الجـــــــــــــــــــمـهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الـشـــــــــعبيـة

**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

وزارة التـــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

جـــــــــــــامعة أبي بكـــر بـلـقـايد – تـلمســـــــــــــان –

Université Aboubakr Belkaïd – Tlemcen –

Faculté de TECHNOLOGIE

# THESE

Présentée pour l'obtention du **grade** de **DOCTORAT 3ème Cycle**

**En** : Automatique

**Spécialité** : Automatique

**Par** : MELLOUK Wafa

**Sujet**

## Reconnaissance multimodale de l'affect par apprentissage profond

Soutenue publiquement, le 23/ 06 /2024, devant le jury composé de :

| | | | |
|---|---|---|---|
| Mr. BENYAHIA Boumediene | Professeur | Université de Tlemcen | Président |
| Mlle. HANDOUZI Wahida | MCA | Université de Tlemcen | Directrice de thèse |
| Mr. BEREKSI REGUIG Fethi | Professeur | Université de Tlemcen | Examinateur 1 |
| Mr. SAIDI-SIEF Ali | MCA | Université de Skikda | Examinateur 2 |
| Mme. GHOUALI Amel | MCA | École Supérieure en Sciences Appliquées de Tlemcen | Examinatrice 3 |

*Décoder les émotions des autres demande de la sensibilité, car 90% d'un message émotionnel est non verbal.*

Daniel Goleman

# Dédicaces

Je dédie cette thèse, réalisée avec beaucoup de dévouement :

— À mon cher père, qui nous a quittés trop tôt. Chaque étape de ce voyage académique est marquée par ton amour dans mon cœur.

— À ma mère, ma source constante de soutien, qui m'a toujours orienté vers le droit chemin. C'est grâce à son amour, son soutien, son courage et ses sacrifices que j'ai atteints ce niveau d'études. Cette thèse est dédiée à toi, en reconnaissance éternelle de tout ce que tu as fait pour moi.

# Remerciements

Tout d'abord, je tiens à exprimer ma profonde remerciements envers ma directrice de thèse, Mlle Handouzi Wahida, Maitre de conférences A à l'Université de Tlemcen pour ses précieux conseils, son expertise et ses encouragements constants tout au long de mes recherches. Son encadrement attentif et sa motivation m'ont donné la confiance nécessaire pour atteindre tous les objectifs et assurer la réussite de ce projet.

Un sincère remerciement à Monsieur Hadj Abdelkader Mohammed Amine, Professeur et Directeur du laboratoire d'Automatique de Tlemcen LAT, pour son précieux soutien inestimable tout au long de ce travail de recherche.

Je tiens à exprimer ma gratitude aux membres du jury pour avoir généreusement consacré leur temps à examiner et évaluer ce travail : M. Bereksi Raguig Fethi, Professeur à l'Université de Tlemcen ; M. Saidi Seif Ali, Maître de Conférences A à l'Université de Skikda ; Et Mme Ghouali Amel, Maître de Conférences A à l'École Supérieure des Sciences Appliquées de Tlemcen, pour avoir accepté d'être les examinateurs de cette thèse.

Je remercie vivement Monsieur Benyahia Boumediène, Professeur d'Université de Tlemcen, pour avoir accepté de présider cette thèse.

Je remercie également mes collègues du Laboratoire de Recherches Automatique de Tlemcen LAT. De plus, je tiens à exprimer mes remerciements envers mes chers amis et collègues du Laboratoire d'Ingénierie Manufacturière de Tlemcen MELT, notamment Djazia Nadjat Sekkal, Besma Zeddam et Nacera Tahraoui, pour leur soutien et les moments privilégiés que nous avons partagés, qui ont contribué pour faciliter le stress du travail.

Enfin, je remercie tous ceux qui ont contribué directement ou indirectement à ce travail. Merci infiniment.

# Résumé

La reconnaissance automatique de l'affect est un domaine de recherche crucial visant à améliorer l'intelligence artificielle afin qu'elle puisse identifier de manière précise et automatique les états affectifs des humains. La complexité de ce domaine réside dans la diversité des expressions affectives, qui se manifestent à travers différents canaux, notamment les modalités physiques et physiologiques. Des études récentes indiquent que l'approche consistant à fusionner différentes modalités permet d'obtenir des résultats plus fiables et de mener à une analyse plus approfondie et complète de ces états affectifs.

L'objectif principal de cette thèse est le développement de méthodes de reconnaissance automatique des émotions et du stress en exploitant deux modalités distinctes : les expressions faciales et le signal iPPG (Photopléthysmographie par imagerie). Les expressions faciales, modalité non verbale aisément acquise, offrent une représentation externe de l'état affectif des individus. D'autre part, le signal iPPG est utilisé comme mesure physiologique qui reflète les changements du rythme cardiaque avec l'état affectif. La fusion de ces deux modalités permet d'intégrer différentes caractéristiques propres à chaque modalité, ce qui représente l'avantage majeur de l'approche multimodale.

Nos recherches se concentrent sur trois axes. En premier, nous avons proposé une nouvelle approche d'étude basée sur la classification des émotions humaines selon deux échelles, valence et arousal, en utilisant des signaux iPPG extraits de vidéos faciales. La mise en œuvre de cette méthode implique plusieurs étapes, telles que la collecte précise des signaux iPPG, leur prétraitement, et enfin la classification. En ce qui concerne la classification, nous avons proposé une architecture d'apprentissage profond combinant un réseau neuronal convolutif unidimensionnel 1D-CNN et un réseau de neurones mémoire à long terme LSTM.

Le deuxième axe est concentré sur la reconnaissance automatique des émotions à partir des expressions faciales. Notre objectif était d'obtenir une classification précise des sept émotions de base, en tenant compte des différentes positions de tête, des regards variés, de l'âge et du sexe. Cette méthode est basée sur deux étapes importantes : le prétraitement des images, qui vise à conserver et clarifier les caractéristiques pertinentes de nos images, et classification par proposition d'une architecture d'apprentissage profond 2D-CNN.

Le troisième axe de cette thèse concerne la conception d'un système multimodal de re-

connaissance automatique du stress, s'appuyant sur les expressions faciales et les signaux iPPG. Une architecture 3D-CNN est proposée pour la classification en utilisant les données des expressions faciales, tandis qu'une architecture 1D-CNN est utilisée avec les signaux iPPG. Après l'extraction des caractéristiques de chaque modalité, une fusion de ces caractéristiques est appliquée, suivie de l'utilisation de couches entièrement connectées du réseau neuronal pour la classification des états de stress ou de non-stress.

Les résultats que nous avons obtenus démontrent la puissance et l'efficacité des méthodes que nous proposons. Nous avons atteint une précision de classification de 73,33 % pour la valence à l'aide des signaux iPPG et de 96,55 % pour les expressions faciales dans différentes poses de la tête. Lesquelles surpassent celles des autres approches récemment proposées par différents chercheurs. De plus, nous avons démontré l'efficacité de la performance de l'approche multimodale par rapport à l'approche unimodale, atteignant une précision de validation de 100%.

## Mots-clés

Emotion, Stress, Expressions faciales, Photopléthysmographie par imagerie iPPG, Vidéo faciale, Variabilité cardiaque, Apprentissage profond, Unimodale, Multimodale

# Abstract

Automatic recognition of affect is a crucial area of research aimed at improving artificial intelligence so that it can accurately and automatically identify human affective states. The complexity of this field lies in the diversity of affective expressions, which manifest themselves through different channels, including physical and physiological modalities. Recent studies indicate that the approach of merging different modalities provides more reliable results and leads to a more thorough and complete analysis of these affective states.

The main objective of this thesis is to develop methods for the automatic recognition of emotions and stress by exploiting two distinct modalities : facial expressions and the iPPG (Imaging Photoplethysmography) signal. Facial expressions, an easily acquired non-verbal modality, provide an external representation of an individual's emotional state. On the other hand, the iPPG signal is used as a physiological measure that reflects changes in heart rate with affective state. Merging these two modalities allows us to integrate different characteristics specific to each modality, which is the major advantage of the multimodal approach.

Our research focuses on three areas. Firstly, we have proposed a new study approach based on the classification of human emotions according to two scales, valence and arousal, using iPPG signals extracted from facial videos. The implementation of this method involves several steps, such as the accurate collection of iPPG signals, their pre-processing and, finally, classification. For classification, we have proposed a deep learning architecture combining a one-dimensional convolutional neural network 1DCNN and a long-term memory neural network LSTM.

The second area focuses on the automatic recognition of emotions based on facial expressions. Our aim was to obtain an accurate classification of the seven basic emotions, taking into account different head positions, different gazes, age and gender. This method is based on two important steps : image pre-processing, which aims to preserve and clarify the relevant features of our images, and classification by proposing a 2D-CNN deep learning architecture.

The third aspect of this thesis concerns the design of a multimodal system for automatic stress recognition, based on facial expressions and iPPG signals. A 3D-CNN architecture is proposed for classification using facial expression data, while a 1D-CNN architecture is used with iPPG signals. After extracting the features of each modality, a fusion of these features is

applied, followed by the use of fully connected layers of the neural network for the classification of stress or non-stress states.

The results we obtained demonstrate the power and effectiveness of the methods we propose. We achieved a classification accuracy of 73.33% for valence using iPPG signals and 96.55% for facial expressions in different head poses. These outperform other approaches recently proposed by various researchers. In addition, we demonstrated the efficiency of the performance of the multimodal approach compared with the unimodal approach, achieving a validation accuracy of 100%.

## keywords

Emotion, Stress, Facial expressions, Facial video, Imaging Photoplethysmography iPPG, Cardiac variability, Deep learning, unimodal, multimodal

# Liste des Publications

## Revues internationales

- W. Mellouk, W. Handouzi, « CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals », Biomed. Signal Process. Control, vol. 85, p. 104907, août 2023, doi : 10.1016/j.bspc.2023.104907.
- W. Mellouk, W. Handouzi, « Facial emotion recognition using deep learning : review and insights », Procedia Comput. Sci., vol. 175, p. 689-694, janv. 2020, doi : 10.1016/j.procs.2020.07.101.

## Communications internationales

- W. Mellouk, W. Handouzi, « Multimodal contactless human stress detection using deep learning », in 2024 The 6th Conference on Computing Systems and Applications (CSA), April 2024 (Accepté et présenté)
- W. Mellouk, W. Handouzi, « Comparison and evaluation of IPPG methods for HR estimation under different face regions », in 2022 19th International Multi-Conference on Systems, Signals & Devices (SSD), mai 2022, p. 1956-1961. doi : 10.1109/SSD54932.2022.9955726.
- W. Mellouk, W. Handouzi, « Convolutional Neural Network for Identifying Human Emotions with Different Head Poses », in Innovations in Smart Cities Applications. Volume 4,M. Ben Ahmed, İ. Rakıp Karaș, D. Santos, O. Sergeyeva, et A. A. Boudhir, Éd, in Networks and Systems, Cham, Springer International Publishing, 2021, pp. 785-796. doi :$10.1007/978-3-030-66840-2-59$.
- W. Mellouk, W. Handouzi, "A proposal for a multimodal emotion recognition system using deep learning", International conference on advanced intelligent system for sustainable development AI2SD'2019, July 2019

## Communication nationale

– W. Mellouk, W. Handouzi« Facial emotion detection, using convolutional neural networks : review and insights». Deep learning INDABAX Algeria, April 2019.

# Table des matières

# Table des figures

# Liste des tableaux

# Liste des abréviations

**AC** Précision (de l'anglais Accuracy)

**CNN** Réseau de neurones convolutifs (de l'anglais Convolutional Neural Network)

**1D-CNN** Réseau de neurones convolutifs à 1 dimension (de l'anglais Convolutional Neural Network 1D)

**3D-CNN** Réseau de neurones convolutifs à 3 dimension (de l'anglais Convolutional Neural Network 3D)

**DL** Apprentissage profond (de l'anglais Deep Learning)

**DNN** Réseau de neurone profond (de l'anglais Deep Neural Network)

**DSP** Densité spectrale de puissance

**EF** Expressions faciales

**EDA** Activité électrodermique

**ECG** Electrocardiogramme

**EMG** Electromyogramme

**EEG** Electroencéphalogramme

**FC** Fréquence Cardiaque

**F1** Score F1

**IA** Intelligence Artificielle

**ICA** Analyse en composantes indépendantes (Independent Component Analysis)

**IPPG** Photopléthysmographie par imagerie

**IVFC** Variabilité de la fréquence cardiaque par imagerie

**GSR** Réponse galvanique cutanée

**LSTM** Réseau de neurones à mémoire à long terme (de l'anglais Long Short-Term Memory)

**ML** Apprentissage automatique (de l'anglais Machine Learning)

**PPG** Photopléthysmogramme

**POS** Plan orthogonal à la peau

**RSP** Respiration

**PR** Précision (de l'anglais Precision)

**ROI** Région d'intérêt (de l'anglais Region of Interest)

**RNN** Réseau de neurones récurrent (de l'anglais Reccurent Neural Network)

**RE** Sensibilité (de l'anglais Recall)

**SKT** Température Cutanée

**SN** Système nerveux

**VFC** Variabilité de la fréquence cardiaque

Chapitre 1

# Introduction

La reconnaissance automatique de l'affect est le processus d'intégration des états affectifs d'une personne, par exemple (émotions ou stress), dans des machines et des appareils par l'intelligence artificielle. Ce domaine de recherche vise à développer des dispositifs permettant de détecter et d'interpréter différents états affectifs chez l'individu, ce qui en fait un domaine de recherche interdisciplinaire lié à la psychologie, au traitement du signal, au traitement d'image et à l'apprentissage automatique.

À cet égard, il a été révélé que lorsque les machines connaissent l'état affectif des individus, elles peuvent leur fournir des services plus pratiques dans plusieurs domaines tels que :

- Marketing : réflexion sur la satisfaction des clients en vue d'améliorer les services marketing ainsi que les produits [1].
- Santé : amélioration du diagnostic des enfants autistes pour développer des méthodes de traitement, ainsi que détection des maladies psychologiques telles que la dépression et le stress, pouvant réduire l'immunité du corps et provoquer de nombreuses maladies [2], [3].
- Éducation : détection du niveau d'apprentissage [4].
- Robotique : amélioration de l'interaction et de la communication entre l'humain et les robots, les voitures et les interfaces intelligentes [5].

En général, les états affectifs des individus se manifestent par des modifications physiques telles que les expressions faciales, les postures et la tonalité de la voix, ainsi que par des changements physiologiques, comme l'électrocardiogramme (ECG), le photopléthysmogramme (PPG) et d'autres signaux physiologiques.

La modalité physiologique a suscité un grand intérêt dans la recherche, car de nombreuses études ont démontré que les changements émotionnels entraînent des changements physiologiques chez les individus [6]. De plus, les données physiologiques sont considérées comme plus fiables que d'autres modalités physiques, qui peuvent être dissimulées ou contrôlées par les individus [7].

Bien que la reconnaissance automatique de l'affect ait été étudiée à travers plusieurs modalités, la modalité des expressions faciales est la plus largement utilisée [8]. En effet, lors d'une expérience émotionnelle, les premiers signes apparaissent sur le visage et sont visibles, ce qui permet de les extraire facilement comme paramètre à étudier [8]. De nombreux travaux dans

la littérature ont obtenu des résultats impressionnants en se concentrant sur les expressions faciales et ont abordé divers aspects tels que les différences d'âge, de sexe, de race, l'occlusion partielle du visage, les variations d'éclairage et les différentes positions de la tête devant les caméras [9].

Actuellement, avec tous les progrès réalisés dans la reconnaissance automatique de l'affect, les chercheurs se tournent vers l'étude des systèmes multimodaux afin d'améliorer la performance de classification. Ces systèmes reposent sur la fusion de deux ou plusieurs modalités, permettant ainsi d'intégrer des données de natures différentes et améliorant la fiabilité de la classification de l'affect par rapport aux systèmes unimodaux existants [10].

## 1.1 Motivations et objectifs

Les émotions et le stress puissent se manifester de divers canaux, englobant des manifestations externes que l'individu peut maîtriser comme les expressions faciales et des changements internes échappant à son contrôle comme les signaux physiologiques, les recherches récentes privilégient une approche multimodale, incluant, audiovisuels [11] et physio-visuels [12], dans le but d'améliorer la fiabilité de la classification des émotions et du stress.

Ces dernières années, les progrès de la recherche dans la reconnaissance automatique des émotions et du stress se sont de plus en plus intéressés vers la modalité physiologique. Les signaux physiologiques sont mesurés à l'aide de capteurs attachés à une partie spécifique du corps humain correspondant à un signal particulier [7]. Cependant, cette méthode d'acquisition de données présente une limitation, car le placement de capteurs dans le corps humain peut susciter la confusion, incommoder l'utilisateur et entraîner des données moins précises.

Récemment, les chercheurs ont démontré que la fréquence cardiaque (FC) peut être estimée à partir de vidéos de visages humains capturés avec une simple caméra [13]. Cette technique se base sur l'extraction du signal photopléthysmographique par imagerie (iPPG) en mesurant les variations de couleur rouge, verte et bleue (RVB) extraites du visage d'un être humain.
L'importance de ce domaine de recherche réside dans la capacité à estimer la FC sans perturber les individus avec des capteurs [14]. De plus, la présence généralisée de caméras intégrées dans les smartphones et les ordinateurs portables offre une opportunité pratique pour l'application

de cette technique.

Avec les progrès dans le domaine de l'extraction des signaux photopléthysmographiques par imagerie (iPPG) et de la fréquence cardiaque par imagerie (iFC), ainsi que l'importance croissante de l'utilisation des signaux physiologiques dans la reconnaissance automatique des émotions et du stress, notre recherche s'est concentrée sur l'utilisation de ces deux aspects pour parvenir à une classification précise des émotions et du stress à partir de vidéos faciales, ce qui constitue le principal objectif de cette thèse.

D'autre part, les expressions faciales sont l'une des modalités les plus étudiées par les chercheurs [8]. De nombreux obstacles ont été surmontés, avec des taux de reconnaissance dépassant les 80% [9]. À cet égard, notre deuxième objectif est de parvenir à un taux de reconnaissance élevé pour la classification des émotions, tout en surmontant plusieurs obstacles tels que les variations de pose de tête, la diversité des regards, ainsi que les différences d'âge, de sexe et de race des individus.

Compte tenu de l'importance de créer des systèmes multimodaux pour améliorer les résultats, nous avons mis en œuvre un système physio-visuel basé sur des vidéos faciales.

Au cours des dernières années, l'apprentissage profond (DL) s'est imposé comme une approche très réussie et efficace, notamment en raison des résultats obtenus grâce à ses architectures permettant l'extraction et la classification automatiques. Parmi les réseaux notables, on retrouve les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN) [15]. Des efforts considérables déployés dans la recherche et le développement d'architectures de réseaux neuronaux profonds ont abouti à des résultats particulièrement satisfaisants dans le domaine de l'informatique affective. De nombreuses études dans la littérature démontrent des performances nettement supérieures avec les techniques d'apprentissage profond par rapport aux méthodes traditionnelles d'apprentissage automatique [16]. Ces constatations ont renforcé notre motivation à les utiliser dans nos propres études.

## 1.2 Structure de la thèse

Cette thèse est organisée en six chapitres dans le but de mettre en lumière deux domaines principaux : la reconnaissance automatique de l'affect et l'extraction de la fréquence cardiaque à partir des vidéos faciales. En outre, elle vise à présenter nos publications réalisées au cours

de cette recherche à travers une exposition claire et précise de nos méthodes et des résultats obtenus.

Dans le deuxième chapitre, nous débuterons en présentant les concepts fondamentaux de l'émotion et du stress, en fournissant des définitions, ainsi qu'en examinant leur représentation et leurs composantes. Ensuite, nous explorerons un aperçu des techniques et des architectures d'apprentissage profond que nous utiliserons dans nos études de recherche. Enfin, nous ferons une synthèse de la littérature scientifique, en mettant l'accent sur la reconnaissance automatique des émotions et du stress, à la fois unimodale et multimodale.

Le chapitre trois explore les techniques de mesure des paramètres cardiaques, en mettant particulièrement l'accent sur le signal iPPG extrait des vidéos faciales. Nous détaillons toutes les étapes du processus d'extraction du signal iPPG, ainsi que la mesure de la fréquence cardiaque à partir de ce signal iPPG. Ensuite, nous offrons une revue de la littérature sur les différentes recherches menées sur l'utilisation de ces signaux dans le domaine de la reconnaissance automatique de l'affect.

Le chapitre quatre présente nos méthodologies expérimentales mises en œuvre dans notre étude, organisé comme suit :

— Dans un premier temps, nous mènerons une étude comparative de quatre algorithmes d'extraction de signal iPPG sur différentes régions du visage. Le but de cette étude sera de choisir la méthode la plus efficace pour extraire des signaux iPPG précis à partir vidéos faciale des bases de données affectifs. Nous fournirons une description détaillée des algorithmes utilisés dans cette étude comparative ainsi que les résultats obtenus.
— La deuxième section, nous continuerons notre étude en utilisant les signaux iPPG dans la classification des émotions selon l'échelle de valence et arousal. Nous présenterons les étapes du prétraitement du signal et l'architecture DL proposée.
— La troisième section de ce chapitre sera consacrée à la présentation d'une étude qui a été menée pour détecter sept émotions de base dans différentes positions de tête, en présentant une architecture d'apprentissage profond 2D-CNN. Chaque partie de cette section détaillera le pipeline d'études qui a été mis en œuvre. Nous commencerons par révéler les étapes de prétraitement des données, puis nous détaillerons l'architecture DL proposée.
— Le dernier volet dédié à la proposition d'un système physio-visuel pour la classification

binaire des états de stress ou de non-stress. Cette étude, axée sur l'aspect multimodal sans contact, repose sur une entrée de données unique, à savoir les vidéos faciales RVB. Nous commencerons par présenter les étapes de prétraitement pour chaque modalité. Nous décrirons ensuite les architectures d'apprentissage profond proposées pour chacune de ces modalités. Enfin, nous allons aborder la présentation de notre système multimodale.

Le chapitre cinq sera consacré à la présentation des résultats de nos études empiriques, y compris des discussions approfondies et une comparaison avec les derniers développements.

Enfin, dans le chapitre six, nous résumerons l'ensemble des travaux réalisés dans cette thèse, discuterons des défis rencontrés, révélerons les limites identifiées et suggérerons des perspectives de travaux futurs.

Chapitre 2

# La reconnaissance affective par apprentissage automatique

## 2.1 Introduction

Ce chapitre propose une revue de littérature sur la reconnaissance automatique des émotions et du stress. La première section abordera les notions générales sur l'émotion et le stress, ainsi que leur représentation et leurs composants. La deuxième section donne un aperçu des différentes techniques d'apprentissage profond. Nous proposons ensuite une revue des recherches récentes sur la détection automatique des émotions et du stress par apprentissage profond, utilisant les expressions faciales et les signaux physiologiques, ainsi que l'approche multimodale qui combine ces deux modalités.

## 2.2 Généralités sur les émotions et le stress

### 2.2.1 Qu'est-ce que l'émotion

Les émotions sont des états motivationnels et multicomponentiel étudiés par les psychologues. En 1872, Charles Darwin s'exprime pour la première fois sur les expressions émotionnelles et leur fondement. Selon la théorie de Darwin, il existe des émotions universelles et des émotions adaptatives en fonction de l'environnement [17], tandis que Caffi et Janney [18] estiment que les émotions sont un phénomène temporaire qui se manifeste à travers le visage, la voix, les changements corporels et les signaux physiologiques.

Chez les êtres humains, les émotions sont des réactions à des événements et des situations externes, telles que la vision d'un objet effrayant ou interne, tels que les souvenirs. En raison de la nature des événements, le type d'émotion est déterminé, mais pour les scientifiques, ils n'ont pas encore déterminé une définition unique et non ambigu pour le terme émotions [19], et cela est expliqué par Fehr et Russel [20] *'Chacun sait ce qu'une émotion, jusqu'à ce qu'on lui demande d'en donner une définition. À ce moment-là, il semble que plus personne ne sache'.*

De nombreuses définitions du concept d'émotion ont été proposées, et en 1980, Kleinginna et al. [21] ont établi un schéma regroupant 92 définitions et neuf déclarations dans 11 catégories.

Les théoriciens définissent les émotions en différents aspects tels que l'aspect expressif, physiologique, cognitive et subjective [17].

L'aspect expressif est relié aux changements physiques tels que les expressions faciales, le

ton de la voix, etc. L'aspect physiologique qui est constitué à des changements physiologiques tels que la température corporelle, fréquence cardiaque … etc [17], [22].

L'aspect cognitif, qui est associé au niveau d'interprétation mentale, de compréhension et d'évaluation de l'individu après le stimulus émotionnel, tandis que l'aspect subjectif englobe les pensées qui circulent dans l'esprit de l'individu et les sentiments qu'il ressent [17], [23].

### 2.2.2 Représentation des émotions

La reconnaissance automatique des émotions est un domaine qui accorde une grande importance à la représentation des émotions. De manière générale, deux types de représentations sont couramment utilisés : la présentation catégorique et la présentation dimensionnelle.

#### A L'approche catégorielle

L'approche catégorielle ou le modèle discret met en évidence certaines émotions considérées comme universelles et innées, et ce sont des émotions de base, primaires qui possèdent des caractéristiques spécifiques. En outre, selon Nugier [17], il existe des émotions secondaires qui sont le résultat d'un mélange d'émotions primaires.

Différents modèles émotionnels ont été proposés et développés par les chercheurs, et chaque auteur fournit une liste d'émotions de base telles que [24] :

- Darwin 1872 : colère, dégoût, joie, peur, tristesse
- James 1884 : amour, chagrin, douleur, peur, rage
- Arnold 1960 : amour, aversion, colère, courage, découragement, désespoir, désir, espoir, haine, peur, tristesse
- Tomkin 1962 : anxiété, colère, dégoût, honte, intérêt, joie, mépris, surprise
- Izard 1971 : colère, culpabilité, dégoût, détresse, intérêt, joie, honte, mépris, peur, surprise
- Plutchik 1980 : apathie, colère, confiance, dégoût, joie, peur, surprise, tristesse
- Ekman 1982 : colère, dégoût, joie, peur, tristesse, surprise
- Fridja 1986 : désir, intérêt, bonheur, surprise
- Oatley 1989 : bonheur, colère, dégoût, inquiétude, tristesse

Dans le domaine de la reconnaissance automatique des émotions, de nombreuses bases de données ont été créées avec des expressions faciales qui catégorisent les émotions de base,

surtout les émotions qui ont proposé par Ekman.

Un exemple de modèle émotionnel catégoriel fourni par Plutchik qui consiste en une roue contenant 8 émotions multidimensionnelles de base, Voir la figure2.1.

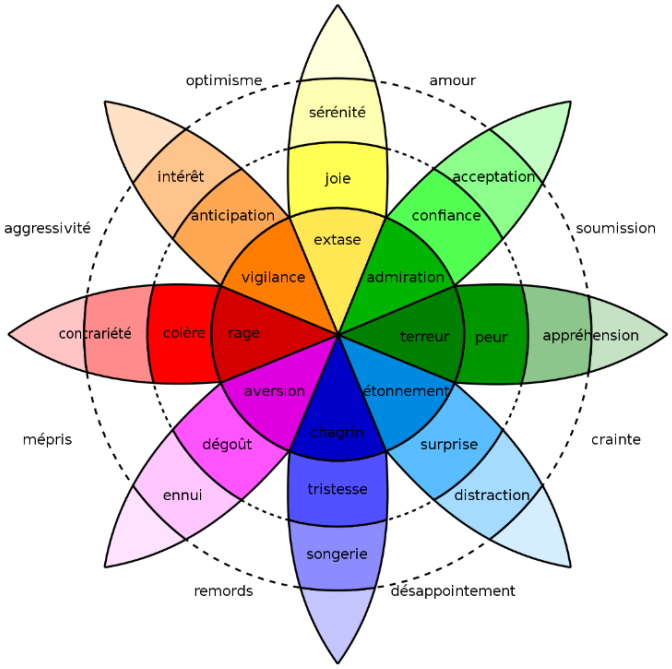



Figure 2.1 – La roue de Plutchik [25]

## B L'approche dimensionnelle

En 1896, Wilhelm Wundt [26] a été le premier à proposer le modèle dimensionnel, il présente les émotions dans une sphère de trois dimensions, chaque dimension labellisée à une propriété commune de toutes les émotions présentées comme plaisir-déplaisir, calme-excitation, relaxation-tension. Ensuite, en 1954, Schlosberg [27], propose un modèle de deux dimensions puis à trois dimensions : colère-bonheur, surprise-peur, sommeil-tension. L'approche dimensionnelle décrit les émotions en utilisant des dimensions indépendantes possédant des propriétés essentielles de l'expérience émotionnelle.

Le modèle de deux dimensions Valence-Arousal est l'un des modèles les plus couramment utilisés dans la reconnaissance automatique des émotions [7].

Valence représente la polarité des émotions, ce qui permet de distinguer les émotions positives telles que la joie des émotions négatives comme la colère. En revanche, Arousal

désigne le niveau d'excitation corporelle associé à une émotion, se manifestant par des réactions physiologiques telles que la fréquence cardiaque. D'autres travaux de recherche ajoutent une troisième dimension, appelée dominance, qui est liée aux efforts déployés par les individus pour contrôler leurs émotions [28]. La figure 2.2 illustre la répartition des émotions le long des deux axes : Valence et Arousal.

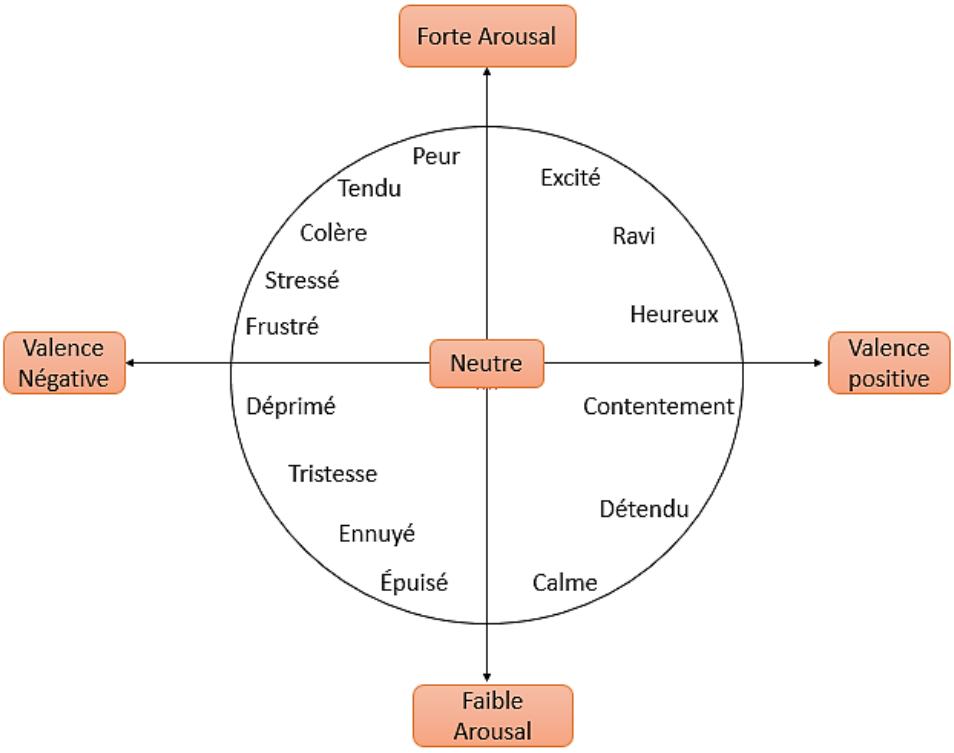



Figure 2.2 – La représentation dimensionnelle des émotions : Valence-Arousal

### 2.2.3 Stress

Il est généralement accepté que tout individu est confronté à diverses pressions dans sa vie quotidienne, ce sentiment influençant souvent leur comportement. Le terme "stress" est couramment employé pour décrire cette sensation d'être sous pression. En règle générale, le stress est déclenché par des facteurs et des stimuli à la fois externes et internes [29].

Tout comme pour l'émotion, il n'existe pas de définition claire et commune du stress. Au début du 20e siècle, Cannon [30] a introduit les termes "homéostasie" et "réponse de combat ou de fuite" ("fight or flight"). Selon Cannon [30], les facteurs de stress sont des menaces qui perturbent l'homéostasie, c'est-à-dire l'état d'équilibre du corps dans lequel les paramètres physiologiques sont stables dans une plage constante.

En 1970, Selye [31] définit le stress comme un ensemble non spécifique de réponses déclenchées, quelle que soit la nature des facteurs du stress. D'un point de vue psychobiologique, le stress est considéré comme un processus de réaction complexe qui incluait des éléments cognitifs, psychologiques et comportementaux [32].

Le degré de réponse au stress dépend principalement de deux choses, premièrement, le type de stress auquel une personne est exposée, et deuxièmement, la capacité du corps à résister à ce stress [29].

Il existe une relation étroite entre le stress et les émotions, car il a été démontré que des niveaux élevés de stress influencent les émotions, et il y a un lien entre le stress et la représentation dimensionnelle des émotions [33], [34]. Valenza et. al [35] ont présenté le stress dans le modèle circumplex avec une valence négative et une arousal élevée. De plus, Schimmack et Rainer [36] présentent que la dimension d'arousal est divisée en arousal tendu (stress prolongé) et arousal énergétique (endormi-actif).

### 2.2.4 Composantes des émotions et du stress

Les émotions et le stress occupent une place centrale dans nos vies, permettant d'exprimer nos réactions face aux diverses situations. Leur manifestation peut prendre une forme verbale, par l'usage de mots, ou non verbale, à travers les expressions faciales, les comportements, ou même les changements physiologiques. Nous présentons deux modalités essentielles de notre étude.

#### A Expressions faciales

D'après les statistiques réalisées par Rouast et al. [8] les expressions faciales sont la modalité non verbale la plus intéressante à étudier par les chercheurs. En psychologie, le visage est considéré comme une source importante d'informations émotionnelles [37]. Selon Charles Darwin, les expressions faciales ont un impact significatif sur la communication non verbale des humains. Étant visibles et facilement détectables, plusieurs bases de données sont collectées afin de les utiliser dans l'amélioration des études sur la détection précise de l'affect.
Lors d'une expérience affective, des changements musculaires s'activent automatiquement et présentent un état affectif spécifique. Ekman est l'un des chercheurs les plus renommés et

actifs dans cette modalité. Il a créé le Facial Action Coding System (FACS), qui associe chaque type d'émotion à des changements musculaires faciaux spécifiques appelés Action Units (AU) [38]. Un exemple d'unités d'action faciale du FACS pour l'émotion de peur est présenté dans la figure 2.3

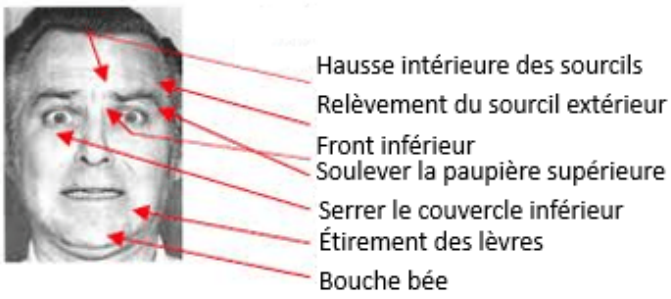



Figure 2.3 – Exemple d'unités d'action faciale du FACS pour l'émotion peur[39]

## B Signaux physiologiques

Les signaux physiologiques sont sensibles aux changements affectifs. Chaque émotion active le système nerveux autonome, entraînant des variations physiologiques spécifiques telles que des changements du rythme cardiaque, de la température corporelle, etc. [7]. En ce qui concerne le stress, de nombreuses études ont démontré que le stress entraîne des changements physiologiques majeurs, tels qu'une augmentation du rythme cardiaque, de la transpiration et du rythme respiratoire [40].
Le principal avantage de la modalité physiologique par rapport aux autres modalités physiques est l'incapacité de l'humain à la cacher et à la contrôler, car elle s'active automatiquement et involontairement lorsque l'individu est exposé à des stimuli [7].

De nombreux signaux physiologiques cités par la suite sont couramment utilisés dans les travaux de recherche liés à cette modalité et atteignent un taux de reconnaissance élevé [7] :

- l'électrocardiographique (ECG) : Signal de l'activité électrique de contraction et de relaxation du muscle cardiaque.
- photopléthysmographique (PPG) : Mesure optique des changements du flux sanguin dans les vaisseaux sanguins associés aux réactions cardiovasculaires et résultant de changements du volume et du diamètre du sang.

Les signaux ECG et PPG sont des mesures liées à l'activité du cycle cardiaque, grâce

auxquelles la fréquence cardiaque et la variabilité de la fréquence cardiaque peuvent être déterminées. Les trois types de signaux sont cruciaux et utiles pour reconnaître les émotions et le stress humaines.

- Signal d'activité électrodermique (EDA) : Mesure de la conductivité cutanée qui fournit les propriétés électriques de la peau dues à l'activité de sécrétion sudorale. Les glandes sudoripares sont réparties sur la peau et s'activent lorsqu'elles reçoivent des signaux du système nerveux, ce qui reflète une réponse spécifique au stimulus.
- Signal Electroencéphalogrammes (EEG) : Mesure de l'activité cérébrale, est l'un des signaux les plus largement utilisés dans la reconnaissance des émotions et du stress en raison de la disponibilité du casque portable et peu coûteux [41], [42].
- Electromyographie (EMG) : Ce signal est associé à l'activité musculaire, et pour les émotions et le stress, ce signal est utilisé pour mesurer l'activité des muscles du visage.
- Température de la peau : Ce signal représente la température corporelle, généralement influencée par l'exposition à des situations stressantes.
- Fréquence respiratoire : Mesure du volume d'air expiré ou inhalé au cours d'un cycle respiratoire. Au repos, le rythme respiratoire est lent et régulier, mais dans un état d'excitation tel que le stress et la peur, le rythme respiratoire augmente.

Tous ces signaux ont été bien utilisés dans le domaine de la reconnaissance automatique des émotions et du stress, mais il ne faut pas négliger que l'inconvénient de cette modalité réside dans la manière dont elle est mesurée, car il est nécessaire de fixer des capteurs sur le corps humain pour enregistrer des signaux physiologiques parallèlement aux réactions affectives, ce qui peut perturber les personnes en provoquant des changements dans leur état psychique. De plus, il n'est ni approprié ni convivial de les utiliser dans des applications en dehors du laboratoire.

## 2.3 Reconnaissance automatique de l'affect

La reconnaissance automatique de l'affect fait partie des sujets importants de l'étude par ordinateur du comportement humain. Il est important de créer des systèmes automatisés robustes et fiables, capables de détecter avec précision les émotions et le stress afin de les appliquer dans des applications et des interfaces du monde réel, dans le but d'améliorer et de faciliter l'interaction entre les humains et les machines automatisées.

En général, la classification automatique de l'affect s'articule autour de trois étapes essentielles : le prétraitement des données, l'extraction des caractéristiques importantes et la classification finale. La structure de base d'un système de reconnaissance automatique de l'affect est illustrée dans la figure 2.4.

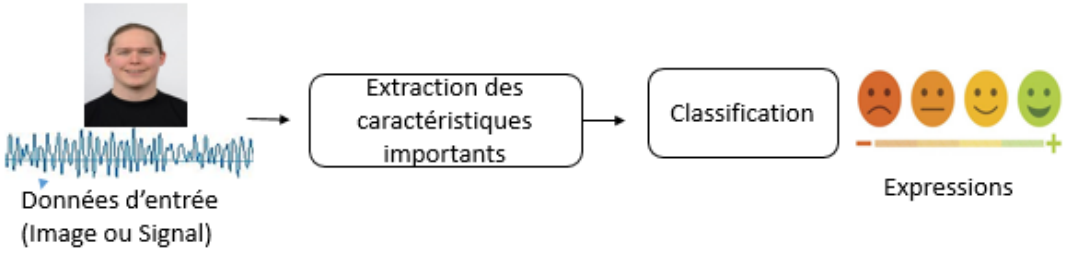



Figure 2.4 – Structure basique d'un système de reconnaissance automatique de l'affect [43]

Au cours de ces dernières années, l'apprentissage profond (DL) a gagné en importance dans plusieurs domaines, tels que le traitement d'images, la vision par ordinateur, la reconnaissance vocale, la reconnaissance des émotions et du stress, et bien d'autres. Cette approche est généralisable et hautement évolutive, car elle peut être appliquée à différentes applications avec divers types de données [15]. Récemment, Les méthodes basées sur l'apprentissage profond ont révolutionné les études de l'état de l'art, remplaçant avec succès les approches conventionnelles [16].

L'apprentissage profond (DL) se compose de plusieurs couches sous forme d'une architecture hiérarchique. Cela permet l'utilisation de nombreuses étapes d'unités de traitement d'information non linéaires pour l'extraction des caractéristiques et la classification. Il s'agit d'une approche universelle capable de résoudre presque tous types de problèmes. Nous faisons usage d'algorithmes d'apprentissage profond dans cette thèse pour élaborer nos systèmes de reconnaissance automatique de l'affect.

Il est important de noter que pour qu'un réseau de neurones profonds fonctionne de manière optimale, deux éléments sont essentiels : une base de données riche et une puissance de calcul élevée, c'est-à-dire un ordinateur puissant [8].

### 2.3.1 Généralités sur les réseaux de neurones profonds

#### A Réseau de neurone profond DNN

Les réseaux de neurones artificiels sont l'un des algorithmes d'une grande importance dans l'apprentissage profond. Inspirés du fonctionnement du réseau neuronal biologique du cerveau humain, les réseaux neuronaux artificiels sont constitués d'une ou plusieurs couches cachées contenant des unités de traitement appelées neurones artificiels [24]

L'entrée du réseau reçoit des données, telles que des images ou des signaux, sous forme de variables indépendantes qui doivent être standardisées et normalisées. En sortie du réseau, une prédiction est générée. Pendant la phase d'apprentissage, les paramètres extraits de la première couche sont transférés vers la deuxième couche, et ainsi de suite jusqu'à la dernière couche de prédiction. Après chaque couche neuronale, une fonction d'activation est appliquée, disposant d'un poids ajustable pour réduire l'erreur de prédiction. Ce mécanisme est connu sous le nom de rétro-propagation [15]. La figure 2.5 présente un aperçu général d'un réseau DNN.

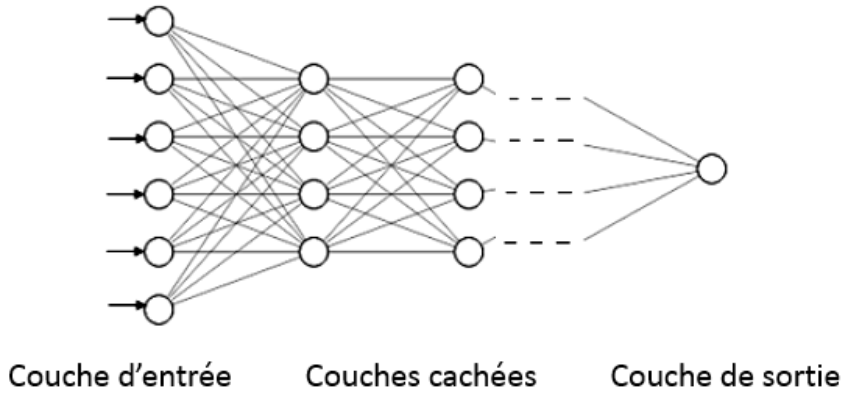



Figure 2.5 – Aperçu d'un réseau neurone profond DNN

#### B Réseau de neurone convolutif CNN

En 1989, pour la première fois, LeCun et al. [44] ont proposé un réseau de neurones convolutif appelé LeNet pour la classification des chiffres manuscrits, obtenant des résultats satisfaisants. En 1998, LeCun et al. [45] ont démontré que la classification par des réseaux de neurones convolutifs 2D-CNN permet d'obtenir de meilleurs résultats par rapport à différentes

méthodes classiques d'apprentissage.

La base de l'architecture CNN se compose de couches convolutives, de couches pooling et de couches entièrement connectées, et parfois des couches dropout et de Batch-normalization sont ajoutées pour éviter le problème de sur-apprentissage [46]. Dans la figure 2.6, vous trouverez un aperçu général d'un réseau CNN.

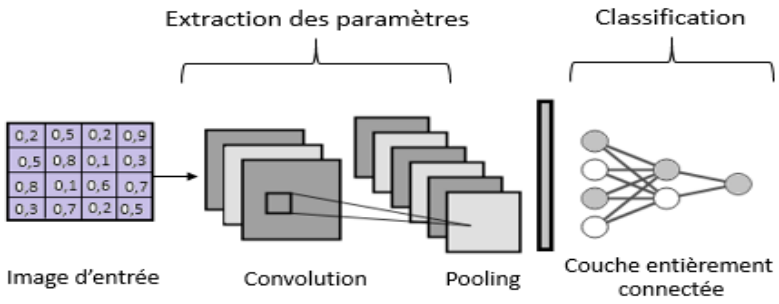



Figure 2.6 – Illustration typique du réseau neuronal convolutif

## B.1 Couche convolutive

La couche convolutive est une composante essentielle d'un réseau neuronal convolutif, car elle extrait les caractéristiques importantes des données. Cette couche se compose de plusieurs filtres qui apprennent automatiquement les propriétés significatives des données en effectuant la somme pondérée entre les filtres et chaque portion des données d'entrée, générant ainsi des cartes des caractéristiques (feature maps) en sortie. Les résultats de cette étape, une fois soumis à une fonction d'activation, sont considérés comme les entrées de la couche suivante du réseau [47].

L'opération de convolution est caractérisée par trois paramètres essentiels : le nombre et la taille du filtre utilisé, le pas de déplacement (stride en anglais) et l'ajout de bordures (Padding en anglais).
Le nombre et la taille des filtres sont fixés lors du processus de convolution dans la couche. Stride est le pas de glissement du filtre et padding est une technique d'ajouter des valeurs nulle aux lignes et colonnes de la matrice de la carte des caractéristiques (feature maps) afin de conserver les mêmes dimensions de la carte de caractéristiques de sortie [47].

## B.2 Pooling

La couche de pooling est une technique de sous-échantillonnage qui vise à identifier les caractéristiques les plus pertinentes, réduisant ainsi la taille de la carte des caractéristiques. Cela aide à diminuer le temps de calcul et à simplifier le processus d'apprentissage [48].

Il existe deux types de pooling : le pooling maximum et le pooling moyen [48]. L'opération de pooling est composée d'une fenêtre glissante sur la carte des caractéristiques produites par l'étape de convolution. Dans ce cas, le pooling maximum prend simplement la valeur la plus élevée de la carte des caractéristiques, supprime le reste des informations. Pour le pooling moyen, génère la valeur moyenne [49].

## B.3 Couche Entièrement connectée

La dernière étape des CNN est la classification, où les scores de chaque classe sont calculés en utilisant les caractéristiques extraites des couches précédentes, comprenant la convolution et le pooling [48]. Selon le type de classification souhaitée, une fonction d'activation est ajoutée à la sortie du réseau. Par exemple, pour une classification binaire, il est préférable d'utiliser la fonction sigmoïde, tandis que pour une classification catégorielle, la fonction softmax est souvent privilégiée.

De nouvelles architectures de réseaux de neurones convolutions CNN ont été développées au fil des années pour traiter des données spécifiques. Le 1D-CNN, tel que décrit par Kiranyaz et al. [50], est utilisé pour les données séquentielles unidimensionnelles. Son principe repose sur l'application de filtres unidimensionnels glissant sur le vecteur de données. À chaque point, une multiplication matricielle locale est effectuée et le résultat est additionné dans la carte de caractéristiques finale. Cette approche réduit la nécessité de procédures complexes d'extraction des caractéristiques et garantit de meilleurs résultats.

D'autre part, les réseaux 3D-CNN, comme décrits par Zhou et al. [51], ont la capacité de modéliser de meilleures informations temporelles en appliquant des filtres 3D aux cubes formés par l'empilement de séquences d'images. Cela permet de conserver l'information temporelle et améliore la capacité des modèles à analyser les données spatio-temporelles.

## C Réseau de neurone récurrent RNN

Un réseau de neurone récurrent RNN était proposé en 1982 par Hopfield Netwark, mais l'idée a été écrite en 1974 [46]. Les réseaux de neurones récurrents sont des types d'architecture d'apprentissage profonds conçus pour des données séquentiel tel que les vidéos, texte et les paroles. Contrairement au CNN, qui repose sur l'indépendance entre les exemples d'entraînement et sur des filtres convolutifs capables d'extraire automatiquement des caractéristiques spatiales, ils ne peuvent pas conserver ces caractéristiques dans des données séquentielles au fil du temps. Ce n'est pas le cas des réseaux de neurones récurrents, dont la structure est composée de nœuds produisant une sortie utilisant l'entrée actuelle et l'état caché des nœuds précédents, les RNN génèrent également un état caché actuel qui sert d'entrée pour le nœud suivant. Ceci contribue à maintenir la séquence des paramètres au fil du temps. De plus, les RNN peuvent modéliser simultanément les dépendances séquentielles et temporelles à plusieurs échelles [46].

### C.1 Long Short-Term Memory LSTM

Un réseau de Long Short-Term Memory LSTM est une architecture spécifique de réseau de neurone récurrent, conçu pour éviter le problème de la fuite du gradient. LSTM était proposé et développé par Sepp Hochreiter [52], et il s'agit d'une solution immédiate pour déterminer le problème de la fuite de gradients.

Contrairement aux RNN traditionnels, LSTM est conçu pour contrôler et préserver le flux d'informations dans le réseau avec sa structure qui se compose principalement de trois portes telles que : la porte d'entrée, la porte de sortie et la porte d'oubli [53].

Le principe de fonctionnement du réseau LSTM dépend de l'entrée $x(t)$, de la sortie du neurone précédent $h(t-1)$ et aussi notamment de l'état de la cellule $z(t)$ qui n'existe pas dans les RNN traditionnels et qui détermine le fonctionnement du LSTM. La cellule LSTM est

exprimée mathématiquement comme suit [54] :

$$f_t = \sigma(W_f h_{(t-1)} + W_f x_t + b_f) \tag{2.1}$$

$$i_t = \sigma(W_i h_{(t-1)} + W_i x_t + b_i) \tag{2.2}$$

$$\widetilde{C}_t = \tanh(W_c h_{(t-1)} + W_c x_t + b_c) \tag{2.3}$$

$$C_t = f_t * C_{(t-1)} + i_c * \widetilde{C}_t \tag{2.4}$$

$$o_t = \sigma(W_o h_{(t-1)} + W_o x_t + b_o) \tag{2.5}$$

$$h_t = o_t * \tanh(C_t) \tag{2.6}$$

Tel que $x_t$ le vecteur d'entrée et $h_{(t-1)}$ est l'état caché au temps $t$. La porte d'oublie $f_t$ permet d'indiquer si les informations doivent être conservées ou supprimées en appliquant la fonction d'activation sigmoïde et la valeur de sortie est comprise entre $0$ et $1$ ce qui signifie que si elle est $1$, elle conserve l'information et si elle est $0$, l'information est supprimée [53].

La porte d'entrée $i_t$ détermine le niveau auquel les nouvelles mémoires affectent les anciennes mémoires, puis détermine la quantité de nouvelles informations à transmettre et la quantité à supprimer pour ajouter d'autres nouvelles informations. Et finalement la porte de sortie $o_t$ qui détermine quelles parties de l'état de la cellule sont transmises comme sortie [53].

L'architecture typique d'une cellule LSTM est représentée dans la figure 2.7.

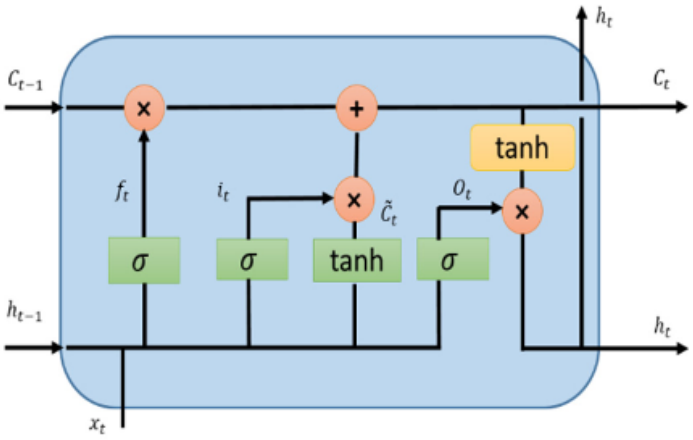



Figure 2.7 – La structure typique du réseau LSTM [46]

### 2.3.2 Reconnaissance automatique de l'affect à travers les signaux physiologiques

Les méthodes d'apprentissage profond ont récemment suscité un vif intérêt dans le domaine de la reconnaissance émotionnelle à partir de divers signaux physiologiques. De nombreuses contributions et techniques d'analyse ont été développées, ainsi que des architectures d'apprentissage profond, dans le but d'améliorer cette approche d'étude.

Dans ce contexte, Yang et al. [55] ont mené une étude où ils ont classé les émotions en trois catégories distinctes : positives, négatives et neutres. Pour ce faire, ils ont exploité des signaux PPG et ECG. Ils ont extrait des caractéristiques à la fois dans le domaine temporel et fréquentiel, puis les ont utilisées comme entrées pour une architecture 1DCNN. Leur approche a permis d'atteindre une précision de classification de 75,44 %.

Nakisa et al. [56] ont proposé une méthode de reconnaissance des émotions basée sur les signaux EEG et PPG. Ils ont utilisé une technique de segmentation avec différentes tailles de fenêtres temporelles. Pour la classification, ils ont combiné une architecture 1D-CNN avec LSTM. Les résultats de leur expérience ont montré que le modèle spatio-temporel proposé, avec une segmentation du signal de 10 secondes, a atteint un taux de reconnaissance significatif de 71,61 %.

Granados et al. [57] ont mené une étude comparative entre les approches traditionnelles d'apprentissage automatique et les approches basées sur l'apprentissage profond DL, utilisant les signaux GSR et ECG de la base de données AMIGOS [58]. Dans leur étude, ils ont choisi d'appliquer l'architecture 1DCNN et ont constaté que cette approche obtenait de meilleurs résultats que les méthodes traditionnelles pour la classification.

Dans une étude comparative menée par Wang et al. [59] ont évalué les performances de deux architectures d'apprentissage profond pour la reconnaissance des émotions à partir de signaux PPG. Ces architectures étaient une architecture BiLSTM (Réseau de neurones LSTM bidirectionnels) seule et une combinaison de ResNet (Réseau de neurones résiduels) et de BiLSTM. Les résultats de l'étude ont montré que l'architecture ResNet-BiLSTM a surpassé l'architecture BiLSTM en termes de performances.

Al Machot et al. [60] ont innové en utilisant exclusivement les signaux d'activité électrodermique (EDA) pour classer les émotions en fonction de leur valence et de leur arousal.

Pour ce faire, ils ont converti ces signaux en matrices pour les traiter avec une architecture 2D-CNN. Leurs résultats ont démontré des performances prometteuses, avec des précisions de classification atteignant 85 % sur la base de données DEAP [61] et 81 % sur la base de données MAHNOB-HCI [62].

Dans une étude menée par Lee et al. [63], un taux de reconnaissance plus élevé a été obtenu en utilisant des signaux PPG à impulsion unique correspondant à 1,1 seconde. Ils ont appliqué une segmentation et une normalisation du signal PPG brut, puis utilisé les résultats obtenus comme entrée dans une architecture 1DCNN. Leurs résultats ont montré une précision de 75,3 % pour la classification de la valence et de 76,2 % pour l'arousal.

Une classification des émotions dans Arousal et Valence par le signal EEG de l'ensemble de données DEAP a été proposée par Ramzan et Dawn [64]. Une nouvelle architecture CNN-LSTM a été proposée, obtenant des résultats de 97,41% pour la classification de valence, 97,33% pour l'éveil et 97,68% pour la dominance.

Li et Liu [65] ont développé deux architectures d'apprentissage profond de type 1D-CNN et DNN pour la classification binaire du stress chez l'humain, utilisant les signaux de l'ECG, de l'EMG, de l'EDA, de la respiration et de la température de la peau pour cette classification. Les résultats obtenus sont très prometteurs, avec une précision de 99,80 % avec l'architecture 1DCNN et 99,65 % pour l'architecture basée sur DNN. Dans une étude similaire, Albertetti et al. [66] ont proposé une architecture RNN utilisant deux signaux physiologiques EDA et BVP, provenant de la base de données SEDY. Les résultats de leur étude ont montré une performance significative, avec un F1 score de 71 %.

Zeeshan et al. [67] ont présenté une technique novatrice basée sur la fusion du spectre du signal ECG et du signal brut pour la classification du stress humain. En utilisant cette approche, ils ont atteint une précision de 72,7 % dans la classification du stress. Cette recherche met en évidence l'efficacité de la fusion des informations spectrales et brutes des signaux ECG pour identifier et classifier le stress chez les individus.

Une nouvelle approche d'étude pour la classification des états de stress ou de non-stress a été proposée par Elzeiny et Qaraque [68]. Dans cette étude, les chercheurs ont converti les intervalles inter-battements extraits de PPG en images spatiales, puis en images du domaine fréquentiel. les images sont ensuite utilisées comme entrée du réseau 2D-CNN pour la

classification. Cette approche permet d'atteindre un taux de reconnaissance de validation de 100%.

D'après les diverses recherches examinées, il est notable de constater l'impact significatif de la modalité physiologique, impliquant une variété de signaux physiologiques, sur la classification des émotions et du stress. De plus, l'utilisation de techniques d'apprentissage profond a considérablement amélioré la précision des résultats, dépassant les 70 %. Il convient également de noter que la majorité des méthodes adoptent des réseaux de neurones convolutifs 1D-CNN et des réseaux de neurones récurrents LSTM pour la classification. En outre, certaines études ont exploré l'utilisation d'architectures 2D-CNN pour traiter des données séquentielles préalablement converties en images bidimensionnelles.

Le tableau 2.1 résume tous les travaux cités, mentionnant l'architecture d'apprentissage profonds utilisée, les bases de données et la précision obtenue.

Il est pertinent de noter que toutes les études mentionnées précédemment ont utilisé des signaux physiologiques mesurés à l'aide de capteurs placés sur le corps des individus exposés à une stimulation affective. Malheureusement, cette approche d'extraction de signaux physiologiques n'est pas pratique et n'est donc pas couramment utilisée dans les applications réelles. C'est pourquoi les chercheurs se tournent désormais vers des études sur la reconnaissance automatique des états affectives utilisant la mesure sans contact de signaux physiologiques. C'est ce que nous allons approfondir et expliquer ultérieurement dans cette thèse.

### 2.3.3 Reconnaissance automatique de l'affect par des expressions faciales

Les expressions faciales restent les indices préférés des chercheurs pour étudier la reconnaissance des émotions et du stress, en raison de leur visibilité et de leur facilité de capture. Elles représentent les premiers signaux manifestant les états affectifs et leurs intensités, comme mis en avant par Li et al. [69]. Actuellement, la mesure des expressions faciales est intégrée dans plusieurs applications et domaines tels que la santé, la réalité virtuelle, l'éducation, et bien d'autres [70], [71].

L'extraction des traits affectifs d'un visage à l'autre représente une tâche complexe, car chaque individu peut exprimer ses émotions et son niveau de stress de manière différente par

Tableau 2.1 – Travaux connexes sur la reconnaissance automatique des émotions et du stress à l'aide de signaux physiologiques

| Auteurs | Base de données | Type de signal | Architecture d'apprentissage profond | Précision% |
|---|---|---|---|---|
| Yang et al. [55] | 47 Sujets | PPG | 1DCNN | 75,4% |
| Granados et al. [57] | AMIGOS | ECG, GSR | 1DCNN | 81%, 71% pour Arousal et 71%, 75% pour valence respectivement pour ECG, GSR |
| Nakisa et al. [56] | 20 sujets | EEG, VFC | 1DCNN-LSTM | 71.61% |
| Wang et Yu. [59] | 40 Sujets | PPG | ResNet-BiLSTM | 89.15% |
| Al machot et al. [60] | DEAP, MAHNOB-HCI | EDA | 2DCNN | 85%, 81% |
| Lee et al. [63] | DEAP | PPG | 1DCNN | 75.3% pour la valence et 76.2% pour Arousal |
| Ramzan et Dawn [64] | DEAP | EEG | CNN-LSTM | 97.41% pour la valence 97.39% pour Arousal |
| Li et Liu [65] | WESAD | ECG, EMG, EDA, SKT, SGR | 1D-CNN, DNN | 99.80%, 99.65% |
| Albertetti et al. [66] | SEDY | EDA, BVP | RNN | F1 score= 71% |
| Zeeshan et al. [67] | 9 participants | ECG | 1DCNN, ResNet_18 | 72.7% |
| Elzeiny et qaraque [68] | WESAD | PPG | 2D-CNN | 99% d'entraînement et 100% validation. |

rapport à un autre. De plus, chaque personne perçoit les choses sous des angles différents, introduisant ainsi une diversité d'expressions faciales. La vision par ordinateur des expressions faciales doit surmonter divers défis tels que l'âge, la race, le sexe, les conditions d'éclairage, la pose de la tête devant les caméras et les occlusions causées par des accessoires comme des lunettes de soleil, des foulards, ainsi que les maladies de la peau.

La reconnaissance des émotions et du stress à travers les expressions faciales en utilisant des techniques d'apprentissage profond a connu des avancées significatives ces dernières années. Cette progression est attribuable à la disponibilité de bases de données riches en données faciales annotées, ainsi qu'à l'amélioration de la puissance de calcul informatique [8].

La recherche sur la reconnaissance automatique de l'affect à travers des EF est divisée en deux parties : utilise des bases de données statiques et dynamiques [69]. De plus, des bases de

données produites en laboratoire dans des conditions contrôlées avec des expressions faciales mimiques et d'autres avec des expressions spontanées [8]. Dans la suite, nous présentons certains travaux récents avec la contribution et l'objectif de chaque article.

Mollahosseini et al. [72] ont développé un réseau de neurones convolutif (CNN) en exploitant plusieurs bases de données disponibles. Après l'extraction des caractéristiques faciales des données, les images ont été redimensionnées à une résolution de 48 x 48 pixels. Ensuite, ils ont appliqué une méthode d'augmentation des données, et ont introduit la possibilité d'utiliser la technique du réseau dans un réseau (Network-in-Network), qui permet d'améliorer les performances locales en appliquant des couches de convolution de manière spécifique à chaque région. Cette approche contribue également à atténuer le problème de sur-ajustement.

Lopes et al. [73] ont étudié de l'impact du prétraitement des données avant l'étape de l'entraînement du réseau. L'augmentation des données, la correction de la rotation, le recadrage, le sous-échantillonnage avec 32x32 pixels et la normalisation de l'intensité des images sont les étapes utilisées. Les auteurs montrent que la combinaison de toutes ces étapes de prétraitement est plus efficace que les appliquer séparément.

Pour le problème de la fuite du gradient, Cai et al. [74] proposent une nouvelle architecture CNN avec des couches de Sparce Batch-Normalization SBN (Normalisation par lots éparse). La propriété de ce réseau est d'utiliser deux couches de convolution successives au début, Suivie par des couches de pooling et SBN, et pour réduire le problème de sur-ajustement, une couche de dropout appliqué au milieu de trois couches entièrement connectés.

Chowdary et al. [75] ont adopté une approche d'apprentissage par transfert, où ils ont utilisé les architectures de réseaux de neurones pré-entraînées telles que ResNet50, MobileNet, VGG19 et InceptionV3. Ces réseaux ont été enrichis en ajoutant de nouvelles couches entièrement connectées, conçues pour capturer les caractéristiques spécifiques aux tâches envisagées. Ils ont obtenu une précision de 96% sur la base de données CK+ [76].

Borgalli et Surve [77] ont proposé un réseau d'apprentissage profond appelé Costum-CNN. La validation croisée K-fold a été utilisée pour former le réseau. Une précision de 86,78 % a été obtenue avec FER 2013 [78], 92,27 % avec CK+ [76] et 91,58 % avec JAFEE [79].

Pour la reconnaissance automatique des émotions des visages masqués, Agarwal et Susan [80] ont utilisé des modèles pré-entraînés tels que : ResNet 50, Inception-V3, Xception, AlexNet

et EfficientNet. Leur étude démontre que le modèle Inception-V3 surpasse les autres en termes de précision.

En 2023, Karnati et al. [81] présente une étude complète sur les méthodes de DL proposées, les différentes techniques de prétraitement d'images, ainsi qu'une discussion et une comparaison sur les performances atteintes ainsi que les différentes bases de données existantes et utilisées. Helaly et al. [82] ont amélioré le modèle ResNet18 basée sur l'apprentissage par transfert, pour obtenir une meilleure précision utilisant les bases de données CK+ [76] et FER2013 [79].

Tous les travaux mentionnés précédemment se basent sur des bases de données statiques, où les taux de précision dépassent souvent les 90%. Cependant, ces approches ne tiennent pas compte des informations temporelles, ce qui est crucial pour les données séquentielles.

Kim et al. [83] étudient la variation des expressions faciales tout au long des états émotionnels. Ils proposent une architecture spatio-temporelle combinant des réseaux CNN et des réseaux de neurones récurrents LSTM. Dans cette approche, le CNN est d'abord utilisé pour apprendre les caractéristiques spatiales des expressions faciales dans les images séquentielles, puis le LSTM est appliqué pour préserver la séquence de ces images. Ils ont obtenu une précision de 78,61% avec la base de données MIMI [84] et de 60,98% avec la base de données CASEMII [85].

Yu et al. [86] ont introduit une nouvelle architecture baptisée "spatio-temporelle Convolutional with Nested LSTM" (STC-NLSTM). Cette architecture repose sur trois sous-réseaux d'apprentissage profond : un 3DCNN pour l'extraction des caractéristiques spatio-temporelles, suivi du T-LSTM pour la préservation de la dynamique temporelle, puis du C-LSTM pour modéliser les fonctionnalités multiniveaux. Leur étude a obtenu une précision de 99,8% avec la base de données CK+ [76].

Une combinaison de VGG-19 et BiLSTM a été proposée par Mohana et al. [87], une précision de 92 % a été obtenue avec la base de données CK+ [76] et de 84 % avec une base de données interne.

Singh et al [88] ont proposé une approche hybride combinant un 3D-CNN et un Conv-LSTM. Après la détection des visages et l'identification des points de repère pour l'alignement des images. Ils ont utilisé un 3D-CNN pour extraire les caractéristiques spatio-temporelles, suivies par un Conv-LSTM pour conserver ces informations dans le temps. Leur expérimentation

a été menée sur les ensembles de données SAFEE [89], CK+ [76] et AFEW [90], obtenant des précisions de 98,83%, 95,10% et 43,86% respectivement.

Pour la classification du stress, Almeida et Rodrigues [91] ont exploré l'utilisation de réseaux neuronaux pré-entraînés tels que VGG16, VGG19 et Inception-ResNetV2. Ils ont cherché à déterminer le meilleur modèle de détection du stress en utilisant deux classifieurs distincts. Après plusieurs expériences, ils ont déterminé que VGG16, associé à un classifieur basé sur une couche convolutive, est le meilleur. Leur étude a abouti à un taux de reconnaissance élevé de 92%.

Dans le même contexte, Zhang et al. [92] ont proposé une nouvelle architecture CNN qui combine des fonctionnalités faibles avec des fonctionnalités élevées. Il convient d'indiquer que des précisions élevées, à savoir 91,2%, 80,4% et 99,3%, ont été obtenues respectivement avec les ensembles de données CK+ [76], Oulu-CASIA [93] et KMU-FED [94].

Jeon et al. [95] ont proposé une nouvelle méthode pour la classification du stress humain en niveaux faibles, élevés et neutres. Ils ont utilisé un réseau résiduel ResNet-18 pour extraire les caractéristiques spatiales, puis ont effectué une couche de pooling moyenne suivie d'une classification à l'aide des couches de neurones entièrement connectées. Un taux de reconnaissance de 51.58% dans le niveau bas de stress et 66.24% avec le haut niveau de stress a été obtenu en utilisant la base de données SWELL-KW [96].

Concernant Giannakakis et al. [97] ont proposé une nouvelle technique de détection du stress humain basée sur la détection automatique des unités d'action faciale. En effet, une fois le processus de détection des visages terminé, ils ont extrait les caractéristiques géométriques et d'apparence. Après ces étapes, un réseau de neurone a été appliqué afin de classer les unités d'action (UA). Les résultats ont montré des changements observables dans certaines UA spécifiques du visage en état de stress, par opposition à l'état neutre.

D'après les travaux cités, l'importance de la modalité de l'expression faciale dans la classification de l'affect est manifeste. De plus, les méthodes existantes se divisent en deux catégories : celles qui exploitent des données statiques et celles qui utilisent des données séquentielles. Il est important de souligner l'impact significatif du prétraitement des données sur l'amélioration de la précision de la classification. Par ailleurs, tandis que les données spatiales sont généralement traitées avec des réseaux CNN, la plupart des études utilisent

une combinaison de CNN et de LSTM pour les données séquentielles. Le tableau 2.2 présente l'ensemble des travaux cités, exposant les architectures d'apprentissage profond employées, les bases de données utilisées, ainsi que les niveaux de précision atteints.

Tableau 2.2 – Travaux connexes sur la reconnaissance automatique des émotions et du stress à l'aide des expressions faciales

| Auteurs | Base de données | Architecture d'apprentissage profond | Précision |
|---|---|---|---|
| Mollahosseini et al. [72] | BMultiPie, MMI, DISFA, FERA, SFEW, CK+ FER2013 | CNN | 94.7%, 77.9%, 55%, 76.7%, 47.7%, 93,2%, 61.1% |
| Lopes et al. [73] | CK+, JAFFE, BU-3DFE | CNN | 96.76% Pour CK+ |
| Cai et al. [74] | JAFFE, CK+ | SBN-CNN | 95.24%, 96.87% |
| Chowdary et al. [75] | Ck+ | Resnet50, MobileNet, VGG19, InceptionV3 | 96% |
| Borgalli et surve [77] | FER2013, CK+, JAFEE | Costum-CNN | 86,78%, 92,27%, 91,58% |
| Helaly et al. [82] | CK+, FER2013 | Resnet50 | 98%, 83% |
| Kim et al. [83] | MMI, CASMEII | CNN-LSTM | 78.61%, 60.98% |
| Yu et al. [98] | CK+, Oulu-CASIA, MMI, BP4D | STC-NLSTM | 99.8%, 93.45%, 84.53% |
| Mohana et al. [87] | CK+, Base de données interne | VGG-19-LSTM | 92%, 84% |
| singh et al [88] | SAFEE, CK+, AFEW | 3DCNN-Conv-LSTM | 98.83%, 95.10%, 43.86% |
| Almeida et Rodrigues [91] | KDFE, CK+ | VGG16-CNN | 92% |
| Zhang et al. [92] | CK+, Oulu-CASIA, KMU-FED | CNN | 91.2%, 80.4%, 99.3% |
| Jeon et al. [95] | SWELL-KW | ResNet | 80.55% Neutre, 51.58% Bas-Stress, 66.24% Haut-Stress. |
| Giannakakis et al. [97] | KDFE, CK+ | VGG16-CNN | 92% |

### 2.3.4 Reconnaissance automatique multimodale de l'affect

Puisque les états affectifs chez les individus se manifestent à travers différentes modalités, plusieurs études ont montré que la fusion de ces modalités peut fournir des informations riches

sur les états affectifs, les rendant ainsi plus représentatifs [99]. Bien que la reconnaissance automatique de l'affect avec une approche unimodale atteigne des performances importantes et un succès remarquable, de nombreuses recherches ont confirmé que la fusion de deux ou plusieurs modalités améliore les performances et la précision par rapport à l'approche unimodale [10].

L'extraction de caractéristiques est une étape cruciale de la reconnaissance de l'affect, qu'il soit unimodale ou multimodale. Le procédé d'extraction des caractéristiques et d'apprentissage pour la reconnaissance de l'affect unimodale peut être utilisé pour la reconnaissance multimodale [10].

En général, il existe deux types de fusion : la fusion précoce et la fusion tardive [100]. La fusion précoce est basée sur la fusion des caractéristiques de chaque modalité en un seul vecteur, tandis que la fusion tardive est basée sur la fusion décisionnelle de chaque modalité, où les caractéristiques de chaque modalité sont extraites et classées séparément [10].

Dans la littérature, les expressions faciales sont la modalité la plus utilisée dans les systèmes multimodaux fusionnant avec les signaux physiologiques [101], la posture [102] et largement utilisées avec la modalité audio [103], [104].

Dans ce contexte, Huang et al. [105] ont réalisé une fusion des signaux EEG et les expressions faciales (EF). Ils ont adopté une approche où les EF étaient classifiés par un réseau de neurones entièrement connecté tandis que les signaux EEG étaient classifiés par un classifieur de type séparateur à vaste marge (Support vector machines SVM). De même, Zhu et al. [106] ont obtenu une précision de 78,47 % pour la valence et de 72,74 % pour l'arousal. Leur approche impliquait l'utilisation d'une architecture de type 1D-CNN pour extraire les caractéristiques des signaux EEG et un réseau de neurones convolutifs en deux dimensions (2D-CNN) pour les EF.

Un système profond en plusieurs étapes a été proposé par Lie et al. [107]. Dans leur travail, ils ont utilisé l'ensemble de données RECOLA [108] et ont obtenu une précision de 59,7% pour la valence et de 61,9% pour l'arousal.

De plus, Saffaryazdi et al. [109] ont étudié la fusion des EF avec différents types de signaux physiologiques tels que l'EEG, le GSR et le PPG. Ils ont fusionné les signaux PPG et GSR au stade précoce, puis ont ajouté les EF et les signaux EEG avant de les fusionner au niveau décisionnel.

Ils ont obtenu une précision de 60,1 % pour la valence et de 60,9 % pour l'arousal en utilisant la base de données DEAP [61].

Une fusion tardive a été proposée par Chang et al. [110] pour classifier quatre émotions en fonction des signaux VFC, GSR et de la température cutanée et les EF. ils ont utilisé deux réseaux neuronaux de quantification vectorielle d'apprentissage (en anglais learning vector quantization LVQ) pour la classification des émotions avant la fusion. Un taux de reconnaissance élevé atteint 95% dans cette étude.

Oh et Kim [101] ont proposé une méthode de classification distinguant les émotions positives et négatives en fusionnant les EF avec les signaux VFC. Ils ont utilisé un 1D-CNN pour classer les émotions à partir des signaux VFC et un 2D-CNN pour les expressions faciales. Les caractéristiques de chaque modalité ont été fusionnées, ce qui a donné une précision de 86,2%.

Pour classifier le niveau de stress, Seo et al. ont développé un système multimodale qui fusionne les EF avec les signaux respiratoires et les signaux ECG. Ils ont utilisé une approche de fusion précoce pour combiner ces modalités d'entrée. Leur système a atteint une précision moyenne de 73,3% dans la classification du niveau de stress [111].

Xiang et al. [112] ont créé une base de données composée de vidéos faciales et de signaux physiologiques tels que VFC, EDA, Temp et le IBI de la VFC pour des conducteurs. Des réseaux de neurones convolutifs spatio-temporels ont été appliqués pour chaque modalité avant la fusion. Les résultats démontrent que la multimodalité améliore considérablement la précision avec des augmentations de précision de 11,28% et 6,83% par rapport à l'utilisation uniquement des vidéos faciale ou de signaux physiologiques, respectivement.

Le tableau présenté 2.3 synthétise les travaux récents de la littérature sur la classification des émotions humaines en fusionnant les expressions faciales et les signaux physiologiques à l'aide de techniques d'apprentissage profond. Il répertorie les différentes méthodes employées, les bases de données utilisées, ainsi que les résultats obtenus en termes de précision. Étant donné notre objectif de développer un système multimodal basé sur les expressions faciales et les signaux physiologiques, notre étude de la littérature met en lumière les différentes techniques récentes de fusion de ces deux modalités. Il est évident que, dans une approche multimodale, l'efficacité de la classification repose en grande partie sur la méthode et la technique d'inté-

gration des caractéristiques extraites de chaque modalité, ce qui influe directement sur les performances globales du système.

Tableau 2.3 – Travaux connexes sur la reconnaissance multimodale de l'affect (Expressions faciales + Signaux physiologiques)

| Auteurs | Modalités | Base de données | Précision |
|---|---|---|---|
| Huang et al. [105] | EF + EEG | MAHNOB-HCI, DEAP | Valence : 75% Arousal : 75.63%, Valence : 80.30% Arousal : 74.23% |
| Zhu et al. [106] | EF + EEG | DEAP | Valence : 78.47% Arousal : 72.74% |
| Lie et al. [107] | EF + ECG, VFC, FC, GSR, SCR, SCL | RECOLA | Valence : 59,7%, Arousal : 61,9% |
| Saffaryazdi et al. [109] | EF+ EEG, GSR, PPG | DEAP | Valence : 60.1%, Arousal : 60,9% |
| Chang et al. [110] | EF+ VFC+ GSR | 06 Participants | 95% |
| Oh et Kim [101] | FE+ VFC | 53 Participants | 86.2% |
| Seo et al. [111] | EF+ECG | 24 Participants | 73.3% |
| Xiang et al. [112] | EF+VFC | 24 Participants | 83.84% |

## 2.4 Conclusion

Ce chapitre a été conçu dans le but d'établir une base solide en présentant de manière générale les concepts fondamentaux liés aux thèmes pertinents pour notre recherche. Nous avons d'abord exploré les notions d'émotion et de stress, mettant en lumière leurs diverses composantes pour une meilleure compréhension.

Par la suite, une vue d'ensemble du domaine de l'apprentissage profond a été abordée, soulignant les différentes architectures que nous avons exploitées dans nos études. Enfin, nous avons passé en revue plusieurs méthodes récentes de la littérature concernant la reconnaissance des émotions et du stress par l'apprentissage profond à travers les expressions faciales et les signaux physiologiques, ainsi que la fusion de ces deux modalités. L'objectif de cette revue était de présenter les diverses contributions et méthodes proposées par différents chercheurs, incluant les techniques de prétraitement des données, les différentes architectures d'apprentissage profond proposées, ainsi que les résultats obtenus. Cette démarche nous offre un aperçu général des développements récents dans ce domaine, afin de les développer dans nos travaux.

# Chapitre 3

# Mesure de la fréquence cardiaque à partir des vidéos faciales : application à la reconnaissance de l'affect

## 3.1 Introduction

Ce chapitre est dédié à l'étude des signaux physiologiques liés à la réaction cardiaque, avec un accent particulier sur l'extraction de la fréquence cardiaque à partir de vidéos faciales RVB. Nous commencerons par une explication du fonctionnement autonome du cœur, suivi d'une exploration des différents paramètres cardiaques mesurables. Ensuite, nous présenterons en détail la technique de mesure du signal photopléthysmographique par imagerie (iPPG) et son utilisation pour calculer la fréquence cardiaque à partir des vidéos faciales. Enfin, nous présenterons les travaux récents effectués dans le domaine de la reconnaissance automatique d'affects, en se basant sur l'utilisation des signaux photopléthysmographiques par imagerie iPPG et de la variabilité de la fréquence cardiaque par imagerie iVFC.

## 3.2 Principe de l'activité autonome du cœur

Le cœur est un organe musculaire complexe qui présente un mécanisme de contraction et de relaxation qui joue un rôle important dans la variation de rythme cardiaque, c'est-à-dire le rythme à laquelle le muscle pompe le sang dans le corps. Le cerveau et le cœur sont connectés et influencés l'un par l'autre via le système nerveux autonome (SNA), et il existe deux types de connexions : la connexion du système nerveux sympathique (SNS) et la connexion du système nerveux parasympathique (SNP) [113].

Au repos et sans stimulation, les centres de contrôle de la fréquence cardiaque du système nerveux fournissent une petite quantité de stimulation au muscle cardiaque et lui donnent un rythme autonome. Cependant, lors de l'excitation, les centres de contrôle du système nerveux (SN) libèrent de neurotransmetteur noradrénaline et provoquent une augmentation de la fréquence cardiaque (FC). Quant à la diminution de la FC, les centres de contrôle du SN libèrent de neurotransmetteur acétylcholine dans le SNP, assurant ainsi un équilibre régulateur des composants fonctionnels, physiologiques, autonomes et parasympathiques [113], [114].

Cette activité cardiaque génère un courant électrique qui est mesuré par des électrodes placées sur la peau du corps humain, plus précisément sur la paroi thoracique. Ce signal est appelé électrocardiogramme (ECG). De plus, pour chaque cycle cardiaque au cours duquel le cœur pompe le sang dans tout le corps, la variation du volume sanguin dans les vaisseaux sanguins est mesurée sous la forme d'un signal appelé photopléthysmogramme (PPG), qui est

réalisée en mesurant la variation de l'intensité de la lumière réfléchie par la peau grâce à des capteurs placés sur l'oreille ou le doigt [113]. La figure 3.1 présente l'activité cardiaque et sa relation avec le signal PPG et ECG.

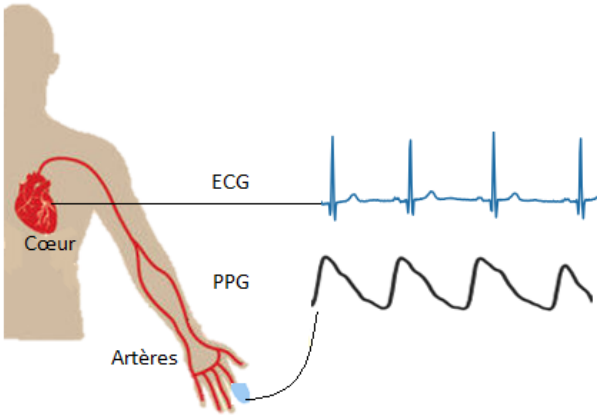



Figure 3.1 – Relation visuelle entre l'électrocardiogramme (ECG) et la photopléthysmographie (PPG) dans le corps humain [115]

## 3.3 Mesure des paramètres cardiaque

La mesure des paramètres cardiaques constitue l'un des aspects essentiels fournissant des informations cruciales sur le fonctionnement cardiaque. Dans ce domaine, de nombreuses techniques ont été élaborées afin d'obtenir ces paramètres physiologiques avec une précision accrue.

### 3.3.1 Électrocardiographie

L'électrocardiographie est un signal électrique qui reflète l'activité du cœur en fonction du temps. La fréquence cardiaque et stimulée par l'interaction des cellules myocardiques, donnant au cœur le potentiel de pomper le sang dans l'organisme.

La dépolarisation des cellules du myocarde de chaque battement cardiaque entraîne des changements électriques qui sont le signal ECG et qui mesurée par des électrodes placées sur la peau. L'impulsion électrique est générée à l'inversion de la polarité électrique de la paroi cellulaire cardiaque, qui est chargée positivement à l'état de repos, ensuite cette inversion produit une négativité sur la surface externe de la paroi cellulaire, qui se propage sous forme d'impulsion au tissu cardiaque adjacent [116].

Le signal ECG est composé de trois composants principaux marqués par les lettres P, Q, R, S, T et parfois, il utilise un autre composant appelé U. L'onde P représente l'activation des cavités supérieures du cœur, les oreillettes, tandis que le complexe QRS et l'onde T représentent l'excitation des ventricules ou de la cavité inférieure du cœur [117]. L'électrocardiogramme peut fournir une indication du rythme de tous les battements cardiaques, ainsi que de la santé du muscle cardiaque. Figure 3.2 présente le tracé d'une impulsion de l'ECG ainsi ses compositions.

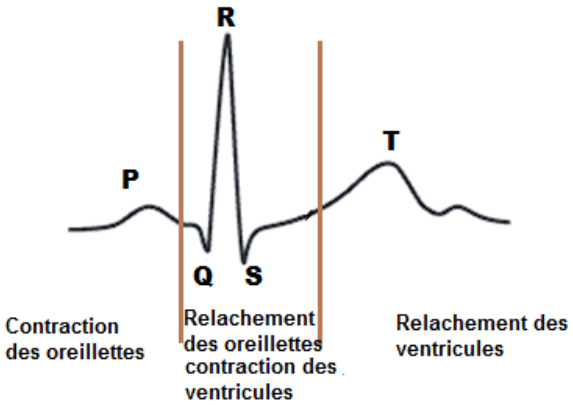



Figure 3.2 – Électrocardiogramme : Représentation graphique et composition

### 3.3.2 Photopléthysmographie

La photopléthysmographie est une technique optique permettant de mesurer l'activité cardiaque en estimant le volume de sang dans les tissus. Elle contient des informations importantes sur le système nerveux, cardiovasculaire et respiratoire. Ces dernières années, la photopléthysmographie est l'un des signaux les plus utilisés pour surveiller l'état de santé cardiovasculaire et respiratoire des patients, ainsi que pour la détection automatique des émotions humaines et du stress en raison de sa capacité à effectuer des analyses continues, peu coûteuses et faciles à réaliser [118].

En 1937, Hertzman et al. [119] ont décrit pour la première fois la mesure de la fluctuation du sang dans le doigt. Puisque le sang absorbe plus de lumière que les tissus, la variation du flux sanguin peut être capturée par des capteurs optiques qui mesurent les différences d'intensité lumineuse. Les variations de volume sont générées par des variations de pression dans les artères, qui se manifestent tout au long du cycle cardiaque.

Le système PPG se compose principalement de composants optoélectroniques qui

éclairent le tissu ainsi que d'un photodétecteur qui convertit les changements radiatifs représentant la fluctuation du volume sanguin artériel qui se produit dans une partie dans le corps comme le doigt [120]. Il existe deux types de fonctionnement des capteurs PPG : la transmission ou la réflexion de la lumière. Le fonctionnement en transmission dans lequel le module d'émission et le photodétecteur sont situés opposés et quand le fonctionnement en réflexion, le module d'émission est situé du même côté avec le photodétecteur [121]. Figure 3.3 montre le système PPG en contact, et la figure 3.4 représente le tracé de signal PPG avec ses compositions.

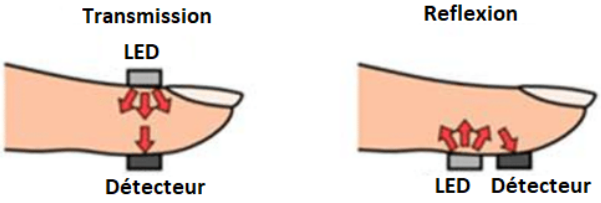



Figure 3.3 – Fonctionnement du capteur de photopléthysmographie [122]

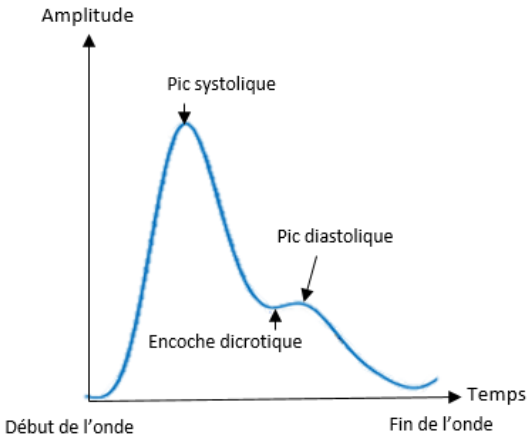



Figure 3.4 – Tracé et composition d'un photopléthysmogramme.

### 3.3.3 Fréquence cardiaque et sa variabilité

La fréquence cardiaque est le nombre de contractions et de relaxations du cœur humain par minute. C'est un indicateur instable et énergétique, et ses fluctuations transmettent des informations importantes sur la santé, l'état émotionnel et physique de la personne [123]. Il a été démontré que la fréquence cardiaque n'est pas constante et qu'il existe plusieurs facteurs qui affectent cet indicateur, tels que : l'âge, le sexe, l'état émotionnel, les maladies et autres. Par

exemple, la fréquence cardiaque au repos chez les adultes en bonne santé se situe entre 60 et 100 battements par minute, ce qui varie par rapport aux enfants qui est plus élevée [114]. La variabilité de la fréquence cardiaque (VFC) correspond à la variation temporelle de la période entre des battements cardiaques successifs. La VFC constitue un indicateur important qui reflète la capacité du cœur à réagir et à s'adapter à des situations inattendues ainsi qu'à divers stimuli [114]. De nombreuses études de recherche ont été menées ces dernières années pour examiner la relation entre le système nerveux autonome (SNA) et la mortalité cardiovasculaire, ainsi que les problèmes cardiaques [124], [125].

Dans le domaine de la reconnaissance automatique des émotions humain, VFC est l'un des indicateurs importants présents des changements dans différentes émotions. Zhu et al. [126] résument et présentent l'interaction entre la VFC et les états émotionnels des humains. Shi et al. [127] ont étudié les différences dans les signaux VFC pour deux émotions : le bonheur et la tristesse, auprès de 48 participants. Les résultats montrent que la fréquence cardiaque moyenne dans le bonheur est plus élevée que dans la tristesse. Selon Valderas et al. [128], qui ont étudié les différences dans les signaux de variabilité de la fréquence cardiaque (VFC), de bonheur, de peur et de calme chez 25 participants, ils ont découvert que la fréquence cardiaque moyenne augmentait avec les émotions négatives.

La VFC est généralement calculée en analysant une série d'intervalles d'impulsions de signaux ECG ou PPG et cela est effectué par des méthodes linéaires dans le domaine fréquentiel et temporel [129].

### A Domaine Temporel

La mesure de la variabilité du rythme cardiaque dans le domaine temporel est largement utilisée par l'analyse de spectrale de l'intervalle R-R à partir du signal ECG et P-P à partir du signal PPG [130]. L'intervalle R-R ou bien P-P est calculé comme l'intervalle de temps entre deux pics successifs comme le montre dans la figure 3.5. Ces intervalles sont abrégés par IBI, ce qui signifie Inter-beat Interval en anglais.

Il existe plusieurs indicateurs statistiques pour le domaine temporel présente l'activité sympathique et parasympathique du cœur, tels que : l'écart type de tous les intervalles SDNN, l'écart type des moyennes des intervalles pour toutes les 5 min SDANN, moyenne de l'écart type des intervalles toutes les 5 min, moyenne quadratique des différences entre les intervalles

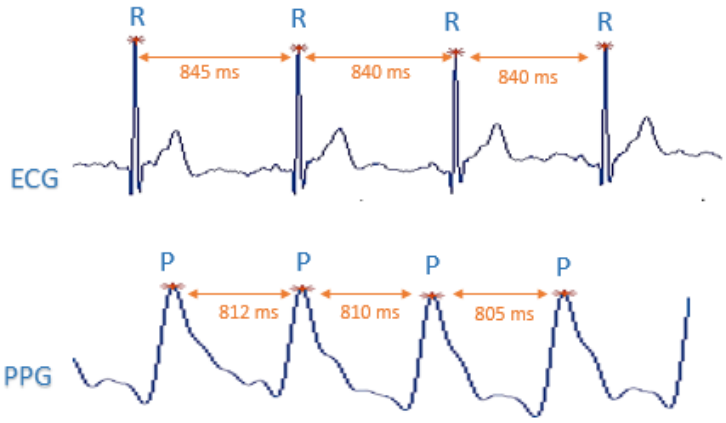



Figure 3.5 – Variabilité de la fréquence cardiaque en fonction de la variation temporelle entre les intervalles R-R pour un signal ECG ou P-P pour un signal PPG.

adjacents rMSSD, et le pourcentage des intervalles adjacents pNN50 [131].

Il existe d'autres méthodes dans le domaine temporel pour mesurer la VFC appelées méthodes géométriques, qui permettent la présentation des intervalles de battements cardiaques systole et diastole du cœur tels que : Index triangulaire (RRtri), Interpolation triangulaire des intervalles R-R (TINN) et graphe de Poincaré [131].

### B Domaine Fréquentiel

Une autre approche linéaire est envisageable dans le domaine fréquentiel pour mesurer la variabilité de la fréquence cardiaque. La densité spectrale de puissance (DSP) est la méthode la plus fréquemment utilisée qui permet d'évaluer la façon dont la puissance est distribuée en fonction de la fréquence. Cette analyse est effectuée en utilisant des algorithmes mathématiques. Le calcul de la DSP est basé sur des méthodes paramétriques et non paramétriques (modèle autorégressif et la technique de Transformée de Fourier) [132].

L'analyse du domaine fréquentiel est divisée en trois bandes de fréquences distinctes, connues sous le nom de composantes spectrales [132].

- De 0.15 Hz à 0.40 Hz est la bande de la haute fréquence ajustée par le système nerveux parasympathique et générée par la respiration.
- De 0.04 Hz à 0.15 Hz est la bande de basse fréquence ajustée par le système nerveux parasympathique et sympathique.
- De 0.003 Hz à 0.04 Hz est la bande de très basse fréquence ajustée par le système nerveux

sympathique.

## 3.4 Mesure de photopléthysmographie à distance

Avec tous les progrès réalisés dans le domaine de l'acquisition de paramètres physiologiques comme la fréquence cardiaque, les techniques présentées dans les sections précédentes posent plusieurs difficultés, dont la gêne des sujets par les capteurs placés sur le corps, ainsi que le problème des mouvements affectant la précision des mesures. L'utilisation de capteurs tels que les oxymètres nécessite un nettoyage après chaque utilisation, et l'oubli de cette étape de prévention peut conduire à des contaminations, peuvent conduire à des maladies. De plus, la mesure du signal ECG peut parfois poser des problèmes avec les nourrissons et les patients brûlés ou avec des peaux fragiles et irrités [133]. Afin de surmonter ces problèmes et depuis 2008, les chercheurs ont adopté une nouvelle approche d'étude qui s'articule autour du développement de nouvelles méthodes d'enregistrement des signaux physiologiques sans aucun contact avec le corps humain à l'aide de caméras simples et peu coûteuses.

La photopléthysmographie par imagerie iPPG, également connue sous le nom de photopléthysmographie sans contact, est une technique qui reprend les mêmes principes que la photopléthysmographie en contact, mais utilisant la lumière ambiante comme source de lumière et une caméra comme photorécepteur [134]. Il existe différents types de caméras, mais afin de rendre la technologie iPPG moins chère et applicable, il est préférable d'utiliser des caméras simples, que l'on retrouve dans diverses technologies disponibles dans le monde.

La technique utilisée est basée sur l'extraction du signal de photopléthysmographie PPG en mesurant les changements de couleur rouge-vert-bleu (RVB) du visage d'une personne au cours du cycle cardiaque [14]. L'extraction du signal iPPG à l'aide d'une simple caméra est considérée comme un développement technologique qui ouvre des horizons prometteurs dans divers domaines.

En médecine, la mesure des paramètres physiologiques constitue l'une des étapes importantes du suivi des patients et de la détection des maladies cardiovasculaires et autres [135]. Ces dernières années, la technologie iPPG a été largement utilisée pour éviter de déranger les personnes avec les fils des capteurs et améliorer le confort des patients [136]. La pandémie de COVID-19 a mis en évidence le rôle essentiel de l'extraction du signal iPPG dans la détection

précoce, le diagnostic et la surveillance de la maladie, en permettant une mesure à distance des fréquences cardiaque et respiratoire [137].

Dans la reconnaissance automatique des émotions, l'activité cardiaque est l'un des indices qui indiquent l'état émotionnel et le stress. Ces dernières années, les chercheurs attirent vers l'utilisation des signaux iPPG dans la reconnaissance automatique des émotions, cette technique est prometteuse et utile dans ce domaine à cause de la facilité d'obtenir les données [138].

Un autre domaine suscitant un intérêt important pour l'utilisation des signaux iPPG est celui des deepfakes ou de l'anti-falsification du visage. Cette forme de fraude implique l'utilisation de faux visages pour commettre des actes illégaux. Le iPPG est un signal calculé à partir des variations de couleur RVB de la peau du visage au fil du temps, ce qui le distingue des vidéos contenant des visages truqués [139]

Les méthodes d'extraction des signaux iPPG reposent généralement sur quatre étapes essentielles : la détection des visages et l'extraction des signaux RVB, la combinaison des signaux RVB pour former le signal iPPG, et enfin le filtrage et l'extraction des paramètres physiologiques tels que la fréquence cardiaque et la fréquence respiratoire [14]. La figure ci-dessous 3.6 illustre comment l'extraction iPPG est réalisée à l'aide d'une caméra.

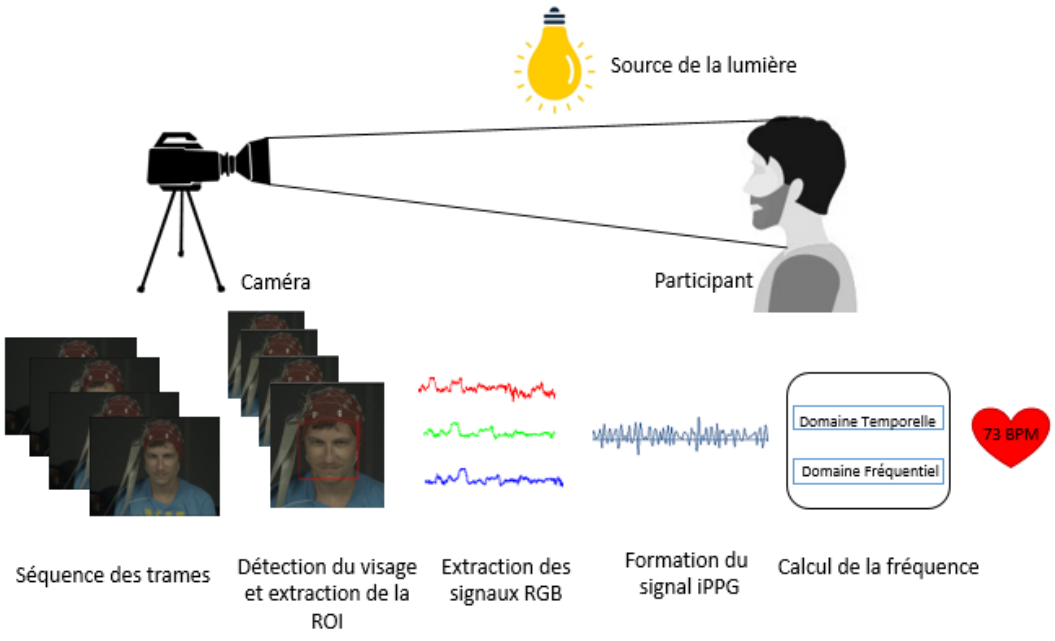



Figure 3.6 – Principe général d'éxtraction du signal iPPG à l'aide d'une caméra

### 3.4.1 Sélection de la région d'intérêt et l'extraction des signaux RVB

Le iPPG repose sur la mesure du volume sanguin en analysant la lumière réfléchie par la peau. Il est donc crucial de déterminer la région d'intérêt (ROI) qui contient les indications majeures de la photopléthysmographie iPPG [140]. Le choix du ROI est la première étape indispensable pour extraire le signal iPPG. Les recherches dans ce domaine sont axées sur l'utilisation du visage comme ROI, car il est le plus visible et rarement caché par des vêtements [141]. Il existe de nombreux algorithmes d'extraction de visages, tels que Viola et Jones [142], MTCNN [143] et Dlib [144] ainsi que des algorithmes de détection des points du visage [145] et des pixels de la peau [146]. Pour surmonter le problème de mouvement de la tête, l'algorithme Kanade-Lucas-Tomasi est utilisé [147]. La détection des visages est effectuée dans chaque trame de la vidéo, suivie de la suppression de l'arrière-plan à l'aide d'une technique de recadrage pour réduire les données de traitement et le bruit.

Il a été démontré dans plusieurs études que l'utilisation du visage entier introduisait du bruit dans le signal iPPG, provenant des lunettes, des sourcils ou même de la barbe, incitant les chercheurs à utiliser une zone du visage contenant que les pixels de la peau au lieu du visage complet [148]. Certaines recherches ont suggéré que le front est la zone la plus propice à l'extraction du signal iPPG, tandis que d'autres ont indiqué que les joues et le nez étaient plus adaptés à cette utilisation [149]. De plus, d'autres études ont montré que les lèvres transportent davantage de signaux iPPG que d'autres parties du visage, mais que ces signaux sont sujets à du bruit lors de mouvements faciaux ou de la parole [150].

Après avoir déterminé la ROI, une analyse spatiale des pixels de peau est appliquée afin d'obtenir trois signaux correspondant à trois couleurs : rouge, vert et bleu RVB. Ensuite, ces signaux sont combinés pour obtenir un seul signal, qui est le signal iPPG.

Avant de calculer le signal iPPG, une étape importante consiste à traiter les signaux RVB bruts afin de réduire le bruit causé par le contraste lumineux, les mouvements de tête ou les expressions faciales. Cette étape vise également à conserver uniquement les informations pertinentes pour le signal iPPG, ce qui permet d'augmenter le rapport signal sur bruit (signal-to-noise ratio SNR) et de fournir un signal de meilleure qualité.

Le filtrage passe-bande est l'une des opérations les plus appliqués par les chercheurs, afin de supprimer toutes les fréquences en dehors de la bande passante $[0.7 \quad 2.5]$ Hz qui correspond

à la plage de fréquence cardiaque [42 150] BPM. Plusieurs filtres passe-bande sont proposés tels que [14] :

- Le filtre moyen mobile (Moving average MA) permet de supprimer les hautes fréquences du signal, tout en éliminant les bruits d'errance et les artefacts de contenu mouvement qui peut provoquer par l'utilisateur.
- Le filtre Butterworth IIR est utilisé pour améliorer les performances de détection des pics.
- Le filtre FIR à base de la fenêtre de Hamming est très performant pour atténuer le bruit haut fréquence.

### 3.4.2 Formation du signal iPPG

Après avoir extrait et filtré les signaux de couleur RVB, une combinaison linéaire est effectuée pour obtenir un signal iPPG unidimensionnel précis avec moins de bruit. Plusieurs méthodes ont été proposées et développées à cet effet, certaines étant basées sur les propriétés statistiques des signaux RVB, tandis que d'autres s'appuient sur leurs propriétés physiques, correspondant à l'interaction de la lumière avec les tissus.

Les algorithmes de séparation aveuglent de sources (BSS) était proposée par Wedekind et al. [151]. L'objectif de ces algorithmes est de séparer le contenu du signal iPPG souhaité du bruit en raison de son indépendance statistique et de sa corrélation [152]. Deux algorithmes BSS qui présentent un grand intérêt à utiliser et à développer par les chercheurs dans ce domaine sont : Analyse en composantes indépendantes (Independent component analysis ICA) et analyse en composantes principales (Principal component analysis PCA).

Il existe aussi des algorithmes d'extraction de signaux iPPG qui impliquent l'utilisation de méthodes exploitant l'interaction entre les tissus et la lumière, par exemple Green, CHROM et POS.

### 3.4.3 Mesure de la fréquence cardiaque

Une fois le signal iPPG extrait, il est possible d'estimer la fréquence cardiaque soit par l'analyse fréquentielle, soit par l'analyse temporelle [153].

Le principe de calcul de FC dans le domaine fréquentiel est basé sur le calcul de la densité spectrale de puissance (DSP) du signal iPPG, qui permet de répartir le signal de puissance dans

le domaine fréquentiel, et la fréquence maximale est considérée comme une $F_C$ [153] :

$$FC = 60F_{\max}\,(\text{bpm}) \tag{3.1}$$

La DSP est généralement calculée par transformation de Fourier discrète (TFD) ou par modélisation autorégressive (AR) [153]. Tandis que, dans le domaine temporel, elle repose sur la détection de pics systoliques dans le signal iPPG qui consiste à déterminer la valeur maximale du signal d'onde de pouls [146]. Après avoir obtenu les pics systoliques, l'intervalle de temps entre les deux pics adjacents IBI (en anglais Inter-Beat-Interval) est déterminé et le signal IBI est obtenu [154]. L'inverse de la moyenne d'IBI est considéré comme FC [154] :

$$FC = \frac{60}{\text{moy}(IBI)}\,(\text{bpm}) \tag{3.2}$$

## 3.5 Reconnaissance automatique de l'affect à l'aide de signaux physiologiques sans contact

Avec les avancées réalisées dans ce domaine, l'extraction des signaux Photopléthysmographie par imagerie iPPG et fréquence cardiaque FC à partir de vidéos faciales a attiré l'attention des chercheurs en vue de l'utilisation de ces signaux dans la reconnaissance automatique de l'affect. Ce domaine de recherche est relativement récent, avec un nombre limité d'études disponibles à ce jour.

À cet égard, Kessler et al. [155] ont utilisé des signaux iPPG bruts pour classer la douleur en quatre niveaux ; ils ont atteint une précision de 65,79 % en utilisant le classifieur Random Forest RF. Récemment, Benezeth et al. [156] ont mené une étude comparative portant sur la variabilité de la fréquence cardiaque (VFC) mesurée sans contact et ses liens avec les états émotionnels. Leur recherche a mis en évidence une corrélation significative entre les états émotionnels et la VFC.

Zheng et ses collègues [157] ont confirmé que leur méthode d'estimation de la fréquence cardiaque sans contact peut fournir des informations émotionnelles pertinentes en visualisant les changements de variabilité de la fréquence cardiaque se produisant lors de la stimulation émotionnelle des participants.

Yu et al.[158] ont proposé une nouvelle méthode pour extraire les signaux iPPG basée

sur des techniques d'apprentissage profond. Ils ont ensuite employé ces signaux iPPG pour classifier les émotions selon deux échelles, la valence et l'arousal, en utilisant un classifieur de type séparateur à vaste marge (Support Vector machines SVM). Leurs résultats ont montré une précision de 46,86% pour la valence et de 44,02% pour l'arousal.

Pour la détection du stress, Maaoui et al. [159], ont utilisé les classifieurs de type séparateur à vaste marge SVM et l'analyse discriminante linéaire (en anglais linear discriminant analysis LDA) avec les signaux iPPG extrait à partir des vidéos faciales des 12 participants. Par ailleurs, Meziati Sabour et al. [160] ont créé une nouvelle base de données UFBC-Phys [160] dans le but de classifier le stress humain. Leur étude a démontré que la méthode plane orthogonale à la peau (POS) peut obtenir des performances remarquables dans l'extraction des signaux iPPG par des vidéos faciales. De plus, ils ont utilisé un classifieur SVM en se basant sur la iVFC, obtenant ainsi un taux de reconnaissance de 85,48%.

Le tableau 3.1 offre une synthèse des travaux référencés, présentant les signaux physiologiques sans contact exploités, les bases de données utilisées, ainsi que les précisions obtenues. L'utilisation de signaux physiologiques extraits des vidéos faciales, tels que l'iPPG et l'iVFC, pour la reconnaissance automatique de l'affect reste peu répandue, malgré l'avantage principal de cette technique qui permet d'extraire ces signaux sans perturber les individus avec des capteurs invasifs. De plus, la plupart des recherches se limitent à l'utilisation de méthodes classiques d'apprentissage automatique pour la classification. Par ailleurs, les résultats obtenus par Yu et al.[158] sont faibles, avec une précision qui ne dépasse pas 50% tant dans la classification de la valence que de l'arousal. Ceci souligne l'importance d'améliorer cette précision dans la classification des émotions dans nos études expérimentales.

Grâce au succès de l'approche multimodale, plusieurs études se concentrent sur le développement de systèmes physio-visuels combinant à la fois les expressions faciales et les signaux physiologiques sans contact extraits des vidéos faciales.

Yu et al. [98] ont utilisé un réseau de neurones convolutif tridimensionnel (3D-CNN) pour extraire des caractéristiques à partir de chaque modalité, à savoir les expressions faciales (EF) et les signaux iPPG. Ils ont ensuite appliqué une fusion précoce suivie de couches de réseau neuronal pour la classification des émotions, en exploitant la base de données DEAP [61]. De même, Ouzar et al. [161] ont proposé un modèle de reconnaissance multimodale des émotions basé sur des clips vidéo faciaux frontaux, en fusionnant les expressions faciales avec les signaux

Tableau 3.1 – Travaux connexes sur la reconnaissance automatique des émotions et du stress à l'aide de signaux physiologiques non contact

| Auteurs | Base de données | Signaux physiologiques sans contact | Méthode de classification | Précisions% |
|---|---|---|---|---|
| Yu et al.[158] | MAHNOB-HCI | iVFC | SVM | 46.86% pour valence 44.02% pour arousal |
| Maaoui et al. [159] | 12 particpants | iPPG | SVM, LDA | 94.40%–91.10% |
| Meziati Sabour et al. [160] | UBFC-Phys | iVFC | SVM | 85.84% |

iPPG et iVFC. Ils ont utilisé un réseau d'attention convolutif à décalage temporel multitâche (MTTS-CAN) pour extraire les signaux iPPG de la base de données multimodale des émotions spontanées BP4D+ [162]. Ces mêmes auteurs ont obtenu un taux de précision de 91,07% dans la classification du stress en combinant les EF avec la VFC sans contact, utilisant la base de données UBFC-Phys [138].

Le tableau 3.2 résume les recherches référencées dans le domaine de la reconnaissance physio-visuelle de l'affect à partir des vidéos faciales. Il est important de noter que ce sujet de recherche est encore relativement récent, avec un nombre limité d'études disponibles. Cependant, l'adoption d'approches multimodales, combinant les données provenant de différentes sources telles que les expressions faciales et les signaux iPPG et iVFC, semble considérablement améliorer la précision des systèmes de reconnaissance par rapport aux approches unimodales, comme le met en évidence le tableau 3.1. Cette augmentation de précision souligne l'importance de l'intégration de multiples modalités pour une classification plus précise de l'affect à partir des vidéos faciales.

Tableau 3.2 – Travaux connexes sur la reconnaissance multimodale de l'affect à partir des vidéos faciales

| Auteurs | Modalités | Base de données | Précision |
|---|---|---|---|
| Yu et al. [98] | FE + iPPG | DEAP | Coefficient de corrélation Valence : 0.25, Arousal : 0.10 |
| Ouzar et al. [161] | FE + iPPG, FE + iVFC | BP4+ | 70.59%, 71.90% |
| Ouzar et al. [138] | FE + iVFC | UBFC-Phys | 91.07% |

## 3.6 Conclusion

Ce chapitre a offert une vue d'ensemble du domaine de l'extraction de la fréquence cardiaque à partir des signaux iPPG extraits des vidéos faciales.

Nous avons débuté en détaillant le fonctionnement du cœur et les diverses mesures disponibles, incluant le signal ECG, le PPG, la FC et leur variabilité VFC.

Ensuite, nous avons exposé en profondeur la principale méthode d'extraction du signal iPPG à partir des vidéos faciales, en mettant en lumière les étapes impliquées.

La dernière partie de ce chapitre est consacrée à une revue de la littérature sur les travaux d'exploitation des signaux sans contact dans le domaine de la reconnaissance automatique d'affects. Ceci donne un aperçu des progrès récents dans la détection des émotions à l'aide de signaux physiologiques extraits sans contact de vidéos faciales, et met en évidence l'importance de ces progrès dans nos études et le potentiel d'amélioration.

Chapitre 4

# Reconnaissance des états affectifs via visage : Utilisation de l'apprentissage profond

## 4.1 Introduction

Ce chapitre est consacré à la présentation de nos études expérimentales menées dans cette thèse, qui visent à la reconnaissance automatique de l'affect (émotion ou stress). Nos expérimentations reposent sur l'utilisation de deux modalités distinctes : les expressions faciales et les signaux iPPG extraits de vidéos faciales, ainsi qu'une combinaison de ces deux modalités. Nos études s'articulent autour de deux axes principaux : le prétraitement des données, suivi de la proposition d'une architecture DL pour la classification.

## 4.2 Bases de données

Le succès des méthodes et algorithmes d'extraction des signaux iPPG à partir de vidéos faciales, ainsi que la détection automatique de l'affect, a largement bénéficié de la disponibilité de nombreuses bases de données.

Dans le cadre de notre étude, il est crucial d'utiliser des bases de données physiologiques et visuelles pour extraire le signal iPPG à partir des vidéos faciales et évaluer ces signaux par rapport aux signaux physiologiques prélevés directement sur le corps humain. En outre, il convient de tenir compte de l'utilisation de bases de données affectives dans le cadre d'une étude portant sur la détection des émotions et du stress par les signaux iPPG.

Dans le domaine de la reconnaissance automatique des émotions humaines basée sur les expressions faciales, plusieurs bases de données sont disponibles, qu'il s'agisse d'images ou de vidéos. La plupart de ces bases de données fournissent des expressions qui couvrent les émotions de base précédemment identifiées par Ekman : colère, tristesse, surprise, peur, dégoût, joie, neutralité. De plus, ils présentent de multiples défis, tels que l'âge, le sexe, la race, la position de la tête, le regard, les problèmes d'occlusion et la variabilité de l'éclairage.

Afin de réaliser nos contributions, nous avons choisi d'utiliser les bases de données suivantes : MAHNOB-HCI [62], UBFC-Phys [160] et RaFD [43].

### 4.2.1 MAHNOB-HCI

La base de données MAHNOB-HCI a été créée en 2011 par Soleymani et al. [62]. Elle représente une base de données multimodale (audio, visuelle et physiologique) développée

pour des recherches en détection des émotions. Cependant, elle est fréquemment utilisée pour évaluer les algorithmes et les techniques d'extraction des signaux iPPG et de FC à partir de vidéos faciales. Cette base de données contient des données de 30 participants de divers sexes et origines ethniques.

Chaque participant est invité à regarder attentivement une série de 20 vidéos sélectionnées spécifiquement pour susciter des émotions, tandis qu'en parallèle, ses réactions sont enregistrées et suivies à l'aide d'une caméra. En outre, des capteurs sont utilisés pour recueillir divers signaux physiologiques tels que le GSR (conductance cutanée), l'EEG (électroencéphalogramme), l'ECG (électrocardiogramme), la température de la peau (SKT) et la respiration (RSP). Les vidéos faciales enregistrées sont ensuite compressées avec une résolution de (780x580) pixels et 61 images par seconde. La figure 4.1 illustre le protocole expérimental utilisé pour construire la base de données MAHNOB-HCI.

Figure 4.1 – Le protocole expérimental utilisé pour construire la base de données MAHNOB-HCI [62]

À la fin de chaque test, les participants sont invités à donner leur avis sur ce qu'ils ont ressenti, par le test SAM (Self Assessment Manikin), sur une échelle de 1 à 9. Dans cette échelle, 1 représente l'arousal le plus bas et la valence la plus négative, et 9 l'arousal le plus élevé et la valence la plus positive. La figure 4.2 illustre cette échelle SAM. [62].

### 4.2.2 UBFC-Phys

La base de données UBFC-phys a été créée en 2021 par Sabour et al. [160]. Il s'agit d'une base de données multimodale (physio-visuelle) comprenant 56 participants de sexe masculin et

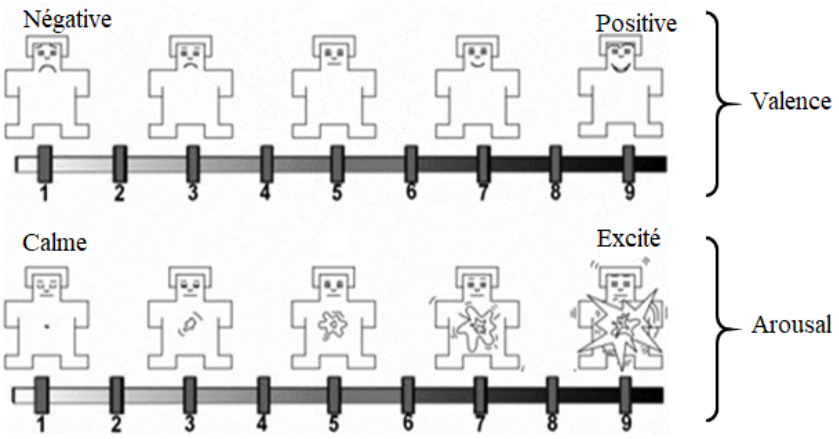



Figure 4.2 – Échelles SAM pour la valence et Arousal [60]

féminin. Les participants sont exposés à une expérience en trois tâches, une tâche de repos T1, une tâche d'expression T2 et une tâche arithmétique T3. Les participants à chaque tâche sont filmés et enregistrés en vidéo à une fréquence d'images de 35 fps et une résolution d'image de 1024 × 1024 pixels. De plus, les signaux physiologiques BVP et EDA sont également enregistrés à l'aide du bracelet Empathic E4.

Un aperçu du protocole expérimental utilisé pour construire la base de données UBFC-Phys est présenté dans la figure 4.3.

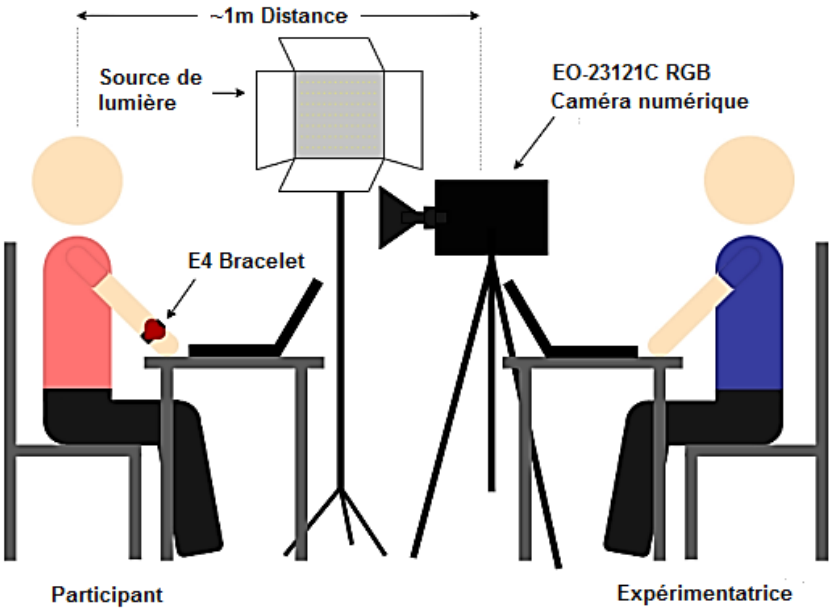



Figure 4.3 – Le protocole expérimental utilisé pour construire la base de données UBFC-Phys [160]

### 4.2.3 RAFD

La base de données RaFD [43] contient 8040 images d'expressions faciales mimiques, réparties en huit émotions : colère, tristesse, surprise, peur, dégoût, joie, neutre et mépris. Elle inclut des individus, qu'ils soient des hommes, des femmes ou des enfants, de diverses races. Images capturées sous différents angles d'exposition (0°, 45°, 90°, 135°, 180°) avec trois directions de regard telles que l'avant, la gauche et la droite. La base de données a été créée dans un environnement contrôlé dans un laboratoire.

Il y a 201 images pour représenter chaque émotion. La figure 4.4 présente un échantillon de la base de données RaFD avec les poses de tête suivantes : 180°, 135°, 90°, 45° et 0°.

Figure 4.4 – Échantillons d'images de la base de données RAFD illustrant différentes poses de tête : 180°, 135°, 90°, 45° et 0°[43]

## 4.3 Étude comparative des méthodes d'extraction de signaux iPPG dans différentes régions d'intérêt

L'objectif principal de cette étude est d'effectuer une analyse comparative des performances de quatre méthodes couramment utilisées pour l'extraction de signaux iPPG. Étant donné que la détection des régions d'intérêt est une étape cruciale dans ces méthodes, notre étude a porté sur deux différentes régions d'intérêt du visage en utilisant deux bases de données : UBFC-Phys et MAHNOB-HCI. Pour chaque base de données, nous avons inclus 10 sujets, et pour chaque sujet, nous avons utilisé une minute de chaque vidéo, ce qui équivaut à dix vidéos par base de données. La figure 4.5 donne un aperçu de notre étude comparative.

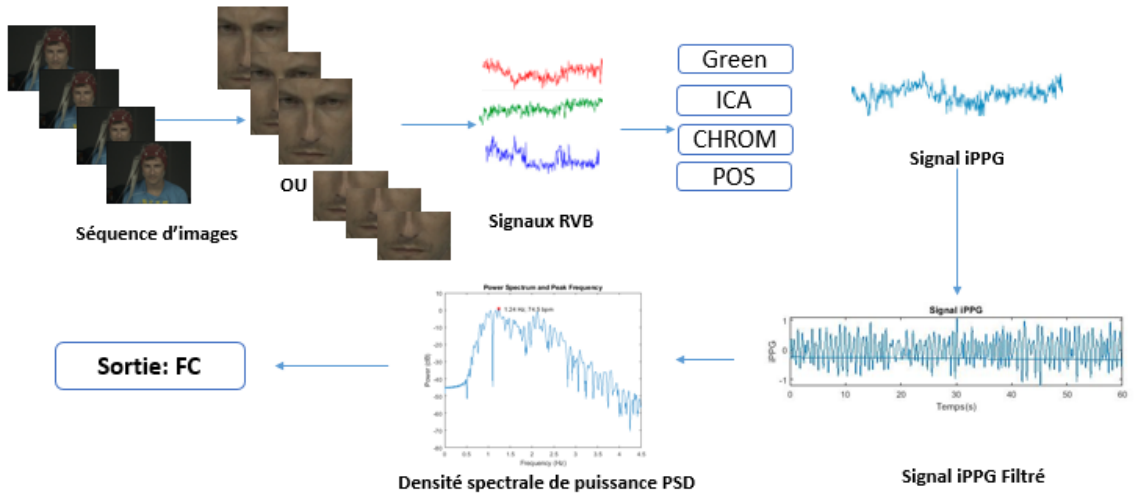



Figure 4.5 – Aperçu de méthode proposée

### 4.3.1 Sélection de la région d'intérêt

La première étape de la plupart des systèmes de mesure de la fréquence cardiaque basés sur la vidéo consiste à extraire la région d'intérêt ROI [134], [163].

Dans notre analyse comparative des méthodes d'extraction de signaux iPPG à partir de vidéos faciales, nous avons amélioré notre approche en incorporant deux types de régions d'intérêt ROI, tout en évaluant l'impact de ces choix sur la précision de la FC dans deux bases de données distinctes. Pour la première région d'intérêt, nous avons restreint l'analyse au visage, excluant ainsi l'arrière-plan et les cheveux. Pour la deuxième région d'intérêt, nous nous sommes concentrés uniquement sur la partie inférieure du visage, excluant les yeux et le front.

Afin d'obtenir la ROI souhaité, nous avons tout d'abord utilisé l'algorithme Viola et Jones [142] pour détecter les visages dans chaque image de la vidéo, puis nous l'avons recadré.

Dans la figure 4.6, on peut observer un échantillon des bases de données MAHNOB-HCI et UBFC-Phys, où l'on peut voir la différence entre les ROIs utilisés dans notre expérience.

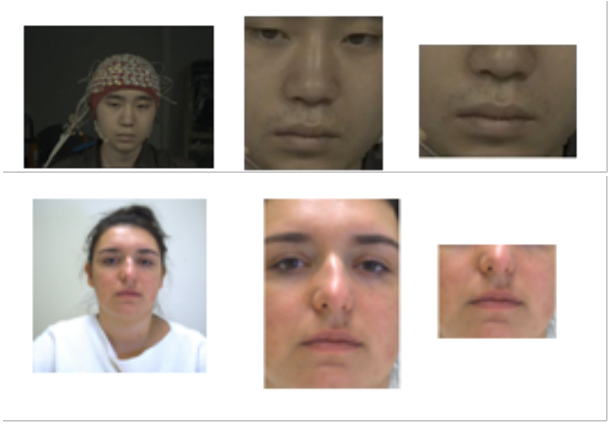


Figure 4.6 – Exemples d'images de la base de données MAHNOB-HCI [62]et UBFC-Phys [160] avec diverses régions d'intérêt (ROIs) utilisés.

### 4.3.2 Description de différents algorithmes d'extraction du signal iPPG utilisée

Après avoir détecté la ROI de chaque image, les signaux RVB sont extraits en calculant la moyenne des pixels des canaux de couleur rouge, vert et bleu. Ces signaux sont ensuite filtrés par un filtre passe-bande Butterworth de 3 ième ordre à phase nulle avec une plage de fréquences de [0.7 2.5] Hz.

Dans cette étude comparative, nous avons utilisé quatre algorithmes d'extraction de signal iPPG les plus couramment utilisés dans ce domaine, à savoir Analyse en composantes indépendantes ICA [134], Green [164], CHROM [165] et Plane Orthogonal-to Skin POS [166]. La mise en œuvre de cette étape de l'étude a été réalisée en utilisant le code MATLAB disponible sur GitHub, présenté par McDuff et Blackford [167].

#### A Analyse en composantes indépendantes ICA

Analyse en composantes indépendantes ICA est une méthode statistique, proposée par Poh et al. [134], qui vise à décomposer un mélange de signaux de couleur source en signaux indépendants. L'idée est basée sur l'hypothèse que le signal cardiaque est mélangé de manière linéaire dans les traces temporelles RVB et que les perturbations de la lumière, du bruit et du mouvement sont statistiquement indépendantes du signal cardiaque. Plusieurs algorithmes sont utilisés pour ICA tels que la maximisation de l'information mutuelle, la maximisation de la non-gaussianité ou des méthodes algébriques qui exploitent la structure des cumulants du quatrième ordre tels que : Diagonalisation d'une combinaison de matrices de covariance JADE

[168] et FastICA [157].

Dans notre étude, nous avons appliqué l'algorithme JADE, puis le signal iPPG obtenu est filtré par un filtre a transformé de Fourier rapide FFT.

### B Green

Proposé par Verkruysse et al. [164], qui utilise uniquement le signal de couleur vert qui est calculé à partir de l'intensité moyenne des pixels extraite de la ROI. Elle est considérée comme l'approche la plus simple et la moins coûteuse en termes de temps de calcul pour l'extraction du signal iPPG.

### C CHROM

En 2013, Haan et Jeanne [165] ont proposé la méthode CHROM, utilisant une combinaison linéaire des signaux de couleur RVB, permettant d'éliminer la composante spéculaire responsable de la présence de la couleur éclairée dans le signal iPPG. En bref, la lumière réfléchie par la peau est composée de deux composants : un élément à réflexion diffuse, dont les variations sont liées au cycle cardiaque, et un élément à réflexion spéculaire, qui montre la couleur de la peau et qui dépendent de l'angle entre la caméra et la peau, la luminosité et les mouvements de la personne devant la caméra, ce qui conduit à une mauvaise précision du signal iPPG extrait.

Les signaux de couleur RVB $X_r X_v X_b$ extraits de la ROI sont filtrés et en divisant par leur valeur moyenne. Les vecteurs de chrominance $Xc$ et $Yc$ sont calculés :

$$Xc = 3 * X_r - 2 * X_v \tag{4.1}$$

$$Yc = 1.5 * X_r + X_v + 1.5 * X_b \tag{4.2}$$

Le signal iPPG est calculé par la formule suivante :

$$\text{ippg} = Xc - \alpha * Yc$$

Tels que : $\alpha = \frac{\delta(Xc)}{\delta(Yc)}$, où $\delta$ représente l'écart type de $Xc$ et $Yc$.

**D POS**

La méthode (Plane Orthogonal-to Skin POS) a été récemment proposée par Wang et al. [166]. Dans le même but que la méthode CHROM. Après la normalisation temporelle des signaux de couleurs RVB $X_r, X_v, X_b$, une projection du signal sur le plan orthogonal à la peau est effectuée et les vecteurs $Xp$ et $Yp$ sont calculés de cette façon :

$$Xp = X_v - X_b \tag{4.3}$$

$$Yp = 2 * X_r + X_v + X_b \tag{4.4}$$

Le signal iPPG est calculé par la formule suivante :

$$\text{ippg} = Xp + \alpha * Yp$$

Tels que : $\alpha = \frac{\delta(Xp)}{\delta(Yp)}$, où $\delta$ représente l'écart type de $Xp$ et $Yp$.

### 4.3.3 Calcul de la fréquence cardiaque et les paramètres d'évaluation de notre étude

Après avoir sélectionné le signal iPPG, il est possible de calculer la fréquence cardiaque. Dans notre étude, nous avons utilisé la technique de densité spectrale de puissance (DSP), où la fréquence maximale du spectre correspond à la fréquence cardiaque. La figure 4.7 montre comment le signal iPPG est représenté et comment la fréquence cardiaque est estimée à partir du signal de densité spectrale de puissance.
Pour mener à bien notre étude comparative sur les différentes méthodes d'extraction de signaux iPPG à partir de vidéos faciales, il convient de faire une comparaison entre la fréquence cardiaque estimée obtenue par le signal iPPG $Fc_{\text{est}}$ et la fréquence cardiaque réelle calculée à partir des signaux capturés par le capteur $Fc_{\text{réel}}$.

Dans la base de données MAHNOB-HCI, les signaux ECG ont été enregistréss à partir de différentes parties du corps humain. Nous avons utilisé les données du canal EXG2, qui correspondent au coin supérieur gauche de la poitrine sous l'os de la clavicule. Les signaux ECG sont disponibles au format BDF avec une fréquence de 256 Hz.

Les signaux PPG sont accessibles dans la base de données UBFC-Phys grâce au capteur

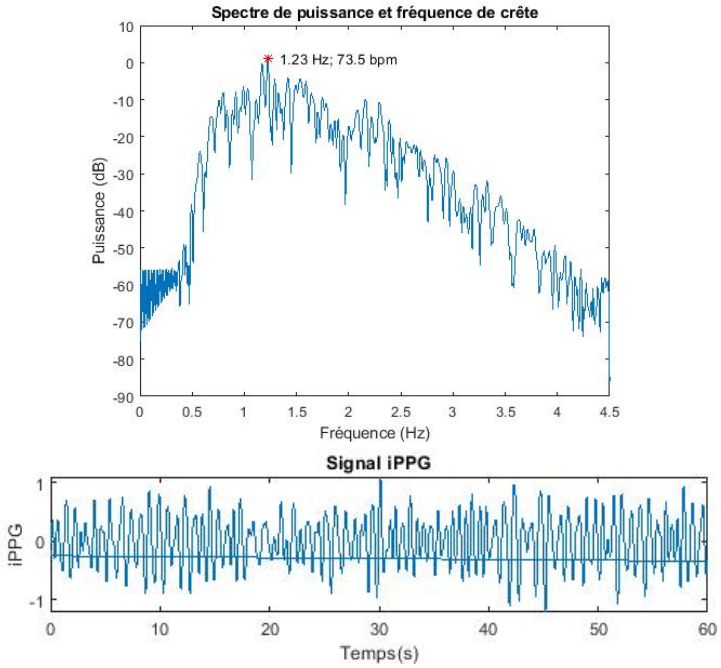



Figure 4.7 – Signal iPPG et FC estimée

Empathy E4, et ils ont une fréquence de 64 Hz.

La FC est calculée en suivant la méthode proposée par McDuff et Blackford [167], qui repose sur la détection des pics pour mesurer la différence moyenne entre deux pics IBI voir section 3.4.3.

Après avoir effectué tous les calculs, la dernière étape de cette expérience est de procéder aux mesures d'évaluation. Trois paramètres, tels que l'erreur absolue moyenne MAE, l'erreur quadratique moyenne RMSE et le rapport signal sur bruit SNR, peuvent être utilisés et sont calculés en fonction de $FC_{réel}$ et de $FC_{est}$. Voici les définitions de ces paramètres :

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|FC_{réel} - FC_{est}| \tag{4.5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(FC_{réel} - FC_{est})^2} \tag{4.6}$$

$$\tag{4.7}$$

ou $N$ : Nombre de vidéos.

Le rapport signal sur bruit $SNR$ est calculé selon la méthode décrite par Haan et Jeanne [165], et il s'agit d'un rapport énergétique qui englobe la fréquence cardiaque fondamentale, la première harmonique et l'énergie restante du spectre.

### 4.3.4 Résultats et discussions

Le but de cette étude est de comparer quatre approches d'extraction du signal iPPG à partir de vidéos faciales, en se concentrant sur deux zones différentes du visage. Dans une première étape, nous avons entamé notre comparaison en excluant la détection du ROI afin d'évaluer son influence sur l'extraction du signal iPPG ainsi que sur la précision de la fréquence cardiaque ($FC$) extraite.

Les tableaux 4.1 et 4.2 présentent les résultats obtenus en utilisant les bases de données MAHNOB-HCI et UBFC-Phys.

Tableau 4.1 – Les résultats obtenus avec la base de données MAHNOB-HCI sans détection du ROI

| Base de données | MAHNOB-HCI | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 19.53 | 26.72 | 20 | 17.16 |
| RMSE/bpm | 21.46 | 27.85 | 26.03 | 19.58 |
| SNR | -3.52 | -2.96 | -3 | -4.53 |

Tableau 4.2 – Les résultats obtenus avec la base de données UBFC-Phys sans détection du ROI

| Base de données | UBFC-Phys | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 11.41 | 2.7 | 3.8 | 3.87 |
| RMSE/bpm | 15.85 | 3.36 | 5.03 | 4.09 |
| SNR | -3.06 | -3.08 | -4.07 | -4.48 |

D'après les résultats présentés dans les tableaux 4.1 et 4.2, on constate clairement qu'avec l'absence de détection de la ROI, les performances sur la base de données UBFC-Phys sont nettement supérieures à celles de la base de données MAHNOB-HCI. Dans la base de données UBFC-Phys, la méthode ICA affiche la meilleure erreur moyenne avec une moyenne de 2,7 BPM, tandis que dans la base de données MAHNOB-HCI, la méthode POS enregistre la meilleure erreur moyenne avec une moyenne de 17,16 BPM. Ces variations significatives dans les résultats peuvent être attribuées à diverses conditions de création des bases de données, telles que le type de caméra utilisée, et les conditions d'éclairage.

Ensuite, nous procédons à une analyse comparative des diverses approches d'extraction des signaux iPPG en tenant compte de l'étape de détection du ROI. Dans la première approche,

nous utilisons la première ROI, englobant l'utilisation du visage tout en éliminant l'arrière-plan et les cheveux. Les résultats obtenus sont présentés dans les tableaux 4.3 et 4.4.

Tableau 4.3 – Les résultats obtenus avec la base de données MAHNOB-HCI avec la première ROI

| Base de données | MAHNOB-HCI | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 17.63 | 2.25 | 3.96 | 10.92 |
| RMSE/bpm | 19.69 | 3.49 | 5.49 | 15.49 |
| SNR | -5.88 | -2.86 | -3.28 | -5.37 |

Tableau 4.4 – Les résultats obtenus avec la base de données UBFC-Phys avec la première ROI

| Base de données | UBFC-Phys | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 4.46 | 1.64 | 1.94 | 1.9 |
| RMSE/bpm | 6.60 | 2.28 | 3.04 | 2.88 |
| SNR | -2.18 | -4.42 | -3.82 | -4.97 |

Nous avons également mené une analyse similaire en utilisant le deuxième ROI, qui cible spécifiquement la partie inférieure du visage, englobant la bouche et le nez. Les résultats de cette analyse sont présentés dans les tableaux 4.5 et 4.6.

Tableau 4.5 – Les résultats obtenus avec la base de données MAHNOB-HCI avec la deuxième ROI

| Base de données | MAHNOB-HCI | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 13.09 | 6.87 | 4.85 | 9.53 |
| RMSE/bpm | 17.18 | 10.39 | 5.90 | 13.35 |
| SNR | -4.95 | -23.16 | -3.22 | -5.25 |

D'après l'analyse statistique présentée dans les tableaux précédents (Voir les tableaux 4.3, 4.4, 4.5, 4.6), il est clairement perceptible qu'il existe un impact significatif de l'utilisation d'une région d'intérêt (ROI). En effet, la détection du ROI améliore la précision de l'extraction du signal iPPG, ce qui se traduit par une précision accrue de la FC. Par exemple, en utilisant la première ROI avec la base de données MAHNOB-HCI, l'erreur moyenne a diminué de 26,72 BPM à 2,25 BPM, et avec la base de données UBFC-Phys, l'erreur moyenne a diminué de 2,7 BPM à 1,64 BPM.

Tableau 4.6 – Les résultats obtenus avec la base de données UBFC-Phys avec la deuxième ROI

| Base de données | UBFC-Phys | | | |
|---|---|---|---|---|
| Méthodes | Green | ICA | CHROM | POS |
| MAE/bpm | 10.48 | 1.7 | 1.94 | 2.24 |
| RMSE/bpm | 16.03 | 3 | 3.04 | 2.11 |
| SNR | -2.23 | -4.58 | -3.87 | -4.7 |

Après avoir analysé les résultats des différents tests utilisant les deux types de ROIs, nous avons conclu que la méthode ICA est la plus efficace par rapport aux autres méthodes d'extraction du signal iPPG. Par ailleurs, en analysant les quatre méthodes, GREEN se distingue comme étant la moins efficace en termes de précision et d'efficacité sur les deux bases de données, que ce soit en utilisant le premier type de ROI ou le deuxième.

Avec la base de données UBFC-Phys, il est notable que les trois méthodes ICA, CHROM et POS présentent des résultats satisfaisants, avec des proportions très proches, que ce soit avec le premier ROI ou le deuxième. Ces résultats soulignent des aspects importants, notamment que les conditions d'imagerie et de création de la base de données UBFC-Phys sont meilleures que celles de la base de données MAHNOB-HCI.

Comme un résultat final, on peut constater que la méthode ICA avec le premier ROI qui consiste à utiliser le visage complet avec l'élimination de l'arrière plan et les cheveux du sujet (Voir la figure 4.6) est la plus efficace, que ce soit en utilisant les bases de données MAHNOB-HCI ou UBFC-Phys. C'est cette constatation qui nous a poussés à adopter cette méthode dans nos études expérimentales concernant l'extraction des signaux iPPG et leur utilisation dans la classification des émotions et du stress.

## 4.4 Reconnaissance automatique de l'émotion humaine à partir des signaux iPPG

Au cours des dernières années, les chercheurs ont évolué de l'utilisation de signaux physiologiques mesurés par des capteurs placés sur le corps humain dans la classification des émotions vers l'utilisation de signaux physiologiques extraits à partir de vidéos RVB des visages humains, tels que l'iPPG et l'iVFC.

L'objectif de cette recherche est de réaliser une classification des émotions humaines

sur deux axes (valence et arousal) à l'aide des signaux iPPG. Dans cette étude, nous avons fait appel à la base de données MAHNOB-HCI et avons privilégié l'utilisation de techniques d'apprentissage profond pour la classification. La figure 4.8 donne un aperçu de la méthode utilisée dans cette étude.

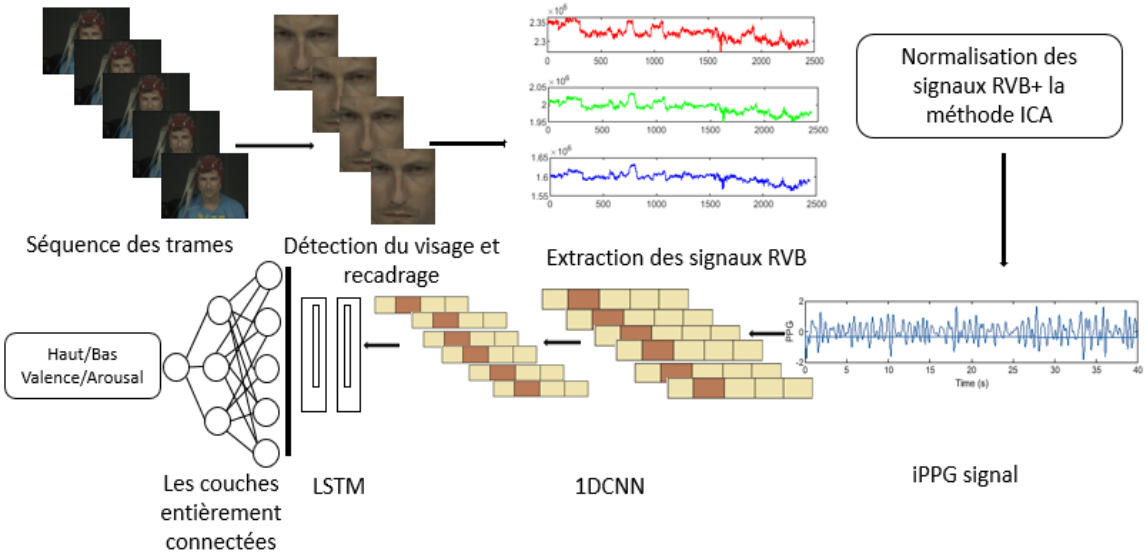



Figure 4.8 – Aperçu de la méthodologie expérimentale proposée

### 4.4.1 Collecte de signaux iPPG

Pour mener à bien cette étude, il a été considéré comme essentiel de créer une nouvelle base de données contenant des signaux iPPG précis extraits de vidéos RVB faciales. Afin d'atteindre cet objectif, nous avons utilisé des études comparatives préalablement réalisées pour obtenir une méthode efficace dans la base de données MAHNOB-HCI.

Boccignone et al. [14] ont réalisé une étude comparative et ont découvert que les méthodes ICA et PCA sont significativement plus efficaces avec la base de données MAHNOB-HCI que les méthodes Green, CHROM, POS, SSR, LGI et PBV.

Wang et al. [169] ont mené une étude comparative de trois méthodes, à savoir ICA, PCA et Green, sur la base de données MAHNOB-HCI. Il s'est avéré que la méthode ICA appliquée sur l'ensemble du visage ou sur le visage sans la bouche ni les yeux donne de meilleurs résultats que lorsqu'elle est appliquée à d'autres régions du visage, comme le front, le menton et les joues. La figure 4.9 présente l'évaluation des performances de chaque méthode d'extraction de

la FC sans contact sur la base de données MAHNOB-HCI étudiée par Wang et al. [169].

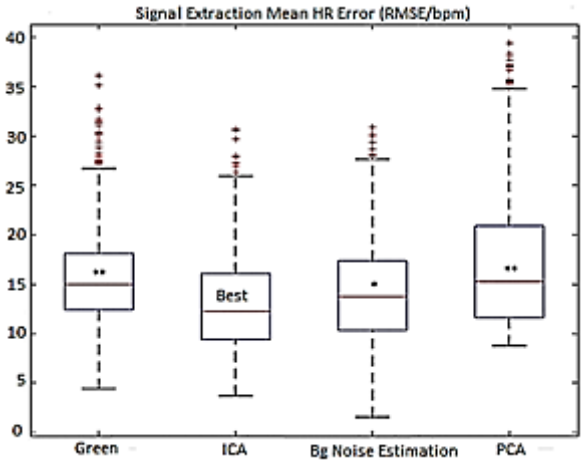



Figure 4.9 – L'erreur quadratique moyenne RMSE obtenue par Wang et al. [169] dans différentes méthodes d'extraction de la FC à distance.

De plus, nous avons mené une étude comparative des quatre méthodes d'extraction de signal iPPG les plus couramment utilisées sur deux régions faciales d'intérêt différentes, en utilisant la base de données MAHNOB-HCI, comme présenté dans la section précédente, 4.3. A partir de notre analyse comparative, nous avons observé que la méthode ICA, lorsqu'elle est appliquée à l'ensemble du visage avec l'élimination de l'arrière-plan et des cheveux, offre de meilleures performances que d'autres méthodes telles que Green, POS et CHROM.

Sur la base de ce qui précède, nous avons choisi d'adopter la même approche de manipulation de la méthode ICA décrite dans la section 4.3 de ce chapitre pour extraire les signaux iPPG.

Il est important de souligner que plusieurs étapes ont été suivies pour mettre en place notre nouvelle base de données, qui intègre les signaux iPPG. Dans un premier temps, dix sujets (participants) ont été téléchargés à partir de la base de données MAHNOB-HCI ; chaque sujet comprenant 20 vidéos de visages frontaux en couleur RVB. Ensuite, l'algorithme ICA a été appliquée à ces vidéos d'une durée de 40 secondes (Voir la section 4.3). La fréquence cardiaque FC obtenue a été comparée à celle calculée à partir du signal ECG disponible dans la base de données MAHNOB-HCI. Il convient de souligner que seuls les signaux iPPG fournissant une FC précise ont été collectés.

À noter que les signaux ECG utilisés issus de la base de données MAHNOB-HCI pro-

viennent du canal EXG2 correspondant au coin supérieur gauche du thorax sous l'os de la clavicule, au format BDF, avec une fréquence de 256 Hz. Deux méthodes différentes ont été appliquées pour calculer la $Fc_{\text{ver}}$. La première méthode utilisait la détection des pics pour calculer la différence de temps moyenne entre deux pics, tandis que la seconde utilisait la Biosppy Toolbox [170]. Il s'est avéré que les deux méthodes donnent des résultats similaires.

À la fin de cette phase de l'étude, à partir de 193 vidéos, nous avons extrait que 153 signaux iPPG précis avec $MAE = 1.55BPM$, $RMSE = 2.18BPM$ et $SNR = -3.70dB$. Les 40 signaux moins précis que nous avons supprimés.

Il est important de noter que l'objectif principal de cette étude est de catégoriser les émotions humaines en utilisant des signaux iPPG. La classification a été effectuée à l'aide de techniques DL. La figure 4.10 donne un aperçu des différentes étapes suivies dans l'étude proposée.


Figure 4.10 – Étapes suivies dans l'étude proposée

### 4.4.2 Définition des classes des émotions et prétraitement des données

Dans la base de données MAHNOB-HCI, après chaque expérience, les participants ont été invités à enregistrer leurs états émotionnels de deux manières différentes. Tout d'abord, ils l'ont fait en termes d'étiquettes discrètes telles que joie, plaisir, peur, colère, anxiété, tristesse, ennui, surprise et neutre. De plus, ils ont fourni une représentation dimensionnelle en termes de valence et d'arousal, selon le test SAM (Self Assessment Manikin), avec une distribution sur une échelle de 1 à 9, comme illustré dans la figure 4.2. La plage de 1 à 4 faits référence à des émotions de valences négatives et arousal faible, et celui de 5 à 9 correspond à une valence positive et Arousal élevé.

Avant d'effectuer la tâche de classification, deux techniques de prétraitement des données ont été utilisées, à savoir la normalisation et la segmentation des données. Le nouvel ensemble de données comprend des signaux iPPG provenant de différentes personnes. Il est nécessaire d'utiliser une technique de normalisation des données pour conserver les données à la même échelle.

La méthode de normalisation des données a été appliquée comme suit :

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{4.8}$$

Où $X$ est le signal iPPG et $X_{\text{norm}}$ est le signal iPPG de normalisé. Notez que l'objectif principal de cette technique est de supprimer la plus petite valeur du vecteur du signal, puis de diviser par l'étendue du vecteur. De cette façon, toutes nos valeurs sont comprises entre 0 et 1 ; la plus petite valeur est zéro et la plus grande est un. Il s'agit d'une étape importante qui a été largement utilisée par plusieurs chercheurs tels que Lee et al. [63], [171].

Chaque signal iPPG a été extrait d'un clip vidéo de 40 secondes, générant 2440 valeurs. Après l'étape de normalisation, le signal a été soumis au processus de segmentation en différentes tailles de 2,4 et 10 seconde. Le but de notre étude est d'avoir évalué la performance de la stratégie de segmentation du signal sur des courts intervalles dans la classification des émotions. Cela permettra à l'avenir de l'utiliser dans des applications émotionnelles en temps réel [63], [56].

### 4.4.3 Architecture proposée

Dans le cadre de la classification des émotions, nous avons proposé une architecture hybride d'apprentissage profond (DL) qui combine des réseaux de neurones convolutifs 1D (1D-CNN) et des réseaux LSTM (Long Short-Term Memory).

En ce qui concerne les données séquentielles, telles que les signaux, le 1DCNN est l'une des architectures d'apprentissage profond les plus appropriées pouvant être utilisées pour la classification, comme cela a déjà été prouvé et confirmé par certains chercheurs [63], [57]. Il a été constaté que lorsque des couches de convolution sont appliquées à des zones successives du signal, des vecteurs de caractéristiques unidimensionnels sont générés. Suite à cette étape, des couches de Pooling sont appliquées pour réduire le nombre de paramètres entraînables. Pour préserver la séquence de fonctionnalités au fil du temps, des LSTM sont utilisées. Enfin, des couches de neuronnes entièrement connectées sont utilisées pour le processus de classification. Un aperçu de l'architecture proposée est présenté sur la figure 4.11.

L'architecture 1DCNN-LSTM a connu une large utilisation dans le domaine de la classification automatique des émotions humaines à partir de signaux physiologiques, comme en

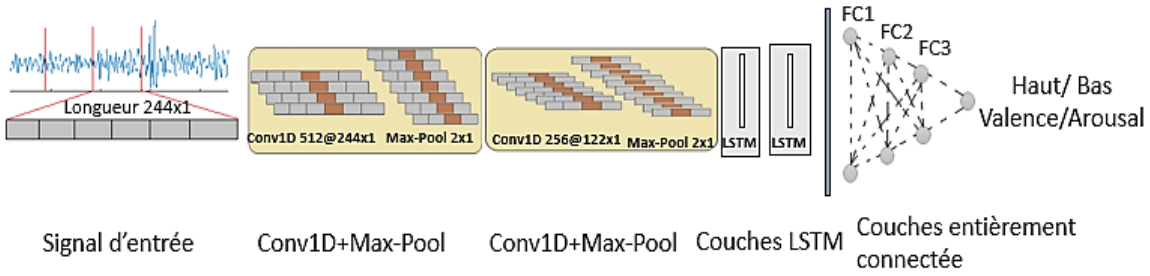



Figure 4.11 – Aperçu de l'architecture CNN-LSTM proposée

témoignent plusieurs études récentes [172], [173]. Dans ce contexte, plusieurs chercheurs ont proposé de nouvelles architectures ; celles-ci variaient en termes de nombre de couches utilisées et de nombre de paramètres. Dans notre cas, le signal iPPG unidimensionnel est introduit dans le réseau d'entrée. Deux couches de convolution sont utilisées. Chaque couche est suivie d'une couche de pooling maximale de taille 2x1. Après cela, deux couches LSTM sont ajoutées et pour la classification, quatre couches entièrement connectées sont utilisées.

Il est important de souligner que la dernière couche contient un neurone activé avec une fonction sigmoïde afin de catégoriser les émotions en fonction de leur valence négative ou positive et de leur degré d'arousal élevé ou faible. Dans le but de résoudre ou d'éviter le problème de overfitting, nous avons utilisé des couches dropout et de batch-normalization. Le tableau ci-dessous 4.9 présente les détails du réseau proposé.

### 4.4.4 Synthèse

L'objectif fondamental de cette étude est de proposer une nouvelle méthode de classification de la valence et de l'arousal des émotions à partir du signal iPPG extrait des vidéos faciales. L'utilisation des signaux physiologiques sans contact dans la classification des émotions représente une approche prometteuse.

Durant la phase d'extraction des signaux iPPG, nous avons exploité la base de données émotionnelle MAHNOB-HCI en appliquant la technique ICA. La détection des visages a été réalisée à l'aide de l'algorithme Viola-Jones, suivi de la technique de recadrage.

De plus, pour la classification, nous avons proposé une architecture d'apprentissage profond CNN-LSTM dans le but de classer les émotions sur l'échelle valence-arousal, après deux étapes importantes de prétraitement des signaux iPPG telles que la normalisation et la

Tableau 4.7 – Détails de l'architecture CNN-LSTM proposée

| Type de couches | Taille des filtres en entrée | Fonction d'activation | Taille des filtres en sortie | Paramètres |
|---|---|---|---|---|
| Conv1D | 521,50 | Relu | 244,512 | 26112 |
| MAX Pooling | 2,1 | \ | 122,512 | 0 |
| Batch Normalization | \ | \ | 122,512 | 2048 |
| Dropout | 0.3 | \ | 122,512 | 0 |
| Conv1D | 256,25 | Relu | 122,256 | 3277056 |
| MAX Pooling | 2,1 | \ | 61,256 | 0 |
| Batch Normalization | \ | \ | 61,256 | 1024 |
| Dropout | 0.3 | \ | 61,256 | 0 |
| LSTM | 256 | Hard-sigmoid | 61,256 | 525312 |
| LSTM | 128 | Hard-sigmoid | 61,256 | 197120 |
| Flatten | \ | \ | 7808 | 0 |
| Fully connected | 256 | Sigmoid | 256 | 1999104 |
| Dropout | 0.5 | \ | 512 | 0 |
| Entièrement connecté | 128 | Sigmoid | 128 | 32896 |
| Entièrement connecté | 64 | Sigmoid | 64 | 8256 |
| Output | 1 | Sigmoid | 1 | 65 |

segmentation.

Après avoir présenté notre étude basée sur la modalité physiologique, dans la section suivante nous présenterons notre étude proposée pour la classification des émotions en utilisant les expressions faciales comme modalité.

## 4.5 Reconnaissance des émotions à partir des expressions faciales et différentes poses de tête

La modalité des expressions faciales représente une modalité fondamentale et importante dans la recherche sur les émotions humaines. Malgré les études significatives réalisées et les résultats obtenus, des défis subsistent dans ce domaine de recherche en raison de la diversité

des expressions émotionnelles propres à chaque individu, ainsi que des variations liées à l'âge, au sexe, l'éclairage, et à la position de la tête... etc.

Cette section vise à mettre en place un système de reconnaissance automatique des émotions humaines basé sur les expressions faciales. Notre étude s'est spécifiquement concentrée sur la classification des sept émotions de base définies par Paul Ekman, à savoir la joie, le dégoût, la tristesse, la surprise, la colère, la peur et neutre. Nous avons choisi d'utiliser l'ensemble de données RaFD qui contient des expressions faciales mimiques. En outre, elle est riche en défis, tels que les différences d'âge, de sexe et d'origine ethnique, ainsi que les différents regards et positions de tête.

Au cours des dernières années, l'utilisation de différentes architectures d'apprentissage profond a abouti à des résultats prometteurs dans le domaine de la reconnaissance automatique des émotions [9]. Dans cette perspective, nous avons conçu deux architectures de réseaux CNN afin d'obtenir des taux de reconnaissance plus élevés, que ce soit avec des images de visages frontales ou avec des poses de la tête différents.

### 4.5.1 Prétraitement des images

Le prétraitement des images est une étape cruciale dans l'analyse des expressions faciales, car elle permet d'identifier les paramètres pertinents, d'améliorer la qualité de l'image et de supprimer les caractéristiques non pertinentes dans l'apprentissage. Cela permet d'améliorer la précision de la classification.

Dans un premier temps, nous réduisons la taille des images de 681x1024 à 200x200 pixels ; cette réduction nous permet d'obtenir une meilleure détection faciale par l'algorithme de Haar-cascade avec la bibliothèque OpenCV [174]. Les images obtenues passent par l'étape de recadrage et de réduction à 48x48 pixels. La figure 4.12 montre toutes les étapes de prétraitement des images effectuées dans notre étude.

Après avoir effectué les étapes de prétraitement, toutes les données traitées ont une résolution de 48x 48 pixels et ont été classées en sept émotions de base, comme indiqué dans le tableau 4.8.

La figure 4.13 montre un échantillon d'images de la base de données RaFD obtenu après étapes de prétraitement, montrant sept émotions de base, avec un visage frontal et différents

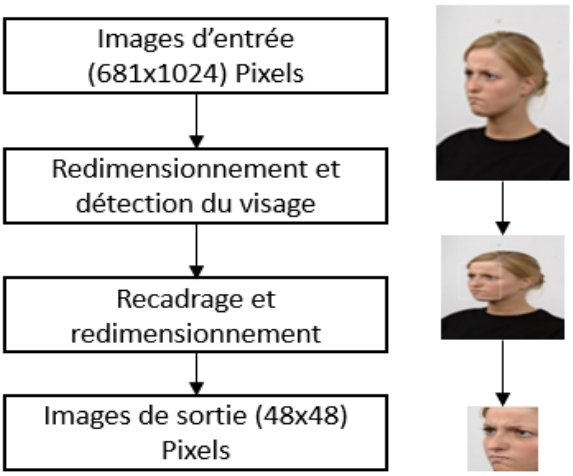



Figure 4.12 – les étapes de prétraitement proposées [43]

Tableau 4.8 – Nombre d'images par émotion dans la base de données après les étapes de prétraitement.

| Émotions | Colère | Dégout | Peur | Joie | Neutre | Triste | surprise |
|---|---|---|---|---|---|---|---|
| Nombres d'images de pose frontale | 200 | 200 | 200 | 200 | 200 | 200 | 200 |
| Nombre d'images de différentes poses de la tête à 45 degrés et 135 degrés | 395 | 374 | 371 | 371 | 399 | 395 | 349 |

regards (frontal, gauche et droit). Et la figure 4.14 présente un échantillon d'images obtenues après les étapes de prétraitement avec différentes poses de la tête et de regards (frontal, gauche et droit).

Après le prétraitement des images, la classification est la prochaine étape à franchir, et nous avons utilisé des techniques d'apprentissage profond dans notre étude.

### 4.5.2 Architecture proposée pour la classification des émotions

Les travaux de recherche sur la reconnaissance automatique des émotions humaines à partir des expressions faciales mettent en évidence l'utilisation d'architectures CNN [9].

Dans cette section, nous présentons notre proposition d'architecture CNN utilisée pour classer les sept émotions de base telles que la colère (AN), le dégoût (DI), la peur (FE), la joie (HA), le neutre (NE), la tristesse (SA), la surprise (SU).

Notre architecture se compose de deux couches de convolution, suivies chacune d'un


Figure 4.13 – Échantillon d'images prétraitées de la base de données RaFD, montrant les émotions avec différents regards. (a) Colère, (b) Dégoût, (c) Peur, (d) Heureux, (e) Neutre, (f) Tristesse, (g) Surprise [43]

Figure 4.14 – Échantillon d'images prétraitées de la base de données RaFD, montrant les émotions avec différents regards, poses de tête et sexes [43]

pooling maximum, dont les sorties sont introduites dans des couches de neuronnes entièrement connectées. La couche de sortie comprend sept neurones activés par la fonction Softmax pour indiquer l'une des sept émotions de base. En outre, pour prévenir le surapprentissage, nous avons adopté la technique du dropout. La figure 4.15 offre un aperçu de notre réseau proposé.

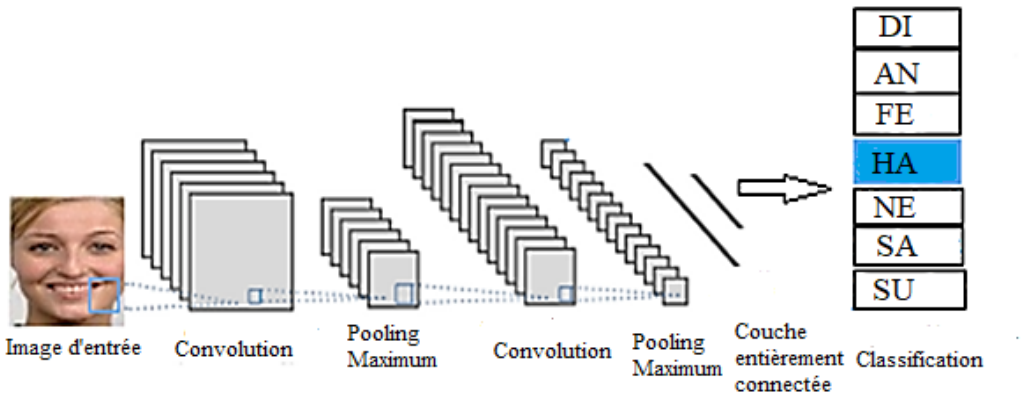



Figure 4.15 – Aperçue générale de notre architecture CNN proposéec [43]

Il convient également de noter que 80% de la base de données ont été utilisés pour l'entrainement, tandis que 10% pour la validation et 10% pour le test. Les détails concernant les réseaux proposés sont résumés dans les tableaux suivants 4.9 et 4.10. Le premier tableau expose l'architecture CNN proposée pour les images de visages frontaux, tandis que le deuxième tableau est destiné à présenter l'architecture CNN pour les images de visages avec différentes poses de tête.

Tableau 4.9 – Détails proposés par CNN pour les images du visage frontal

| Type des couches | Tailles des filtres en entrée | Fonction d'activation | Tailles des filtres en sortie |
|---|---|---|---|
| Conv2D | 32x3x3 | Relu | 46x46x32 |
| Max Pooling | 2,2 | Relu | 23x23x32 |
| Conv2D | 64x3x3 | Relu | 21x21x64 |
| Max Pooling | 2,2 | Relu | 10x10x64 |
| Dropout | 0.5 | \ | 10x10x64 |
| Flatten | \ | \ | \ |
| Entièrement connecté | 128 | Relu | 128 |
| Entièrement connecté | 64 | Relu | 64 |
| Output | 7 | Softmax | 7 |

Tableau 4.10 – Détails du CNN proposé pour la classification des images faciales sous différentes poses de la tête

| Type des couches | Tailles des filtres en entrée | Fonction d'activation | Tailles des filtres en sortie |
|---|---|---|---|
| Conv2D | 32x3x3 | Relu | 46x46x32 |
| Max Pooling | 2,2 | Relu | 23x23x32 |
| Dropout | 0.5 | \ | 23x23x32 |
| Conv2D | 64x3x3 | Relu | 21x21x64 |
| Max Pooling | 3,3 | Relu | 7x7x64 |
| Dropout | 0.5 | \ | 7x7x64 |
| Flatten | \ | \ | \ |
| Entièrement connecté | 256 | Relu | 256 |
| Dropout | 0.5 | \ | 256 |
| Output | 7 | Softmax | 7 |

### 4.5.3 Synthèse

Dans cette partie, nous avons étudié un système unimodale qui se focalise sur les expressions faciales. Pour cette étude, nous avons opté pour la base de données RaFD en raison de ses multiples défis tels que les différentes positions du regard et de la tête, les variations d'âge, de race et de sexe.
Avant de commencer la phase de classification, nous avons réalisé une étape cruciale : le prétraitement des images. Cette phase permet à nos réseaux d'acquérir uniquement les caractéristiques essentielles avec une durée d'entrainement réduit. Deux architectures CNN ont été proposées pour effectuer une classification en sept émotions de base.

Il convient de noter que la base de données utilisée dans cette étude est statique et présente des expressions faciales mimique. Suite à cette étude, nous avons avancé en adoptant une approche multimodale qui combine les expressions faciales avec les signaux iPPG extraits de vidéos faciales RVB. Cette approche nous exige d'utiliser des données vidéo qui montrent des expressions spontanées au fil du temps. Ce qui nous présenterons dans la section suivante.

## 4.6 Reconnaissance multimodale du stress

Aujourd'hui, la détection automatique du stress humain suscite un intérêt croissant de la part des chercheurs en raison de son importance cruciale. Cette section est consacrée à la présentation de notre étude, qui vise à développer une architecture d'apprentissage profond pour la classification binaire de l'état de stress ou de non-stress en fusionnant l'expression faciale et les signaux iPPG. L'originalité de cette étude réside dans l'utilisation de signaux iPPG extraits à distance à partir de vidéos faciales RVB. La figure 4.16 donne un aperçu général de notre étude.

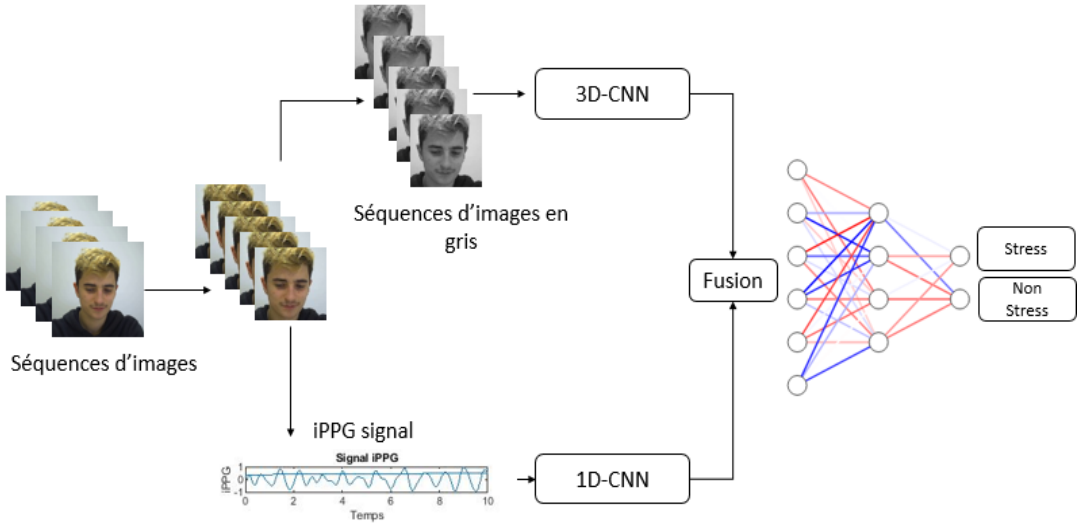



Figure 4.16 – Aperçu de la méthode suggérée.

Dans cette contribution, différents types d'architectures DL sont introduits, en utilisant deux modalités différentes, telles que le réseau neuronal convolutif 3D-CNN pour la reconnaissance des expressions faciales et le réseau neuronal convolutif 1D-CNN pour la reconnaissance des émotions à l'aide de signaux iPPG sans contact extraits de vidéos faciales. Les caractéristiques cruciales de chaque méthode sont ensuite fusionnées.

### 4.6.1 Préparation des données

Notre étude se décompose en deux catégories de données différentes : les expressions faciales et les signaux iPPG. En raison de la limitation de la puissance de calcul, nous avons utilisé 10 secondes par vidéo pour 12 participants, ce qui nous a permis d'obtenir une séquence

de 350 images par vidéo.

### A Préparation des données pour les expressions faciales

Il convient de souligner que les vidéos faciales RVB de la base de données UBFC-Phys présentent une qualité supérieure, avec une fréquence d'images de 35 fps et une résolution de 1024 x 1024 pixels. La technique de recadrage a été appliquée à chaque image. Le visage a été privilégié dans cette étude, permettant de le détecter tout en éliminant l'arrière-plan [175]. Les images, initialement en couleur RVB, ont été converties en noir et blanc (niveaux de gris). Enfin, les images utilisées comme entrée du réseau d'apprentissage profond ont été redimensionnées à 75 x 75 pixels en niveaux de gris. La figure 4.17 montre les étapes utilisées dans le prétraitement d'image pour la reconnaissance des expressions faciales.

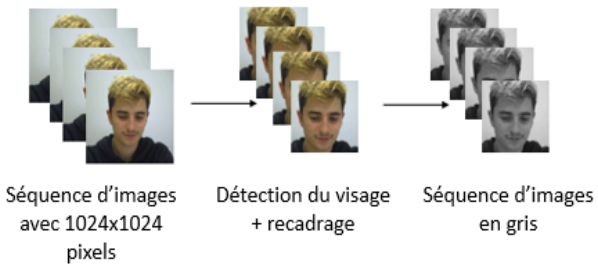



Figure 4.17 – Étapes de prétraitement proposées pour la reconnaissance des expressions faciales

### B Collecte de signaux iPPG sans contact

En deuxième point, cette étude vise à présenter une classification du stress ou du non-stress basée sur des signaux iPPG sans contact.

Il convient de rappeler que, dans notre travail précédent, une étude comparative a été menée sur quatre méthodes d'extraction de signaux iPPG, sous deux régions faciales différentes (voir la section 4.3).
Le but de cette étape est de trouver l'algorithme le plus efficace pour fournir des signaux iPPG plus précis, ce qui nous permet d'obtenir des informations précises sur les réactions cardiaques.

Pour rappel, les méthodes d'extraction du signal iPPG utilisées dans notre étude comparative sont : Green, ICA, CHROM et POS sous deux zones différentes du visage. La première consiste à utiliser tout le visage, en éliminant le fond et les cheveux de la tête. Tandis que La

deuxième zone consiste à utiliser la partie inférieure du visage, qui contient les joues, la bouche et le menton. Voir la section 4.3.

Afin d'évaluer les résultats obtenus, nous avons calculé trois paramètres, tels que l'erreur absolue moyenne MAE, l'erreur quadratique moyenne RMSE et le rapport signal sur bruit SNR. Les paramètres sont calculés en tenant compte de la fréquence cardiaque réelle $FC_{réel}$ et de la fréquence cardiaque estimée $FC_{est}$.
La fréquence cardiaque $FC_{réel}$ est calculée à partir de signal PPG disponible de la base de donnée UBFC-Phys utilisant la méthode de la différence moyenne entre deux pics IBI. Voir section 3.4.3.

De ces résultats comparatifs, nous concluons que la méthode ICA avec le premier ROI est la meilleure avec MAE= 1,64, RMSE= 2,28 et SNR=-4,42. Ce qui nous a incité à l'utiliser pour collecter les signaux iPPG et les utiliser dans le domaine de la classification l'état de stress ou de non-stress humains.

### 4.6.2 Réseaux d'Apprentissage Profond proposés

Cette section vise principalement à présenter les architectures d'apprentissage profond proposées pour la classification du stress ou du non-stress. Nous commençons notre étude en explorant d'abord l'approche unimodale, avant de passer à l'approche multimodale.

#### A Expressions faciales

Après avoir effectué le prétraitement des données faciales, nous passons à l'étape d'extraction des paramètres et à la classification. Étant donné que nous travaillons avec des vidéos, l'utilisation d'un réseau spatio-temporel 3D-CNN a été considérée comme appropriée pour classifier les états de stress et de non-stress chez l'humain. Cette architecture d'apprentissage profond 3D-CNN s'est révélée être la plus adaptée, car elle permet d'analyser simultanément un ensemble de trames séquentielles, préservant ainsi des caractéristiques importantes au fil du temps [176]. Le tableau 4.11 présente l'architecture du 3D-CNN proposée pour la détection du stress humain.

Tableau 4.11 – Détails de l'architecture du réseau 3D-CNN proposée - Expressions Faciales

| Type de couches | Tailles des filtres en entrée | Tailles des filtres en sortie | Paramètres |
|---|---|---|---|
| Conv3D | (16,(3x3x50)) | (None, 16, 73, 73, 301) | 7216 |
| Max Pooling | (3,3,3) | (None, 16, 24, 24, 100) | 0 |
| Dropout | 0.5 | (None, 16, 24, 24, 100) | 0 |
| Conv3D | (32, (3,3,25)) | (None, 32, 22, 22, 76) | 115232 |
| Max Pooling | (3,3,3) | (None, 32, 7, 7, 36) | 0 |
| Dropout | 0.5 | (None, 32, 7, 7, 25) | 0 |
| Conv3D | (64, (3,3,10)) | (None, 64, 5, 5, 16) | 184384 |
| Max Pooling | (2,2,2) | (None, 64, 2, 2, 5) | 0 |
| Dropout | 0.5 | (None, 64, 2, 2, 5) | 0 |
| Flatten | \ | (None, 320) | 0 |
| Entièrement connecté | 1024 | (None, 1024) | 328704 |
| Dropout | 0.5 | (None, 1024) | 0 |
| Entièrement connecté | 256 | (None, 256) | 524800 |
| Dropout | 0.5 | (None, 256) | 0 |
| Entièrement connecté | 128 | (None, 128) | 131328 |
| Dropout | 0.5 | (None, 128) | 0 |
| Output | 2 | (None, 2) | 258 |

## B signaux iPPG

Après avoir acquis les iPPG nécessaires, notre attention s'est tournée vers la classification basée sur les données extraites. En accord avec le concept exposé dans notre travail, tel que présenté dans la section 4.4, l'utilisation du modèle 1D-CNN pour la classification des signaux

s'est avérée plus appropriée après leur normalisation et leur segmentation.

Il est pertinent de noter que les signaux iPPG sans contact extraits précédemment ont une dimension de 1054x1.

Lors de l'étape de segmentation, le signal iPPG a été divisé en deux signaux d'une taille de 527x1 correspondant chacun à cinq secondes. Ces deux signaux plus petits (527x1) ont ensuite été utilisés à l'entrée du réseau 1D-CNN.

Le tableau 4.12 expose l'architecture 1D-CNN proposée pour la détection du stress ou du non-stress à l'aide de signaux iPPG.

Tableau 4.12 – Détails de l'architecture du réseau 1D-CNN proposée

| Type de couches | Tailles des filtres en entrée | Tailles des filtres en sortie | Paramètres |
|---|---|---|---|
| Conv1D | (32,527) | (None,527,32) | 16896 |
| Max Pooling | (1,3) | (None, 263,32) | 0 |
| Batch Normalization | \ | (None, 263,32) | 128 |
| Dropout | 0.5 | (None, 263,32) | 0 |
| Conv1D | (64, 250)) | (None,263,32) | 256032 |
| Max Pooling | (1,2) | (None, 131.32) | 0 |
| Batch Normalization | \ | (None, 131,32) | 128 |
| Dropout | 0.5 | (None, 131.32) | 0 |
| Flatten | \ | (None, 4192) | 0 |
| Entièrement connecté | 1024 | (None, 1024) | 4293632 |
| Entièrement connecté | 256 | (None, 256) | 262400 |
| Entièrement connecté | 128 | (None, 128) | 32896 |
| Entièrement connecté | 64 | (None, 64) | 8256 |
| Output | 1 | (None, 1) | 65 |

## C Reconnaissance multimodale du stress

En complément de notre étude, cette section met en place un système multimodal pour classifier les états de stress ou de non-stress. La figure 4.18 présente le système multimodal proposé pour notre étude. Le modèle proposé est fondé sur les approches de traitement des données et de modèle d'apprentissage profond présentées précédemment. L'étape d'extraction des caractéristiques pour chaque méthode a été réalisée en premier lieu. Ensuite, les caractéristiques sont combinées en un seul module, suivi d'une phase de classification où des couches neuronales entièrement connectées ont été utilisées pour classer l'état de stress humain ou non. Cette méthode de fusion est connue sous le nom de fusion précoce.

Le modèle proposé a été entraîné pendant 50 époques en utilisant l'algorithme d'optimisation Adam.

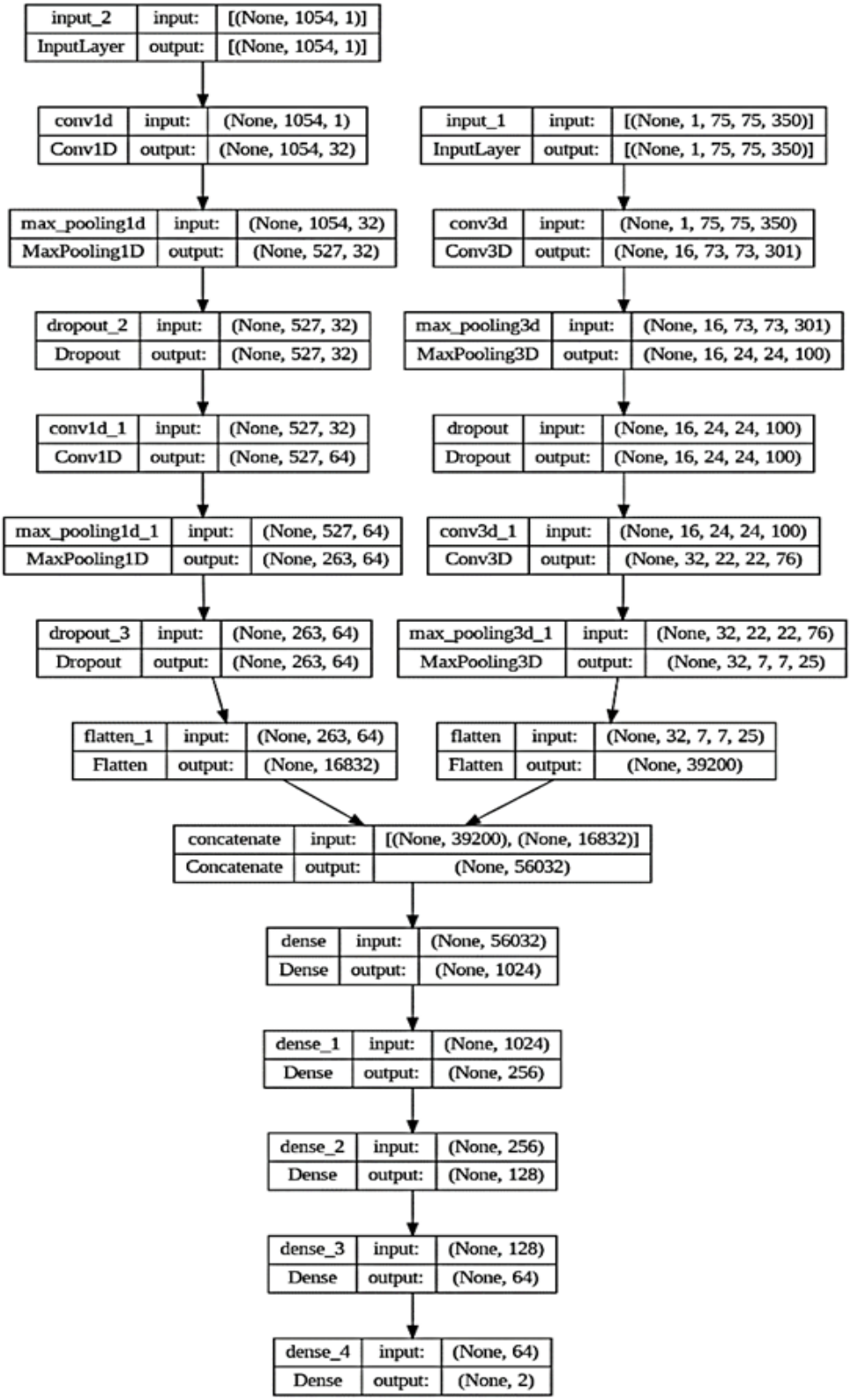



Figure 4.18 – Modèle d'apprentissage profond multimodale proposé pour classification du stress ou non.

### 4.6.3 Synthèse

Dans cette section, nous avons proposé un système multimodal basé sur la fusion des expressions faciales et des signaux iPPG en utilisant des techniques d'apprentissage profond.

La première phase de notre étude a consisté à explorer la classification entre les états de stress et de non-stress à partir des expressions faciales, en adoptant des techniques d'apprentissage profond telles que le 3D-CNN.

La deuxième étape de l'étude s'est concentrée sur un domaine de recherche crucial, à savoir l'extraction des signaux iPPG à partir de vidéos faciales RVB et l'utilisation de ces signaux comme deuxième modalité pour la classification entre les états de stress et de non-stress, en utilisant le réseau 1D-CNN.

La dernière étape de cette étude consiste à fusionner les caractéristiques de chaque modalité et à effectuer la classification à l'aide d'un réseau neuronal entièrement connecté.

## 4.7 Conclusion

Ce chapitre a été conçu pour présenter notre étude expérimentale réalisée au cours de cette thèse, se focalisant sur la reconnaissance automatique de l'affect.

Nous avons utilisé deux modalités distinctes : les expressions faciales et les signaux iPPG extrait à partir de vidéos de visages humains RVB. Afin d'optimiser l'exploitation de ces signaux, une étude comparative des méthodes d'extraction du signal iPPG les plus répandues a été entreprise, visant à identifier la méthode la plus performante fournissant des signaux plus précis. Suite à cette étude comparative, nos conclusions ont mis en évidence que la méthode ICA, associée à une détection précise du visage, éliminant l'arrière-plan et la partie des cheveux de la tête des participants, a conduit aux meilleurs résultats.

Après cette phase, nous avons appliqué ces signaux à la classification des émotions sur deux échelles valence et arousal.

Une autre facette de notre étude s'est concentrée sur l'approche de classification des émotions en utilisant les expressions faciales sous différents pose de la tête.

Le dernier objectif ultime atteint dans ce chapitre est de proposer un système multimodale

de classification du stress, basé à la fois sur les expressions faciales et les signaux iPPG.

Pour chaque étude présentée dans ce chapitre, nous avons exposé en détail nos étapes d'étude, accompagnées d'illustrations des différentes étapes et des techniques employées dans le prétraitement des données. Ensuite, nous avons abordé les architectures d'apprentissage profond qui ont été utilisées.

Dans le chapitre suivant, nous présentons les résultats obtenus pour chaque expérience réalisée, suivis d'une discussion de ces résultats.

Chapitre 5

# Résultats et discussions

## 5.1 Introduction

Ce chapitre vise à présenter les résultats obtenus au cours des expérimentations menées dans le chapitre précédent (voir Chapitre 3.6). Il est à noter que notre travail a impliqué trois études distinctes dans le domaine de la classification des émotions et du stress humain.

Dans un premier temps, nous avons proposé un système pour le développement de la reconnaissance automatique des émotions humaines basé les signaux iPPG extraits à partir des vidéos faciales.

La deuxième étude expérimentale visait à développer un système améliorant la reconnaissance de sept émotions de base en se basant sur les expressions faciales.

La troisième étude se concentrait sur la proposition d'un système physio-visuel de reconnaissance automatique du stress, basé sur l'analyse des vidéos faciales.

Tous les résultats obtenus seront comparés avec d'autres travaux réalisés dans le même domaine, suivis d'une discussion approfondie de ces résultats.

## 5.2 Résultats et Discussions : Classification des émotions via signaux iPPG

Dans cette section, nous présenterons les résultats obtenus dans l'étude réalisée pour la classification binaire des émotions sur les deux échelles de valence et d'arousal. Cette étude est détaillée dans la section 4.4.

Nous rappelons que notre méthode commence par la collecte des signaux iPPG extraits à partir des vidéos de la base de données MAHNOB-HCI, en utilisant l'algorithme ICA. Ensuite, les signaux obtenus passent par la technique de normalisation et de segmentation en différentes tailles. Pour la classification, nous avons proposé une architecture d'apprentissage profond CNN-LSTM.

### 5.2.1 Implémentation et résultats

Le modèle présenté dans la section 4.4.3 a été entraîné sur 50 époques en utilisant l'algorithme d'optimisation Adam avec un learning rate de 0,0001, en utilisant la fonction de

perte de cross-entropie binaire. Il est également pertinent de mentionner que 90% de la base de données ont été utilisés pour l'entraînement, tandis que le reste a été réservé à la validation. Cette expérience a été réalisée en utilisant des packages Python sur un ordinateur performant équipé d'un processeur Intel Core i7-7700K, cadencé à 4,20 GHz, et de 16 Go de RAM.

Pour évaluer la méthode proposée, nous avons décidé de calculer quatre métriques importantes : la précision (Accuracy AC), la précision (Presicion PR), le score F1 (F1 score) et la sensibilité (Recall RE). Depuis les années 1950, les chercheurs ont largement exploité ces paramètres pour évaluer la performance de leurs méthodes [177]. Ces paramètres peuvent être évalués à l'aide des formules ci-dessous :

$$\mathrm{AC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{5.1}$$

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5.2}$$

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.3}$$

$$\mathrm{F1} = \frac{2\mathrm{PR} * \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}} \tag{5.4}$$

FP, FN sont utilisés pour désigner les faux positifs et les faux négatifs, tandis que TP et TN sont utilisés pour désigner le vrai positif et le vrai négatif, respectivement. Premièrement, une classification des émotions a été réalisée sur l'échelle de valence. Une classification binaire des valences positives et négatives a été réalisée et le signal iPPG a été segmenté en différentes tailles. Le tableau 5.1 présente les performances obtenues lors de notre expérience, tandis que les figures 5.1 et 5.2 illustrent la matrice de confusion et les courbes de perte associées.

Tableau 5.1 – Résultats obtenus sur l'échelle de valence

| Taille de signal iPPG | AC | PR | RE | F1 |
|---|---|---|---|---|
| 2sec | 60.67% | 61.29% | 48.5% | 55.52% |
| 4sec | 73.33% | 80.43% | 61.66% | 69.81% |
| 10sec | 70.83% | 83.78% | 51.66% | 63.91% |

Les résultats obtenus montrent que la meilleure précision a été obtenue avec une segmentation du signal en fenêtres de 4 seconde, avec AC=73,33%, PR=80,43%, RE=61,66% et F1=69,81%.

La matrice de confusion présentée sur la figure 5.1 indique clairement que le modèle

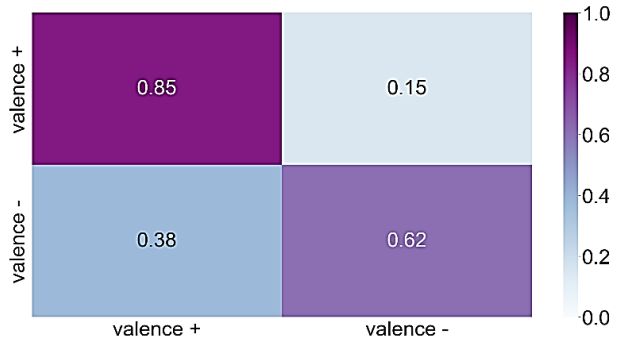



Figure 5.1 – Matrice de confusion obtenue dans l'échelle de valence

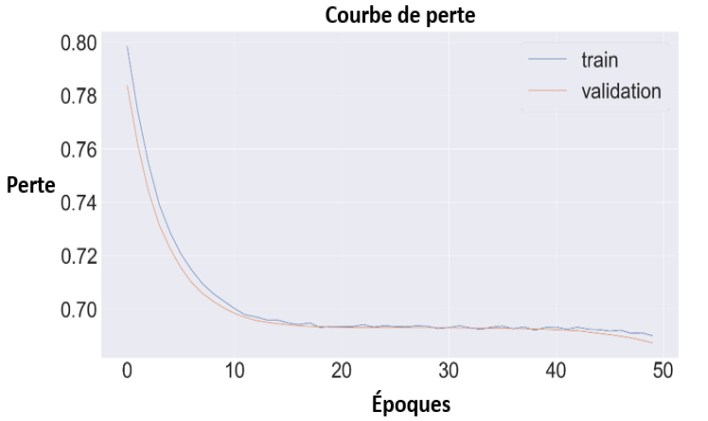



Figure 5.2 – Courbes de perte obtenues dans l'échelle de valence

proposé peut atteindre un taux de reconnaissance plus élevé sur les émotions positives que sur les émotions négatives.

Nous avons réitéré les étapes de l'expérience que nous avons effectuée sur l'échelle de valence dans l'échelle d'arousal. Les résultats obtenus sont présentés dans le tableau 5.2.

Tableau 5.2 – Résultats obtenus selon l'échelle d'Arousal

| Taille de signal iPPG | AC | PR | RE | F1 |
|---|---|---|---|---|
| 2sec | 57.5% | 56.839% | 62.5% | 59.53% |
| 4sec | 60.00% | 58.33% | 70.5% | 63.63% |
| 10sec | 50.63% | 50.74% | 42.5% | 46.5% |

En se basant sur ces résultats, on peut constater que notre réseau proposé obtient des performances de classification inférieures sur l'échelle d'arousal par rapport à l'échelle de valence, avec un score de AC = 60 %, PR = 56,89 %, RE = 82,5% et F1 = 67,34%.

Comme pour la valence, notre réseau proposé atteint un meilleur taux de reconnaissance

avec une segmentation du signal de quatre secondes. La figure 5.3 et 5.4 présente la matrice de confusion et les courbes de perte obtenues respectivement. Par ailleurs, notre modèle obtient de meilleurs résultats lorsque l'excitation est négative que lorsqu'elle est positive.

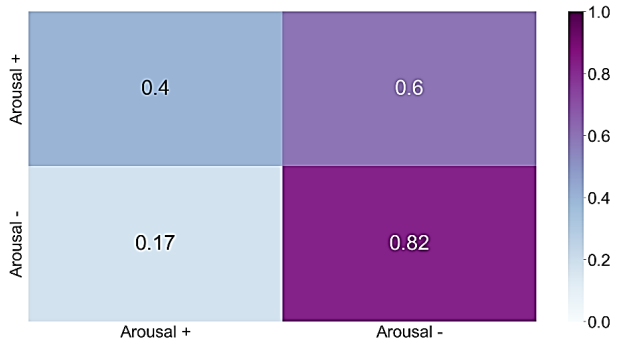



Figure 5.3 – Matrice de confusion obtenue dans l'échelle de Arousal

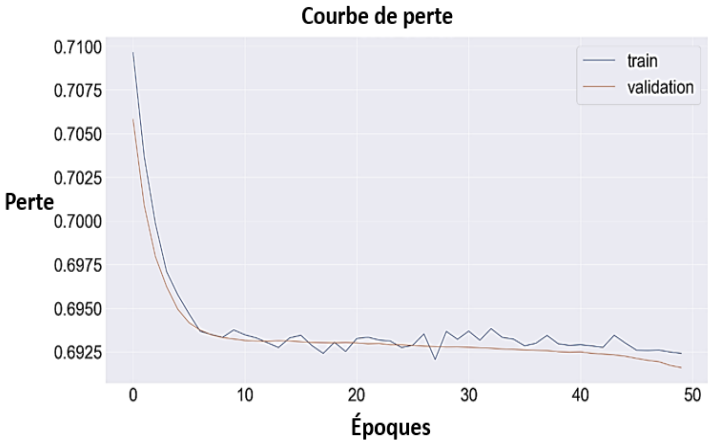



Figure 5.4 – Courbes de perte obtenues dans l'échelle de Arousal

### 5.2.2 Discussions et comparaison des résultats

La présente contribution vise principalement à identifier et à classer les émotions humaines à l'aide de signaux iPPG. Notre étude est menée sur les deux échelles de valence et d'arousal. Nous avons obtenu une précision de 73,33 % en valence et de 60 % en arousal en utilisant des signaux segmentés de quatre secondes. Ces résultats prometteurs sont significatifs dans ce domaine d'étude et démontrent l'efficacité de notre approche.

Le tableau ci-dessous 5.3 présente les résultats de recherches récentes sur la classification des émotions humaines, en utilisant des signaux physiologiques extraits de vidéos RVB.

D'après les résultats présentés dans le tableau, notre méthode obtient de bien résultats (précision de 73,33 % en valence et 60 % en arousal) que ceux rapportés précédemment par

Tableau 5.3 – Comparaison des résultats de classification des émotions à l'aide de signaux physiologiques sans contact.

| Auteurs | Méthodes | Signaux physiologiques sans contact | Émotions | Base de données | Précisions% |
|---|---|---|---|---|---|
| Maaoui et al. [159] | SVM, LDA | iPPG | Stress | 12 participants | 94.40%, 91.10% |
| Meziati Sabour et al. [160] | SVM | iVFC | Stress | UBFC-Phys | 85.48% |
| Yang et al. [178] | 1D-CNN | iPPG | Pain | BioVid | 58.92% |
| Ouzar et al. [161] | MLP | iVFC, iPPG | Bonheur, peur, douleur, embarras | BP4D+ | 53.59%, 55.33% |
| Lampier et al. [179] | SVM, KNN | iPPG | Valence | \ | 42%, 38% |
| Yu et al. [158] | SVM | iVFC | Valence, Arousal | MAHNOB-HCI | 46.86%, 44.02% |
| Notre Méthode | 1DCNN-LSTM | iPPG | Valence, Arousal | MAHNOB-HCI | 73.33%, 60% |

Lampier et al. [179], Ouzar et al. [161], Yu et al. [158] et Yang et al. [178]. De plus, il est important de souligner que cette approche d'étude est novatrice et qu'il existe un manque de recherches dans ce domaine.

En ce qui concerne la classification du stress humain, de nombreux chercheurs ont atteint des taux de reconnaissance élevés en raison des forts changements se produisant dans les réactions du cœur et des vaisseaux sanguins lorsqu'une personne est confrontée à un stress.

De plus, il est important de noter que notre étude diffère des études précédemment menées par Maaoui et al. [159] et Meziati Sabour et al. [160], car nous avons effectué une classification des différents types d'émotions sur l'échelle de valence et d'arousal.

Ouzar et al. [161] ont utilisé l'une des techniques d'apprentissage en profondeur les plus puissantes dans le but d'extraire des signaux iPPG plus précis à partir de vidéos faciales (MTTS-CAN). De même, Yu et al. [158] Ont proposé le réseau PhysNet128-3DCNN-ED. Il convient de souligner que notre méthode a atteint un taux de reconnaissance plus élevé, principalement attribuable à la classification basée sur les réseaux d'apprentissage profond proposés dans notre étude.

En conclusion, notre architecture 1DCNN-LSTM proposée, utilisant uniquement des segments de quatre secondes, peut atteindre une meilleure précision de classification que celle obtenue avec d'autres méthodes d'apprentissage automatique telles que le SVM proposé par Yu

et al. [158], et le réseau feed-forward MLP proposé par Ouzar et al. [161]. Ce qui soulignant la puissance de notre méthodologie.

## 5.3 Résultats et Discussions : Classification des émotions à travers les expressions faciales sous différentes poses de la tête

Dans cette section, nous exposons en détail les résultats issus de l'étude présentée dans la section 4.5, les comparant avec les résultats récents avancés par divers chercheurs.

Il est important de souligner que notre méthode vise à classer sept émotions en fonction des différentes poses de la tête, notamment avec des angles de 45°, 90° et 180°, en utilisant la base de données RaFD.

Les images extraites de la base de données RaFD ont été soumises à plusieurs étapes de prétraitement, comprenant le redimensionnement, la détection du visage, et le recadrage, suivi d'un redimensionnement final en 48x48 pixels. Pour la classification des émotions, deux architectures d'apprentissage profond de type 2D-CNN ont été développées : l'une dédiée aux visages frontaux et l'autre aux trois poses différentes de la tête.

### 5.3.1 Reconnaissance des émotions avec pose de la tête frontale

Au début, nous avons commencé notre expérience avec seulement des images faciales frontales. La figure 5.5 illustre la précision de l'entrainement par rapport à la précision de la validation (train accuracy Vs validation accuracy), ainsi que la perte de validation par rapport à la perte de l'entrainement (Loss validation Vs Loss train) obtenue dans notre expérience. Nous avons utilisé l'algorithme d'optimisation Adam pour entraîner notre modèle pendant 50 époques. Le tableau 5.4 montre la matrice de confusion que nous avons obtenue.

La figure 5.5 montre que notre modèle présentée dans le tableau 4.9 à une précision d'entrainement de 99 % et une précision de validation de 98 %, avec une perte de validation de 5 %. Cela démontre l'efficacité de notre méthode proposée, tant au niveau du pré-traitement des images de la base de données qu'au niveau de l'architecture 2D-CNN proposée.

De plus, sur la base de la matrice de confusion (Voir le tableau 5.4), notre méthode est

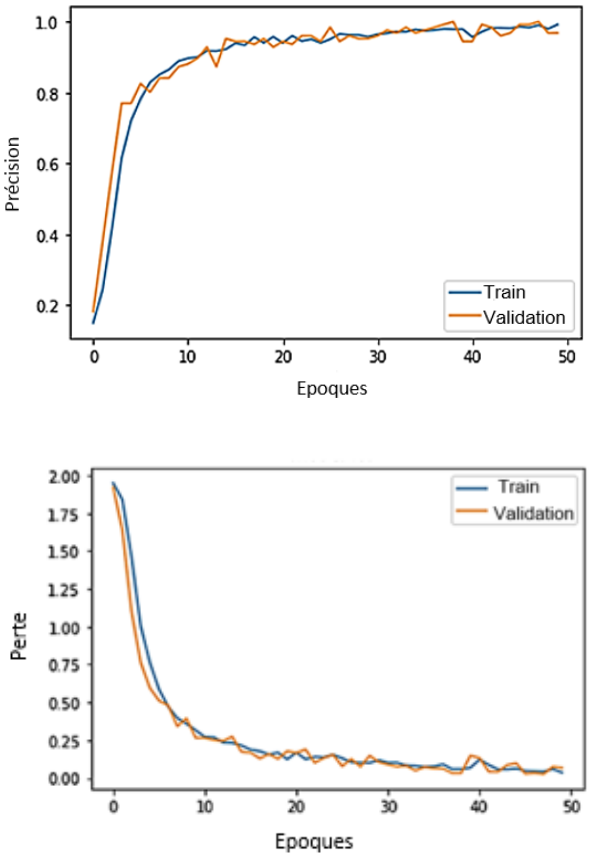



Figure 5.5 – Performance du modèle : précision et perte avec des images du visage frontal

plus efficace pour identifier le dégoût, la peur, la joie, la neutralité et la tristesse, mais nous avons remarqué une confusion entre l'émotion surprise et la peur et la colère avec le dégoût.

Une comparaison des études récentes sur la classification des émotions humaines à partir des expressions faciales, en utilisant la même base de données RaFD, est présentée dans le tableau 5.5.

En se basant sur les résultats de comparaison présentés dans ce tableau 5.5, nous pouvons affirmer que notre méthode atteint une précision compétitive par rapport aux autres méthodes.

Dans cette partie de notre étude, nous avons mis en évidence la capacité de notre architecture 2D-CNN à détecter avec plus de précision les émotions sur des visages captés en frontal. C'est pourquoi nous avons choisi d'utiliser cette architecture dans la section suivante, où nous travaillerons avec des données plus complexes contenant des images capturées dans différentes positions de tête.

Tableau 5.4 – Matrice de confusion obtenue avec des visages frontaux.

| Émotions | AN | DI | FE | HA | NE | SA | SU |
|---|---|---|---|---|---|---|---|
| AN | 18 | 1 | 0 | 0 | 0 | 0 | 0 |
| DI | 0 | 15 | 0 | 0 | 0 | 0 | 0 |
| FE | 0 | 0 | 23 | 0 | 0 | 0 | 0 |
| DI | 0 | 0 | 0 | 14 | 0 | 0 | 0 |
| NE | 0 | 1 | 0 | 0 | 23 | 0 | 0 |
| SA | 0 | 0 | 0 | 0 | 0 | 23 | 0 |
| SU | 0 | 0 | 1 | 0 | 0 | 0 | 22 |

Tableau 5.5 – Comparaison des méthodes de classification des émotions avec des visages frontales utilisant de la base de données RafD

| Auteurs | Précisions |
|---|---|
| Mavani et al. [180] | 95.71% |
| Fathallah et al. [181] | 93.33% |
| Sun et al. [182] | 99.17% |
| Yolcu et al. [183] | 94.44% |
| Notre méthode | 98.57% |

### 5.3.2 Reconnaissance des émotions avec trois poses de la tête

Dans cette partie, nous avons ajouté des images faciales provenant de deux poses de tête différentes (45° et 135°) aux images frontales. Tout d'abord, nous avons choisi de commencer notre expérience en utilisant l'architecture CNN proposée présenté dans le tableau 4.9. La figure 5.6 présente les résultats obtenus, illustrant les courbes de précision et de perte. Nous avons également réussi à atteindre un taux de reconnaissance élevé, avec un score de 98% pour la précision de l'entraînement et un taux de 96,30% pour la précision du test. Malheureusement, notre système a rencontré un problème de sur-ajustement après 40 époques.

Le sur-ajustement se produit lorsque notre modèle devient trop ajusté à un ensemble de données spécifique et ne peut pas appliquer ce qu'il a appris aux nouvelles entrées [184]. Pour résoudre ce problème, plusieurs techniques sont disponibles telles que l'augmentation des données, l'apprentissage par transfert et les couches de dropout [8].

Afin d'améliorer nos performances, nous avons apporté des modifications à l'architecture CNN en utilisant le dropout après la première couche (convolution-pooling). Ensuite, nous avons changé la taille de la deuxième couche max-pooling. Le tableau 4.10 fournit des renseignements

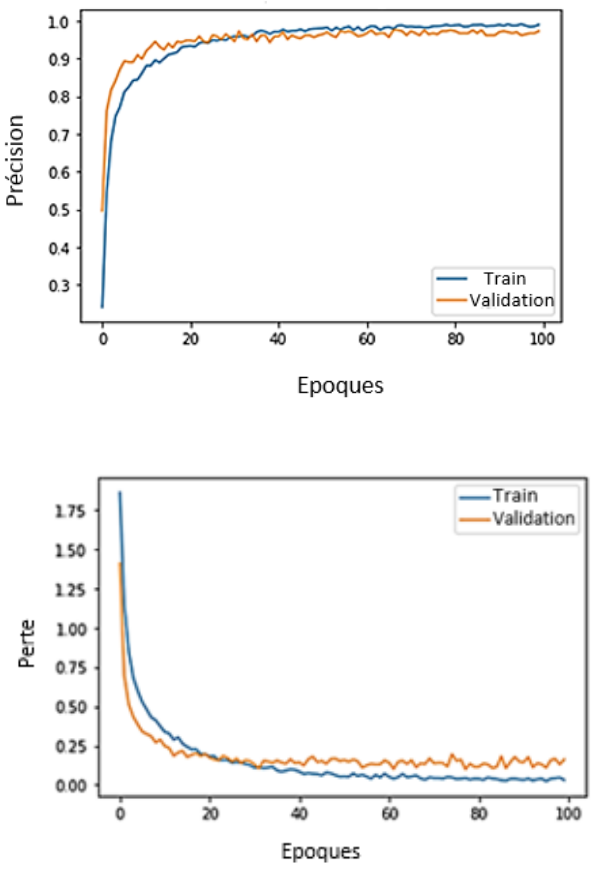



Figure 5.6 – Évaluation de la performance du premier modèle : précision et perte avec des images du visage sous différentes poses de la tête

supplémentaires sur le deuxième CNN proposé, tandis que la figure 5.7 présente les résultats obtenus. Notre modèle CNN a été entraîné sur 200 époques et nous avons choisi d'utiliser l'algorithme d'optimisation Adam.

Les résultats indiquent que notre modèle obtient un taux de reconnaissance plus élevé sans aucun problème de sur-ajustement, avec une précision de validation de 96,55 %. Cela suggère que le dropout est une technique robuste pour résoudre le problème de sur-apprentissage. Dans le tableau 5.6, vous pouvez voir la matrice de confusion obtenue en utilisant l'architecture CNN suggérée dans le tableau 4.10. Nos résultats indiquent que notre architecture est plus performante pour détecter les émotions de dégoût, mais il y a encore des erreurs dans la prédiction d'autres types d'émotions. Il se peut que cela soit causé par le fait que certaines émotions ont des expressions communes et que les expressions ne soient pas totalement claires en fonction des différentes positions de la tête.

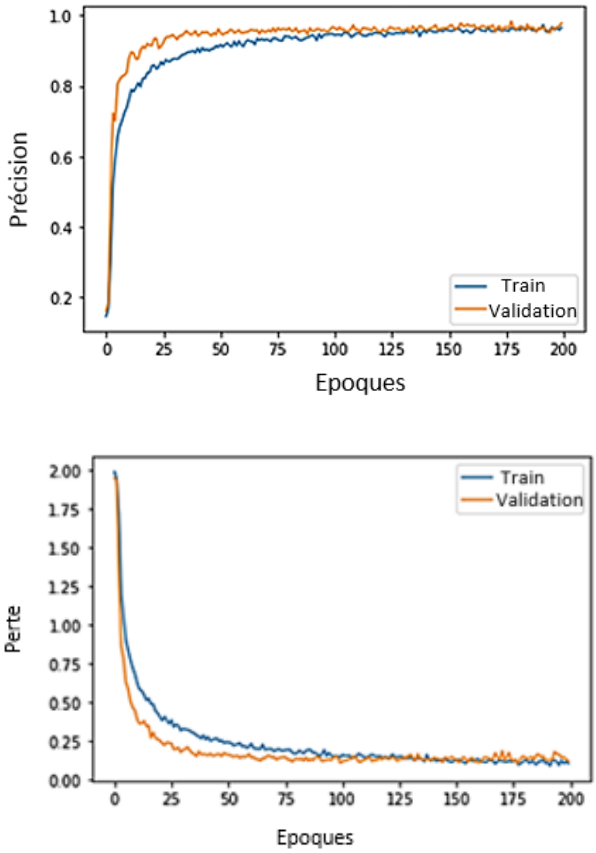



Figure 5.7 – Évaluation de la performance du deuxième modèle : précision et perte avec des images du visage sous différentes poses de la tête

Dans le tableau 5.7, nous exposons la précision obtenue par les chercheurs en utilisant la base de données RaFD. Il est remarquable que notre méthode présente des performances supérieures de 0,28% à celles de Wu et Lin [185].

Pour conclure notre étude, nous avons passé en revue les performances de notre méthode sur les nouvelles images de la base de données LFW (Labeled Faces in the Wild) [185]. Cette base de données contient des photos de célébrités qui expriment des émotions spontanées et non étiquetées. Suite au processus de prétraitement, nous avons réalisé des prédictions émotionnelles sur certaines images. Les résultats que nous avons obtenus sont illustrés dans le tableau 5.8.

Après avoir effectué plusieurs tests, nous pouvons affirmer que notre modèle est capable pour prédire l'état émotionnel des images provenant de l'extérieur de la base de données RaFD. Malgré les différences de conditions entre les images testées et celles de la base de données RaFD, notre modèle 2D-CNN a réussi à reconnaître les émotions avec précision. Cette performance

Tableau 5.6 – Matrice de confusion utilisant des images avec trois poses de tête différentes.

| Émotions | AN | DI | FE | HA | NE | SA | SU |
|---|---|---|---|---|---|---|---|
| AN | 64 | 0 | 0 | 0 | 1 | 0 | 0 |
| DI | 0 | 57 | 0 | 0 | 0 | 0 | 0 |
| FE | 0 | 0 | 48 | 0 | 1 | 0 | 1 |
| DI | 0 | 1 | 0 | 71 | 0 | 0 | 0 |
| NE | 0 | 0 | 0 | 0 | 56 | 2 | 0 |
| SA | 0 | 0 | 1 | 0 | 3 | 43 | 0 |
| SU | 0 | 0 | 4 | 0 | 0 | 0 | 53 |

Tableau 5.7 – Comparaison des méthodes de classification des émotions en utilisant des images de différentes poses de tête provenant de la base de données RafD

| Auteurs | Précisions |
|---|---|
| Wu et lin. [185] | 96.27% |
| Notre méthode | 96.55% |

exceptionnelle est attribuable à l'efficacité des étapes spécifiques de prétraitement des images, ainsi qu'à la puissance de l'architecture 2D-CNN que nous avons proposé.

## 5.4 Résultats et Discussions : Reconnaissance multimodale du stress

Cette étude vise à développer un système multimodale basé sur la fusion des expressions faciales et des signaux iPPG extraits à partir de vidéos faciales de la base de donnée UBFC-Phys. Dans cette section, nous exposons les résultats obtenus en adoptant une approche unimodale, puis en évoluant vers une approche multimodale.

Il est important de souligner que cette étude se divise en trois parties essentielles. La première partie est consacrée à la modalité de l'expression faciale, la deuxième à la modalité des signaux iPPG, et enfin, la troisième porte sur la fusion de ces deux modalités. Pour chaque modalité, des techniques de prétraitement ont été appliquées avant d'intégrer les données dans des réseaux d'apprentissage profond. Les détails de cette expérience sont présentés de manière approfondie dans la section 4.6.

### 5.4.1 Expressions faciales

L'architecture 3D-CNN proposée (consultez le tableau 4.11) a été entraînée sur 100 époques à l'aide de l'algorithme d'optimisation Adam, avec un learning rate de 0,001 et une fonction de perte d'entropie croisée catégorique (en anglais categorical cross-entropy loss). Le modèle développé dans le cadre de cette étude a démontré un taux de reconnaissance remarquable, avec une précision de validation de 80,33 % et une perte de validation de 23 %. La figure 5.8 illustre la matrice de confusion résultante de cette expérience.

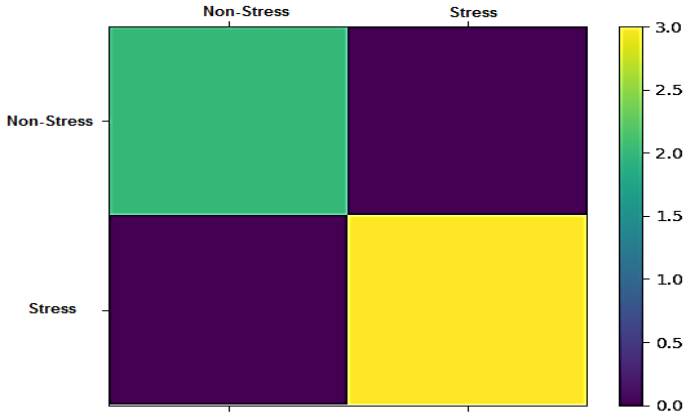



Figure 5.8 – Matrice de confusion - Expressions Faciales

### 5.4.2 Signaux iPPG

Dans le cadre de cette étude, le modèle d'apprentissage profond adopté est un réseau neuronal convolutif 1D-CNN, comme indiqué dans le tableau 4.12. Le modèle proposé a été entraîné sur une période de 50 époques en utilisant l'optimiseur Adam, avec un taux d'apprentissage fixé à 0,01. Pour évaluer l'efficacité du modèle, quatre paramètres fondamentaux ont été calculés : la précision (Accuracy AC), le score F1, la sensibilité (Recall RE), et la précision (Precision PR). Ces paramètres sont détaillés dans la section 5.2.1.

Une fois la phase d'entraînement terminée, les résultats suivants ont été obtenus : AC=75%, PR+ 75% , RE= 50% et F1= 66%. La figure 5.9 illustre respectivement la matrice de confusion ainsi que les courbes de perte de validation et d'entraînement.

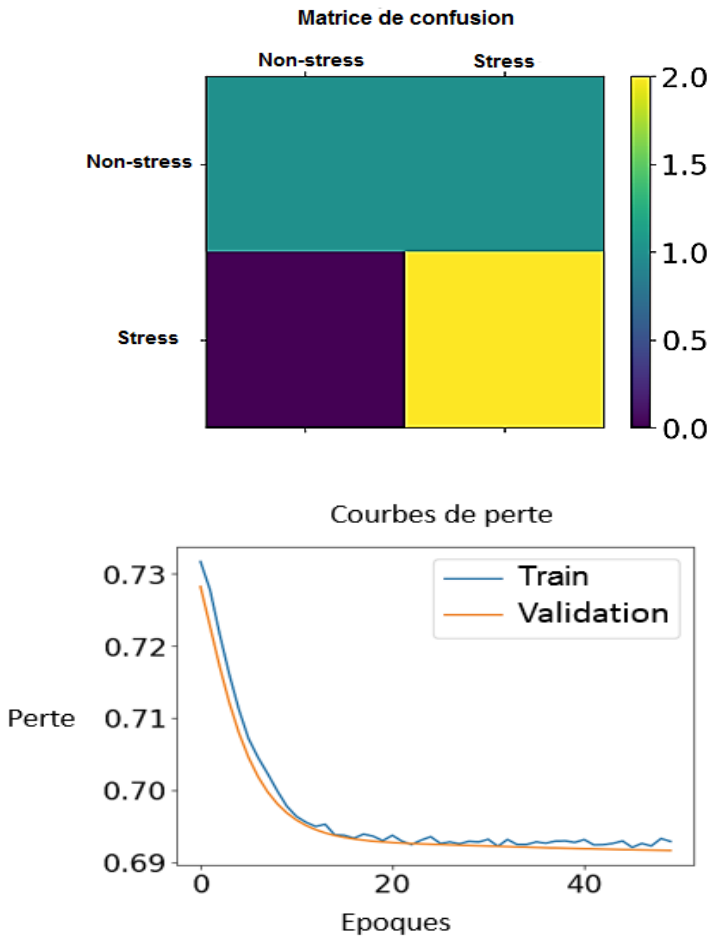



Figure 5.9 – Matrice de confusion et courbes de perte - Signaux iPPG.

### 5.4.3 Système Multimodale

Une fois que les caractéristiques de chaque modalité ont été extraites, on procède à leur fusion et à leur classification. Dans notre cas, nous avons choisi d'utiliser des couches de neurones entièrement connectées. La figure 4.18 présente en détail les détaillants du modèle de fusion proposé.

Ce modèle a été entraîné sur 50 époques en utilisant l'algorithme d'optimisation Adam. Un taux de reconnaissance élevé a été atteint, avec une précision de validation de 100 %. La figure 5.10 présente la matrice de confusion obtenue.

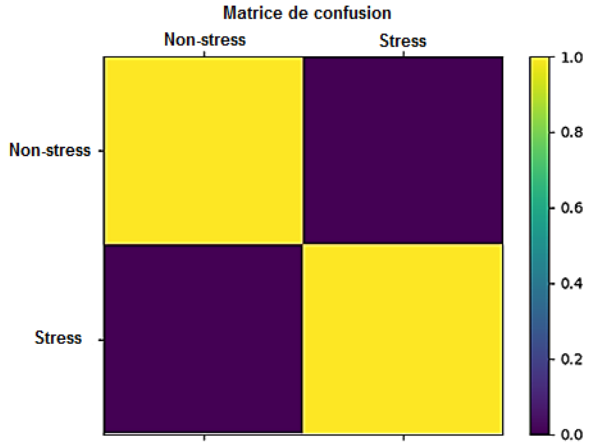



Figure 5.10 – Matrice de confusion pour classification multimodale du stress

### 5.4.4 Discussions et comparaison des résultats

Cette étude propose un système multimodal intégrant à la fois les expressions faciales et les signaux iPPG extraits des vidéos faciales. En raison du nombre limité de recherches dans ce domaine, nous avons comparé nos résultats à ceux de deux autres études évaluées dans la même base de données.

Le tableau ci-dessous (5.9) présente un résumé des divers résultats obtenus par les chercheurs dans la base de données UBFC-PHys.

Tableau 5.9 – Comparaison des performances de la détection du stress humain en utilisant des signaux physiologiques sans contact, des expressions faciales et l'approche multimodale avec l'ensemble de données UBFC-Phys

| Auteurs | Modalité | Méthodes | Précisions |
|---|---|---|---|
| Ouzar et al. [138] | Signaux iPPG utilisant MTTS-CAN | RF | 62.4% |
| | Expressions faciales | VGG16 | 82.48% |
| | Multimodale | MLP | 91.07% |
| Meziati Sabour et al. [160] | Signaux iVFC utilisant POS | SVM | 85.48% |
| Notre Méthode | Signaux iPPG utilisant ICA | 1D-CNN | 75% |
| | Expressions Faciales | 3D-CNN | 83.33% |
| | Multimodale | MLP | 100% |

Nos méthodes, comparées à celles rapportées dans d'autres travaux, se révèlent similaires, voire meilleures. Il est important de noter que notre recherche s'est concentrée sur l'utilisation de seulement 12 participants de la base de données UBFC-Phys, en raison de contraintes de capacité de calcul. Cette approche diffère de celle des autres études présentées dans le tableau

5.9, qui ont utilisé l'ensemble des données de tous les participants de cette base de données.

Il est important de souligner que les modèles 3D-CNN proposés pour les expressions faciales et les modèles 1D-CNN proposés pour les signaux iPPG dans cette étude ont le potentiel d'effectuer des tâches de classification. De plus, il a été constaté que le système multimodal améliore les performances par rapport aux systèmes unimodaux, avec une précision de validation atteignant 100% par rapport à l'approche unimodale atteignant 83.33% et 75 % utilisant les expressions faciales et signaux iPPG respectivement.

## 5.5 Conclusion

Dans ce chapitre, nous avons exposé les résultats de toutes les études expérimentales menées dans le cadre de cette thèse, comme présentées dans le chapitre 3.6.

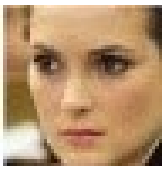
Dans un premier temps, nous avons évalué les résultats de notre étude portant sur la classification des émotions selon les échelles de valence et d'arousal. Cette étude propose une approche novatrice pour la reconnaissance automatique de l'affect, en exploitant des signaux iPPG sans contact. Nous avons obtenu une précision de 73,33 % sur l'échelle de valence et de 60 % sur l'échelle d'arousal.

Ensuite, nous avons présenté les résultats de notre deuxième étude, qui s'est concentrée sur la classification de sept émotions de base en tenant compte de différentes positions de la tête, qui a également obtenu des résultats élevés, atteignant une précision de 96,55 %.

Enfin, la dernière partie de ce chapitre a abordé les résultats de l'étude expérimentale visant à combiner deux modalités, à savoir les expressions faciales et les signaux iPPG. À travers cette approche multimodale, nous avons exploré l'impact bénéfique de fusionner les caractéristiques de chaque modalité pour améliorer les performances de classification atteignant une précision de 100 %.

Nos résultats témoignent de la robustesse de nos études, illustrant l'efficacité de la méthodologie adoptée tout au long de notre recherche. Que ce soit dans les techniques de prétraitement des données ou dans les architectures de réseaux d'apprentissage profond que nous avons développé, notre approche s'est avérée efficace. En outre, une comparaison approfondie avec d'autres travaux récents a permis de mettre en lumière les points forts de nos méthodes.

Tableau 5.8 – Prédictions émotionnelles sur les nouvelles images de la base de données LFW

| | | |
|---|---|---|
|  |  |  prediction = AN |
|  |  |  prediction = HA |
|  |  |  prediction = NE |
|  |  |  prediction = FE |
|  |  |  prediction = SU |

Chapitre 6

# Conclusions et perspectives

## 6.1 Conclusions

La reconnaissance automatique de l'affect est un domaine de recherche interdisciplinaire qui englobe différentes modalités. Avec l'évolution des architectures d'apprentissage profond, le domaine de reconnaissance de l'affect est devenu plus répandu et plus développé. Dans le cadre de cette thèse, nous avons mené une étude sur la reconnaissance automatique des émotions et du stress en utilisant des techniques d'apprentissage profond.

Basé sur l'idée que le visage humain est le premier moyen visible de communication émotionnelle. La méthode des expressions faciales a fait l'objet d'un grand intérêt de la part des scientifiques dans le domaine de la reconnaissance automatique des émotions humaines. Cependant, ces dernières années, un large éventail d'approches ont été proposées, utilisant diverses autres modalités, telles que la modalité physiologique. Les chercheurs sont de plus en plus attirés par cette modalité en raison de la fiabilité des données physiologiques, qui permet une détection plus précise de l'affect.

Bien que la modalité physiologique ait connu un succès remarquable dans la reconnaissance automatique de l'affect, elle présente des limites dans l'acquisition des données. Les individus peuvent percevoir l'utilisation de capteurs portés sur le corps comme inconfortable et peu pratique, ce qui peut entraîner une dégradation de la qualité des données. De plus, ces dispositifs sont souvent inappropriés pour une utilisation dans des situations réelles. Par conséquent, de récentes recherches proposent une nouvelle approche d'étude basée sur l'utilisation de signaux physiologiques sans contact pour la reconnaissance automatique de l'affect.

Les signaux physiologiques sans contact font référence à des types de signaux qui sont extraits à partir de vidéos faciales capturées par une simple caméra, tels que le iPPG et le iVFC. Il existe peu de recherches portant sur l'utilisation de ces signaux iPPG et iVFC sans contact dans le domaine de la reconnaissance automatique de l'affect.

Compte tenu de la variété des manifestations émotionnelles et du stress, les recherches récentes prônent une approche intégrant différentes modalités. Cette approche vise à améliorer les performances de la classification en combinant les données de différents types. La fusion de plusieurs modalités permet d'obtenir des informations plus riches sur les états affectifs humains, ce qui conduit à des résultats plus précis.

Cette thèse a été marquée par de nombreux défis de recherche auxquels nous avons été confrontés. Tout d'abord, une compréhension approfondie du domaine de la reconnaissance automatique de l'affect à travers différentes modalités s'est avérée nécessaire. Ensuite, l'exploration du domaine de l'extraction des signaux physiologiques à partir de vidéos faciales en couleurs RVB a constitué une étape cruciale. Nous avons également dû maîtriser les méthodes et les algorithmes pour extraire ces signaux physiologiques sans contact en utilisant des bases de données affectifs. Enfin, nous avons plongé dans le domaine de l'apprentissage profond, qui suscite un intérêt croissant dans le domaine de la reconnaissance automatique de l'affect.

Dans le deuxième chapitre de ce manuscrit, nous avons présenté une étude bibliographique sur les différents axes pertinents de notre recherche, revisitant les concepts d'émotion et de stress ainsi que leurs composantes. Ensuite, nous avons exposé les progrès récents réalisés dans le domaine de la reconnaissance automatique des émotions et du stress utilisant des réseaux d'apprentissage profond.

Le troisième chapitre offre un aperçu du domaine de l'extraction de signaux physiologiques à partir de vidéos faciales et de ses applications dans le domaine de la reconnaissance automatique de l'affect.

Après cela, dans le chapitre quatre, nous avons passé à la présentation de l'ensemble des expériences réalisées. Le premier travail consiste à proposer une étude comparative des algorithmes d'extraction de signaux iPPG les plus populaires, basés sur des bases de données affective, afin de trouver le meilleur algorithme de transmettre des signaux plus précis. Ensuite, nous avons passé à la réalisation d'un système de classification des émotions robustes et précis utilisant les signaux iPPG. Le principal avantage réside dans le développement du domaine de la reconnaissance automatique des émotions, en utilisant des signaux extraits à distance sans perturber les individus avec des capteurs placés sur le corps.

Ensuite, nous avons abordé l'utilisation des expressions faciales, considérée comme la modalité la plus accessible pour la collecte de données, bien qu'elle présente de nombreux défis. Nous avons proposé une étude dans le but d'obtenir un taux de reconnaissance élevé avec des images capturées dans différentes positions de tête et regards et avec des sujets de sexe et d'âge différents.

Le but ultime étudié dans cette thèse est de créer un système multimodale basé sur les

expressions faciales et les signaux iPPG. Dans ce cas, les expressions faciales utilisées sont spontanées et en temps réel.

Toutes nos études ont été mises en œuvre à l'aide de techniques d'apprentissage profond, car nous avons proposé plusieurs architectures différentes adaptées à divers types de données séquentielles et dynamiques. De plus, des étapes de prétraitement de données ont été appliquées avant la classification, garantissant ainsi la qualité et la clarté des données utilisées dans nos études.

Dans le cinquième chapitre, nous avons exposé l'ensemble des résultats obtenus lors des expériences menées dans cette thèse. Ces résultats mettent en lumière la puissance et les performances de la méthodologie que nous avons employée tout au long de nos travaux. Nous avons atteint une précision significative dans la classification des affects en exploitant les signaux iPPG extraits de vidéos faciales, ce qui constitue une approche novatrice et prometteuse pour faire évoluer le domaine de l'informatique affective. En outre, avec la modalité d'expressions faciales, nous avons réussi à atteindre un taux de reconnaissance élevé en surmontant plusieurs défis, que ce soit avec des données mimiques et dynamiques, ou avec des données séquentielles et des expressions spontanées. Avec notre approche multimodale, nos résultats démontrent clairement l'efficacité de cette approche par rapport à l'approche unimodale, atteignant un taux de précision de validation de 100%.

## 6.2 Perspectives

Les résultats présentés dans ce manuscrit de thèse sont prometteurs, grâce à la qualité des travaux réalisés. Toutefois, ils renferment plusieurs limites et défis qui ouvrent de nouvelles perspectives pour améliorer et développer la recherche dans le domaine de l'informatique affective à l'avenir.

Pour le premier travail, qui vise à réaliser une étude comparative de différentes méthodes d'extraction de signal iPPG dans différentes régions d'intérêt ; Il est très intéressant de développer ce type d'étude en ajoutant d'autres méthodes basées sur des architectures de DL pour l'extraction du signal iPPG. Ce qui nous donne un aperçu des différentes performances obtenues par rapport aux méthodes traditionnelles.
De plus, il est préférable d'ajouter d'autres types de régions d'intérêt, ce qui rend l'étude plus

complète.

Concernant les travaux sur la reconnaissance automatique des émotions et du stress par les signaux iPPG, la première limite est de trouver un algorithme robuste pour extraire des signaux iPPG précis. Dans notre cas, nous avons choisi l'algorithme ICA avec détection de visage et suppression de l'arrière-plan. Cependant, cette méthode ne donne pas de résultats efficaces avec certaines vidéos, et nous avons dû supprimer les signaux inexacts avant classification. Il s'agit de la première limite de notre étude, car dans un scénario d'application réel, on ne peut pas compter sur les capteurs portables pour éliminer les mauvais signaux iPPG.
C'est pourquoi il sera nécessaire à l'avenir de développer ou de sélectionner des méthodes plus puissantes pour extraire le signal iPPG dans toutes les vidéos. En plus, notre méthode est performante avec des données générées en conditions contrôlées en laboratoire, et il sera nécessaire à l'avenir de développer l'étude à partir de vidéos prises en conditions réelles.

Pour notre troisième étude sur la classification des sept émotions de base à partir d'expressions faciales, nous avons surmonté plusieurs défis majeurs, notamment la diversité des poses de tête et des regards. Ces résultats prometteurs ont été obtenus en utilisant des images représentant des expressions faciales mimiques, enregistrées dans des conditions contrôlées en laboratoire. Cela suggère qu'à l'avenir, cette étude pourrait être développée à l'aide de données sur des expressions faciales spontanées capturées dans des situations réelle et naturelle.

Le dernier point atteint dans cette thèse est la création d'un système multimodale combinant les signaux iPPG et les expressions faciales. Il s'agit d'un nouveau cadre d'étude qui permet une détection fiable des émotions avec uniquement une simple caméra. Une suite intéressante de cette étude consiste à ajouter d'autres modalités telles que la parole et les gestes corporels. Ce qui rend l'étude plus complète et le système plus puissant à adapter aux applications réelles et aux interfaces homme-machines.

# Bibliographie


[1] Shaldon Wade Naidoo et al. “Computer Vision: The Effectiveness of Deep Learning for Emotion Detection in Marketing Campaigns”. In : *International Journal of Advanced Computer Science and Applications* 135, p. 100. ISSN : 21565570, 2158107X. DOI : 10.14569/IJACSA.2022.01305100. URL : http://thesai.org/Publications/ViewPaper?Volume=13&Issue=5&Code=IJACSA&SerialNo=100.

[2] Sanna Kuusikko et al. “Emotion Recognition in Children and Adolescents with Autism Spectrum Disorders”. In : *Journal of Autism and Developmental Disorders*, p. 938-945. ISSN : 1573-3432. DOI : 10.1007/s10803-009-0700-0. URL : https://doi.org/10.1007/s10803-009-0700-0.

[3] Faisal Muhammad Shah et al. “Early Depression Detection from Social Network Using Deep Learning Techniques”. In : *IEEE Region 10 Symposium (TENSYMP)*, p. 823-826. DOI : 10.1109/TENSYMP50017.2020.9231008.

[4] Solange Denervaud et al. “Emotion recognition development: Preliminary evidence for an effect of school pedagogical practices”. en. In : *Learning and Instruction* 69, p. 101353. ISSN : 0959-4752. DOI : 10.1016/j.learninstruc.2020.101353. URL : https://www.sciencedirect.com/science/article/pii/S0959475219306383.

[5] Juanpablo Heredia et al. “Adaptive Multimodal Emotion Detection Architecture for Social Robots”. In : *IEEE Access* 10, p. 20727-20744. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2022.3149214.

[6] Patrícia J. Bota et al. “A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals”. In : *IEEE Access* 7, p. 140990-141020. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2019.2944001. URL : https://ieeexplore.ieee.org/document/8849996.

[7] Lin Shu et al. “A Review of Emotion Recognition Using Physiological Signals”. en. In : *Sensors* 187, p. 2074. DOI : 10.3390/s18072074. URL : https://www.mdpi.com/1424-8220/18/7/2074.

[8] Philipp V. Rouast, Marc T. P. Adam et Raymond Chiong. “Deep Learning for Human Affect Recognition: Insights and New Developments”. In : *IEEE Transactions on Affective Computing* 122, p. 524-543. ISSN : 1949-3045. DOI : 10.1109/TAFFC.2018.2890471.

[9] Wafa Mellouk et Wahida Handouzi. “Facial emotion recognition using deep learning: review and insights”. en. In : *Procedia Computer Science* 175, p. 689-694. ISSN : 1877-0509. DOI : 10.1016/j.procs.2020.07.101. URL : https://www.sciencedirect.com/science/article/pii/S1877050920318019.

[10] Bei Pan et al. “A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods”. In : *Neurocomputing* 561, p. 126866. ISSN : 0925-2312. DOI : 10.1016/j.neucom.2023.126866. URL : https://www.sciencedirect.com/science/article/pii/S092523122300989X.

[11] Soujanya Poria et al. “A review of affective computing: From unimodal analysis to multimodal fusion”. en. In : *Information Fusion* 37, p. 98-125. ISSN : 1566-2535. DOI : 10.1016/j.inffus.2017.02.003. URL : https://www.sciencedirect.com/science/article/pii/S1566253517300738.

[12] Jianhua Zhang et al. “Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review”. In : *Information Fusion* 59, p. 103-126. ISSN : 1566-2535. DOI : 10.1016/j.inffus.2020.01.011. URL : https://www.sciencedirect.com/science/article/pii/S1566253519302532.

[13] Ismoil Odinaev et al. “Robust Heart Rate Variability Measurement from Facial Videos”. en. In : *Bioengineering* 107, p. 851. ISSN : 2306-5354. DOI : 10.3390/bioengineering10070851. URL : https://www.mdpi.com/2306-5354/10/7/851.

[14] Giuseppe Boccignone et al. “An Open Framework for Remote-PPG Methods and Their Assessment”. In : *IEEE Access* 8, p. 216083-216103. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3040936.

[15] Yann LeCun, Yoshua Bengio et Geoffrey Hinton. “Deep learning”. en. In : *Nature* 5217553, p. 436-444. ISSN : 1476-4687. DOI : 10.1038/nature14539. URL : https://www.nature.com/articles/nature14539.

[16] Felipe Zago Canal et al. “A survey on facial emotion recognition techniques: A state-of-the-art literature review”. In : *Information Sciences* 582, p. 593-617. ISSN : 0020-0255. DOI : 10.1016/j.ins.2021.10.005. URL : https://www.sciencedirect.com/science/article/pii/S0020025521010136.

[17] Armelle Nugier. “Histoire et grands courants de recherche sur les émotions”. fr. In.

[18] Claudia Caffi et Richard W. Janney. “Toward a pragmatics of emotive communication”. In : *Journal of Pragmatics* 223, p. 325-373. ISSN : 0378-2166. DOI : 10.1016/0378-2166(94)90115-5. URL : https://www.sciencedirect.com/science/article/pii/0378216694901155.

[19] Lisa Feldman Barrett. “Solving the Emotion Paradox: Categorization and the Experience of Emotion”. en. In : *Personality and Social Psychology Review* 101, p. 20-46. ISSN : 1088-8683. DOI : 10.1207/s15327957pspr1001_2. URL : https://doi.org/10.1207/s15327957pspr1001_2.

[20] Beverley Fehr et James A. Russell. “Concept of emotion viewed from a prototype perspective”. In : *Journal of Experimental Psychology: General* 1133, p. 464-486. ISSN : 1939-2222. DOI : 10.1037/0096-3445.113.3.464.

[21] Paul R. Kleinginna et Anne M. Kleinginna. “A categorized list of emotion definitions, with suggestions for a consensual definition”. en. In : *Motivation and Emotion* 54, p. 345-379. ISSN : 1573-6644. DOI : 10.1007/BF00992553. URL : https://doi.org/10.1007/BF00992553.

[22] William James. “What is an Emotion?” In : *Mind* 934, p. 188-205. ISSN : 0026-4423. URL : https://www.jstor.org/stable/2246769.

[23] C. L. Lisetti. “Affective computing”. en. In : *Pattern Analysis and Applications* 11, p. 71-73. ISSN : 1433-755X. DOI : 10.1007/BF01238028. URL : https://doi.org/10.1007/BF01238028.

[24] Caroline Etienne. “Apprentissage profond appliqué à la reconnaissance des émotions dans la voix”. fr. Thèse de doct. Université Paris Saclay (COmUE). URL : https://theses.hal.science/tel-02479126.

[25] Marie-Claire Lemarchand-Chauvin. “L’influence des émotions sur la prise de parole en classe des enseignants stagiaires d’anglais de l’académie de Créteil”. fr. In : *SHS Web of Conferences* 81, p. 02001. ISSN : 2261-2424. DOI : 10.1051/shsconf/20208102001. URL : https://www.shs-conferences.org/articles/shsconf/abs/2020/09/shsconf_icodoc2019_02001/shsconf_icodoc2019_02001.html.

[26] Arthur L. Blumenthal. *A reappraisal of Wilhelm Wundt.* Evolving perspectives on the history of psychology. Washington, DC, US : American Psychological Association. ISBN : 978-1-55798-882-9. DOI : 10.1037/10421-004.

[27] Harold Schlosberg. “Three dimensions of emotion”. In : *Psychological Review* 612, p. 81-88. ISSN : 1939-1471. DOI : 10.1037/h0054570.

[28] James A. Russell. “How shall an emotion be called?” In : *Circumplex models of personality and emotions.* Washington, DC, US : American Psychological Association, p. 205-220. ISBN : 978-1-55798-380-0. DOI : 10.1037/10261-009.

[29] Philip Schmidt et al. *Wearable affect and stress recognition: A review*. doi : 10.48550/arXiv.1811.08854. url : http://arxiv.org/abs/1811.08854.

[30] Walter B. Cannon. *Bodily changes in pain, hunger, fear and rage: An account of recent researches into the function of emotional excitement*. Bodily changes in pain, hunger, fear and rage: An account of recent researches into the function of emotional excitement. New York, NY, US : D Appleton & Company. doi : 10.1037/10013-000.

[31] Hans Selye. "Confusion and Controversy in the Stress Field". In : *Journal of Human Stress* 12, p. 37-44. issn : 0097-840X. doi : 10.1080/0097840X.1975.9940406. url : https://doi.org/10.1080/0097840X.1975.9940406.

[32] Bo Zhang. "Reconnaissance de stress à partir de données hétérogènes". These de doctorat. Université de Lorraine. url : https://www.theses.fr/2017LORR0113.

[33] Suja Sreeith Panicker et Prakasam Gayathri. "A survey of machine learning techniques in physiology based mental stress detection systems". In : *Biocybernetics and Biomedical Engineering* 392, p. 444-469. issn : 0208-5216. doi : 10.1016/j.bbe.2019.01.004. url : https://www.sciencedirect.com/science/article/pii/S020852161830367X.

[34] Philip Schmidt et al. "Wearable-Based Affect Recognition—A Review". en. In : *Sensors* 1919, p. 4079. issn : 1424-8220. doi : 10.3390/s19194079. url : https://www.mdpi.com/1424-8220/19/19/4079.

[35] Gaetano Valenza et al. "Revealing Real-Time Emotional Responses: a Personalized Assessment based on Heartbeat Dynamics". en. In : *Scientific Reports* 41, p. 4998. issn : 2045-2322. doi : 10.1038/srep04998. url : https://www.nature.com/articles/srep04998.

[36] Ulrich Schimmack et Reisenzein Rainer. "Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation". In : *Emotion* 24, p. 412-417. issn : 1931-1516. doi : 10.1037/1528-3542.2.4.412.

[37] Craig A. Smith et Heather S. Scott. "A Componential Approach to the meaning of facial expressions". In : *The Psychology of Facial Expression*. Sous la dir. de James A. Russell et José Miguel Fernández-Dols. Studies in Emotion and Social Interaction. Cambridge : Cambridge University Press, p. 229-254. isbn : 978-0-521-58796-9. doi : 10.1017/CBO9780511659911.012. url : https://www.cambridge.org/core/books/psychology-of-facial-expression/componential-approach-to-the-meaning-of-facial-expressions/ADFE698CF3302CB88626E397B34AAB0D.

[38] Paul Ekman. "Facial expression and emotion". In : *American Psychologist* 484, p. 384-392. issn : 1935-990X. doi : 10.1037/0003-066X.48.4.384.

[39] Anna Tcherkassof. "Le sens dessus dessous des expressions faciales des émotions : vers un nouveau tournant paradigmatique". thesis. Université Grenoble Alpes ; CS 40700, 38058 Grenoble. url : https://hal.science/tel-01868279.

[40] Siao Zheng Bong, M. Murugappan et Sazali Yaacob. "Methods and approaches on inferring human emotional stress changes through physiological signals: a review". In : *International Journal of Medical Engineering and Informatics* 52, p. 152-162. issn : 1755-0653. doi : 10.1504/IJMEI.2013.053332. url : https://www.inderscienceonline.com/doi/abs/10.1504/IJMEI.2013.053332.

[41] Rateb Katmah et al. "A Review on Mental Stress Assessment Methods Using EEG Signals". en. In : *Sensors* 2115, p. 5043. issn : 1424-8220. doi : 10.3390/s21155043. url : https://www.mdpi.com/1424-8220/21/15/5043.

[42] Md. Mustafizur Rahman et al. "Recognition of human emotions using EEG signals: A review". In : *Computers in Biology and Medicine* 136, p. 104696. issn : 0010-4825. doi : 10.1016/j.compbiomed.2021.104696. url : https://www.sciencedirect.com/science/article/pii/S001048252100490X.

[43] Oliver Langner et al. “Presentation and validation of the Radboud Faces Database”. In : *Cognition and Emotion* 248, p. 1377-1388. ISSN : 0269-9931. DOI : 10.1080/02699930903485076. URL : https://doi.org/10.1080/02699930903485076.

[44] Yann LeCun et al. “Handwritten Digit Recognition with a Back-Propagation Network”. In : *Advances in Neural Information Processing Systems*. T. 2. Morgan-Kaufmann. URL : https://proceedings.neurips.cc/paper_files/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html.

[45] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In : *Proceedings of the IEEE* 8611, p. 2278-2324. ISSN : 1558-2256. DOI : 10.1109/5.726791. URL : https://ieeexplore.ieee.org/abstract/document/726791.

[46] Md Zahangir Alom et al. “A State-of-the-Art Survey on Deep Learning Theory and Architectures”. en. In : *Electronics* 83, p. 292. ISSN : 2079-9292. DOI : 10.3390/electronics8030292. URL : https://www.mdpi.com/2079-9292/8/3/292.

[47] Keiron O’Shea et Ryan Nash. *An Introduction to Convolutional Neural Networks*. DOI : 10.48550/arXiv.1511.08458. URL : http://arxiv.org/abs/1511.08458.

[48] Iuliana Tabian, Hailing Fu et Zahra Sharif Khodaei. “A Convolutional Neural Network for Impact Detection and Characterization of Complex Composite Structures”. en. In : *Sensors* 1922, p. 4933. ISSN : 1424-8220. DOI : 10.3390/s19224933. URL : https://www.mdpi.com/1424-8220/19/22/4933.

[49] Florentin Bieder, Robin Sandkühler et Philippe C. Cattin. *Comparison of Methods Generalizing Max- and Average-Pooling*. DOI : 10.48550/arXiv.2103.01746. URL : http://arxiv.org/abs/2103.01746.

[50] Serkan Kiranyaz et al. “1D convolutional neural networks and applications: A survey”. In : *Mechanical Systems and Signal Processing* 151, p. 107398. ISSN : 0888-3270. DOI : 10.1016/j.ymssp.2020.107398. URL : https://www.sciencedirect.com/science/article/pii/S0888327020307846.

[51] Xiaoyan Zhou Yiand Sun, Zheng-Jun Zha et Wenjun Zeng. “MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition”. In : p. 449-458. URL : https://openaccess.thecvf.com/content_cvpr_2018/html/Zhou_MiCT_Mixed_3D2D_CVPR_2018_paper.html.

[52] Sepp Hochreiter et Jürgen Schmidhuber. “Long Short-Term Memory”. In : *Neural Computation* 98, p. 1735-1780. ISSN : 0899-7667. DOI : 10.1162/neco.1997.9.8.1735.

[53] Benjamin Lindemann et al. “A survey on anomaly detection for technical systems using LSTM networks”. In : *Computers in Industry* 131, p. 103498. ISSN : 0166-3615. DOI : 10.1016/j.compind.2021.103498. URL : https://www.sciencedirect.com/science/article/pii/S0166361521001056.

[54] Yong Yu et al. “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures”. In : *Neural Computation* 317, p. 1235-1270. ISSN : 0899-7667. DOI : 10.1162/neco_a_01199. URL : https://doi.org/10.1162/neco_a_01199.

[55] Cheng-Jie Yang et al. “A Convolution Neural Network Based Emotion Recognition System using Multimodal Physiological Signals”. In : *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, p. 1-2. DOI : 10.1109/ICCE-Taiwan49838.2020.9258341.

[56] Bahareh Nakisa et al. “Automatic Emotion Recognition Using Temporal Multimodal Deep Learning”. In : *IEEE Access* 8, p. 225463-225474. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3027026.

[57] Luz Santamaria-Granados et al. “Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)”. In : *IEEE Access* 7, p. 57-67. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2018.2883213.

[58] Juan Abdon Miranda-Correa et al. “AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups”. In : *IEEE Transactions on Affective Computing* 122, p. 479-493. ISSN : 1949-3045. DOI : 10.1109/TAFFC.2018.2884461. URL : https://ieeexplore.ieee.org/abstract/document/8554112.

[59] Shao-Wei Wang et Sung-Nien Yu. “Emotion Recognition Based on Photoplethysmography Using ResNet and BiLSTM Networks”. In : *2021 International Conference on e-Health and Bioengineering (EHB)*, p. 1-4. DOI : 10.1109/EHB52898.2021.9657742. URL : https://ieeexplore.ieee.org/abstract/document/9657742.

[60] Fadi Al Machot et al. “A Deep-Learning Model for Subject-Independent Human Emotion Recognition Using Electrodermal Activity Sensors”. en. In : *Sensors* 197, p. 1659. DOI : 10.3390/s19071659. URL : https://www.mdpi.com/1424-8220/19/7/1659.

[61] Sander Koelstra et al. “DEAP: A Database for Emotion Analysis ;Using Physiological Signals”. In : *IEEE Transactions on Affective Computing* 31, p. 18-31. ISSN : 1949-3045. DOI : 10.1109/T-AFFC.2011.15.

[62] Mohammad Soleymani et al. “A Multimodal Database for Affect Recognition and Implicit Tagging”. In : *IEEE Transactions on Affective Computing* 31, p. 42-55. ISSN : 1949-3045. DOI : 10.1109/T-AFFC.2011.25.

[63] Min Seop Lee et al. “Fast Emotion Recognition Based on Single Pulse PPG Signal with Convolutional Neural Network”. en. In : *Applied Sciences* 916, p. 3355. DOI : 10.3390/app9163355. URL : https://www.mdpi.com/2076-3417/9/16/3355.

[64] Munaza Ramzan et Suma Dawn. “Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals”. In : *International Journal of Neuroscience* 1336, p. 587-597. ISSN : 0020-7454. DOI : 10.1080/00207454.2021.1941947. URL : https://doi.org/10.1080/00207454.2021.1941947.

[65] Russell Li et Zhandong Liu. “Stress detection using deep neural networks”. en. In : *BMC Medical Informatics and Decision Making* 2011, p. 285. ISSN : 1472-6947. DOI : 10.1186/s12911-020-01299-4. URL : https://doi.org/10.1186/s12911-020-01299-4.

[66] Fabrizio Albertetti, Alena Simalastar et Aïcha Rizzotti-Kaddouri. “Stress Detection with Deep Learning Approaches Using Physiological Signals”. en. In : *IoT Technologies for HealthCare*. Sous la dir. de Rossitza Goleva, Nuno Ricardo da Cruz Garcia et Ivan Miguel Pires. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham : Springer International Publishing, p. 95-111. ISBN : 978-3-030-69963-5. DOI : 10.1007/978-3-030-69963-5_7.

[67] Zeeshan Ahmad et al. “Multilevel Stress Assessment From ECG in a Virtual Reality Environment Using Multimodal Fusion”. In : *IEEE Sensors Journal* 2323, p. 29559-29570. ISSN : 1558-1748. DOI : 10.1109/JSEN.2023.3323290. URL : https://ieeexplore.ieee.org/abstract/document/10286334.

[68] Sami Elzeiny et Marwa Qaraqe. “Stress Classification Using Photoplethysmogram-Based Spatial and Frequency Domain Images”. en. In : *Sensors* 2018, p. 5312. ISSN : 1424-8220. DOI : 10.3390/s20185312. URL : https://www.mdpi.com/1424-8220/20/18/5312.

[69] Shan Li et Weihong Deng. “Deep Facial Expression Recognition: A Survey”. In : *IEEE Transactions on Affective Computing*, p. 1-1. ISSN : 1949-3045. DOI : 10.1109/TAFFC.2020.2981446.

[70] Jean Xavier et al. “A Multidimensional Approach to the Study of Emotion Recognition in Autism Spectrum Disorders”. In : *Frontiers in Psychology* 6. ISSN : 1664-1078. URL : https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01954.

[71] Javier Marín-Morales et al. “Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors”. en. In : *Scientific Reports* 81, p. 13657. ISSN : 2045-2322. DOI : 10.1038/s41598-018-32063-4. URL : https://www.nature.com/articles/s41598-018-32063-4.

[72] Ali Mollahosseini, David Chan et Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks". In : *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 1-10. DOI : 10.1109/WACV.2016.7477450. URL : https://ieeexplore.ieee.org/abstract/document/7477450.

[73] André Teixeira Lopes et al. "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order". In : *Pattern Recognition* 61, p. 610-628. ISSN : 0031-3203. DOI : 10.1016/j.patcog.2016.07.026. URL : https://www.sciencedirect.com/science/article/pii/S0031320316301753.

[74] Jun Cai et al. "Facial Expression Recognition Method Based on Sparse Batch Normalization CNN". In : *2018 37th Chinese Control Conference (CCC)*, p. 9608-9613. DOI : 10.23919/ChiCC.2018.8483567. URL : https://ieeexplore.ieee.org/abstract/document/8483567.

[75] M. Kalpana Chowdary, Tu N. Nguyen et D. Jude Hemanth. "Deep learning-based facial emotion recognition for human–computer interaction applications". en. In : *Neural Computing and Applications*. ISSN : 1433-3058. DOI : 10.1007/s00521-021-06012-8. URL : https://doi.org/10.1007/s00521-021-06012-8.

[76] Patrick Lucey et al. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, p. 94-101. DOI : 10.1109/CVPRW.2010.5543262. URL : https://ieeexplore.ieee.org/document/5543262.

[77] Mr Rohan Appasaheb Borgalli et Dr Sunil Surve. "Deep learning for facial emotion recognition using custom CNN architecture". en. In : *Journal of Physics: Conference Series* 22361, p. 012004. ISSN : 1742-6596. DOI : 10.1088/1742-6596/2236/1/012004. URL : https://dx.doi.org/10.1088/1742-6596/2236/1/012004.

[78] Ian J. Goodfellow et al. "Challenges in Representation Learning: A Report on Three Machine Learning Contests". en. In : *Neural Information Processing*. Sous la dir. de Minho Lee et al. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer, p. 117-124. ISBN : 978-3-642-42051-1. DOI : 10.1007/978-3-642-42051-1_16.

[79] Michael Lyons, Miyuki Kamachi et Jiro Gyoba. *The Japanese Female Facial Expression (JAFFE) Dataset*. eng. DOI : 10.5281/zenodo.3451524. URL : https://zenodo.org/records/3451524.

[80] Ashi Agarwal et Seba Susan. "Emotion Recognition from Masked Faces using Inception-v3". In : *2023 5th International Conference on Recent Advances in Information Technology (RAIT)*, p. 1-6. DOI : 10.1109/RAIT57693.2023.10126777. URL : https://ieeexplore.ieee.org/abstract/document/10126777.

[81] Mohan Karnati et al. "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey". In : *IEEE Transactions on Instrumentation and Measurement* 72, p. 1-31. ISSN : 1557-9662. DOI : 10.1109/TIM.2023.3243661. URL : https://ieeexplore.ieee.org/abstract/document/10041168.

[82] Rabie Helaly et al. "DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18". en. In : *Signal, Image and Video Processing* 176, p. 2731-2744. ISSN : 1863-1711. DOI : 10.1007/s11760-023-02490-6. URL : https://doi.org/10.1007/s11760-023-02490-6.

[83] Dae Hoe Kim et al. "Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition". In : *IEEE Transactions on Affective Computing* 102, p. 223-236. ISSN : 1949-3045. DOI : 10.1109/TAFFC.2017.2695999. URL : https://ieeexplore.ieee.org/abstract/document/7904596.

[84] M. Pantic et al. "Web-based database for facial expression analysis". In : *2005 IEEE International Conference on Multimedia and Expo*, 5 pp.–. DOI : 10.1109/ICME.2005.1521424. URL : https://ieeexplore.ieee.org/abstract/document/1521424.

[85] Wen-Jing Yan et al. “CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation”. en. In : *PLOS ONE* 91, e86041. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0086041. URL : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086041.

[86] Zhenbo Yu et al. “Spatio-temporal convolutional features with nested LSTM for facial expression recognition”. In : *Neurocomputing* 317, p. 50-57. ISSN : 0925-2312. DOI : 10.1016/j.neucom.2018.07.028. URL : https://www.sciencedirect.com/science/article/pii/S0925231218308634.

[87] Mohana M., P. Subashini et M. Krishnaveni. “Emotion Recognition from Facial Expression using Hybrid CNN-LSTM Network”. In : *International Journal of Pattern Recognition and Artificial Intelligence* 37. DOI : 10.1142/S0218001423560086.

[88] Rajesh Singh et al. “Facial expression recognition in videos using hybrid CNN & ConvLSTM”. en. In : *International Journal of Information Technology* 154, p. 1819-1830. ISSN : 2511-2112. DOI : 10.1007/s41870-023-01183-0. URL : https://doi.org/10.1007/s41870-023-01183-0.

[89] Sanaul Haq et Philip Jackson. *Speaker-Dependent Audio-Visual Emotion Recognition.*

[90] Jean Kossaifi et al. “AFEW-VA database for valence and arousal estimation in-the-wild”. In : *Image and Vision Computing* 65, p. 23-36. ISSN : 0262-8856. DOI : 10.1016/j.imavis.2017.02.001. URL : https://www.sciencedirect.com/science/article/pii/S0262885617300379.

[91] José Almeida et Fátima Rodrigues. “Facial Expression Recognition System for Stress Detection with Deep Learning:” en. In : *Proceedings of the 23rd International Conference on Enterprise Information Systems*. Online Streaming, — Select a Country — : SCITEPRESS - Science et Technology Publications, p. 256-263. ISBN : 978-989-758-509-8. DOI : 10.5220/0010474202560263. URL : https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010474202560263.

[92] Jin Zhang et al. “Detecting Negative Emotional Stress Based on Facial Expression in Real Time”. In : *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, p. 430-434. DOI : 10.1109/SIPROCESS.2019.8868735.

[93] Guoying Zhao et al. “Facial expression recognition from near-infrared videos”. In : *Image and Vision Computing* 299, p. 607-619. ISSN : 0262-8856. DOI : 10.1016/j.imavis.2011.07.002. URL : https://www.sciencedirect.com/science/article/pii/S0262885611000515.

[94] Mira Jeong et Byoung Chul Ko. “Driver’s Facial Expression Recognition in Real-Time for Safe Driving”. en. In : *Sensors* 1812, p. 4270. ISSN : 1424-8220. DOI : 10.3390/s18124270. URL : https://www.mdpi.com/1424-8220/18/12/4270.

[95] Taejae Jeon et al. “Deep-Learning-Based Stress Recognition with Spatial-Temporal Facial Information”. en. In : *Sensors* 2122, p. 7498. ISSN : 1424-8220. DOI : 10.3390/s21227498. URL : https://www.mdpi.com/1424-8220/21/22/7498.

[96] Saskia Koldijk et al. “The SWELL Knowledge Work Dataset for Stress and User Modeling Research”. In : *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI ’14. New York, NY, USA : Association for Computing Machinery, p. 291-298. ISBN : 978-1-4503-2885-2. DOI : 10.1145/2663204.2663257. URL : https://doi.org/10.1145/2663204.2663257.

[97] Giorgos Giannakakis et al. “Automatic stress analysis from facial videos based on deep facial action units recognition”. en. In : *Pattern Analysis and Applications* 253, p. 521-535. ISSN : 1433-755X. DOI : 10.1007/s10044-021-01012-9. URL : https://doi.org/10.1007/s10044-021-01012-9.

[98] Wenying Yu et al. “Emotion Recognition from Facial Expressions and Contactless Heart Rate Using Knowledge Graph”. In : *2020 IEEE International Conference on Knowledge Graph (ICKG)*, p. 64-69. DOI : 10.1109/ICBK50248.2020.00019. URL : https://ieeexplore.ieee.org/abstract/document/9194536.

[99] Nicu Sebe et al. “Multimodal approaches for emotion recognition: a survey”. In : *Internet Imaging VI*. T. 5670. SPIE, p. 56-67. DOI : 10.1117/12.600746. URL : https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5670/0000/Multimodal-approaches-for-emotion-recognition-a-survey/10.1117/12.600746.full.

[100] Carlos Busso et al. “Analysis of emotion recognition using facial expressions, speech and multimodal information”. In : *Proceedings of the 6th international conference on Multimodal interfaces*. ICMI ’04. New York, NY, USA : Association for Computing Machinery, p. 205-211. ISBN : 978-1-58113-995-2. DOI : 10.1145/1027933.1027968. URL : https://doi.org/10.1145/1027933.1027968.

[101] SeungJun Oh et Dong-Keun Kim. “Comparative Analysis of Emotion Classification Based on Facial Expression and Physiological Signals Using Deep Learning”. en. In : *Applied Sciences* 123, p. 1286. ISSN : 2076-3417. DOI : 10.3390/app12031286. URL : https://www.mdpi.com/2076-3417/12/3/1286.

[102] Chaudhary Aqdus et al. *Deep Emotion Recognition through Upper Body Movements and Facial Expression*. DOI : 10.5220/0010359506690679.

[103] Panagiotis Tzirakis et al. “End-to-End Multimodal Emotion Recognition Using Deep Neural Networks”. In : *IEEE Journal of Selected Topics in Signal Processing* 118, p. 1301-1309. ISSN : 1941-0484. DOI : 10.1109/JSTSP.2017.2764438. URL : https://ieeexplore.ieee.org/abstract/document/8070966.

[104] Asif Iqbal Middya, Baibhav Nag et Sarbani Roy. “Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities”. In : *Knowledge-Based Systems* 244, p. 108580. ISSN : 0950-7051. DOI : 10.1016/j.knosys.2022.108580. URL : https://www.sciencedirect.com/science/article/pii/S0950705122002593.

[105] Yongrui Huang et al. “Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition”. en. In : *Computational Intelligence and Neuroscience* 2017, e2107451. ISSN : 1687-5265. DOI : 10.1155/2017/2107451. URL : https://www.hindawi.com/journals/cin/2017/2107451/.

[106] Qingyang Zhu, Guanming Lu et Jingjie Yan. “Valence-Arousal Model based Emotion Recognition using EEG, peripheral physiological signals and Facial Expression”. In : *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*. ICMLSC ’20. New York, NY, USA : Association for Computing Machinery, p. 81-85. ISBN : 978-1-4503-7631-0. DOI : 10.1145/3380688.3380694. URL : https://doi.org/10.1145/3380688.3380694.

[107] Minjia Li et al. “Multistep Deep System for Multimodal Emotion Detection With Invalid Data in the Internet of Things”. In : *IEEE Access* 8, p. 187208-187221. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3029288. URL : https://ieeexplore.ieee.org/abstract/document/9216023.

[108] Fabien Ringeval et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”. In : *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, p. 1-8. DOI : 10.1109/FG.2013.6553805. URL : https://ieeexplore.ieee.org/abstract/document/6553805.

[109] Nastaran Saffaryazdi et al. “Using Facial Micro-Expressions in Combination With EEG and Physiological Signals for Emotion Recognition”. In : *Frontiers in Psychology* 13. ISSN : 1664-1078. URL : https://www.frontiersin.org/articles/10.3389/fpsyg.2022.864047.

[110] Chuan-Yu Chang et al. “Emotion recognition with consideration of facial expression and physiological signals”. In : *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, p. 278-283. DOI : 10.1109/CIBCB.2009.4925739. URL : https://ieeexplore.ieee.org/abstract/document/4925739.

[111] Wonju Seo et al. “Deep Learning Approach for Detecting Work-Related Stress Using Multimodal Signals”. In : *IEEE Sensors Journal* 2212, p. 11892-11902. ISSN : 1558-1748. DOI : 10.1109/JSEN.2022.3170915. URL : https://ieeexplore.ieee.org/abstract/document/9764756.

[112] Guoliang XIANG et al. “A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals”. In : *Engineering Applications of Artificial Intelligence* 130, p. 107772. ISSN : 0952-1976. DOI : 10.1016/j.engappai.2023.107772. URL : https://www.sciencedirect.com/science/article/pii/S0952197623019565.

[113] Muhammad Anas HASNUL et al. “Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review”. In : *Sensors (Basel, Switzerland)* 2115, p. 5015. ISSN : 1424-8220. DOI : 10.3390/s21155015. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8348698/.

[114] U. RAJENDRA ACHARYA et al. “Heart rate variability: a review”. en. In : *Medical and Biological Engineering and Computing* 4412, p. 1031-1051. ISSN : 1741-0444. DOI : 10.1007/s11517-006-0119-0. URL : https://doi.org/10.1007/s11517-006-0119-0.

[115] Thijs VANDENBERK et al. “Clinical Validation of Heart Rate Apps: Mixed-Methods Evaluation Study”. EN. In : *JMIR mHealth and uHealth* 58, e7254. DOI : 10.2196/mhealth.7254. URL : https://mhealth.jmir.org/2017/8/e129.

[116] Muhammad WASIMUDDIN et al. “Stages-Based ECG Signal Analysis From Traditional Signal Processing to Machine Learning Approaches: A Survey”. In : *IEEE Access* 8, p. 177782-177803. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3026968. URL : https://ieeexplore.ieee.org/document/9206538?denied=.

[117] C SARITHA, V SUKANYA et Y Narasimha MURTHY. “ECG Signal Analysis Using Wavelet Transforms”. en. In.

[118] R. Gayathri PRIYADARSHINI et al. “Review of PPG signal using Machine Learning Algorithms for Blood Pressure and Glucose Estimation”. en. In : *IOP Conference Series: Materials Science and Engineering* 10841, p. 012031. ISSN : 1757-899X. DOI : 10.1088/1757-899X/1084/1/012031. URL : https://dx.doi.org/10.1088/1757-899X/1084/1/012031.

[119] Alrick B. HERTZMAN. “Photoelectric Plethysmography of the Fingers and Toes in Man”. en. In : *Proceedings of the Society for Experimental Biology and Medicine* 373, p. 529-534. ISSN : 0037-9727. DOI : 10.3181/00379727-37-9630. URL : https://journals.sagepub.com/doi/abs/10.3181/00379727-37-9630.

[120] R. Gayathri PRIYADARSHINI et al. “Review of PPG signal using Machine Learning Algorithms for Blood Pressure and Glucose Estimation”. en. In : *IOP Conference Series: Materials Science and Engineering* 10841, p. 012031. ISSN : 1757-899X. DOI : 10.1088/1757-899X/1084/1/012031. URL : https://dx.doi.org/10.1088/1757-899X/1084/1/012031.

[121] Jermana L. MORAES et al. “Advances in Photopletysmography Signal Analysis for Biomedical Applications”. en. In : *Sensors* 186, p. 1894. ISSN : 1424-8220. DOI : 10.3390/s18061894. URL : https://www.mdpi.com/1424-8220/18/6/1894.

[122] Toshiyo TAMURA et Yuka MAEDA. “Photoplethysmogram”. en. In : *Seamless Healthcare Monitoring: Advancements in Wearable, Attachable, and Invisible Devices*. Sous la dir. de Toshiyo TAMURA et Wenxi CHEN. Cham : Springer International Publishing, p. 159-192. ISBN : 978-3-319-69362-0. DOI : 10.1007/978-3-319-69362-0_6. URL : https://doi.org/10.1007/978-3-319-69362-0_6.

[123] Harald M. STAUSS. “Heart rate variability”. In : *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 2855, R927-R931. ISSN : 0363-6119. DOI : 10.1152/ajpregu.00452.2003. URL : https://journals.physiology.org/doi/full/10.1152/ajpregu.00452.2003.

[124] Christine PERRET-GUILLAUME, Laure JOLY et Athanase BENETOS. “Heart Rate as a Risk Factor for Cardiovascular Disease”. In : *Progress in Cardiovascular Diseases* 521, p. 6-10. ISSN : 0033-0620. DOI : 10.1016/j.pcad.2009.05.003. URL : https://www.sciencedirect.com/science/article/pii/S0033062009000322.

[125] Alexander Haensel et al. “The relationship between heart rate variability and inflammatory markers in cardiovascular diseases”. In : *Psychoneuroendocrinology* 3310, p. 1305-1312. ISSN : 0306-4530. DOI : 10.1016/j.psyneuen.2008.08.007. URL : https://www.sciencedirect.com/science/article/pii/S0306453008002126.

[126] Jianping Zhu, Lizhen Ji et Chengyu Liu. “Heart rate variability monitoring for emotion and disorders of emotion”. en. In : *Physiological Measurement* 406, p. 064004. ISSN : 0967-3334. DOI : 10.1088/1361-6579/ab1887. URL : https://dx.doi.org/10.1088/1361-6579/ab1887.

[127] Hongyu Shi et al. “Differences of Heart Rate Variability Between Happiness and Sadness Emotion States: A Pilot Study”. en. In : *Journal of Medical and Biological Engineering* 374, p. 527-539. ISSN : 2199-4757. DOI : 10.1007/s40846-017-0238-0. URL : https://doi.org/10.1007/s40846-017-0238-0.

[128] María Teresa Valderas et al. “Human emotion recognition using heart rate variability analysis with spectral bands based on respiration”. In : *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, p. 6134-6137. DOI : 10.1109/EMBC.2015.7319792. URL : https://ieeexplore.ieee.org/abstract/document/7319792.

[129] M. Bolanos, H. Nazeran et E. Haltiwanger. “Comparison of Heart Rate Variability Signal Features Derived from Electrocardiography and Photoplethysmography in Healthy Individuals”. In : *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 4289-4294. DOI : 10.1109/IEMBS.2006.260607. URL : https://ieeexplore.ieee.org/abstract/document/4462749.

[130] Wan-Hua Lin et al. “Comparison of Heart Rate Variability from PPG with That from ECG”. en. In : *The International Conference on Health Informatics*. Sous la dir. d’Yuan-Ting Zhang. IFMBE Proceedings. Cham : Springer International Publishing, p. 213-215. ISBN : 978-3-319-03005-0. DOI : 10.1007/978-3-319-03005-0_54.

[131] Fred Shaffer et J. P. Ginsberg. “An Overview of Heart Rate Variability Metrics and Norms”. English. In : *Frontiers in Public Health* 5. ISSN : 2296-2565. DOI : 10.3389/fpubh.2017.00258. URL : https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2017.00258/full.

[132] Galya Georgieva-Tsaneva et Evgeniya Gospodinova. “Comparative Heart Rate Variability Analysis of ECG, Holter and PPG Signals”. In : *International Journal of Advanced Computer Science and Applications* 12. DOI : 10.14569/IJACSA.2021.0121261.

[133] Duncan Luguern. “Nouvelle approche pour l’estimation du rythme respiratoire basée sur la photoplethysmographie sans contact”. fr. Thèse de doct. Université Bourgogne Franche-Comté. URL : https://u-bourgogne.hal.science/tel-03169483.

[134] Ming-Zher Poh, Daniel J. McDuff et Rosalind W. Picard. “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.” EN. In : *Optics Express* 1810, p. 10762-10774. ISSN : 1094-4087. DOI : 10.1364/OE.18.010762. URL : https://www.osapublishing.org/oe/abstract.cfm?uri=oe-18-10-10762.

[135] Jingjing Luo et al. “Auxiliary Assessment of Cardiovascular Health Using High-Dimensional Characteristics of Camera-Based iPPG Monitoring”. In : *IEEE Sensors Journal* 2321, p. 26587-26596. ISSN : 1558-1748. DOI : 10.1109/JSEN.2023.3309994. URL : https://ieeexplore.ieee.org/abstract/document/10241319.

[136] Ge Xu et al. “Rational selection of RGB channels for disease classification based on IPPG technology”. EN. In : *Biomedical Optics Express* 134, p. 1820-1833. ISSN : 2156-7085. DOI : 10.1364/BOE.451736. URL : https://opg.optica.org/boe/abstract.cfm?uri=boe-13-4-1820.

[137] Abdoullah Bella et al. “Monitoring of Physiological Signs and Their Impact on The Covid-19 Pandemic: Review”. en. In : *E3S Web of Conferences* 229, p. 01030. ISSN : 2267-1242. DOI : 10.1051/e3sconf/202122901030. URL : https://www.e3s-conferences.org/articles/e3sconf/abs/2021/05/e3sconf_iccsre2021_01030/e3sconf_iccsre2021_01030.html.

[138] Yassine Ouzar et al. “Multimodal Stress State Detection from Facial Videos Using Physiological Signals and Facial Features”. en. In : *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*. Sous la dir. de Jean-Jacques Rousseau et Bill Kapralos. Lecture Notes in Computer Science. Cham : Springer Nature Switzerland, p. 139-150. ISBN : 978-3-031-37745-7. DOI : 10.1007/978-3-031-37745-7_10.

[139] Peipeng Yu et al. “A Survey on Deepfake Video Detection”. en. In : *IET Biometrics* 106, p. 607-624. ISSN : 2047-4946. DOI : 10.1049/bme2.12031. URL : https://onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12031.

[140] Dae-Yeol Kim, Kwangkee Lee et Chae-Bong Sohn. “Assessment of ROI Selection for Facial Video-Based rPPG”. en. In : *Sensors* 2123, p. 7923. ISSN : 1424-8220. DOI : 10.3390/s21237923. URL : https://www.mdpi.com/1424-8220/21/23/7923.

[141] Xun Chen et al. “Video-Based Heart Rate Measurement: Recent Advances and Future Prospects”. In : *IEEE Transions on Instrumentation and Measurement* 6810, p. 3600-3615. ISSN : 1557-9662. DOI : 10.1109/TIM.2018.2879706. URL : https://ieeexplore.ieee.org/abstract/document/8552414.

[142] P. Viola et M. Jones. “Rapid object detection using a boosted cascade of simple features”. In : *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* T. 1, p. I-I. DOI : 10.1109/CVPR.2001.990517.

[143] Ning Zhang, Junmin Luo et Wuqi Gao. “Research on Face Detection Technology Based on MTCNN”. In : *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, p. 154-158. DOI : 10.1109/ICCNEA50255.2020.00040. URL : https://ieeexplore.ieee.org/abstract/document/9239720.

[144] Nataliya Boyko, Oleg Basystiuk et Nataliya Shakhovska. “Performance Evaluation and Comparison of Software for Face Recognition, Based on Dlib and Opencv Library”. In : *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, p. 478-482. DOI : 10.1109/DSMP.2018.8478556. URL : https://ieeexplore.ieee.org/abstract/document/8478556.

[145] Wenjin Wang, Sander Stuijk et Gerard de Haan. “A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation”. In : *IEEE Transactions on Biomedical Engineering* 639, p. 1974-1984. ISSN : 1558-2531. DOI : 10.1109/TBME.2015.2508602. URL : https://ieeexplore.ieee.org/abstract/document/7355301.

[146] R. M. Fouad, Osama A. Omer et Moustafa H. Aly. “Optimizing Remote Photoplethysmography Using Adaptive Skin Segmentation for Real-Time Heart Rate Monitoring”. In : *IEEE Access* 7, p. 76513-76528. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2019.2922304. URL : https://ieeexplore.ieee.org/abstract/document/8735861.

[147] Nawaf Hazim Barnouti, Mohanad Hazim Nsaif Al-Mayyahi et Sinan Sameer Mahmood Al-Dabbagh. “Real-Time Face Tracking and Recognition System Using Kanade-Lucas-Tomasi and Two-Dimensional Principal Component Analysis”. In : *2018 International Conference on Advanced Science and Engineering (ICOASE)*, p. 24-29. DOI : 10.1109/ICOASE.2018.8548818. URL : https://ieeexplore.ieee.org/abstract/document/8548818.

[148] M. A. Hassan et al. “Heart rate estimation using facial video: A review”. In : *Biomedical Signal Processing and Control* 38, p. 346-360. ISSN : 1746-8094. DOI : 10.1016/j.bspc.2017.07.004. URL : https://www.sciencedirect.com/science/article/pii/S1746809417301362.

[149] Puneet Singh Lamba et Deepali Virmani. “Contactless heart rate estimation from face videos”. In : *Journal of Statistics and Management Systems* 237, p. 1275-1284. ISSN : 0972-0510. DOI : 10.1080/09720510.2020.1799584. URL : https://doi.org/10.1080/09720510.2020.1799584.

[150] Lujia Yang et al. “Integration Model of Deep Forgery Video Detection Based on rPPG and Spatiotemporal Signal”. en. In : *Green, Pervasive, and Cloud Computing*. Sous la dir. d’Hai Jin et al. Lecture Notes in Computer Science. Singapore : Springer Nature, p. 113-127. ISBN : 978-981-9998-93-7. DOI : 10.1007/978-981-99-9893-7_9.

[151] Daniel Wedekind et al. “Assessment of blind source separation techniques for video-based cardiac pulse extraction”. In : *Journal of Biomedical Optics* 223, p. 035002. ISSN : 1083-3668, 1560-2281. DOI : 10.1117/1.JBO.22.3.035002. URL : https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-22/issue-3/035002/Assessment-of-blind-source-separation-techniques-for-video-based-cardiac/10.1117/1.JBO.22.3.035002.full.

[152] Smera Premkumar et Duraisamy Jude Hemanth. “Intelligent Remote Photoplethysmography-Based Methods for Heart Rate Estimation from Face Videos: A Survey”. en. In : *Informatics* 93, p. 57. ISSN : 2227-9709. DOI : 10.3390/informatics9030057. URL : https://www.mdpi.com/2227-9709/9/3/57.

[153] Anton M. Unakafov. “Pulse rate estimation using imaging photoplethysmography: generic framework and comparison of methods on a publicly available dataset”. en. In : *Biomedical Physics & Engineering Express* 44, p. 045001. ISSN : 2057-1976. DOI : 10.1088/2057-1976/aabd09. URL : https://doi.org/10.1088/2057-1976/aabd09.

[154] Qi Zhang et Shuang Song. “Heart Rate Variability Parameters Extraction Based on Facial Video”. In : *2018 IEEE International Conference on Information and Automation (ICIA)*, p. 586-590. DOI : 10.1109/ICInfA.2018.8812485. URL : https://ieeexplore.ieee.org/abstract/document/8812485.

[155] Viktor Kessler et al. “Pain recognition with camera photoplethysmography”. In : *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, p. 1-5. DOI : 10.1109/IPTA.2017.8310110.

[156] Yannick Benezeth et al. “Remote heart rate variability for emotional state monitoring”. In : *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, p. 153-156. DOI : 10.1109/BHI.2018.8333392.

[157] Kun Zheng et al. “Non-Contact Heart Rate Detection When Face Information Is Missing during Online Learning”. en. In : *Sensors* 2024, p. 7021. DOI : 10.3390/s20247021. URL : https://www.mdpi.com/1424-8220/20/24/7021.

[158] Zitong Yu, Xiaobai Li et Guoying Zhao. “Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks”. In : *arXiv:1905.02419 [cs]*. URL : http://arxiv.org/abs/1905.02419.

[159] Choubeila Maaoui, Frederic Bousefsaf et Alain Pruski. “Automatic human stress detection based on webcam photoplethysmographic signals”. In : *Journal of Mechanics in Medicine and Biology* 1604, p. 1650039. ISSN : 0219-5194. DOI : 10.1142/S0219519416500391. URL : https://www.worldscientific.com/doi/abs/10.1142/S0219519416500391.

[160] Rita Meziati Sabour et al. “UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress”. In : *IEEE Transactions on Affective Computing*, p. 1-1. ISSN : 1949-3045. DOI : 10.1109/TAFFC.2021.3056960.

[161] Yassine Ouzar et al. “Video-Based Multimodal Spontaneous Emotion Recognition Using Facial Expressions and Physiological Signals”. en. In : p. 2460-2469. URL : https://openaccess.thecvf.com/content/CVPR2022W/ABAW/html/Ouzar_Video-Based_Multimodal_Spontaneous_Emotion_Recognition_Using_Facial_Expressions_and_Physiological_CVPRW_2022_paper.html.

[162] Zheng Zhang et al. “Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis”. In : p. 3438-3446. URL : https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Zhang_Multimodal_Spontaneous_Emotion_CVPR_2016_paper.html.

[163] Xuesong Niu et al. “RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation”. In : *IEEE Transactions on Image Processing* 29, p. 2409-2423. ISSN : 1941-0042. DOI : 10.1109/TIP.2019.2947204. URL : https://ieeexplore.ieee.org/abstract/document/8879658.

[164] Wim Verkruysse, Lars O. Svaasand et J. Stuart Nelson. “Remote plethysmographic imaging using ambient light.” EN. In : *Optics Express* 1626, p. 21434-21445. ISSN : 1094-4087. DOI : 10.1364/OE.16.021434. URL : https://www.osapublishing.org/oe/abstract.cfm?uri=oe-16-26-21434.

[165] Gerard de Haan et Vincent Jeanne. “Robust Pulse Rate From Chrominance-Based rPPG”. In : *IEEE Transactions on Biomedical Engineering* 6010, p. 2878-2886. ISSN : 1558-2531. DOI : 10.1109/TBME.2013.2266196.

[166] Wenjin Wang et al. “Algorithmic Principles of Remote PPG”. In : *IEEE Transactions on Biomedical Engineering* 647, p. 1479-1491. ISSN : 1558-2531. DOI : 10.1109/TBME.2016.2609282.

[167] Daniel McDuff et Ethan Blackford. “iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox”. In : *arXiv:1901.04366 [cs]*. URL : http://arxiv.org/abs/1901.04366.

[168] Xuejun Liao et L. Carin. “A new algorithm for independent component analysis with or without constraints”. In : *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, p. 413-417. DOI : 10.1109/SAM.2002.1191072. URL : https://ieeexplore.ieee.org/abstract/document/1191072.

[169] Chen Wang, Thierry Pun et Guillaume Chanel. “A Comparative Survey of Methods for Remote Heart Rate Detection From Frontal Face Videos”. In : *Frontiers in Bioengineering and Biotechnology* 6, p. 33. ISSN : 2296-4185. DOI : 10.3389/fbioe.2018.00033. URL : https://www.frontiersin.org/article/10.3389/fbioe.2018.00033.

[170] Carlos Carreiras et al. “Biosppy: Biosignal processing in python”. In : *Accessed on* 328, p. 2018. URL : https://github.com/PIA-Group/BioSPPy/.

[171] Min Lee et al. “PPG and EMG Based Emotion Recognition using Convolutional Neural Network:” en. In : *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics*. Prague, Czech Republic : SCITEPRESS - Science et Technology Publications, p. 595-600. ISBN : 978-989-758-380-3. DOI : 10.5220/0007797005950600. URL : http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007797005950600.

[172] Muhammad Najam Dar et al. “CNN and LSTM-Based Emotion Charting Using Physiological Signals”. en. In : *Sensors* 2016, p. 4551. ISSN : 1424-8220. DOI : 10.3390/s20164551. URL : https://www.mdpi.com/1424-8220/20/16/4551 (visité le 14/10/2022).

[173] Nazmun Nahar et al. “A Hybrid CNN-LSTM-Based Emotional Status Determination using Physiological Signals”. en. In : *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering*. Sous la dir. de M. Shamim Kaiser et al. Lecture Notes in Networks and Systems. Singapore : Springer Nature, p. 149-161. ISBN : 9789811675973. DOI : 10.1007/978-981-16-7597-3_12.

[174] Ayushi Sharma et al. “Object Detection using OpenCV and Python”. In : *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, p. 501-505. DOI : 10.1109/ICAC3N53548.2021.9725638. URL : https://ieeexplore.ieee.org/abstract/document/9725638.

[175] Andreas Rössler et al. *FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces*. DOI : 10.48550/arXiv.1803.09179. URL : http://arxiv.org/abs/1803.09179.

[176] Jad Haddad, Olivier Lezoray et Philippe Hamel. “3D-CNN for Facial Emotion Recognition in Videos”. en. In : *Advances in Visual Computing*. Sous la dir. de George Bebis et al. Lecture Notes in Computer Science. Cham : Springer International Publishing, p. 298-309. isbn : 978-3-030-64559-5. doi : 10.1007/978-3-030-64559-5_23.

[177] Nesime Tatbul et al. “Precision and Recall for Time Series”. In : *Advances in Neural Information Processing Systems*. T. 31. Curran Associates, Inc. url : https://proceedings.neurips.cc/paper_files/paper/2018/hash/8f468c873a32bb0619eaeb2050ba45d1-Abstract.html.

[178] Ruijing Yang et al. “Non-contact Pain Recognition from Video Sequences with Remote Physiological Measurements Prediction”. In : *arXiv:2105.08822 [cs]*. url : http://arxiv.org/abs/2105.08822.

[179] L. C. Lampier et al. “A Preliminary Approach to Identify Arousal and Valence Using Remote Photoplethysmography”. en. In : *XXVII Brazilian Congress on Biomedical Engineering*. Sous la dir. de Teodiano Freire Bastos-Filho, Eliete Maria de Oliveira Caldeira et Anselmo Frizera-Neto. IFMBE Proceedings. Cham : Springer International Publishing, p. 1659-1664. isbn : 978-3-030-70601-2. doi : 10.1007/978-3-030-70601-2_242.

[180] Viraj Mavani, Shanmuganathan Raman et Krishna P. Miyapuram. “Facial Expression Recognition Using Visual Saliency and Deep Learning”. In : p. 2783-2788. url : https://openaccess.thecvf.com/content_ICCV_2017_workshops/w40/html/Mavani_Facial_Expression_Recognition_ICCV_2017_paper.html.

[181] Abir Fathallah, Lotfi Abdi et Ali Douik. “Facial Expression Recognition via Deep Learning”. In : *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, p. 745-750. doi : 10.1109/AICCSA.2017.124. url : https://ieeexplore.ieee.org/document/8308363.

[182] Ning Sun et al. “Deep spatial-temporal feature fusion for facial expression recognition in static images”. In : *Pattern Recognition Letters* 119, p. 49-61. issn : 0167-8655. doi : 10.1016/j.patrec.2017.10.022. url : https://www.sciencedirect.com/science/article/pii/S0167865517303902.

[183] Gozde Yolcu et al. “Facial expression recognition for monitoring neurological disorders based on convolutional neural network”. In : *Multimedia tools and applications* 7822, p. 31581-31603. issn : 1380-7501. doi : 10.1007/s11042-019-07959-6. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9181900/.

[184] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In : *Journal of Machine Learning Research* 1556, p. 1929-1958. issn : 1533-7928. url : http://jmlr.org/papers/v15/srivastava14a.html.

[185] Bing-Fei Wu et Chun-Hsien Lin. “Adaptive Feature Mapping for Customizing Deep Learning Based Facial Expression Recognition Model”. In : *IEEE Access* 6, p. 12451-12461. issn : 2169-3536. doi : 10.1109/ACCESS.2018.2805861. url : https://ieeexplore.ieee.org/abstract/document/8291717.