected cell by highlighting a yellow small circle corresponding to the parasite. This demonstrates LIME's ability to isolate critical features influencing the model's prediction. Grad-CAM also highlighted this key region, marking it with intense red and yellow hues to signify its high importance in the model's decision. However, Grad-CAM also revealed activations in adjacent areas, which appear in blue and purple, indicating lesser relevance. Other parts of the cell, particularly on the left, were marked in red and yellow, reflecting significant activations, although less precise than those detected by LIME. For the non-infected cell, LIME primarily highlighted the cell's outline in yellow, suggesting that the model focused on the edges to make its decision. On the other hand, Grad-CAM displayed a broader distribution of activations: the areas at the top and bottom of the cell were marked in red and yellow, while the central regions, on the left and right, appeared in blue and purple. This broader activation pattern indicates that Grad-CAM captures more general areas of interest, while LIME provides a more targeted and specific interpretation. These results highlight the similarities and differences between the two explainability methods. Both techniques identify influential regions for predictions, but their approaches differ: LIME provides a localized and precise interpretation, making it well-suited for analyzing specific features, while Grad-CAM offers a global perspective by visualizing activations across the entire image. Both methods converge on the parasitic region for the infected cell, although Grad-CAM also highlights additional activations in the surrounding areas. For the non-infected cell, Grad-CAM reveals more diffuse activations, while LIME focuses explicitly on the contours of the cell. In conclusion, the combined use of Grad-CAM and LIME provides a complementary interpretation of the decisions made by the Stacked-LSTM with attention mechanism model. Grad-CAM provides an overview by visualizing the activations across the image, while LIME adds local granularity by identifying the most influential features. This complementarity enhances our understanding of the model's decision-making mechanisms and validates its robustness and reliability in the context of malaria detection.

5. Conclusion, limitations and perspectives

In this study, we evaluated standard CNN architectures, achieving notable accuracy with VGG-16 (0.9802) and VGG-19 (0.9601) for malaria detection from microscopic blood smear images. More significantly, we introduced a novel approach using a Stacked-LSTM model enhanced with an attention mechanism, which outperformed traditional CNNs, reaching an accuracy of 0.9912. The attention mechanism effectively increased performance by enabling the model to prioritize critical spatial regions within the image sequences.

To enhance the transparency of our proposed model and facilitate its clinical acceptance, we employed Explainable AI methods, particularly Grad-CAM and LIME, to visualize and interpret model decisions. Grad-CAM provided global visualizations of regions influencing model predictions, while LIME generated detailed and localized pixel-level explanations. This complementary use of XAI techniques significantly enhanced the interpretability of our model, allowing healthcare professionals to trust and validate AI-generated diagnoses with greater confidence.

While our model has demonstrated excellent performance on the benchmark dataset, there are several limitations to consider. First, the dataset is not from a multi-center clinical environment, which limits generalizability. Second, although attention mechanisms improve model interpretability, their integration into real-time clinical workflows may require lightweight alternatives and validating their robustness on diverse clinical datasets, and considering real-time deployment scenarios, particularly on mobile or edge devices. Future work will include validation across diverse clinical settings and the deployment of a mobile-compatible inference engine. Additionally, recent advances in ViT and LLM-based models represent promising directions for further improving performance and interpretability in medical image analysis tasks. Despite these limitations, the model presents a promising tool for early and interpretable malaria screening from blood smears.

CRediT authorship contribution statement

Adil Gaouar: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Souaad Hamza Cherif:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Abdellatif Rahmoun:** Writing – review & editing, Validation, Supervision, Resources, Project administration. **Mostafa El Habib Daho:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization.

Ethical statement

This study utilizes publicly available data from the NIH-NLM-Thin Blood Smears Pf dataset, which was collected following ethical guidelines. The dataset was obtained from malaria patients at Chittagong Medical College Hospital, Bangladesh, and was de-identified before public release. The Institutional Review Board (IRB) at the National Library of Medicine (NLM), National Institutes of Health (NIH), approved the study (IRB#12972). No additional ethical approval was required for our study, as no direct patient interaction or new data collection was involved.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Chow JCL. Quantum computing and machine learning in medical decision-making: A comprehensive review. Algorithms 2025;18:156. http://dx.doi.org/10.3390/a18030156.

[2] Pervez K, Sohail SI, Parwez F, Zia MA. Towards trustworthy AI-driven leukemia diagnosis: A hybrid hierarchical federated learning and explainable AI framework. Inform Med Unlocked 2025;2025:101618. http://dx.doi.org/10.1016/j.imu.2025.101618.

[3] Ortigossa ES, Gonçalves T, Nonato LG. Explainable artificial intelligence (XAI)—From theory to methods and applications. IEEE Access 2024;12:80799–846. http://dx.doi.org/10.1109/ACCESS.2024.3409843.

[4] Kong A, Wilson SA, Yong AH, Nace D, Rogier E, Aidoo M. HRP2 and HRP3 cross-reactivity and implications for HRP2-based RDT use in regions with Plasmodium falciparum hrp2 gene deletions. Malar J 2021;20:207. http://dx.doi.org/10.1186/s12936-021-03739-6.

[5] Yang H, Zhang Y, Zhang Y. Automated malaria detection using YOLO for P. vivax parasite identification in thin blood smear images. IEEE Access 2020;8:195104–14. http://dx.doi.org/10.1109/ACCESS.2020.3034866.

[6] Kassim YM, Palaniappan K, Yang F, Poostchi M, Palaniappan N, Maude RJ, et al. [Dataset] clustering-based dual deep learning architecture for detecting red blood cells in malaria diagnostic smears. IEEE J Biomed Heal Inf 2021;25:1735–46. http://dx.doi.org/10.1109/JBHI.2020.3034863.

[7] Li Y, El Habib Daho M, Conze PH, Zeghlache R, Le Boité H, Tadayoni R, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Comput Biol Med 2024;177:108635. http://dx.doi.org/10.1016/j.compbiomed.2024.108635.

[8] Eisemann N, Bunk S, Mukama T, Baltus H, Elsner SA, Gomille T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med 2025. http://dx.doi.org/10.1038/s41591-024-03408-6.

[9] El Habib Daho M, Li Y, Zeghlache R, Le Boité H, Deman P, Borderie L, et al. Discover: 2-D multiview summarization of optical coherence tomography angiography for automatic diabetic retinopathy diagnosis. Artif Intell Med 2024;149:102803. http://dx.doi.org/10.1016/j.artmed.2024.102803.

[10] Lilda SD, Jayaparvathy R. Enhancing cardiovascular disease classification in ECG spectrograms by using multi-branch CNN. Comput Biol Med 2025;186:109737. http://dx.doi.org/10.1016/j.compbiomed.2025.109737.

[11] Amin J, Anjum MA, Ahmad A, Sharif MI, Kadry S, Kim J. Microscopic parasite malaria classification using best feature selection based on generalized normal distribution optimization. PeerJ Comput Sci 2024;10:e1744. http://dx.doi.org/10.7717/peerj-cs.1744.

[12] Islam MR, Nahiduzzaman M, Goni MOF, Sayeed A, Anower MS, Ahsan M, et al. Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images. Sens 2022;22:4358. http://dx.doi.org/10.3390/s22124358.

[13] Khan A, Gupta KD, Venugopal D, Kumar N. Cidmp: Completely interpretable detection of malaria parasite in red blood cells using lower-dimensional feature space. In: Int jt conf neural netw. IJCNN, Glasgow, UK; 2020, p. 1–8. http://dx.doi.org/10.1109/IJCNN48605.2020.9284834.

[14] Alanazi AF, Alaerjan MA. Bi-LSTM network for classification of medical images: Application in malaria diagnosis. Comput Mater Contin 2023;73:1019–35. http://dx.doi.org/10.32604/cmc.2023.022462.

[15] Qadri AM, Raza A, Eid F, Abualigah L. A novel transfer learning-based model for diagnosing malaria from parasitized and uninfected red blood cell images. Decis Anal J 2023;9:100352. http://dx.doi.org/10.1016/j.dajour.2023.100352.

[16] Kim D, Lee Y, Chin K, Mago PJ, Cho H, Zhang J. Implementation of a long short-term memory transfer learning (LSTM-TL)-based data-driven model for building energy demand forecasting. Sustain 2023;15:2340. http://dx.doi.org/10.3390/su15032340.

[17] Chow JCL. Nanomaterial-based molecular imaging in cancer: Advances in simulation and AI integration. Biomol 2025;15:444. http://dx.doi.org/10.3390/biom15030444.

[18] Minarno AE, Izzah TN, Munarko Y, Basuki S. Classification of malaria using convolutional neural network method on microscopic image of blood smear. JOIV Int J Inf Vis 2024;8:1469. http://dx.doi.org/10.62527/joiv.8.3.2154.

[19] Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ 2018;6:e4568. http://dx.doi.org/10.7717/peerj.4568.

[20] Shang W, Sohn K, Almeida D, Lee H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In: Proc 33rd int conf mach learn. ICML 2016, vol. 48, Int Soc Mach Learn (ISML); 2016, p. 2217–25. http://dx.doi.org/10.48550/arXiv.1603.05201.

[21] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58. http://dx.doi.org/10.5555/2627435.2670313.

[22] LeCun Y, Bengio Y, Hinton G. Deep learning. Nat 2015;521:436–44. http://dx.doi.org/10.1038/nature14539.

[23] Botev A, Lever G, Barber D. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In: Proc int jt conf neural netw. IJCNN 2017, 2017, p. 189–903. http://dx.doi.org/10.1109/IJCNN.2017.7966082.

[24] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305. http://dx.doi.org/10.5555/2188385.2188395.

[25] Grignaffini F, Simeoni P, Alisi A, Frezza F. Computer-aided diagnosis systems for automatic malaria parasite detection and classification: A systematic review. Electron 2024;13(16):3174. http://dx.doi.org/10.3390/electronics13163174.

[26] Wisit L, Sako LU. Image classification of malaria using hybrid algorithms: convolutional neural network and method to find appropriate K for K-nearest neighbor. Indones J Electr Eng Comput Sci 2019;16:382–8. http://dx.doi.org/10.11591/ijeecs.v16.i1.pp382-388.

[27] Zamil YK, Ali SA, Naser MA. Spam image email filtering using K-NN and SVM. Int J Electr Comput Eng 2019;9:245–54. http://dx.doi.org/10.11591/ijece.v9i1.pp245-254.

[28] Koirala A, Jha M, Bodapati S, Mishra A, Chetty G, Sahu PK. Deep learning for real-time malaria parasite detection and counting using YOLO-mp. IEEE Access 2022;1. http://dx.doi.org/10.1109/ACCESS.2022.3208270.

[29] Pereira-Ferrero VH, Valem LP, Pedronette DCG. Feature augmentation based on manifold ranking and LSTM for image classification. Expert Syst Appl 2023;213:118995. http://dx.doi.org/10.1016/j.eswa.2022.118995.

[30] Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attention in vision models. Adv Neural Inf Process Syst 2019;32. http://dx.doi.org/10.48550/arXiv.1906.05909.

[31] Bello I, Zoph B, Vaswani A, Shlens J, Le QV. Attention augmented convolutional networks. In: Proc IEEE/CVF int conf comput vis. 2019, p. 3286–95. http://dx.doi.org/10.48550/arXiv.1904.09925.

[32] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30. http://dx.doi.org/10.48550/arXiv.1706.03762.

[33] Al-Shabi M, Shak K, Tan M. Procan: Progressive growing channel attentive non-local network for lung nodule classification. Pattern Recognit 2022;122:108309. http://dx.doi.org/10.1016/j.patcog.2021.108309.

[34] Sang DV, Chung TQ, Lan PN, Hang DV, Long D, Thuy NT. Ag-curesnest: A novel method for colon polyp segmentation. 2021, http://dx.doi.org/10.48550/arXiv.2105.00402, arXiv preprint arXiv:2105.00402.

[35] Yao C, Tang J, Hu M, Wu Y, Guo W, Li Q, et al. Claw u-net: A unet variant network with deep feature concatenation for scleral blood vessel segmentation. In: CAAI int conf artif intell. 2021, p. 67–78. http://dx.doi.org/10.1007/978-3-030-93049-3_6.

[36] Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In: Eur conf comput vis. 2020, p. 251–66. http://dx.doi.org/10.1007/978-3-030-66415-2_16.

[37] Bozorgpour A, Azad R, Showkatian E, Sulaiman A. Multi-scale regional attention deeplab3+: Multiple myeloma plasma cells segmentation in microscopic images. In: MICCAI workshop comput pathol. 2021, p. 47–56. http://dx.doi.org/10.48550/arXiv.2105.06238.

[38] Gonçalves T, Rio-Torto I, Teixeira LF, Cardoso JS. A survey on attention mechanisms for medical applications: are we moving towards better algorithms? IEEE Access 2022. http://dx.doi.org/10.1109/ACCESS.2022.3206449.

[39] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv preprint arXiv:2010.11929.

[40] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. 2020, http://dx.doi.org/10.48550/arXiv.2010.04159, arXiv preprint arXiv:2010.04159.

[41] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. 2021, http://dx.doi.org/10.48550/arXiv.2102.04306, arXiv preprint arXiv:2102.04306.

[42] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A video vision transformer. In: Proc IEEE/CVF int conf comput vis. 2021, p. 6836–46. http://dx.doi.org/10.1109/ICCV48922.2021.00676.

[43] Chen X, Wang X, Zhou J, Qiao Y, Dong C. Activating more pixels in image super-resolution transformer. In: 2023 IEEE/CVF conf comput vis pattern recognit. CVPR, IEEE; 2023, p. 22367–77. http://dx.doi.org/10.1109/cvpr52729.2023.02142.

[44] Qian S, Zhu Y, Li W, Li M, Jia J. What makes for good tokenizers in vision transformer? IEEE Trans Pattern Anal Mach Intell 2022;1–13. http://dx.doi.org/10.1109/TPAMI.2022.3231442.

[45] Ma J, Bai Y, Zhong B, Zhang W, Yao T, Mei T. Visualizing and understanding patch interactions in vision transformer. IEEE Trans Neural Netw Learn Syst 2023;1–10. http://dx.doi.org/10.1109/TNNLS.2023.3270479.

[46] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2023;45:87–110. http://dx.doi.org/10.1109/TPAMI.2022.3152247.

[47] Al-Hammuri K, Gebali F, Kanan A, Chelvan IT. Vision transformer architecture and applications in digital health: a tutorial and survey. Vis Comput Ind Biomed Art 2023;6:14. http://dx.doi.org/10.1186/s42492-023-00140-9.

[48] Aladhadh S, Alsanea M, Aloraini M, Khan T, Habib S, Islam M. An effective skin cancer classification mechanism via medical vision transformer. Sens 2022;22:4008. http://dx.doi.org/10.3390/s22114008.

[49] Yang G, Luo S, Greer P. A novel vision transformer model for skin cancer classification. Neural Process Lett 2023. http://dx.doi.org/10.1007/s11063-023-11204-5.

[50] Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X. An improved transformer network for skin cancer classification. Comput Biol Med 2022;149:105939. http://dx.doi.org/10.1016/j.compbiomed.2022.105939.

[51] Rajab S, Nakatumba-Nabende J, Ggaliwango M. Interpretable machine learning models for predicting malaria. In: Proc ICSTSN. 2023, p. 1–6. http://dx.doi.org/10.1109/ICSTSN57873.2023.10151538.

[52] Goni MOF, Mondal MNI, Islam SR, Nahiduzzaman M, Islam MR, Anower MS. Diagnosis of malaria using double hidden layer extreme learning machine algorithm with CNN feature extraction and parasite inflator. IEEE Access 2023;11:4117–30. http://dx.doi.org/10.1109/ACCESS.2023.3234279.

[53] Attai K, Ekpenyong M, Amannah C, Asuquo D, Ajuga P, Obot O. Enhancing the interpretability of malaria and typhoid diagnosis with explainable AI and large language models. Trop Med Infect Dis 2024;9:216. http://dx.doi.org/10.3390/tropicalmed9090216.

[54] Chow JCL, Li K. Ethical considerations in human-centered AI: Advancing oncology chatbots through large language models. JMIR Bioinform Biotechnol 2024;5:e64406. http://dx.doi.org/10.2196/64406.

[55] Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: Current trends, challenges, and future directions in large language model-enabled medical chatbots. BioMedInformatics 2024;4:837–52. http://dx.doi.org/10.3390/biomedinformatics4010047.

[56] Rajaraman S, Jaeger S, Antani SK. Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ 2019;7:e6977. http://dx.doi.org/10.7717/peerj.6977.

[57] Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J. Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE J Biomed Heal Inf 2020. http://dx.doi.org/10.1109/JBHI.2019.2939121.

[58] Yang F, Quizon N, Yu H, Silamut K, Maude RJ, Jaeger S. Cascading YOLO: Automated malaria parasite detection for plasmodium vivax in thin blood smears. In: Proc SPIE 11314, med imaging 2020: Comput-aided diagn. vol. 11314, 2020, 113141Q. http://dx.doi.org/10.1117/12.2550143.

[59] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 2019;31:1235–70. http://dx.doi.org/10.1162/neco_a_01199.

[60] Kassim YM, Yang F, Yu H. Diagnosing malaria patients with plasmodium falciparum and vivax using deep learning for thick smear images. Diagn 2021;11:1994. http://dx.doi.org/10.3390/diagnostics11111994.

[61] Garrido-Cardenas JA, González-Cerón L, Manzano-Agugliaro F, Mesa-Valle C. Plasmodium genomics: an approach for learning about and ending human malaria. Parasitol Res 2019;118:535–48. http://dx.doi.org/10.1007/s00436-018-6127-9.

[62] Baer K, Klotz C, Kappe SHI, Schnieder T, Frevert U. Release of hepatic plasmodium yoelii merozoites into the pulmonary microvasculature. PLoS Pathog 2007;3:e171. http://dx.doi.org/10.1371/journal.ppat.0030171.

[63] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[64] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017, http://dx.doi.org/10.48550/arXiv.1702.08608, arXiv preprint arXiv:1702.08608.

[65] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc IEEE int conf comput vis. 2017, p. 618–26. http://dx.doi.org/10.1109/ICCV.2017.74.

[66] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proc 22nd ACM SIGKDD int conf knowl discovery data mining. 2016, p. 1135–44. http://dx.doi.org/10.1145/2939672.2939778.

[67] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv preprint arXiv:1409.1556.

[68] Zou Y, Wu L, Zuo C, Chen L, Zhou B, Zhang H. White blood cell classification network using MobileNetv2 with multiscale feature extraction module and attention mechanism. Biomed Signal Process Control 2025;99:106820. http://dx.doi.org/10.1016/j.bspc.2024.106820.

[69] Iscen A, Avrithis Y, Tolias G, Furon T, Chum O. Fast spectral ranking for similarity search. In: Proc IEEE comput soc conf comput vis. 2018, p. 7632–41. http://dx.doi.org/10.1109/CVPR.2018.00796.

[70] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[71] Lalapura VS, Amudha J, Satheesh HS. Recurrent neural networks for edge intelligence: A survey. ACM Comput Surv 2021;54:1–38. http://dx.doi.org/10.1145/3448974.

[72] Mienye ID, Sun Y. A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access 2023;11:30628–38. http://dx.doi.org/10.1109/ACCESS.2023.3262020.

[73] Mienye ID, Sun Y. A machine learning method with hybrid feature selection for improved credit card fraud detection. Appl Sci 2023;13:7254. http://dx.doi.org/10.3390/app13127254.

[74] Kiperwasser E, Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Trans Assoc Comput Linguist 2016;4:313–27. http://dx.doi.org/10.1162/tacl_a_00101.

[75] Long F, Zhou K, Ou W. Sentiment analysis of text based on bidirectional LSTM with multi-head attention. IEEE Access 2019;7:141960–9. http://dx.doi.org/10.1109/ACCESS.2019.2942614.

[76] Lanjewar MG, Panchbhai KG, Patle LB. Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images. Comput Biol Med 2024;169:107914. http://dx.doi.org/10.1016/j.compbiomed.2023.107914.

[77] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014, http://dx.doi.org/10.48550/arXiv.1409.0473, ArXiv preprint.

[78] Abbasian Ardakani A, Airom O, Khorshidi H, Bureau NJ, Salvi M, Molinari F. Interpretation of artificial intelligence models in healthcare: A pictorial guide for clinicians. J Ultrasound Med 2024;43:1789–818. http://dx.doi.org/10.1002/jum.16524.

[79] Yu H, Yang F, Rajaraman S, Ersoy I, Moallem G, Poostchi M. Malaria screener: a smartphone application for automated malaria screening. BMC Infect Dis 2020;20:825. http://dx.doi.org/10.1186/s12879-020-05453-1.