الجمهورية الجزائرية الديمقراطية الشعبية
PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
وزارة التعليم العالي والبحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة أبي بكر بلقايد - تلمسان -
UNIVERSITY ABOUBAKR BELKAÏD - TLEMCEN -
FACULTY OF TECHNOLOGY
DEPARTMENT OF BIOMEDICAL ENGINEERING

THESIS

prepared under Decree No. 1275
to obtain the Master's Degree in Biomedical Engineering

AI-Powered Biobank Infrastructure for Parasitic Diseases:

Focus on Gastrointestinal Parasites to Enhance Sustainable Health Systems

Prepared by:

FRIDJAT Alaa Khadidja
Biomedical and Hospital Informatics

DIB Esmae
Medical Imaging

Jury Members

Name	Grade	Institution	Role
M. CHIKH Mohammed Amine	Pr.	University of Tlemcen	President
M. TALEB Tariq	MCA	University of Tlemcen	Examiner
M. KADRI Benamar	Pr.	University of Tlemcen	Supervisor
M. BIA Mohamed AL Mubarek	Dr.	Faculty of Medicine, Chungbuk National University, South Korea	Co-Supervisor
M. BENADDA Lotfi	MCA	University of Tlemcen	I2E Expert
M. MOHAMED CHERIF Abdellah	MCA	University of Tissemsilt	Technical Expert

ACADEMIC YEAR
2025 - 2026

Acknowledgment

First and foremost, we would like to express our deepest gratitude and sincere appreciation to our supervisor, **Pr. Kadri Benamar**, for his invaluable guidance, insightful advice, continuous support, and meticulous supervision throughout this project. His expertise, patience, and humility greatly contributed to the successful completion of this academic work.

We would also like to extend our sincere thanks to **Dr. Mohammed Mebarek Bia**, supervisor of the biological aspect of this project, for his valuable assistance, scientific expertise, and continuous guidance. Despite the geographical distance and the numerous requirements related to the biological component, he was always available to support and assist us throughout this work.

We also express our sincere gratitude to **Dr. Barakaeli Abdieli Ndossi** for his valuable support and contribution to the parasitological aspect of this project.

Our sincere thanks are also extended to **Pr. Seongjun Choe** for his support, guidance, and valuable contribution to this work.

We would also like to express our sincere gratitude to our international collaborators from the **International Parasite Resource Bank (iPRB), Chungbuk National University, Cheongju, South Korea**, and to our collaborating partners in parasitology research for their valuable support, expertise, and contribution to this project

Our heartfelt gratitude also goes to all the professors of the **Faculty of Technology, Department of Biomedical Engineering**, for the knowledge, guidance, and encouragement they provided throughout our academic journey. We would particularly like to acknowledge **Pr. Chikh Mohamed El Amine**, Dean of the Faculty of Technology, whose support and contributions greatly enriched our education in the field of **Biomedical and Hospital Informatics**.

We are equally honored to express our gratitude to the members of the examination committee for accepting to evaluate this work, **Pr. Chikh Mohamed El Amine and** Dr. **Tarek Taleb**, as well as all members of the entrepreneurial support program **I2E** for their valuable contributions and encouragement.

Special thanks are also extended to Pr. Bachar Hassan and Dr. Lina Hassnaoui for their support, assistance, and encouragement throughout our university journey.

Finally, we would like to thank everyone who contributed, directly or indirectly, to the successful completion of this academic work. Their support, guidance, and encouragement have been greatly appreciated and will always be remembered.

May Allah reward all those who supported us and contributed to this achievement.

BY: FRIDJAT Alaa khadidja

Acknowledgment

First and foremost, I would like to express my sincere gratitude to Almighty Allah for granting me the strength, patience, and determination to complete this work successfully.

I would like to express my deepest and most sincere gratitude to my supervisor, **Pr. Kadri Benamer**, for his invaluable guidance, constant support, constructive advice, and encouragement throughout the realization of this project. His expertise and dedication have been of great importance in the completion of this work.

We would also like to express our sincere gratitude to our international collaborators from the **International Parasite Resource Bank (iPRB), Chungbuk National University, Cheongju, South Korea**, and to our collaborating partners in parasitology research for their valuable support, expertise, and contribution to this project

I would also like to extend my heartfelt appreciation to my co-supervisor,

Dr. Mohamed Bia Al Mubarek, for his generous assistance, valuable recommendations, and continuous support. Despite the distance, he was always available to guide and help us, and his kindness and dedication are deeply appreciated.

We also express our sincere gratitude to **Dr. Barakaeli Abdieli Ndossi** for his valuable support and contribution to the parasitological aspect of this project.

Our sincere thanks are also extended to **Pr. Seongjun Choe** for his support, guidance, and valuable contribution to this work.

My sincere thanks also go to all the professors and staff members of the **Biomedical Engineering Department**, especially those of the Medical Imaging specialty, for the knowledge and experience they have shared throughout my academic journey.

I would also like to express my gratitude to the medical and technical staff of the institution where I carried out my internship for their cooperation and support.

Finally, I would like to express my heartfelt appreciation to my family and friends for their endless love, patience, encouragement, and unwavering support throughout my studies. Without them, this achievement would not have been possible.

By: Dib Esmae

الإهداء

الحمد لله رب العالمين، له الحمد والمنة على توفيقه ونعمه، فلولا فضل الله وبركته ولطفه ما سلكنا هذا الدرب وأتممناه الا بفضله و لطفه

إلى حبيب القلب، إلى من ناداني بـ "المهندسة" منذ أول يوم في الكلية، **جدي "بشير بيه"**، رحمه الله كنت أنتظر هذا اليوم لتكون معي في يوم تخرجي، لكنك اليوم في جنات الخلد بإذن الله. منك تعلمت معنى الشغف بالعلم والتعلم، فحتى قبل وفاتك بأيام، كنت تقضي وقتك بين الكتب. واليوم، وأنا أبلغ نهاية هذا المسار الهندسي و أدرك أن الهندسة لم تكن مجرد تخصص أدرسه، بل رحلة صبر وبناء، جمعت بين العلم، والاجتهاد، والإيمان بأن لكل حلم أساسًا يُشيّد خطوةً بخطوة.

إلى والديّ العزيزين، إلى أمي **"بيه وداد"**، التي غمرتني حبًا ودعاءً ودعمًا، وتفهمت تعبي وظروفي، فكان هذا الإنجاز ثمرة جهدكِ أنتِ أولًا. وإلى **والدي المهندس عبد الوهاب فريجات**، سندي الأول وداعمي في كل المراحل والقرارات. كنتَ دائمًا مصدر ثقتي وقوتي، أجد في نصائحك الحكمة وفي دعمك الدافع للاستمرار مهما كانت الصعوبات. **وإلى إخوتي عبد الرحمن وعبد النور، والتوأم جاد وإياد.**

إلى المشجع الدائم ومستشاري الأول **خالي العزيز الدكتور محمد المبارك بيه.**

وإلى كل أفراد عائلتي: جدتي **عائشة فيلالي، والدكتور حمزة بيه، والمهندس عدنان بيه**. وإلى جدي **فريجات التارزي**

وإلى عمتي الراحلة، ومعلمتي للرياضيات منذ الصغر **فريجات الزهرة**، رحمها الله

وإلى المؤسسات الخيرية التي أنتمي إليها بكل فخر: **جمعية الإرشاد والإصلاح الجزائرية**، وبالخصوص الأستاذ فردي عبد القادر وعائلته، **ومؤسسة النهضة بالقرآن الكريم**، فمن خلالهم تعلمت الصبر والانضباط والالتزام والإخلاص

إلى صديقتي العزيزة ورفيقة البحث العلمي، **ديب أسماء**

وإلى كل **صديقاتي الحبيبات**، كان وجودكنّ دعمًا وحضنًا وأمانًا، فشكرًا لكنّ من القلب

كما أهدي هذا العمل إلى **الشعب الفلسطيني**، وإلى أهلنا الصامدين في **غزة** على وجه الخصوص، الذين أثبتوا للعالم معاني الصبر والثبات والكرامة رغم كل ما يواجهونه من معاناة

راجيةً أن يكون هذا الإنجاز المتواضع رسالة تضامن، وتذكيرًا بأن العلم لا يحمل المعرفة فحسب، بل يحمل أيضًا الرحمة والأمل والالتزام بالعدالة. وسيبقى صمودهم وشجاعتهم مصدر إلهام لنا وللأجيال القادمة

أهدي هذا الإنجاز إليكم جميعًا

فريجات آلاء خديجة

الإهداء

الحمد لله الذي بنعمته تتم الصالحات، والذي وفقني وأعانني على إتمام هذا العمل، والصلاة والسلام على سيدنا محمد صلى الله عليه وسلم

أهدي ثمرة هذا الجهد المتواضع إلى أعز الناس على قلبي، والديَّ الكريمين **أمي بن صديق رحمة** وأبي **ديب مصطفى**، اللذين كانا ولا يزالان مصدر دعمي وقوتي، واللذين لم يبخلا عليَّ يومًا بحبهما وتضحياتهما وتشجيعهما المستمر. فلهما مني كل الشكر والعرفان والامتنان

إلى جدتي الغالية **الحاجة عائشة**، التي كانت دعواتها الصادقة وتشجيعاتها الدائم خير سند لي في مسيرتي الدراسية، والتي غرست في نفسي الأمل والعزيمة والإصرار على النجاح

إلى **أخواتي العزيزات وأخي الصغير**، الذين كانوا مصدر فرح وسعادة ودعم في مختلف مراحل حياتي، أشكركم على محبتكم ومساندتكم الدائمة

إلى صديقتي وأختي التي لم تلدها أمي وزميلتي في إنجاز هذا المشروع آلاء، شكرًا لكِ على كل لحظات التعاون والتشجيع والدعم التي جمعتنا خلال هذه الرحلة الأكاديمية، وعلى كل الذكريات الجميلة التي ستبقى راسخة في الذاكرة

كما أهدي هذا العمل إلى كل من وقف إلى جانبي، وآمن بقدراتي، وشجعني على مواصلة طريقي نحو النجاح

وأهدي هذا الإنجاز إلى نفسي، تقديرًا لكل جهد بذلته، ولكل صعوبة تجاوزتها، ولكل لحظة صبر واجتهاد كانت سببًا في الوصول إلى هذه المرحلة

وأسأل الله أن يجعل هذا العمل نافعًا ومباركًا، وأن يكون بداية لمزيد من النجاح والتوفيق في المستقبل

أسماء ديب

Abstract

Intestinal parasites remain a significant public health issue requiring accurate diagnosis and structured medical data management. This project aims to design and implement an **AI-powered biobank platform** dedicated to **gastrointestinal parasites**. The system combines digital sample management with artificial intelligence techniques for microscopic image analysis and parasite egg classification. Deep learning, self-supervised learning, and semi-supervised learning approaches were explored to address the limitation of labeled data and support AI-assisted diagnosis. The platform enables sample registration, image upload, expert validation, result storage, and digital report generation. It also falls within the framework of **Ministerial Decree No. 1275** as an innovative graduation project with startup potential in digital health and medical AI. This work represents a step toward a future physical biobank supporting diagnosis, research, and sustainable healthcare systems in Algeria.

Keywords: AI-powered biobank, gastrointestinal parasites, deep learning, medical AI, digital health, Ministerial Decree No. 1275.

الملخص

تُعدّ الطفيليات المعوية من المشكلات الصحية المهمة التي تتطلب تشخيصًا دقيقًا وتنظيمًا فعالًا للبيانات الطبية. يهدف هذا المشروع إلى تصميم وتنفيذ **منصة بنك حيوي مدعومة بالذكاء الاصطناعي** موجهة لدراسة **وتشخيص الطفيليات المعوية**. يجمع النظام المقترح بين الإدارة الرقمية للعينات وتقنيات الذكاء الاصطناعي من أجل تحليل الصور المجهرية وتصنيف بيوض الطفيليات. كما تم توظيف التعلم العميق، والتعلم الذاتي الإشراف، والتعلم شبه الخاضع للإشراف لمعالجة محدودية البيانات الموسومة ودعم التشخيص المساعد بالذكاء الاصطناعي ;تتيح المنصة تسجيل العينات، رفع الصور، مراجعة النتائج من طرف الخبير، تخزين النتائج، وتوليد التقارير الرقمية. كما يندرج المشروع ضمن إطار **القرار الوزاري رقم 1275** باعتباره مشروع تخرج ابتكاريًا قابلًا للتطوير كمؤسسة ناشئة في مجال الصحة الرقمية والذكاء الاصطناعي الطبي. ويمثل هذا العمل خطوة نحو إنشاء بنك حيوي فعلي مستقبلي يدعم التشخيص، البحث العلمي، والأنظمة الصحية المستدامة في الجزائر

الكلمات المفتاحية: البنك الحيوي المدعوم بالذكاء الاصطناعي، الطفيليات المعوية، التعلم العميق، الذكاء الاصطناعي الطبي، الصحة الرقمية، القرار الوزاري رقم 1275.

Résumé

Les parasites intestinaux représentent un problème important de santé publique nécessitant un diagnostic précis et une gestion structurée des données médicales. Ce projet vise à concevoir et mettre en œuvre une **plateforme de biobanque intelligente**, alimentée par l'intelligence artificielle, dédiée à l'étude et au diagnostic des **parasites gastro-intestinaux**. Le système proposé combine la gestion numérique des échantillons avec des techniques d'intelligence artificielle pour l'analyse des images microscopiques et la classification des œufs parasitaires. Des approches d'apprentissage profond, d'apprentissage auto-supervisé et d'apprentissage semi-supervisé ont été explorées afin de répondre à la limitation des données annotées et de soutenir le diagnostic assisté par IA. La plateforme permet l'enregistrement des échantillons, le téléversement des images, la validation des résultats par un expert, le stockage des diagnostics et la génération de rapports numériques. Le projet s'inscrit également dans le cadre de l'arrêté ministériel n°1275, en tant que projet de fin d'études innovant et susceptible d'être développé en startup dans le domaine de la santé numérique et de l'IA médicale. Ce travail constitue une étape vers la création future d'une biobanque physique capable de soutenir le diagnostic, la recherche scientifique et les systèmes de santé durables en Algérie.

Mots-clés : biobanque intelligente, parasites gastro-intestinaux, apprentissage profond, IA médicale, santé numérique, arrêté ministériel n°1275.

TABLE OF CONTENTS

الإهداء ..... **IV**

Abstract ..... **VI**

الملخص ..... **VII**

Résumé ..... **VIII**

General Introduction **1**

1. Context and Motivation **2**

2. Problem Statement **2**

3. Research Objectives **3**

4. Dissertation Structure **3**

Chapter 1: **5**

Biological, Epidemiological, and Biobank-Based Perspectives on Gastrointestinal Parasite Research **5**

1.1 Introduction **6**

1.2 Gastrointestinal Parasites and Their Public Health Impact **6**

1.2.1 Definition of Parasites and the Concept of Parasitism 6

1.2.1.1 Definition of Parasites 6

1.2.1.2 Classification of Parasites 7

1.2.1.3 Life Cycle of Parasites 7

1.2.2 Gastrointestinal Parasites: Main Focus of the Study 8

1.2.2.1 Protozoa and Helminths: The Major Groups of Gastrointestinal Parasites 9

1.2.2.1.1 Protozoa 9

1.2.2.1.2 Helminths 10

1.2.3 Transmission and Risk Factors 11

1.2.3.1 Mode of Transmission 11

1.2.3.1.1 Protozoan Parasites 11

1.2.3.1.2 Helminths 12

1.2.3.2 Factors Influencing the Transmission of Intestinal Parasites 12

1.2.4 Clinical Manifestations and Complications 12

1.2.4.1 Common Symptoms 12

1.2.4.2 Severe Complications and Vulnerable Populations 13

1.3 Current Challenges in the Management and Study of Gastrointestinal Parasites within Sustainable Health Systems **14**

1.3.1 Classic Diagnostic Methods 14

1.3.2 Limitations and Challenges of Traditional Diagnosis 15
1.3.3 Impact on Sustainable Health Systems 15
1.4 Biobanks and Their Role in Parasitic Disease Research 16
1.4.1 Definition and Importance of Biobanks 16
1.4.2 Historical Evolution of Biobanks 17
1.4.3 Theoretical and Functional Roles of Biobanks 18
1.4.4 Biobanks in the Context of Parasitic Diseases 18
1.4.5 Toward an AI-Powered Gastrointestinal Parasite Biobank 18
1.4.5.1 Limitations of Conventional Data Management 19
1.4.5.2 Need for Digital and AI-Based Infrastructure 19
1.4.5.3 Potential Contributions to Diagnosis, Surveillance, and Research 20
1.5 Conclusion 21
Chapter 2: 23
Modern AI Approaches For Medical Image Analysis 23
2.1 Introduction 24
2.2 Artificial Intelligence in Medical Image Analysis 24
2.3 Theoretical and Technical Foundations of CNN and ResNet Architectures 24
2.3.1 Convolutional Neural Networks 24
2.3.2 Residual Networks 25
2.4 Learning paradigms for AI-based medical image analysis 25
2.4.1 Supervised Learning 25
2.4.2 Unsupervised Learning 26
2.4.3 Semi-Supervised Learning 27
2.4.4 Self-Supervised Learning 28
2.5 Supervised Learning and Image Classification: 29
2.6 The Data Annotation Problem in Medical Imaging 30
2.7 Representation Learning and Automatic Feature Extraction 31
2.8 Latent Space and Visual Embeddings 32
2.9 Contrastive Learning 33
2.9.1 SimCLR 33
2.9.2 MoCo 33
2.9.3 BYOL 34
2.9.4 SimSiam 34
2.9.5 Triplet Learning Triplet 34
2.10 Clustering in Latent Space for Image Organization 35
2.11 Pseudo-Labeling and Semi-Supervised Learning 36

2.12 State of the Art on Representation Learning and Limited-Annotation Medical Image Analysis for AI-Based Gastrointestinal Parasite Biobanking 37

2.13 AI Model Evaluation: Validation, Testing, Overfitting, and Generalization 37

2.14 Conclusion 38

CHAPTER 3: 40

Methodology of the Proposed AI-Powered Biobank Framework for Gastrointestinal Parasite Image Analysis under Limited Annotation Conditions 40

3.1 Introduction 41

3.2 General Methodological Overview 41

3.3.1 Motivation and Problem Definition 42

3.3.2 Unlabeled Five-Group Dataset Description 43

3.3.3 Image Preprocessing and Augmentation 43

3.3.4 Supervised Representation Learning 43

3.3.5 Embedding Extraction and Latent Space Projection 44

3.3.6 Clustering Strategy 45

3.3.6.1 SimCLR with K-Means 45

3.3.6.2 SimCLR with Agglomerative Clustering 47

3.3.6.3 BYOL with K-Means 47

3.3.6.4 BYOL with Agglomerative Clustering 49

3.3.6.5 Overall Comparison between SimCLR and BYOL 50

3.3.7 Latent Space Visualization 51

3.3.8 Pseudo-Label Generation from Clusters 51

3.4 Validation Across Five Groups and Eleven Parasite Classes 52

3.4.1 Eleven-Class Dataset Description 53

3.4.2 Experimental Design with Different Label Ratios 53

3.4.2.1 Zero-Label Scenario 54

3.4.2.2 Limited-Label Scenarios: 5%, 10%, and 50% 55

- Description of the ResNet50 Classifier Architecture 55

3.4.2.3 Fully Supervised Baseline: 100% Labels 57

3.4.2.4 Comparative Analysis of Label Ratios 58

3.5 Evaluation Methodology 58

3.5.1 Clustering Evaluation Metrics 59

3.5.2 Classification Evaluation Metrics 59

3.5.3 Internal Validation 59

3.5.4 External Testing 59

3.6 Return to the Initial Five-Group Dataset for Expert-Based Cluster Validation 59

3.6.1 Experimental Context 60
3.6.2 Cluster-Level Findings 60
3.6.3 Conclusion of Cluster-Level Findings 61
3.6.4 General Interpretation 61
3.6.5 Conclusion of Expert-Based Cluster Validation 62
3.7 Conceptual Integration with the AI-Powered Biobank Framework 62
3.8 Conclusion 62
Chapter 4: 63
Design and Implementation of the AI-Powered Biobank Platform 63
4.1 Introduction 64
4.2 Development Environment and Technology Choices 64
4.2.1 PHP and Laravel Framework 64
4.2.2 MySQL Database Management System 65
4.2.3 Python and FastAPI for the AI Diagnosis Service 65
4.2.4 Flutter for Mobile Application Development 66
4.3 System Architecture 66
4.5 Web Platform Implementation 69
4.5.1 Public Home Page 69
4.5.3 Dashboard Interface 70
4.5.4 Sample Management Interface 72
4.5.5 AI Diagnosis Results Interface 74
4.5.6 Reports Interface 74
4.5.7 Epidemiological Spread Map Interface 75
4.5.8 Community Interface 76
4.5.9 Learning Center Interface 77
4.6 AI Diagnosis Service Integration 78
4.6.1 Autoencoder-Based Parasite / No-Parasite Screening 78
4.6.2 ResNet50 11-Class Parasite Classification 79
4.6.3 Final AI Workflow in the Platform 80
4.7 Mobile Application Implementation 80
4.8 Conclusion 83
Chapter 5: 84
Entrepreneurial and Business Study of the AI-Powered Biobank Project 84
5.1 Introduction 85
5.2 General Presentation of the AI Biobank Project 85
5.3 Project Objectives 85

5.3.1 Short-term objectives 85
5.3.2 Medium-term objectives 85
5.3.3 Long-term objectives 86
5.4 Problem Identification and Market Need 86
5.5 Entrepreneurial Opportunity Derived from the Identified Need 86
5.6 Proposed Solution: Product and Services 87
5.7 Value Proposition 87
5.8 Target Market and Beneficiaries 88
5.9 Competitive Advantage and Innovative Aspects 89
5.10 Strategic Market Analysis 89
5.10.1 Market Segmentation 89
5.10.2 Market Competition Analysis 90
5.10.2.1 Direct Competitors 90
5.10.2.2 Indirect Competitors 90
5.10.2.3 Identified Strengths and Weaknesses of Competitors Strengths 90
5.11 Production and Organization Plan 91
5.11.1 Operational Workflow 91
5.11.2 Organization of Roles 92
5.11.3 Future Real Biobank Organization 92
5.12 Legal, Ethical, and Regulatory Framework 92
5.13 Financial Plan 93
5.13.1 Costs and Expenses 93
5.13.1.1 Initial Investment Costs 93
5.13.1.2 Human Resources and Salary Plan 94
5.13.2 Funding Sources 95
5.13.3 Revenue 96
5.13.4 Projected Income Statement 97
5.13.5 Strategies for Reducing the Working Capital Requirement (WCR / BFR) 97
5.13.6 Cash Flow Plan – Year 1 98
5.13.7 Financial Sustainability 99
5.14 Prototype of the Proposed Solution 99
5.14.1 Prototype Overview 99
5.14.2. Main Functional Modules of the Prototype 99
5.14.3 Integrated Prototype Components and Future Development 100
5.14.4. Materials and Cost Estimation 101
5.14.4.1. Materials and Equipment for the Digital Platform Phase 101

5.14.4.2. Materials and Resources for the Future Real Biobank 101
5.14.4.3. Biological and Digital Resources Needed to Fill the Real Biobank 102
5.14.4.4. Summary of Estimated Material Costs 103
5.15. Business Model Canvas 103
5.16 Conclusion 104

LIST OF FIGURES

Figure 1.1: General overview of parasite life cycles and their classification into direct and indirect life cycles **8**
Figure 1.2: Life Cycle of Parasitic Protozoa (Giardia DUODENALIS). (CDC) **10**
Figure 1.3: Microscopic image of helminth eggs — shown here using Schistosoma species for morphological illustration purposes. (Photomicrograph depicting eggs of Schistosoma japonicum, S. mansoni, and S. haematobium, CDC/Dr. Brodsky, public domain)..... **11**
Figure 1.4: Major Complications and Vulnerable Populations in Intestinal Parasitic Infections **14**
Figure 1.5: Impact of Intestinal Parasitic Infection Challenges on Sustainable Health Systems **16**
Figure 1.6: General structure of a biobank..... **17**
Figure 1.7: Conceptual framework of an AI-powered gastrointestinal parasite biobank..... **21**
Figure 2.1: Overview of Supervised Learning in Medical Image Classification..... **26**
Figure 2.2: General Workflow of Unsupervised Learning for Medical Images..... **27**
Figure 2.3: Semi-Supervised Learning for Microscopic Parasite Image Classification..... **28**
Figure 2.4: Overview of Self-Supervised Learning in Medical Image Analysis..... **29**
Figure 2.5: Schematic illustration of common challenges in microscopic medical images..... **30**
Figure 2.6: Data annotation bottleneck in medical imaging..... **31**
Figure 2.7: Representation learning process from image to embedding vector..... **32**
Figure 2.8: Visualization of image embeddings in latent space..... **32**
Figure 2.9: SimCLR framework for self-supervised visual representation learning..... **33**
Figure 2.10: BYOL self-supervised learning architecture..... **34**
Figure 2.11: Comparison of major self-supervised learning methods **35**
Figure 2.12: Clustering of learned embeddings in latent space **35**
Figure 2.13: Pseudo-label generation from clustered image embeddings **36**
Figure 2.14: Dataset split and overfitting behavior in deep learning evaluation **38**
Figure 3.1: General Workflow of the Proposed Methodology..... **42**
Figure 3.2: Embedding extraction and latent-space projection process..... **44**
Figure 3.3: SimCLR-ResNet50 clustering results summary **45**
Figure 3.4: SimCLR-ResNet50 architecture used for self-supervised representation learning..... **46**
Figure 3.5: t-SNE visualization of SimCLR clusters using K-Means and Agglomerative Clustering..... **47**
Figure 3.6: BYOL-ResNet50 architecture used for self-supervised representation learning..... **48**
Figure 3.7: BYOL-ResNet50 clustering results and t-SNE visualization..... **50**
Figure 3.8: Overall comparison between SimCLR and BYOL clustering results..... **51**
Figure 3.9: K-Means-Based Pseudo-Label Generation and Expert Review Summary **52**
Figure 3.11: Zero-label SimCLR-ResNet50 clustering results using PCA100 and K-Means..... **55**
Figure 3.12: Performance results under limited-label scenarios **57**
Figure 3.13: Fully supervised baseline performance using 100% labeled data **57**
Figure 3.14: Performance trend and internal–external gap across label ratios **58**
Figure 4.1: General Architecture of the Proposed AI-Powered Biobank Platform..... **67**
Figure 4.2: Final Conceptual Data Model of the AI-Powered Gastrointestinal Parasite Biobank Platform **68**
Figure 4.3: Public Home Page and Registration Access of the Web Platform..... **70**
Figure 4.6: Regular User Dashboard Interfaces **72**
Figure 4.7: Sample Management Interface and Sample Registration Form..... **73**
Figure 4.8: Sample Details and Microscope Image Interface..... **73**
Figure 4.9: Diagnosis Results Interface..... **74**
Figure 4.10: Reports Center and Generated Diagnosis PDF Report Interface..... **75**
Figure 4.11: Epidemiological Parasite Spread Map Interface..... **76**
Figure 4.12: Research Community Interface: Overview, Post Creation, Discussion Details, and Comments **77**

Figure 4.14 Architecture of the convolutional autoencoder used for Parasite / No-Parasite screening..... **79**
Figure 4.15: Mobile Application Core Interfaces **81**
Figure 4.16: Mobile Results and Reports Interfaces..... **82**
Figure 4.17: Mobile Community and Assistant Interfaces **83**
Figure 5.1: AI-Powered Biobank Workflow for Gastrointestinal Parasites..... **87**

LIST OF TABLES

Table 3.1: Distribution of images in the unlabeled five-group dataset **43**
Table 3.2: Class distribution of the labeled eleven-class parasite **53**
Table 3.3: Cross-cluster comparison. Yellow cells highlight the most notable findings: hookworm near-exclusion in Clusters 2–3, and highest Ascaris purity in Cluster 3. ..... **61**
Table 5.1: Estimated Initial Investment for the AI-Powered Biobank Project **94**
Table 5.2: Estimated Staffing Plan and Annual Salary Costs for the AI-Powered Biobank Project **95**
Table 5.3: Proposed Financing Plan for the AI-Powered Biobank Project **96**
Table 5.4: Estimated Annual Revenue Forecast for the AI-Powered Biobank Project **97**
Table 5.5: Estimated First-Year Financial Result of the AI-Powered Biobank Project **97**
Table 5.6: Estimated Monthly Cash Flow Forecast for the AI-Powered Biobank Project **98**
Table 5.7: Prototype Modules, Main Functions, and Entrepreneurial ValueTable **100**
Table 5.8: Digital Platform Materials, Purpose, and Estimated Costs **101**
Table 5.9: Future Real Biobank Materials, Purpose, and Estimated Costs **102**
Table 5.10: Key Resources, Description, and Purpose **for the Future Real Biobank** **103**
Table 5.11: Estimated Material Costs for the Digital Platform and Future Real Biobank **103**
Table 5.12: Business Model Canvas of the AI-Powered Biobank Project **104**

General Introduction

1. Context and Motivation

Gastrointestinal parasitic diseases remain an important public health concern, especially in regions where access to advanced diagnostic tools, organized medical data, and specialized expertise may be limited. These diseases are commonly diagnosed through microscopic examination of biological samples, a method that remains essential but depends strongly on the experience of specialists, the quality of microscopic images, and the availability of well-organized diagnostic information.

In many healthcare and laboratory environments, data related to parasitic diseases are often fragmented. Microscopic images, sample information, diagnostic results, and patient-related data may be stored separately or managed manually. This situation limits sample traceability, reduces the possibility of reusing data for research, and makes epidemiological monitoring more difficult. At the same time, the absence of a structured biobank specialized in gastrointestinal parasites reduces the availability of organized biological and digital resources for scientific studies and artificial intelligence development.

Recent progress in artificial intelligence, medical image analysis, and biomedical informatics has created new possibilities for improving diagnostic workflows. Deep learning models can assist in the analysis of microscopic images and support specialists by providing preliminary diagnostic suggestions. However, the development of such systems requires structured data, expert validation, and appropriate digital infrastructure. This motivates the development of an AI-powered biobank-oriented platform capable of linking biological sample information, microscopic images, diagnostic results, and research data in one organized system.

Therefore, this project aims to contribute to the improvement of gastrointestinal parasite diagnosis and data management by proposing a digital platform that combines sample traceability, AI-assisted diagnosis, expert validation, digital reporting, and research support. The platform represents a first step toward the future creation of a real biobank specialized in gastrointestinal parasitic diseases.

2. Problem Statement

Despite the importance of gastrointestinal parasitic diseases, their diagnosis and data management still face several limitations. First, microscopic diagnosis is time-consuming and depends on expert interpretation, which may lead to delays or variability in results. Second, microscopic images and sample-related data are often not stored in a structured and reusable way. Third, there is a lack of specialized digital systems that connect samples, images, diagnostic results, and research information within a unified platform.

Another major challenge is the limited availability of expert-labeled microscopic datasets. Artificial intelligence models generally require large amounts of labeled data to achieve reliable performance, but in parasitology, such datasets are difficult to obtain because labeling requires medical or biological expertise. This limitation makes it necessary to explore learning strategies that can use limited labeled data or benefit from unlabeled data.

In addition, the absence of a dedicated biobank infrastructure for gastrointestinal parasites limits the preservation, traceability, and scientific exploitation of biological resources. Without an organized system, it becomes difficult to support long-term research, epidemiological studies, AI model improvement, and collaboration between laboratories, healthcare institutions, and researchers.

Based on these challenges, the main problem addressed in this thesis can be formulated as follows:

How can artificial intelligence, biomedical informatics, and biobank-oriented data organization be combined to improve the management, analysis, and traceability of gastrointestinal parasite samples and microscopic images?

3. Research Objectives

The main objective of this thesis is to design and develop an AI-powered biobank-oriented platform dedicated to gastrointestinal parasites. The proposed solution aims to support sample data management, microscopic image storage, AI-assisted diagnosis, expert validation, digital reporting, and research use.

The specific objectives of this work are:

1. To study the biological, diagnostic, and epidemiological context of gastrointestinal parasitic diseases and highlight the importance of structured data organization.
2. To explore artificial intelligence methods for microscopic image analysis, with particular attention to the challenge of limited labeled data.
3. To investigate self-supervised learning and limited-label classification strategies for parasite image analysis.
4. To develop an AI-assisted diagnosis pipeline capable of providing preliminary parasite classification results from microscopic images.
5. To design and implement a digital platform for managing organizations, users, samples, microscopic images, diagnostic results, reports, and collaboration features.
6. To integrate the artificial intelligence component into the platform through an independent API-based service.
7. To propose a biobank-oriented framework that can serve as a foundation for the future establishment of a real biobank specialized in gastrointestinal parasites.
8. To analyze the entrepreneurial potential of the project as a startup initiative, including its value proposition, target users, business model, financial plan, and future development perspectives.

4. Dissertation Structure

This dissertation is organized into five main chapters.

Chapter 1 presents the biological, epidemiological, and biobank-based background of gastrointestinal parasite research. It introduces the medical importance of gastrointestinal parasites, their diagnostic context, and the role of biobanks in organizing biological and digital resources for healthcare and research.

Chapter 2 focuses on artificial intelligence and medical image analysis. It introduces the main concepts of machine learning, deep learning, convolutional neural networks, and their applications in biomedical image processing. It also discusses the relevance of AI-assisted diagnosis in parasitology.

Chapter 3 presents the artificial intelligence experiments conducted in this work. It describes the datasets, preprocessing steps, learning strategies, self-supervised approaches, limited-label scenarios, classification models, and evaluation metrics used for gastrointestinal parasite image analysis.

Chapter 4 describes the design and implementation of the proposed platform. It presents the web application, mobile application, database structure, user roles, sample management system, diagnosis results, reports, and the integration of the AI diagnosis service through an API.

Chapter 5 discusses the entrepreneurial dimension of the project. It presents the startup approach, problem and market need, value proposition, target market, competitive advantage, financial plan, prototype value, business model canvas, and the future evolution of the digital platform toward a real AI-powered biobank infrastructure.

Finally, a general conclusion summarizes the main contributions of the thesis, discusses the limitations of the proposed work, and presents future perspectives for improving the platform, expanding the dataset, strengthening expert validation, and developing a real biobank for gastrointestinal parasitic diseases.

Chapter 1:

Biological, Epidemiological, and Biobank-Based Perspectives on Gastrointestinal Parasite Research

1.1 Introduction

Gastrointestinal parasitic infections remain among the most prevalent infectious diseases worldwide, affecting more than one billion people, particularly in low- and middle-income countries. The intersection of poverty, inadequate sanitation, and limited healthcare access creates persistent cycles of infection and morbidity that conventional approaches have failed to fully resolve [1].

Intestinal parasites, primarily protozoa and helminths, impose a significant burden on public health through their complex life cycles, diverse transmission routes, and wide-ranging clinical consequences. In resource-limited settings such as Algeria and North Africa, these challenges are compounded by diagnostic limitations, particularly the low sensitivity and expert dependency of traditional microscopy-based methods.

Addressing these gaps requires not only improved diagnostic tools, but also organized, standardized, and digitally supported infrastructures capable of preserving biological samples, managing related data, and supporting long-term surveillance and research. Biobanks, particularly when integrated with artificial intelligence, represent a promising step in this direction.

This work situates the proposed AI-powered gastrointestinal parasite biobank within the broader One Health framework, recognizing that human, animal, and environmental health are deeply interconnected in parasitic disease dynamics.

1.2 Gastrointestinal Parasites and Their Public Health Impact

To understand the public health importance of gastrointestinal parasites, it is necessary to first introduce the basic concepts related to parasites and parasitism, as these concepts form the foundation for understanding their transmission, pathogenicity, and clinical impact.

1.2.1 Definition of Parasites and the Concept of Parasitism

This section provides a general definition of parasites and explains the concept of parasitism as a biological interaction between the parasite and its host, which is essential for understanding the mechanisms of parasitic infections.

1.2.1.1 Definition of Parasites

Parasites are organisms that live inside or on a host organism, from which they obtain nutrients at the host's expense. This relationship, known as parasitism, is characterized by its asymmetric nature: the parasite benefits while the host suffers varying degrees of harm, ranging from subclinical disturbance to severe and life-threatening disease [2].

Among the diverse categories of parasitic organisms, those colonizing the gastrointestinal tract are of particular public health significance. Intestinal parasites interfere directly with digestive function, nutrient absorption, and mucosal immunity, producing a spectrum of clinical consequences that are often underdiagnosed due to the limitations of conventional laboratory methods.

Understanding the nature of parasitism — including how parasites exploit host biology, evade immune defenses, and persist within host tissues — is therefore essential not only for clinical management, but also for the development of accurate diagnostic tools and organized research infrastructures such as the AI-powered biobank proposed in this study.

1.2.1.2 Classification of Parasites

Parasite classification provides the taxonomic and biological framework necessary for understanding transmission dynamics, diagnostic approaches, and disease management strategies. For the purposes of this study, classification is considered primarily in relation to parasites affecting the gastrointestinal tract.

Gastrointestinal parasites are broadly divided into two major groups based on their cellular organization. Protozoa are unicellular organisms capable of completing their entire life cycle within a single host, while helminths are multicellular worms that may require one or more hosts to reach maturity. Both groups are classified as endoparasites, meaning they reside and reproduce within internal host tissues, particularly the intestinal tract [3].

A further clinically relevant distinction concerns the location of the parasite within host cells. Obligate intracellular parasites such as *Cryptosporidium* spp. reside inside intestinal epithelial cells, making their detection and treatment fundamentally different from extracellular parasites such as *Giardia lamblia* or intestinal helminths, which inhabit the intestinal lumen [4].

This classification directly informs the diagnostic strategies, image characteristics, and data organization principles that underpin the proposed AI-powered gastrointestinal parasite biobank.

1.2.1.3 Life Cycle of Parasites

The life cycle of a parasite describes the sequence of developmental stages it passes through during growth, reproduction, and transmission. These stages — which may include eggs, cysts, oocysts, larvae, trophozoites, or adult forms — are not only biologically significant but also diagnostically critical, as many parasites are identified precisely through the microscopic recognition of these forms in clinical samples.

Gastrointestinal parasite life cycles are generally classified into two types. In a **direct life cycle**, the parasite completes its full development within a single host, with transmission occurring through contaminated food, water, soil, or direct contact. In an **indirect life cycle**, the parasite requires one or more intermediate hosts before reaching its definitive host, adding complexity to both transmission dynamics and control strategies.

Environmental factors such as temperature, humidity, and sanitation conditions directly influence the survival and infectivity of parasite stages in the external environment. Understanding these life cycle characteristics is therefore fundamental for identifying transmission routes, designing prevention measures, and — critically for this study — recognizing the morphological features of each developmental stage that serve as the basis for microscopic diagnosis and AI-based image classification.

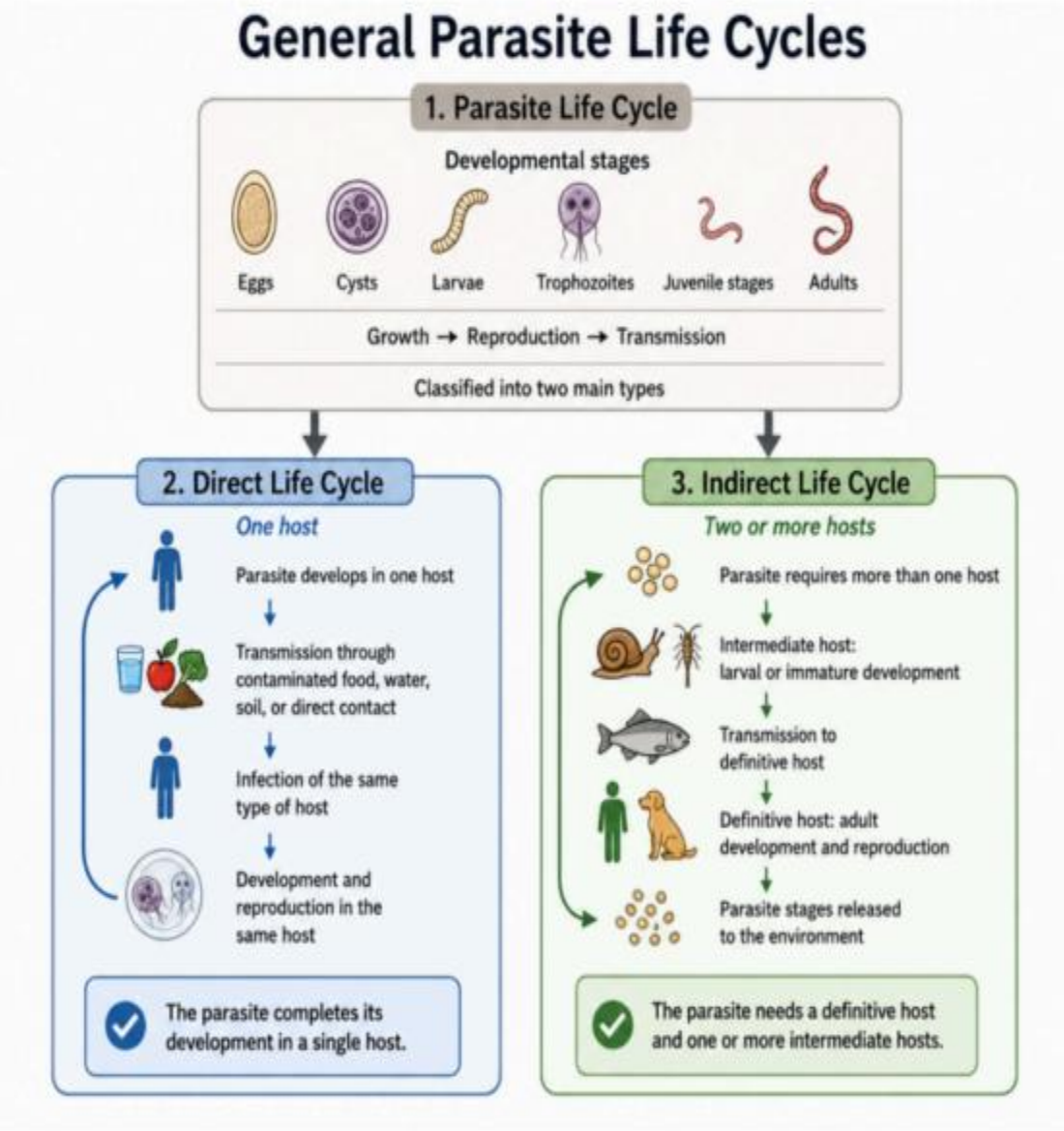


Figure 1.1: General overview of parasite life cycles and their classification into direct and indirect life cycles

1.2.2 Gastrointestinal Parasites: Main Focus of the Study

Gastrointestinal parasites are organisms that colonize the digestive tract, particularly the small and large intestines, where they feed, reproduce, and interfere with normal intestinal function. Their presence may lead to digestive disorders, nutrient malabsorption, chronic inflammation, and impaired host immunity. Transmission occurs primarily through the fecal-oral route, via contaminated food, water, soil, or inadequate hygiene practices, making these infections closely linked to socioeconomic and environmental conditions. Gastrointestinal parasitic infections remain a major public health problem worldwide, with a disproportionate burden in regions with limited sanitation infrastructure and diagnostic resources, including Algeria and North Africa.

1.2.2.1 Protozoa and Helminths: The Major Groups of Gastrointestinal Parasites

Gastrointestinal parasites are mainly divided into protozoa and helminths. Protozoa are single-celled parasites, whereas helminths are multicellular worms. This classification is important because each group differs in morphology, life cycle, transmission, and diagnostic approach.

1.2.2.1.1 Protozoa

Protozoa are unicellular parasites capable of infecting the gastrointestinal tract and causing a wide range of intestinal diseases. Common gastrointestinal protozoa include *Giardia lamblia*, *Entamoeba histolytica*, and *Cryptosporidium* spp. These organisms exist in different morphological forms depending on their life cycle stage: the trophozoite is the active feeding form, while the cyst or oocyst is the environmentally resistant stage responsible for transmission. This ability to shift between forms allows protozoa to survive outside the host and persist in contaminated water, food, and soil.

From a public health perspective, intestinal protozoa are significant causes of diarrhea, abdominal pain, malabsorption, and chronic gastrointestinal disorders, particularly in children and immunocompromised individuals [5]. Their impact is especially pronounced in regions with inadequate water treatment and sanitation infrastructure, where waterborne transmission is difficult to control. *Cryptosporidium* oocysts, for example, are resistant to standard chlorination and can survive in treated water supplies, making waterborne outbreaks a persistent public health challenge [6].

From a diagnostic standpoint, protozoa present particular challenges because their microscopic forms are small and can be difficult to distinguish from artifacts or non-pathogenic organisms. Accurate identification therefore requires high image quality, standardized laboratory procedures, and experienced interpretation — limitations that the proposed AI-powered biobank directly aims to address.

Blastocystis hominis represents an emerging concern among intestinal protozoa. Although its pathogenicity remains debated, it is among the most frequently detected parasites in human stool samples worldwide, with prevalence rates exceeding 30% in some developing countries [7], underscoring the need for organized surveillance and standardized diagnostic data.

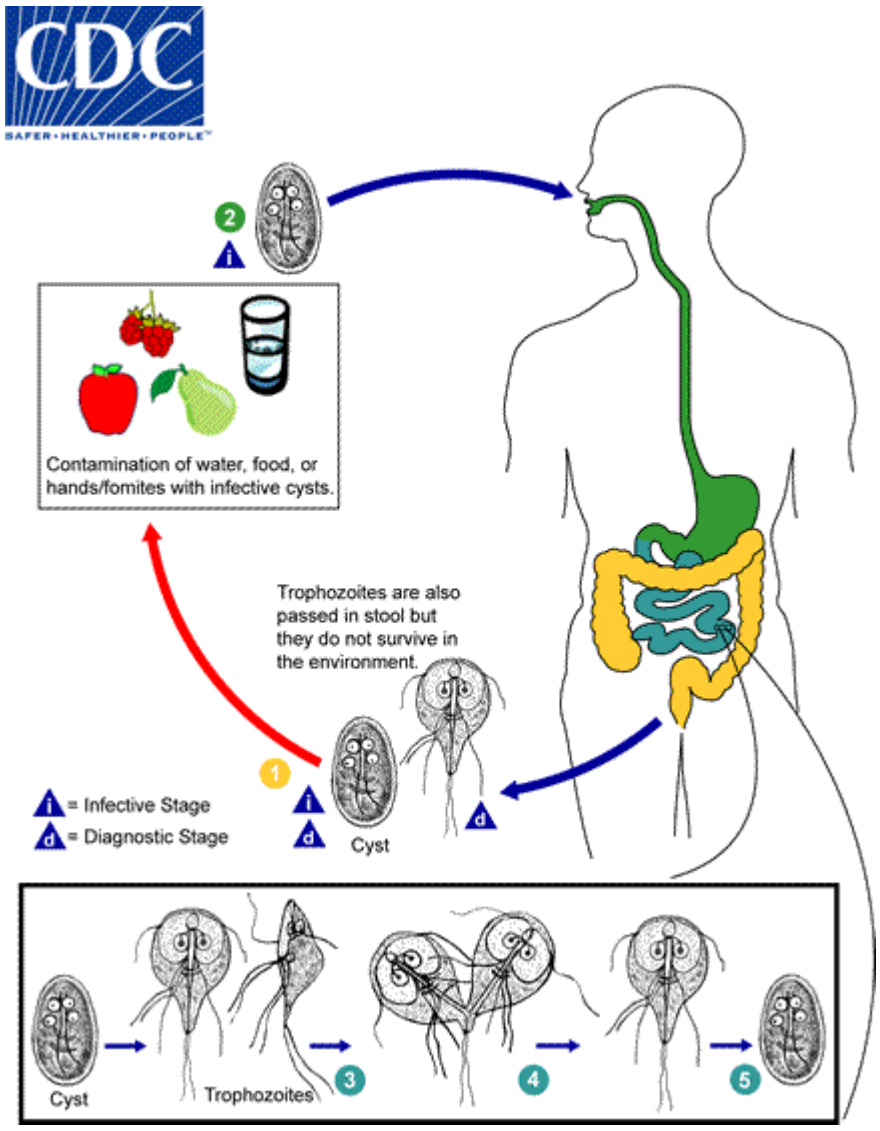


Figure 1.2: Life Cycle of Parasitic Protozoa (Giardia DUODENALIS). (CDC)

This diagram illustrates the main stages of the life cycle of intestinal protozoa, including the active trophozoite stage and the environmentally resistant cyst stage, using Giardia spp as key examples.

1.2.2.1.2 Helminths

Helminths are multicellular parasitic worms that colonize the gastrointestinal tract and cause chronic intestinal diseases in humans. The main groups of intestinal helminths include nematodes such as *Ascaris lumbricoides*, *Trichuris trichiura*, and hookworms (*Necator americanus* and *Ancylostoma duodenale*), as well as certain trematodes that affect the digestive tract.

Their developmental cycle passes through distinct stages — eggs, larvae, and adult worms — each with specific morphological characteristics that are diagnostically relevant. Many intestinal helminths are transmitted through contaminated soil, food, or water, particularly in environments with poor sanitation. Some species additionally require intermediate hosts to complete specific stages of their development.

Helminth infections constitute a major public health burden, particularly in endemic and resource-limited regions. They are associated with abdominal pain, diarrhea, anemia, malnutrition, and growth impairment in children. Soil-transmitted helminths (STHs) — primarily *Ascaris lumbricoides*, *Trichuris trichiura*, and hookworms — are

among the most prevalent neglected tropical diseases globally, causing an estimated 5.2 million disability-adjusted life years (DALYs) annually, with the greatest burden in sub-Saharan Africa, South Asia, and Southeast Asia [8].

From a diagnostic perspective, helminth identification relies on the microscopic detection of eggs, larvae, or adult forms in stool samples. The morphology, size, and shell structure of helminth eggs vary between species and represent key identification criteria in laboratory diagnosis. These morphological features also constitute the visual data that AI-based classification systems must learn to recognize accurately [9].

Beyond their direct gastrointestinal effects, helminths are known to modulate host immune responses, which may influence infection severity and susceptibility to reinfection — factors relevant to clinical management and long-term epidemiological monitoring [10].



Figure 1.3: Microscopic image of helminth eggs — shown here using Schistosoma species for morphological illustration purposes. (Photomicrograph depicting eggs of Schistosoma japonicum, S. mansoni, and S. haematobium, CDC/Dr. Brodsky, public domain)

1.2.3 Transmission and Risk Factors

1.2.3.1 Mode of Transmission

Intestinal parasites are transmitted through different routes depending on their biological group, life cycle, and environmental conditions. In general, transmission occurs when infective stages such as cysts, oocysts, eggs, or larvae enter the human body through contaminated food, water, soil, or hands.

1.2.3.1.1 Protozoan Parasites

Intestinal protozoa such as *Giardia lamblia*, *Entamoeba histolytica*, *Blastocystis hominis*, and *Cryptosporidium* spp. are mainly transmitted through the fecal-oral route, via contaminated drinking water, unsafe food, or poor hand hygiene. Zoonotic transmission also plays a role, particularly for *Cryptosporidium*, where animal reservoirs such as bovine calves contribute significantly to environmental contamination. The resistance of *Cryptosporidium* oocysts to standard water chlorination makes waterborne outbreaks especially difficult to control [6].

1.2.3.1.2 Helminths

Intestinal helminths such as *Ascaris lumbricoides*, *Trichuris trichiura*, and hookworms are primarily transmitted through contact with contaminated soil, food, or water. Eggs and larvae typically require a period of development in the environment before becoming infective. Hookworm larvae are unique in their ability to penetrate intact skin through barefoot contact with contaminated soil, representing a transmission route distinct from the fecal-oral pathway and requiring targeted prevention strategies [11].

1.2.3.2 Factors Influencing the Transmission of Intestinal Parasites

The transmission of gastrointestinal parasitic infections is shaped by a combination of environmental, behavioral, and socioeconomic factors. Poor sanitation, limited access to safe drinking water, and contamination of food and soil represent the primary environmental drivers of infection risk. Inadequate hand hygiene further facilitates fecal-oral transmission, particularly in communities with limited hygiene infrastructure.

Socioeconomic conditions play a central role in determining vulnerability. Communities affected by poverty, food insecurity, and inadequate housing face a disproportionately high burden of intestinal parasitic infections. Population displacement and internal migration can further accelerate transmission by introducing infections into previously unaffected areas and overwhelming local healthcare capacity [12].

In Algeria, rapid urbanization in peri-urban areas has created new pockets of vulnerability, where growing populations lack adequate sanitation infrastructure despite their proximity to urban centers — a context that directly motivates the need for improved diagnostic and surveillance systems such as the one proposed in this study.

The main risk factors can be summarized as follows:

- Poor sanitation and inadequate waste management
- Limited access to safe drinking water
- Contaminated food, soil, and water
- Poor hand hygiene and fecal-oral transmission
- Direct skin contact with contaminated soil
- Repeated exposure and reinfection in endemic areas
- Socioeconomic vulnerability and population displacement

1.2.4 Clinical Manifestations and Complications

The clinical impact of gastrointestinal parasitic infections varies from asymptomatic carriage to significant intestinal and systemic complications, influenced by the parasite type, parasite load, and the immune status of the host.

1.2.4.1 Common Symptoms

Gastrointestinal parasitic infections present with a wide spectrum of clinical manifestations, ranging from asymptomatic carriage to severe intestinal disease, depending on the parasite species, infectious burden, and immune status of the host. It is important to recognize that asymptomatic infection does not equate to the absence of harm — subclinical infections may still cause occult blood loss, impaired nutrient absorption, and immune dysregulation, particularly in children experiencing repeated reinfection [13].

Common clinical manifestations include:

- Abdominal pain, cramps, bloating, and intestinal discomfort

- Diarrhea, which may be watery in giardiasis and cryptosporidiosis, or dysenteric in invasive amebiasis
- Nausea, vomiting, flatulence, and foul-smelling stools
- Fever, fatigue, and weight loss
- Malabsorption and nutritional deficiencies in chronic or repeated infections
- Iron-deficiency anemia, particularly in hookworm infection due to chronic intestinal blood loss [14].

In invasive infections, clinical presentation may extend beyond the gastrointestinal tract. *Entamoeba histolytica*, for example, can cause dysentery and, in severe cases, extraintestinal complications such as liver abscesses. Heavy *Ascaris lumbricoides* infections may lead to intestinal obstruction, particularly in young children.

1.2.4.2 Severe Complications and Vulnerable Populations

While many gastrointestinal parasitic infections follow a mild or self-limiting course, certain populations face a significantly elevated risk of severe and long-term complications. Three groups are particularly vulnerable: children, immunocompromised individuals, and pregnant women.

Children are disproportionately affected by the consequences of chronic and repeated parasitic infections. Persistent infection contributes to growth retardation, developmental delays, reduced school performance, and weakened immunity. Cognitive impairment and reduced educational attainment have been documented in school-age children with heavy helminth burdens, reflecting the broader developmental impact of untreated infection [15].

Immunocompromised individuals face heightened risk of severe disease. In HIV-positive patients, intestinal parasitic co-infections are associated with accelerated immune deterioration, higher rates of diarrhea-related mortality, and reduced antiretroviral therapy efficacy [16].*Cryptosporidium* infection in particular can become a chronic and life-threatening condition in patients without adequate antiretroviral access, as no fully effective treatment currently exists for cryptosporidiosis in severely immunocompromised hosts.

Pregnant women represent another high-risk group. Helminth infections during pregnancy are associated with maternal anemia, low birth weight, and increased risk of preterm delivery — consequences that extend beyond the mother to affect neonatal health outcomes [17].

These clinical realities underscore the importance of early and accurate diagnosis. The limitations of conventional diagnostic methods in detecting low-intensity infections and distinguishing between parasite species directly motivate the need for improved diagnostic infrastructure, including AI-assisted tools and organized biobank resources as proposed in this study [18].

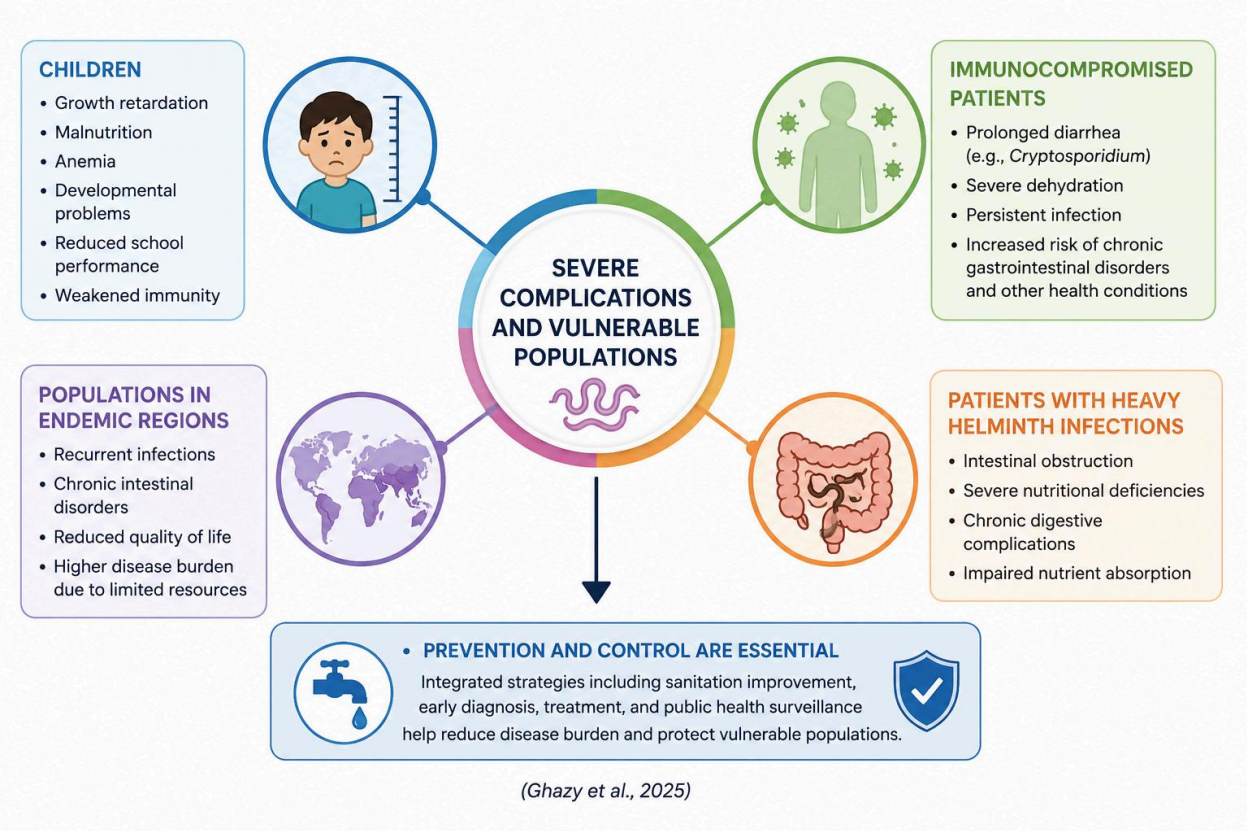


Figure 1.4: Major Complications and Vulnerable Populations in Intestinal Parasitic Infections

1.3 Current Challenges in the Management and Study of Gastrointestinal Parasites within Sustainable Health Systems

The management of gastrointestinal parasitic infections faces several diagnostic, operational, and public health challenges. These limitations affect early detection, disease monitoring, and the ability of health systems to respond effectively, especially in resource-limited settings.

1.3.1 Classic Diagnostic Methods

The diagnosis of gastrointestinal parasitic infections has historically relied on the microscopic examination of stool samples. This approach enables the detection of diverse parasite stages — including trophozoites, cysts, oocysts, eggs, and larvae — and remains the most widely practiced diagnostic method in routine laboratory settings, particularly in resource-limited environments where it offers accessibility and relatively low cost.

Several techniques are employed to improve detection sensitivity. Direct wet mount examination provides a rapid initial assessment of stool samples, while concentration methods such as formalin-ethyl acetate sedimentation increase the likelihood of detecting low-burden infections. Specific staining procedures further enhance the visualization of particular parasite forms: modified Ziehl-Neelsen staining is essential for detecting *Cryptosporidium* oocysts and *Cyclospora cayetanensis*, while trichrome staining is widely used for identifying protozoan trophozoites and cysts.

For helminth infections, the Kato-Katz thick smear technique remains the standard method for quantifying soil-transmitted helminth egg counts and estimating infection intensity. This quantitative information is important for guiding treatment decisions and monitoring the outcomes of deworming programs at the population level [19].

Where greater diagnostic precision is required, molecular techniques such as multiplex PCR offer higher sensitivity and improved species-level identification. However, their application remains limited in resource-constrained settings due to cost, infrastructure requirements, and the need for specialized technical expertise [20].

1.3.2 Limitations and Challenges of Traditional Diagnosis

Despite its widespread use, microscopy-based diagnosis presents several important limitations that affect diagnostic accuracy, reproducibility, and reliability — particularly in low-intensity infections and resource-limited settings.

Sensitivity is a fundamental concern. Microscopy may fail to detect parasites when the infectious burden is low or when eggs, cysts, or oocysts are shed intermittently in stool samples. Studies have reported sensitivity rates for microscopic diagnosis of giardiasis as low as 50–70% in single-sample examinations, highlighting the risk of false-negative results in routine screening [21].

Diagnostic accuracy is also strongly dependent on human expertise. Parasite stages must be distinguished from debris, artifacts, and morphologically similar non-pathogenic organisms — a task that requires substantial training and experience. This dependence introduces significant inter-observer variability, both between laboratories and within the same laboratory over time. A particularly important consequence is the inability to reliably distinguish between pathogenic and non-pathogenic species based on morphology alone, such as *Entamoeba histolytica* versus *Entamoeba dispar* — a distinction with direct implications for treatment decisions [22].

Operational constraints further compound these challenges. Sample quality, transportation conditions, staining consistency, and microscope availability all influence the final diagnostic result. In mixed infections or cases with low parasite concentrations, diagnostic performance may be further reduced.

These limitations have driven interest in complementary and alternative approaches. Molecular platforms such as multiplex PCR improve sensitivity and species differentiation but require laboratory infrastructure that is often unavailable in endemic settings [20]. Emerging technologies including loop-mediated isothermal amplification (LAMP) and CRISPR-based diagnostic tools offer high sensitivity and specificity with reduced infrastructure demands, making them promising candidates for point-of-care testing in resource-limited environments [23].The integration of AI-assisted image analysis into existing microscopy workflows represents a further avenue for addressing the limitations of human expertise while preserving the accessibility advantages of microscopy — an approach that forms a central component of the biobank system proposed in this study.

1.3.3 Impact on Sustainable Health Systems

The diagnostic and management challenges associated with gastrointestinal parasitic infections extend beyond individual patient care, placing a cumulative burden on health systems — particularly in endemic and resource-limited settings. This burden manifests across several interconnected dimensions.

Diagnostic limitations directly undermine health system efficiency. Missed or delayed diagnoses resulting from low microscopy sensitivity lead to repeated testing, unnecessary treatments, and prolonged patient morbidity. Fragmented and non-standardized data further reduce the capacity of health authorities to monitor parasite prevalence, detect outbreaks, and evaluate the effectiveness of control programs.

At a broader level, the chronic nature of gastrointestinal parasitic infections generates sustained demand for healthcare resources. Recurrent infections require repeated medical consultations, treatments, and follow-up, increasing long-term healthcare costs. The scarcity of trained parasitologists and laboratory technicians in

endemic countries creates a structural bottleneck that limits the impact of even well-designed diagnostic and control programs.

The socioeconomic consequences are equally significant. Intestinal parasitic infections reduce work capacity and educational productivity, contributing to broader cycles of poverty and underdevelopment in affected communities [2].Effective and sustainable management therefore requires not only improved diagnostic tools, but also organized data infrastructures capable of supporting epidemiological surveillance, research, and evidence-based public health decision-making [24].

These converging challenges establish the need for a specialized, digitally supported gastrointestinal parasite biobank — one capable of organizing diagnostic data, preserving biological samples, and integrating AI-based tools to strengthen both clinical diagnosis and population-level surveillance.

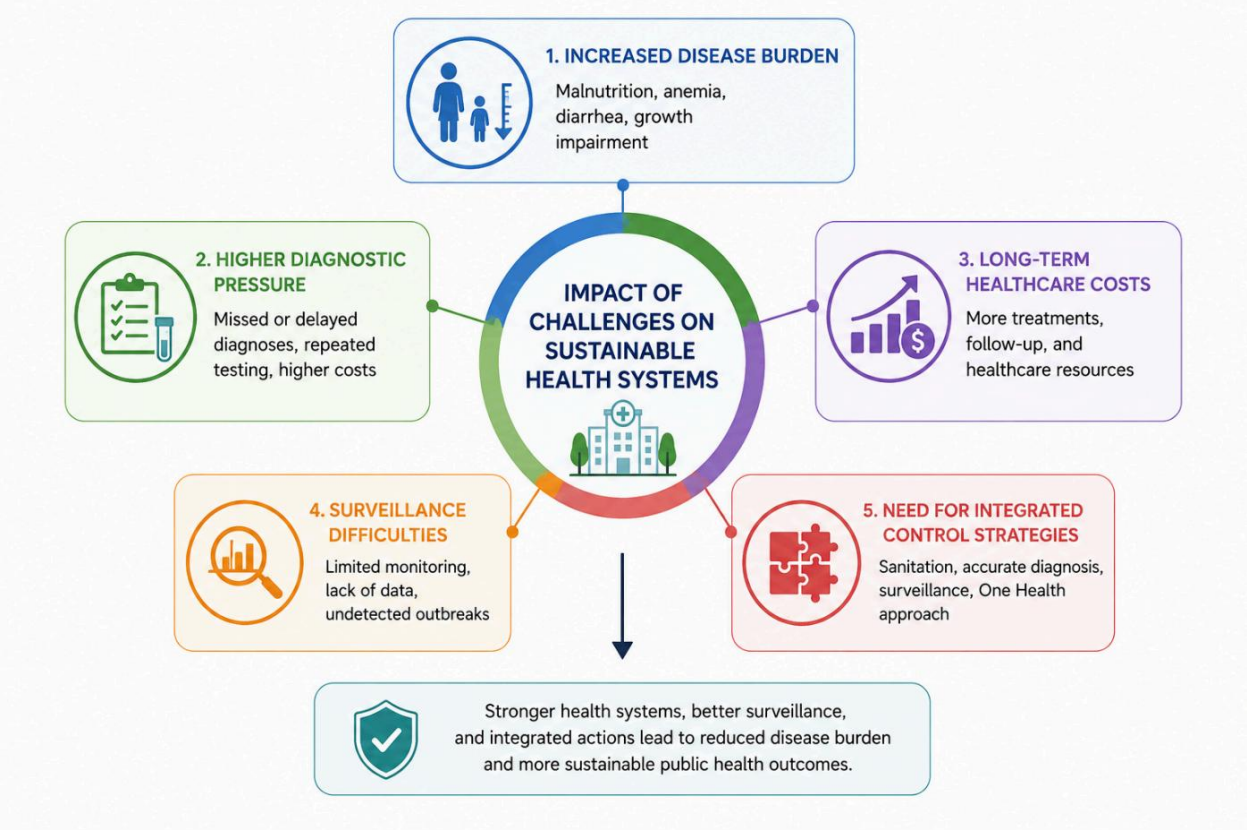


Figure 1.5: Impact of Intestinal Parasitic Infection Challenges on Sustainable Health Systems

1.4 Biobanks and Their Role in Parasitic Disease Research

1.4.1 Definition and Importance of Biobanks

A biobank is a structured research infrastructure that collects, processes, stores, and manages biological samples together with related clinical, epidemiological, and biological data. These samples may include blood, tissue, DNA, urine, stool, or other biological materials. By systematically linking samples with reliable and standardized data, biobanks support biomedical research, biomarker discovery, disease monitoring, and long-term epidemiological studies [25].

The scientific value of a biobank depends not only on the quality of its biological collections, but also on the governance frameworks that regulate their use. International bodies such as the International Society for Biological and Environmental Repositories (ISBER) and the Organisation for Economic Co-operation and

Development (OECD) have established guidelines emphasizing sample quality, data integrity, informed consent, and ethical oversight. These principles are particularly critical in low- and middle-income country settings, where regulatory frameworks may be less developed and where building community trust in research institutions requires deliberate and sustained effort [26].,[27].

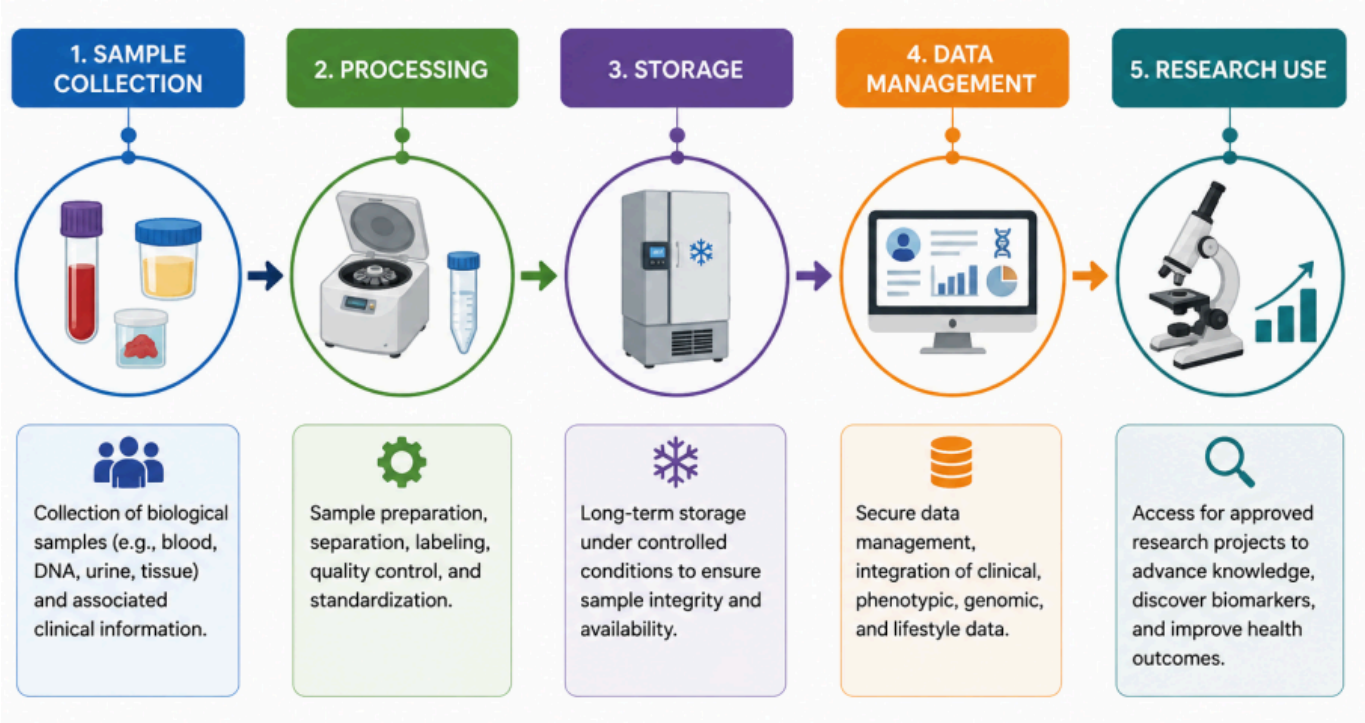


Figure 1.6: General structure of a biobank

1.4.2 Historical Evolution of Biobanks

Biobanks have undergone substantial transformation since their origins as simple biological sample repositories. Early collections focused primarily on preserving materials for genetic and epidemiological research. The subsequent development of high-throughput sequencing technologies, bioinformatics platforms, and digital health systems enabled modern biobanks to integrate biological samples with clinical, genomic, lifestyle, and environmental data at an unprecedented scale.

Large-scale initiatives such as the UK Biobank — which recruited approximately 500,000 participants and linked biological samples with deep phenotypic and genotypic data — illustrate the transformative research potential of well-organized biobanking infrastructure [28].

In the domain of infectious and parasitic diseases, coordinated initiatives such as the MalariaGEN Project and TropNet have demonstrated how standardized biobanks can accelerate drug resistance surveillance, vaccine development, and genomic epidemiology across multiple countries and research institutions [29].Today, biobanks are recognized not merely as storage facilities, but as organized, ethically governed, and data-driven research infrastructures that underpin precision medicine, population health studies, and evidence-based public health policy.

1.4.3 Theoretical and Functional Roles of Biobanks

Biobanks fulfill several interconnected roles within the biomedical research ecosystem. At the most fundamental level, they preserve high-quality biological samples and organize associated data in ways that make them accessible and reusable for multiple research purposes. This standardization enables large-scale studies, inter-institutional collaboration, and the longitudinal follow-up of disease trends over time.

From a theoretical perspective, biobanks function as nodes within a broader translational research cycle. Samples collected and stored in biobanks generate data that inform new scientific hypotheses, which in turn drive further sample collection and analysis. This iterative process — described as the "biobank-enabled research cycle" — accelerates scientific knowledge generation far beyond the capacity of individual studies[30].In applied terms, biobanks contribute to biomarker discovery, diagnostic tool development, risk stratification, and disease surveillance, forming a critical link between basic laboratory science and clinical or public health application.

1.4.4 Biobanks in the Context of Parasitic Diseases

In parasitic disease research, biobanks support the collection and preservation of diverse sample types including stool, blood, tissue, parasite isolates, microscopic images, and associated metadata. These resources are essential for diagnosis, epidemiological monitoring, drug resistance surveillance, pathogen characterization, and outbreak investigation.

For gastrointestinal parasites specifically, a specialized biobank provides a means of organizing parasite-related samples and data in a standardized and retrievable format — a capacity that is particularly valuable in endemic and resource-limited regions where data are frequently fragmented, short-term, or stored in non-digital formats. Existing international resources such as the BEI Resources Repository and WHO parasite reference collections have demonstrated the feasibility of centralized parasitological sample management. However, these platforms are designed primarily for reference laboratory and research purposes and are not adapted for local or regional epidemiological monitoring. A specialized gastrointestinal parasite biobank tailored to the Algerian and North African context would therefore address a critical infrastructural gap, enabling local sample preservation, regional surveillance, and integration with international research networks.

1.4.5 Toward an AI-Powered Gastrointestinal Parasite Biobank

The diagnostic and surveillance limitations described in the preceding sections point toward the need for a new type of research infrastructure — one that combines the organizational principles of biobanking with the analytical capabilities of artificial intelligence. The AI-powered gastrointestinal parasite biobank proposed in this study is designed to address this need.

Unlike conventional biobanks focused primarily on physical sample storage, the proposed system centers on the collection, organization, and AI-assisted analysis of microscopic parasite images, together with associated clinical, diagnostic, and epidemiological metadata. This focus reflects the nature of gastrointestinal parasite diagnosis, which depends fundamentally on the visual identification of developmental stages — eggs, cysts, oocysts, larvae, trophozoites, and adult forms — in microscopic stool preparations.

Recent advances in deep learning have demonstrated the potential of convolutional neural networks (CNNs) — particularly architectures such as ResNet and EfficientNet — to perform parasite detection and classification with accuracy comparable to or exceeding that of expert microscopists, when trained on sufficiently large and well-annotated datasets [31].The proposed biobank is designed to generate and curate precisely such datasets, creating

a self-reinforcing infrastructure in which accumulated diagnostic data continuously improves AI model performance.

In the present study, the proposed system begins as a digital biobank — a first operational step toward a future physical biobank infrastructure. The digital layer organizes parasite images, diagnostic outputs, sample metadata, and analytical results in a structured and reusable format. This foundation is designed to be scalable, supporting future extension to physical sample storage, standardized laboratory workflows, and national-level surveillance integration.

The core components of the proposed system include:

- Collection and structured organization of gastrointestinal parasite microscopic images
- Storage of sample metadata, clinical information, and diagnostic results
- AI-assisted parasite detection and species classification
- Expert review and diagnostic validation tools
- Epidemiological monitoring dashboards and data visualization
- Reusable annotated datasets for AI model development and research
- Ethical governance, data privacy protections, and regulatory compliance mechanisms

1.4.5.1 Limitations of Conventional Data Management

In most gastrointestinal parasite diagnostic settings, data management relies on conventional approaches — paper records, isolated spreadsheets, disconnected laboratory reports, and unstructured image storage. While adequate for basic documentation, these methods do not support the integration of biological samples, microscopic images, diagnostic results, clinical data, and epidemiological information within a unified and searchable system.

This fragmentation carries significant consequences. It limits the ability to trace samples longitudinally, compare cases across time periods or geographic areas, share data between laboratories, or reuse collected information for AI model development and research. The microscopic images generated during routine parasite diagnosis represent a particularly underutilized resource: when stored without standardized metadata and structured organization, their long-term scientific and diagnostic value is largely lost.

The scale of this problem is substantial. Studies of laboratory information management systems in low- and middle-income country settings have found that fewer than 30% of clinical laboratories in sub-Saharan Africa and North Africa use any form of digital data management, with the majority depending on paper-based records that are vulnerable to loss, damage, and transcription errors [32].This data management deficit directly undermines disease surveillance capacity, quality assurance, and research potential in precisely the regions where gastrointestinal parasitic infections impose the greatest burden.

1.4.5.2 Need for Digital and AI-Based Infrastructure

Addressing these limitations requires a digital infrastructure capable of organizing parasite diagnostic data in a structured, secure, and interoperable format. Such a system must go beyond simple data storage to actively connect microscopic images with sample metadata, clinical records, diagnostic outputs, and expert annotations — creating a coherent and queryable knowledge base.

Artificial intelligence integration amplifies the value of this infrastructure significantly. AI models can support automated image classification, diagnostic assistance, data quality control, pattern recognition across large datasets, and the preparation of curated training sets for future model development. Federated learning frameworks offer a particularly promising pathway for building robust AI models without centralizing sensitive

patient data, enabling collaborative model training across multiple laboratory sites while preserving data privacy and institutional autonomy [33].

This digital biobank infrastructure is especially relevant in the Algerian and North African context, where diagnostic capacity is uneven, data are fragmented, and the need for scalable and cost-effective surveillance solutions is acute. By establishing a structured digital foundation now, the proposed system creates the conditions for future expansion into a full physical biobank with standardized laboratory protocols, national registry integration, and international research collaboration capabilities.

1.4.5.3 Potential Contributions to Diagnosis, Surveillance, and Research

The proposed AI-powered gastrointestinal parasite biobank is designed to generate impact across three interconnected domains: clinical diagnosis, epidemiological surveillance, and scientific research.

In **diagnosis**, the system supports laboratory specialists by organizing microscopic images, providing AI-assisted parasite classification suggestions, storing diagnostic results, and enabling expert review and validation. The objective is not to replace human expertise, but to extend its reach — reducing the impact of inter-observer variability, supporting less experienced technicians, and improving diagnostic consistency across laboratory sites. Clinical validation studies will be essential to prospectively establish the sensitivity, specificity, and predictive values of the AI classification system under real-world laboratory conditions, ensuring its utility beyond benchmark dataset performance [34].

In **surveillance**, the platform enables the systematic collection and analysis of diagnostic data stratified by parasite species, geographic location, patient demographics, and sample date. Real-time epidemiological dashboards can serve as early warning systems for outbreak detection and transmission pattern monitoring — functions analogous to those implemented successfully within the WHO Integrated Disease Surveillance and Response framework for other infectious diseases, and now extended here to gastrointestinal parasites in resource-limited settings.

In **research**, the biobank provides curated, annotated, and reusable datasets that support the development of more accurate AI classification models, comparative studies between parasite species, image quality analysis, and longitudinal epidemiological trend assessment. Over time, these datasets represent a growing scientific asset — one that can support evidence-based health policy, inform national control programs, and contribute locally contextualized data to international parasitological research networks.

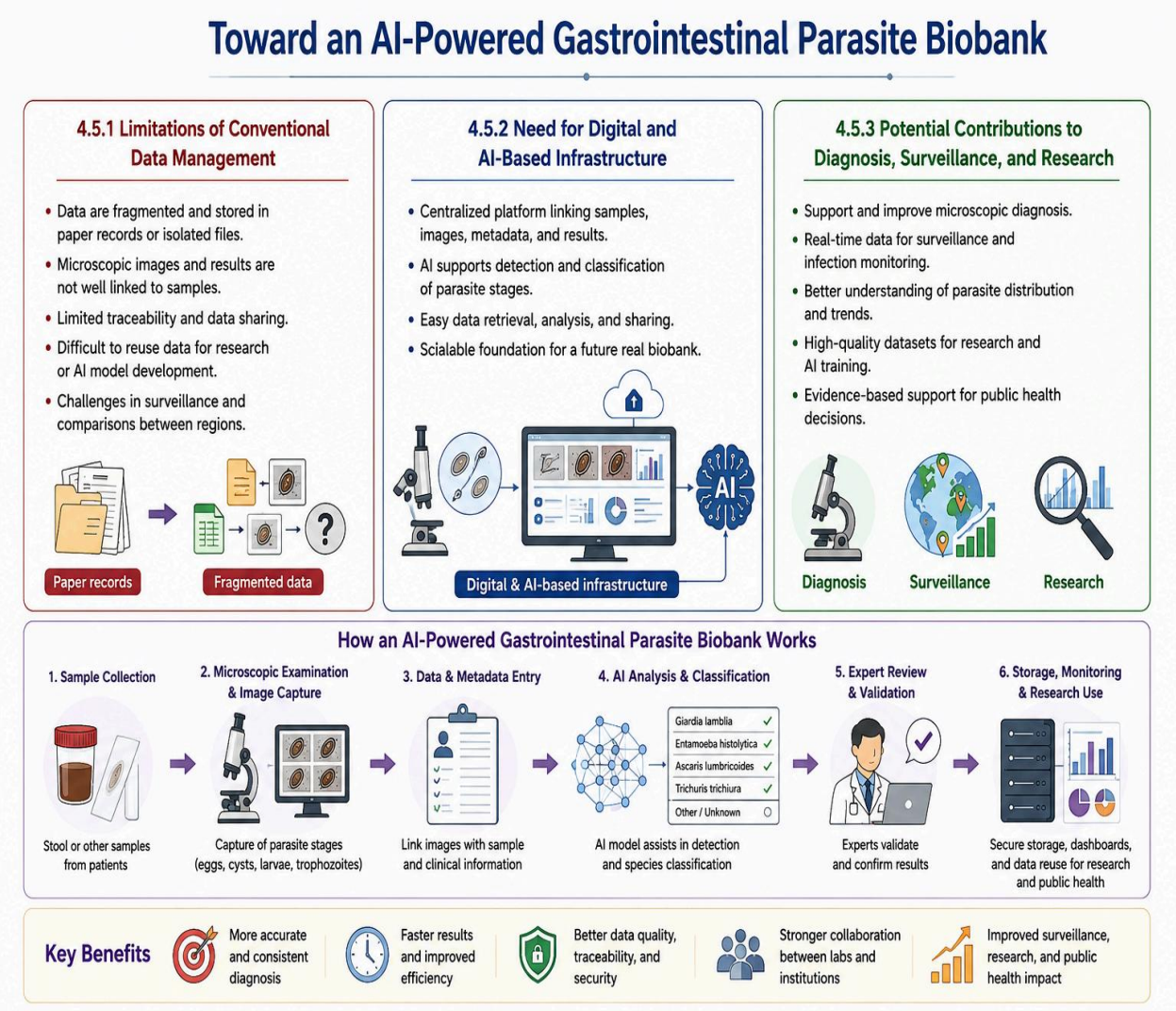


Figure 1.7: Conceptual framework of an AI-powered gastrointestinal parasite biobank

1.5 Conclusion

Gastrointestinal parasitic infections represent a persistent and multidimensional public health challenge, shaped by the biological complexity of the parasites involved, the environmental and socioeconomic conditions that drive transmission, and the diagnostic limitations that impede timely and accurate detection. The biological diversity of intestinal protozoa and helminths — reflected in their distinct morphologies, life cycles, and transmission routes — demands diagnostic approaches capable of distinguishing between species and developmental stages with precision and consistency.

In resource-limited settings such as Algeria and North Africa, these challenges are compounded by sanitation gaps, limited laboratory infrastructure, workforce shortages, and fragmented data management systems. Conventional microscopy-based diagnosis, while accessible and widely practiced, suffers from sensitivity limitations, inter-observer variability, and an inability to reliably differentiate between morphologically similar species — deficiencies that carry direct consequences for patient management and epidemiological surveillance.

The concept of biobanking offers a structured response to these infrastructural gaps. As biobanks have evolved from simple sample repositories into ethically governed, digitally integrated research platforms, their potential to support parasitic disease research — through standardized sample preservation, data organization, and AI-assisted analysis — has grown substantially. A specialized AI-powered gastrointestinal parasite biobank, designed for the Algerian and North African context, represents a concrete step toward bridging the gap between conventional parasitological diagnosis and the demands of modern evidence-based public health.

The following chapter builds on this foundation by presenting the artificial intelligence and medical image analysis principles that underpin the proposed biobank system, with particular attention to the challenges of applying deep learning to microscopic parasite images in resource-limited diagnostic environments.

Chapter 2:

Modern AI Approaches For Medical Image Analysis

2.1 Introduction

Artificial intelligence has become a central component of modern medical image analysis. In domains such as radiology, pathology, dermatology, endoscopy, and microscopy, AI systems are increasingly used to extract visual information, support diagnosis, reduce interpretation time, and assist clinicians in decision-making. This importance is even more evident in microscopic image analysis, where the target structures may be small, visually similar, noisy, and difficult to interpret without expert knowledge.

In the context of gastrointestinal parasite detection, these challenges are particularly significant because parasite eggs and microscopic structures often present high morphological similarity, variations in illumination, image noise, and limited availability of expert annotations.

Traditional supervised learning approaches require large annotated datasets. However, in medical imaging, expert annotation is expensive, time-consuming, and sometimes inconsistent. This limitation motivates modern approaches such as representation learning, self-supervised learning, semi-supervised learning, clustering, and pseudo-labeling. These methods allow models to exploit unlabeled data and learn meaningful visual representations before or alongside supervised classification.

2.2 Artificial Intelligence in Medical Image Analysis

Medical image analysis refers to the computational processing and interpretation of medical images in order to extract clinically meaningful information. Artificial intelligence contributes to this field by enabling automatic feature extraction, image classification, lesion detection, segmentation, prognosis prediction, and clinical decision support. Deep learning models have shown strong performance because they can learn hierarchical visual representations directly from raw images instead of relying only on handcrafted features.

AI has been applied in several medical imaging domains, including:

- Radiology, where it supports the interpretation of X-ray, CT, MRI, and ultrasound images.
- Pathology, oncology, and dermatology, where it helps detect abnormal tissues, classify tumor patterns, and analyze skin lesions.
- Microscopic parasitology, where it can assist in the analysis of parasite eggs and larvae by extracting meaningful morphological patterns.

However, AI systems in healthcare require careful validation because their performance can be affected by image quality, acquisition conditions, patient variability, and dataset bias.

2.3 Theoretical and Technical Foundations of CNN and ResNet Architectures

CNN and ResNet architectures provide the main deep learning foundation for this work. They allow microscopic parasite images to be transformed into meaningful visual representations that can be used for classification, clustering, and self-supervised learning.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks are widely used in image analysis because they can automatically extract visual features from images. They use convolutional filters to detect patterns such as edges, textures, shapes, and

structures. In medical imaging, CNNs support classification, detection, and segmentation tasks. In microscopic parasitology, CNNs can learn important visual characteristics such as parasite shape, texture, boundaries, and contrast. This reduces the need for manual feature engineering. However, very deep CNNs may suffer from training difficulties such as vanishing gradients.

2.3.2 Residual Networks

Residual Networks were introduced to improve the training of deep CNNs. Their main idea is the use of skip connections, which allow information and gradients to flow more easily through the network. This makes it possible to train deeper models more effectively. ResNet architectures, such as ResNet18 and ResNet50, are commonly used in medical image analysis. In this work, ResNet is important as an encoder for extracting meaningful representations from microscopic parasite images, which can later be used for classification, self-supervised learning, clustering, and pseudo-label generation.

2.4 Learning paradigms for AI-based medical image analysis

Learning paradigms define the training strategy of an artificial intelligence model according to the availability of labels and the type of supervision used during optimization. In medical image analysis, this distinction is essential because annotated images are often limited, while unlabeled images may be more accessible. Therefore, supervised, unsupervised, semi-supervised, and self-supervised learning represent complementary approaches that support the development of robust medical image analysis systems.

2.4.1 Supervised Learning

Based on this classification, supervised learning can first be introduced as the classical training paradigm in medical image analysis, especially when reliable expert annotations are available.

Supervised learning is the most traditional and widely used paradigm in medical image classification. In this approach, each training image is associated with a known label or diagnostic category. The model receives an input image and learns to predict its corresponding class by minimizing the difference between the predicted output and the ground-truth label. For example, in microscopic parasite image analysis, a labeled image may correspond to a specific parasite species or diagnostic class.

In deep learning, supervised learning often uses architectures such as CNNs and ResNet. These models automatically learn visual features and link them to predefined classes. Supervised learning can achieve high performance when data are large, balanced, and well annotated. However, in medical imaging, its main limitation is the need for expert labels, which require specialized knowledge and time.

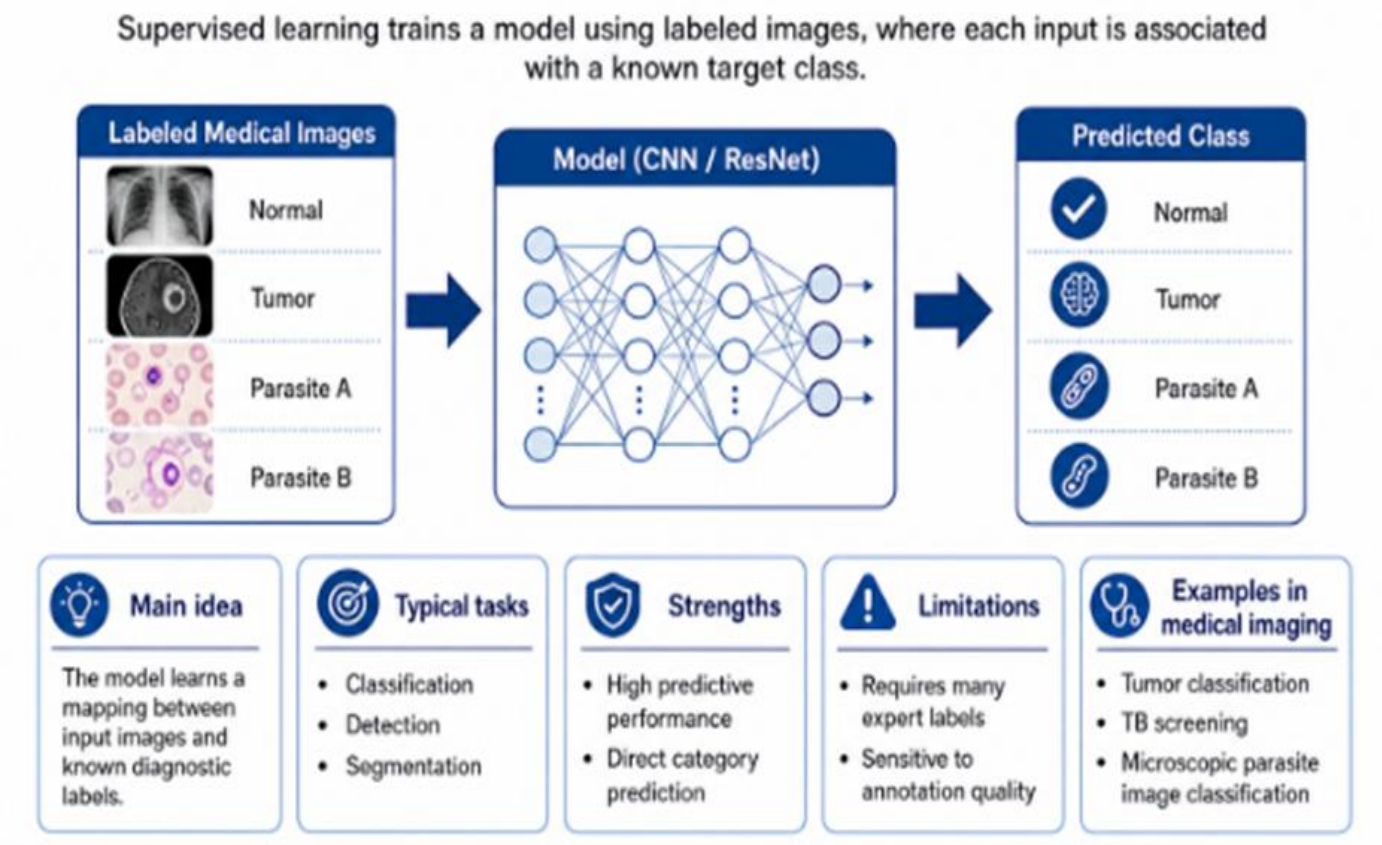


Figure 2.1: Overview of Supervised Learning in Medical Image Classification

2.4.2 Unsupervised Learning

Unsupervised learning uses images without predefined labels. Instead of learning a direct mapping between images and diagnostic categories, the model attempts to identify similarities, structures, or patterns within the data. This paradigm is useful when expert annotations are unavailable or when the objective is to explore the organization of a dataset before classification.

In medical image analysis, unsupervised learning is often associated with dimensionality reduction, latent space visualization, anomaly discovery, and clustering. For example, after extracting embeddings from microscopic images, clustering algorithms can group visually similar samples based on morphology, texture, or structural patterns. Although unsupervised methods do not directly provide clinical labels, they can support data exploration, sample selection, and the discovery of meaningful visual groups that may later be reviewed by experts.

Although unsupervised learning is useful for exploring hidden structures in unlabeled data, it does not directly use available diagnostic labels. Semi-supervised learning therefore offers a compromise by combining a small labeled subset with a larger unlabeled dataset.

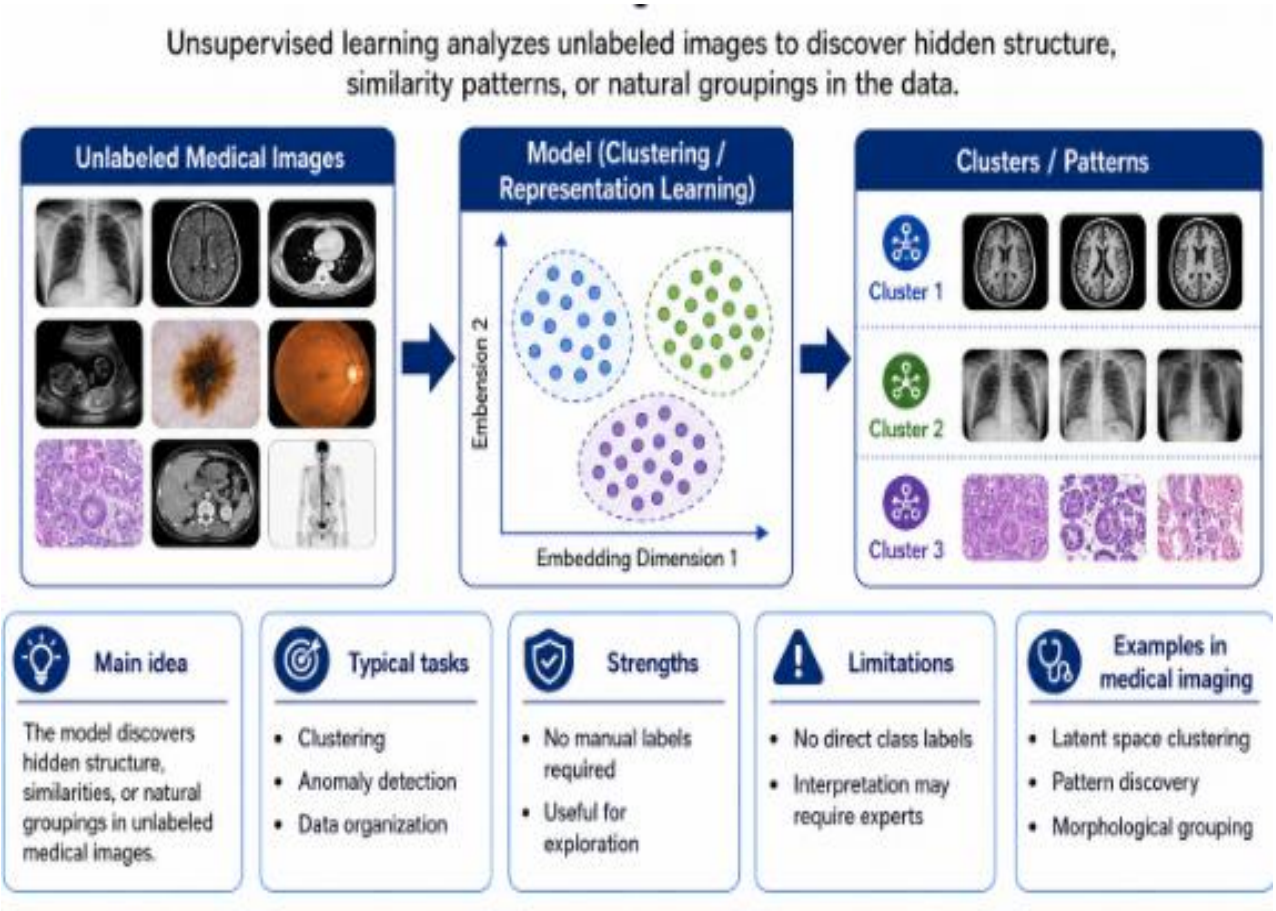


Figure 2.2: General Workflow of Unsupervised Learning for Medical Images

2.4.3 Semi-Supervised Learning

Semi-supervised learning represents an intermediate strategy between supervised and unsupervised learning. It uses a limited amount of labeled data together with a larger amount of unlabeled data. This approach is particularly relevant in medical imaging because expert-labeled datasets are usually scarce, while unlabeled images can often be collected more easily from clinical or laboratory workflows.

The objective of semi-supervised learning is to improve model performance by exploiting the information contained in unlabeled samples. Common strategies include pseudo-labeling, consistency regularization, and teacher-student training. In pseudo-labeling, a model predicts labels for unlabeled images, and the most confident predictions are used as additional training targets. In the context of parasite image analysis, this paradigm is useful because it reduces dependence on full manual annotation while still allowing the model to benefit from available expert-labeled examples.

While semi-supervised learning still depends on a limited number of expert labels, self-supervised learning goes further by learning visual representations directly from unlabeled data. This makes it particularly suitable for medical image datasets where manual annotation is scarce.

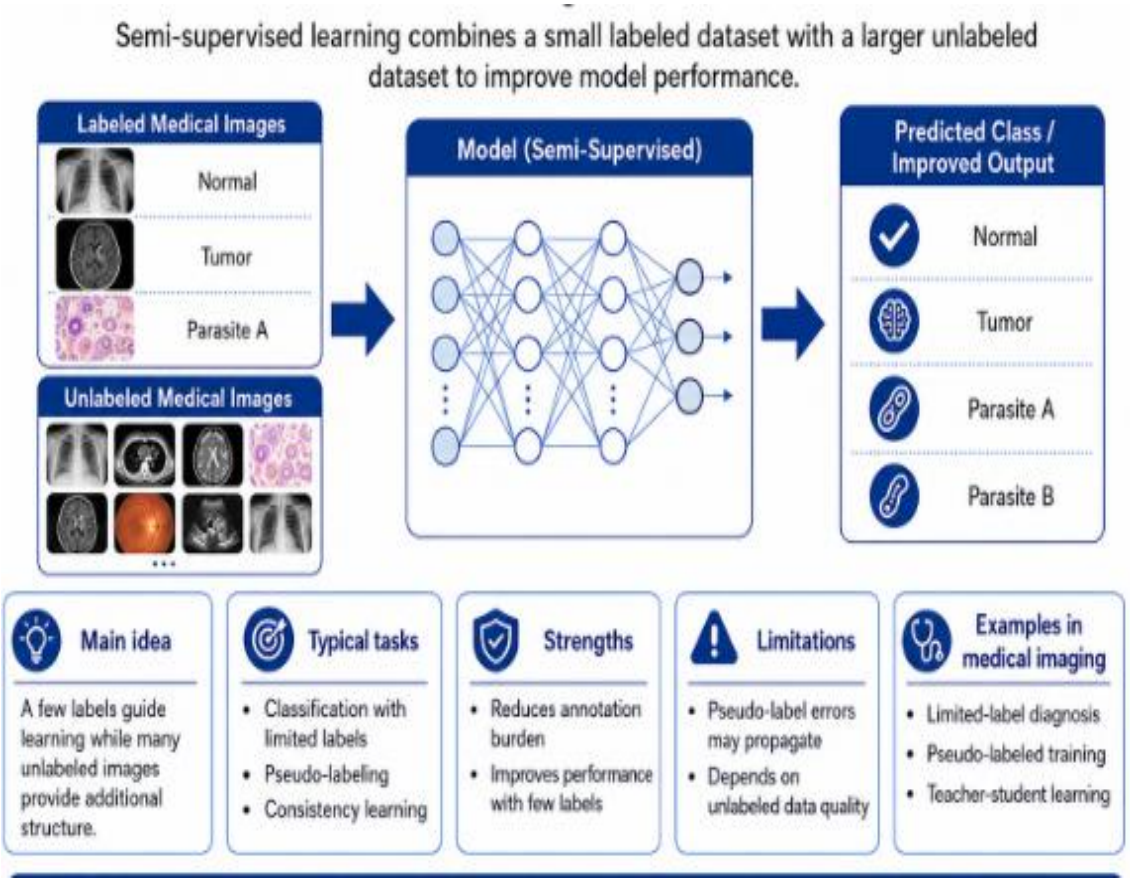


Figure 2.3: Semi-Supervised Learning for Microscopic Parasite Image Classification

2.4.4 Self-Supervised Learning

Self-supervised learning is a modern learning paradigm that aims to learn useful visual representations from unlabeled data. Unlike supervised learning, it does not require expert-provided labels. Instead, the model generates its own learning signal from the data through a pretext task. In image analysis, this is commonly achieved by applying different transformations or augmentations to the same image. The model is then trained to produce stable and consistent representations despite these changes. In this way, self-supervised learning allows the network to learn meaningful visual features before using any manual annotation.

Several methods have been proposed for self-supervised representation learning, including SimCLR, MoCo, BYOL, SimSiam, and Triplet Learning. Although these methods differ in their training strategies, they share the same general objective: learning embeddings that preserve important visual similarities between images.

For microscopic parasite images, self-supervised learning is particularly useful because many datasets are unlabeled or only partially annotated. The model can learn morphological features such as parasite shape, boundary structure, internal texture, contrast, and visual organization directly from the images. These learned embeddings can later be used for several downstream tasks, including classification, clustering, pseudo-label generation, and semi-supervised training.

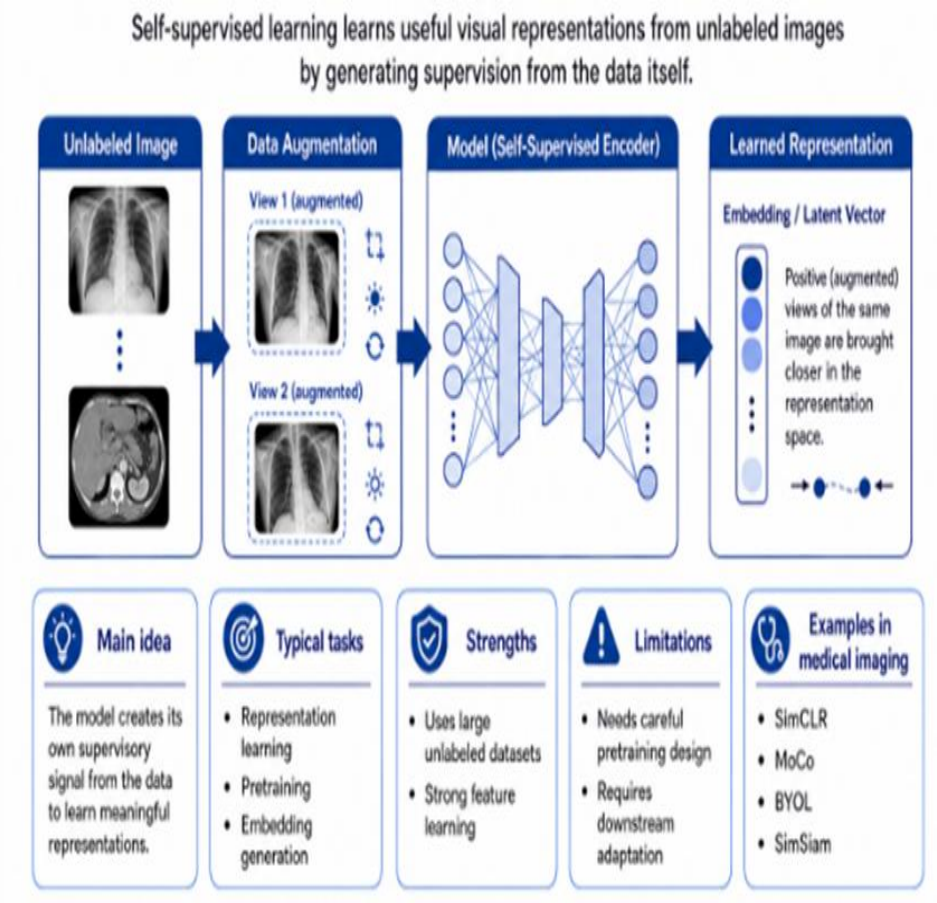


Figure 2.4: Overview of Self-Supervised Learning in Medical Image Analysis

Overall, the four learning paradigms provide a coherent foundation for AI-based medical image analysis:

- Supervised learning is effective when reliable expert labels are available.
- Unsupervised learning helps discover hidden structures in unlabeled datasets.
- Semi-supervised learning combines a small amount of labeled data with larger unlabeled datasets.
- Self-supervised learning enables representation learning without manual annotation.

This progression highlights the importance of addressing the annotation problem in medical imaging, especially in microscopic parasitology where expert labeling is costly and time-consuming. Therefore, representation learning, latent space organization, clustering, and pseudo-labeling become important strategies for preparing and exploiting medical image datasets with limited annotations.

2.5 Supervised Learning and Image Classification:

Supervised learning is the classical learning approach used when labeled images are available. In this paradigm, each image is associated with a known label, and the model learns to classify it into predefined categories by comparing its predictions with the ground-truth labels. After training, the model is evaluated on validation and test sets in order to measure its ability to generalize to unseen images.

In medical imaging, supervised CNN-based models are commonly applied in several tasks, such as:

- Disease classification and radiological diagnosis.

- Tumor detection and histopathology analysis.
- Tuberculosis screening and other diagnostic support tasks.

When medical datasets are limited, transfer learning is often used. This technique adapts a pretrained model, such as an ImageNet-based CNN, to a specific medical imaging task. Supervised learning can achieve strong performance when the dataset is large, diverse, and accurately annotated. However, its effectiveness depends strongly on the quality and quantity of expert labels. Small, biased, or incorrectly labeled datasets may lead to overfitting and poor generalization.

This limitation is especially important in microscopic parasite image analysis, where expert annotation is difficult, time-consuming, and limited. Therefore, the annotation problem motivates the use of alternative learning strategies that can exploit unlabeled or partially labeled medical images.

2.6 The Data Annotation Problem in Medical Imaging

A central challenge in medical image analysis is the scarcity of high-quality annotated data. Unlike natural image datasets, medical images require labeling by trained specialists who have the necessary clinical or laboratory expertise. This process is expensive, slow, and sometimes subjective, especially when visual differences between classes are subtle or when image quality varies.

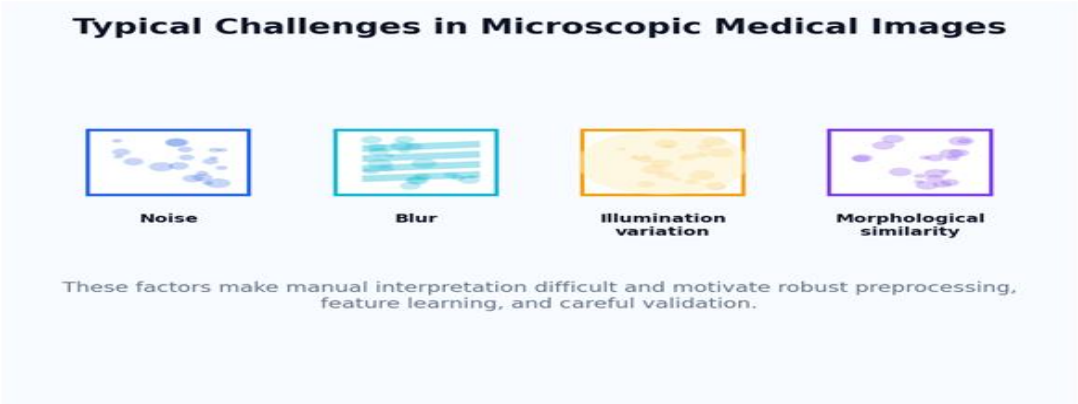


Figure 2.5: Schematic illustration of common challenges in microscopic medical images

In microscopic parasitology, the annotation problem is even more critical because parasite eggs or larvae may present similar shapes, textures, and morphological structures. In addition, image acquisition conditions can vary according to staining quality, focus, illumination, magnification, microscope type, and camera settings. These variations make manual annotation more difficult and can also affect the performance of supervised learning models.

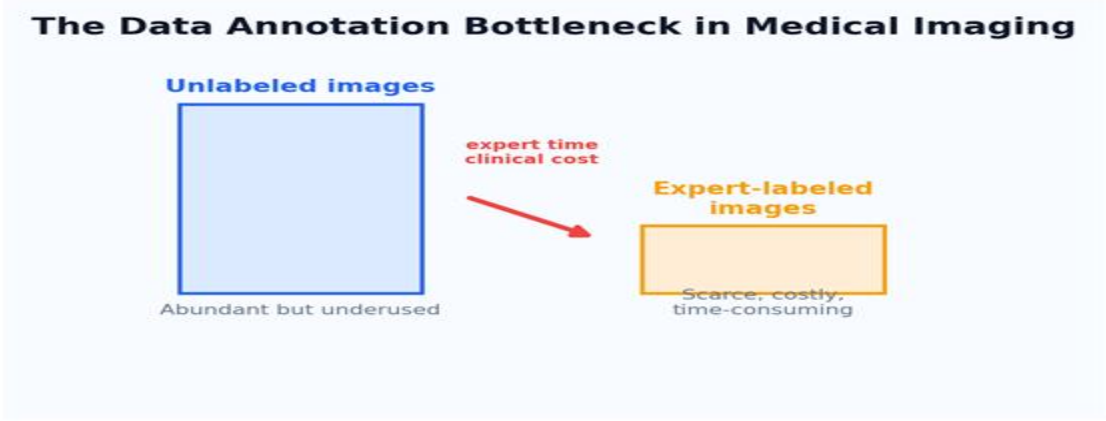


Figure 2.6: Data annotation bottleneck in medical imaging

The annotation bottleneck limits the scalability of purely supervised approaches. In many medical contexts, large quantities of unlabeled images can be collected from laboratories or clinical workflows, while only a small subset can be accurately labeled by experts. As a result, relying only on fully supervised learning may restrict the development of robust and generalizable AI systems.

This challenge motivates the use of methods that can learn from unlabeled or partially labeled data, such as unsupervised learning, semi-supervised learning, and self-supervised representation learning. These approaches aim to reduce dependence on manual annotation while still extracting meaningful visual features that can support classification, clustering, pseudo-labeling, and expert-assisted diagnosis.

2.7 Representation Learning and Automatic Feature Extraction

Representation learning refers to the ability of a model to automatically learn meaningful features from data. In image analysis, the input image is transformed into a numerical feature vector, also called an embedding, which summarizes important visual information. These learned representations can then be used for several downstream tasks, including classification, clustering, visualization, retrieval, and pseudo-label generation.

The main advantage of representation learning is that it reduces the need for handcrafted features. In classical image processing, experts manually define descriptors such as shape, texture, color, or edge-based features. In contrast, deep learning models learn these features directly from images. This is particularly useful in medical imaging, where relevant visual patterns may be subtle, complex, and difficult to describe manually.

In microscopic parasite image analysis, representation learning can help capture morphological characteristics such as roundness, boundary structure, internal texture, transparency, and visual similarity between parasite categories. When these representations are learned from large unlabeled datasets, they can provide a strong foundation for clustering, pseudo-labeling, and semi-supervised classification.

After learning these visual representations, it becomes necessary to understand how they are organized inside the model. This organization is commonly analyzed through the concept of latent space, where each image embedding can be represented as a point according to its visual similarity with other images.

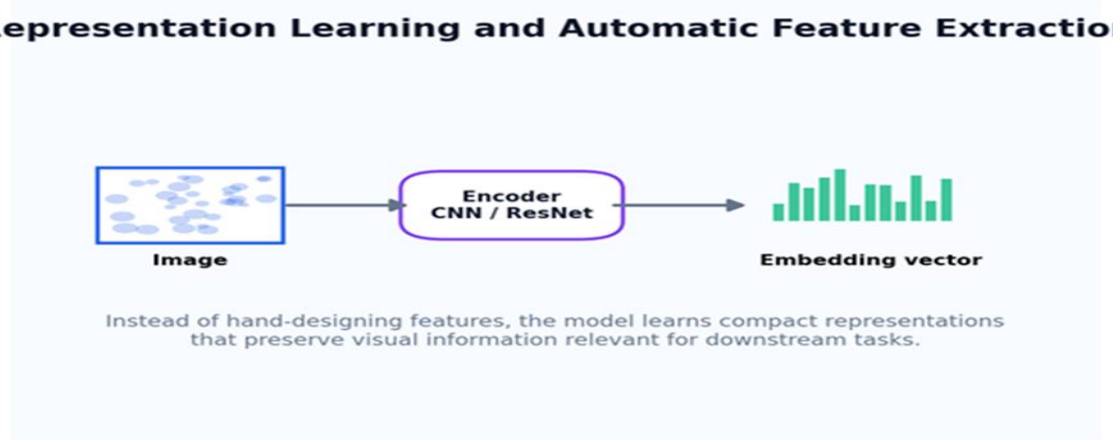


Figure 2.7: Representation learning process from image to embedding vector

2.8 Latent Space and Visual Embeddings

A latent space is a representation space in which each image is encoded as a point or vector. Images with similar visual characteristics are expected to appear close to each other, while visually different images are expected to be located farther apart. Therefore, latent space provides a useful way to analyze the internal structure of a dataset and identify hidden visual patterns.

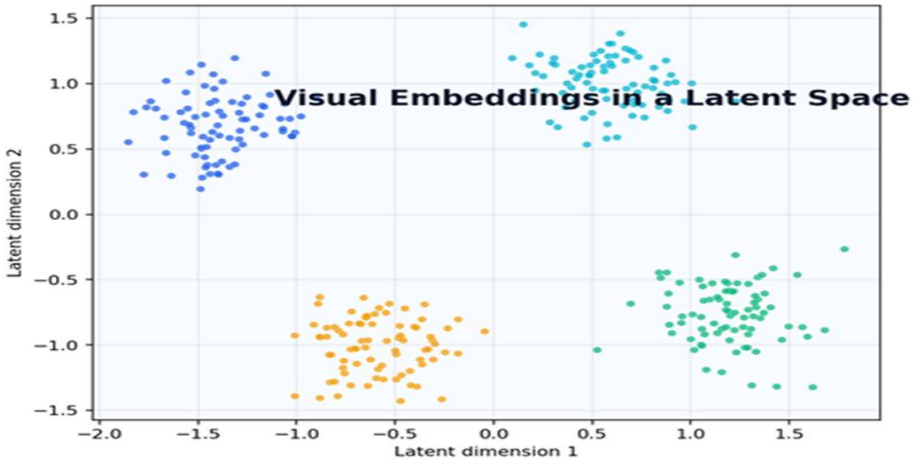


Figure 2.8: Visualization of image embeddings in latent space

If the latent space is well structured, similar parasite images tend to form compact groups. This organization can support pseudo-label generation, expert review, and the training of downstream classification models. In this way, latent space analysis connects representation learning with the later methodological steps of clustering and semi-supervised classification.

Latent space analysis is particularly important in self-supervised and semi-supervised learning. After training an encoder, the learned embeddings can be visualized using dimensionality reduction methods such as PCA, t-SNE, or UMAP. They can also be grouped using clustering algorithms. If the embeddings are well structured, similar images form compact clusters, which can later support pseudo-labeling and classifier training [35], [36].After

explaining how image embeddings are organized in latent space, it is important to describe the learning strategies that help build such structured representations. Contrastive learning is one of the most important approaches for this purpose, because it trains the model to bring similar images closer together and push different images farther apart in the embedding space.

2.9 Contrastive Learning

Contrastive learning is a self-supervised representation learning strategy that aims to organize images in the embedding space according to visual similarity. The main idea is to train the model so that different augmented views of the same image, or visually similar images, are represented close to each other, while dissimilar images are pushed farther apart. In this way, the model learns meaningful visual representations without requiring manual labels.

In medical image analysis, contrastive learning is useful because it can exploit unlabeled image collections and learn discriminative features from visual patterns. For microscopic parasite images, this may help the model capture similarities related to shape, texture, boundaries, and internal structures. These learned embeddings can later support clustering, pseudo-labeling, expert review, and downstream classification.

Since contrastive learning provides the general principle of learning similarity-based embeddings, the next section presents the main self-supervised methods that implement this idea in different ways. These methods include SimCLR, MoCo, BYOL, SimSiam, and Triplet Learning, which differ in their use of positive pairs, negative samples, momentum encoders, or Siamese architectures.

2.9.1 SimCLR

SimCLR learns representations by applying two random augmentations to the same image and treating them as a positive pair. These two views are passed through a shared encoder and a projection head, then optimized using a contrastive loss. The model learns invariance to transformations such as cropping, color variation, blur, and rotation, which can be useful for medical images acquired under variable conditions.

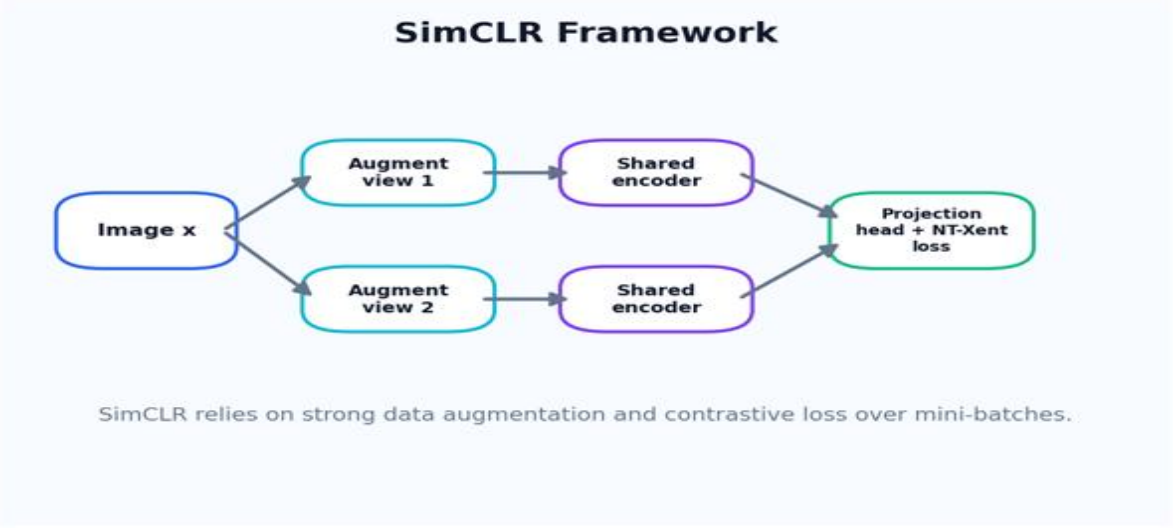


Figure 2.9: SimCLR framework for self-supervised visual representation learning

2.9.2 MoCo

Momentum Contrast (MoCo) is another contrastive framework that uses a dynamic dictionary of negative samples and a momentum-updated encoder. This design allows the model to compare an image representation with a large set of stored representations, improving training stability. MoCo has been used in different medical imaging applications where unlabeled data are available but labels are limited [37].

2.9.3 BYOL

Bootstrap Your Own Latent (BYOL) is a self-supervised method that does not require explicit negative pairs. It uses two networks: an online network and a target network. The online network predicts the representation produced by the target network for another augmented view of the same image. The target network is updated gradually using an exponential moving average of the online network. This mechanism enables representation learning without directly comparing against negative examples [38].

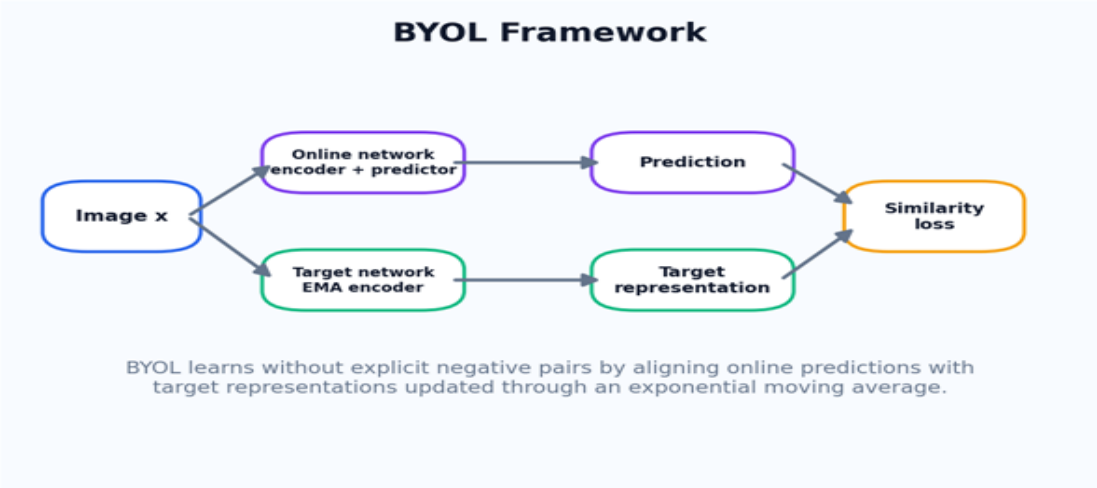


Figure 2.10: BYOL self-supervised learning architecture

2.9.4 SimSiam

SimSiam is a simplified Siamese self-supervised method that also avoids explicit negative samples. It relies on a stop-gradient operation and a prediction head to prevent representation collapse. Its simplicity makes it attractive for representation learning, especially when computational resources or dataset size are limited [39].

2.9.5 Triplet Learning Triplet

learning is based on three samples: an anchor, a positive sample similar to the anchor, and a negative sample different from the anchor. The objective is to reduce the distance between the anchor and positive embeddings while increasing the distance between the anchor and negative embeddings. This method is useful for learning metric spaces where similarity relationships are important[40].

The following comparison summarizes the main self-supervised learning methods according to their core idea, use of explicit negative pairs, and typical output. This helps clarify the differences between contrastive methods such as SimCLR and MoCo, non-contrastive methods such as BYOL and SimSiam, and metric-learning approaches such as Triplet Learning.

Comparison of Major Self-Supervised Learning Methods

Method	Core idea	Explicit negative pairs	Typical output
SimCLR	two augmented views	Yes	encoder features
MoCo	momentum encoder and queue	Yes	stable features
BYOL	online-target alignment	No	robust features
SimSiam	stop-gradient Siamese learning	No	compact features
Triplet	anchor-positive-negative samples	Yes	metric space

Figure 2.11: Comparison of major self-supervised learning methods

2.10 Clustering in Latent Space for Image Organization

Clustering is an unsupervised learning technique that groups samples according to similarity. When applied to learned embeddings, clustering can reveal the internal organization of an image dataset. Instead of clustering raw pixels, which are high-dimensional and noisy, the model clusters compact representations extracted by a trained encoder.

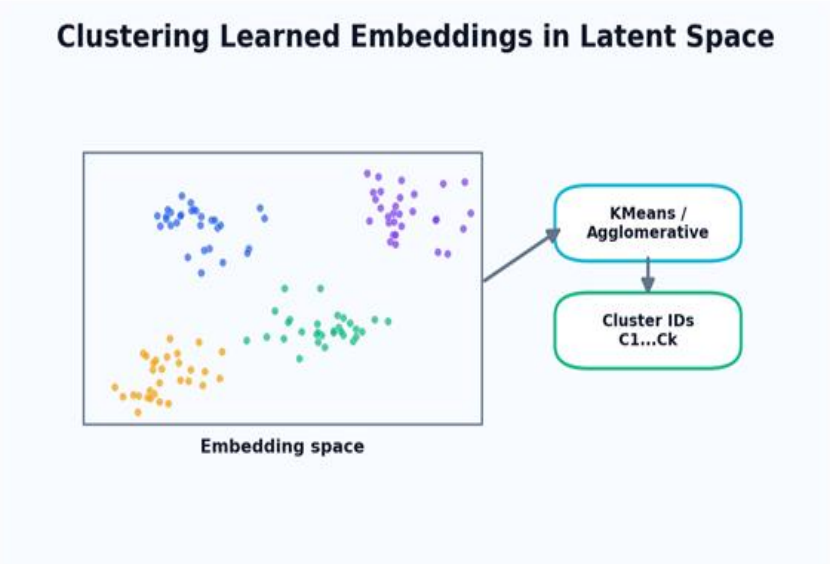


Figure 2.12: Clustering of learned embeddings in latent space

Common clustering algorithms include KMeans, Agglomerative Clustering, Gaussian Mixture Models, DBSCAN, and Spectral Clustering. KMeans groups samples around centroids, while Agglomerative Clustering builds a hierarchical structure by progressively merging similar samples. In medical image analysis, clustering can support data exploration, anomaly detection, cohort discovery, and pseudo-label generation.

For microscopic images, clustering may help organize visually similar samples, even when expert labels are unavailable. If the self-supervised encoder learns meaningful morphological features, the resulting clusters may correspond to visual groups that can later be inspected by experts or used as pseudo-labels [32].

Although clustering can organize unlabeled images into visually similar groups, these groups become more useful when they are converted into training signals. This transition leads to pseudo-labeling, where cluster assignments or model predictions are used as temporary labels to support semi-supervised learning.

2.11 Pseudo-Labeling and Semi-Supervised Learning

Pseudo-labeling is a technique that transforms unlabeled data into usable training data by assigning automatic labels based on model predictions or clustering results. These labels are not equivalent to expert annotations, but they can provide useful supervision when manually labeled data are limited. In this context, pseudo-labeling represents a bridge between latent-space clustering and downstream classifier training.

A typical pseudo-labeling pipeline begins by training an encoder on unlabeled images, often using self-supervised learning. The encoder extracts visual embeddings that summarize the morphological characteristics of each image. These embeddings can then be reduced in dimension using methods such as PCA and grouped using clustering algorithms such as KMeans or Agglomerative Clustering. The resulting cluster identifiers may be used as pseudo-labels, especially when the clusters contain visually coherent samples.



Figure 2.13: Pseudo-label generation from clustered image embeddings

In microscopic parasite image analysis, pseudo-labeling can help reduce dependence on full manual annotation. For example, visually similar parasite images can be grouped together in the latent space, and representative samples from each cluster can be inspected by experts. After this review, the pseudo-labels can support the training of a supervised classifier or be combined with a small set of expert-labeled images in a semi-supervised learning strategy.

However, the reliability of pseudo-labels depends strongly on the quality of the learned representations and the structure of the clusters. If the embeddings are poorly organized, the generated pseudo-labels may introduce noise and reduce model performance. Therefore, pseudo-labeling should be evaluated carefully using internal clustering metrics, external validation, visual inspection, or expert review whenever possible.

Overall, pseudo-labeling provides an important methodological link between self-supervised representation learning, clustering, and semi-supervised classification. It allows unlabeled microscopic images to contribute to model training while still leaving space for expert validation and correction.

After explaining how pseudo-labeling can transform unlabeled images into useful training signals, it is important to situate these methods within the broader research context. The following section therefore presents the state of the art on representation learning and limited-annotation medical image analysis, with a specific focus on their relevance to AI-based gastrointestinal parasite biobanking.

2.12 State of the Art on Representation Learning and Limited-Annotation Medical Image Analysis for AI-Based Gastrointestinal Parasite Biobanking

Recent research in medical image analysis has increasingly focused on reducing dependence on large expert-labeled datasets. Although CNN and ResNet architectures have achieved strong performance in image classification, detection, and segmentation [41],[42].,most supervised approaches still require high-quality annotations, which are often difficult and time-consuming to obtain in medical imaging. This limitation is more critical in microscopic gastrointestinal parasite analysis, where parasite eggs and larvae may present similar shapes, textures, boundaries, and internal structures [43].

To address this challenge, several learning strategies have been explored:

- Representation learning, which transforms images into embeddings that capture meaningful visual information.
- Self-supervised learning, using methods such as SimCLR, MoCo, BYOL, and SimSiam to learn from unlabeled images [44]. [33]. [34]. [35]..
- Semi-supervised learning, which combines a small labeled dataset with larger unlabeled data through pseudo-labeling, consistency regularization, and teacher-student models [45].
- Latent-space clustering, using methods such as K-Means or Agglomerative Clustering to organize embeddings, support pseudo-label generation, guide expert review, and identify ambiguous samples.

Within this context, an AI-powered gastrointestinal parasite biobank is not limited to image classification. It can also organize microscopic images, metadata, and diagnostic information while supporting representation learning, clustering, pseudo-labeling, and classification. However, limited-annotation microscopic parasite image analysis remains less explored compared with radiology, pathology, dermatology, and oncology, and most parasite-related studies still rely mainly on supervised learning. Therefore, this work addresses this gap by combining self-supervised representation learning, latent-space clustering, pseudo-labeling, and classification within an AI-powered gastrointestinal parasite biobank framework. After reviewing the state of the art, the next section discusses model evaluation, with particular attention to overfitting, generalization, and performance metrics.

2.13 AI Model Evaluation: Validation, Testing, Overfitting, and Generalization

Evaluation is essential for determining whether an AI model is reliable. Training accuracy alone is not sufficient because a model may memorize the training data without learning patterns that generalize to new images. Therefore, datasets are commonly divided into training, validation, and testing subsets. The training set is used to

optimize model parameters, the validation set is used for model selection and hyperparameter tuning, and the test set is used for final performance estimation.

Overfitting occurs when a model performs very well on training data but poorly on unseen data. This problem is common in medical imaging when datasets are small or unbalanced. Regularization techniques, data augmentation, early stopping, transfer learning, and external validation can help reduce overfitting. For medical AI systems, generalization is particularly important because images may come from different devices, hospitals, acquisition protocols, or patient populations.

In addition to accuracy, evaluation should include metrics such as precision, recall, F1-score, confusion matrix analysis, and class-wise performance. For clustering, internal metrics such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz score can be used, while external metrics such as purity, adjusted Rand index, normalized mutual information, and Hungarian accuracy can be computed when ground-truth labels are available. These metrics help assess whether the learned representations and clusters are meaningful [46].

After presenting the main evaluation principles and metrics used to assess classification and clustering performance, the chapter can now be concluded by summarizing the theoretical foundations that support the proposed methodology.



Figure 2.14: Dataset split and overfitting behavior in deep learning evaluation

2.14 Conclusion

This chapter presented the main artificial intelligence approaches used in medical image analysis, including deep learning, representation learning, self-supervised learning, clustering, pseudo-labeling, and model evaluation.

To address this limitation, self-supervised methods such as SimCLR, MoCo, BYOL, SimSiam, and Triplet Learning can learn meaningful representations from unlabeled images. These representations can be visualized in latent space, clustered, and used for pseudo-label generation, which is particularly useful in microscopic parasite image analysis where expert labels are scarce.

Finally, reliable evaluation is necessary to ensure model generalization through validation, testing, overfitting analysis, clustering metrics, and external evaluation. Therefore, the next chapter translates these theoretical concepts into the practical methodology adopted in this work, including dataset preparation, representation learning, clustering, pseudo-labeling, classification, and evaluation.

CHAPTER 3:

Methodology of the Proposed AI-Powered Biobank Framework for Gastrointestinal Parasite Image Analysis under Limited Annotation Conditions

3.1 Introduction

This chapter presents the methodological framework adopted in our study for gastrointestinal parasite image analysis within the proposed AI-powered biobank. Building on the theoretical foundations discussed in the previous chapter, it explains how self-supervised learning, clustering, pseudo-labeling, expert validation, and supervised classification were practically implemented.

We describe the complete experimental workflow, starting from the organization of unlabeled microscopic parasite images, followed by representation learning using SimCLR and BYOL, latent-space clustering using K-Means and Agglomerative Clustering, pseudo-label generation, and expert-based validation. It also presents the second experimental stage based on an eleven-class labeled dataset, where different supervision levels were evaluated, including 0%, 5%, 10%, 50%, and 100% labeled data.

3.2 General Methodological Overview

The proposed methodology was designed as a progressive pipeline for gastrointestinal parasite image analysis under limited annotation conditions. It combines self-supervised learning, clustering, pseudo-label generation, expert validation, and supervised classification to exploit both unlabeled and labeled microscopic image datasets.

The workflow was organized into two main phases. The first phase used an initial microscopic parasite image dataset organized into five groups without reliable diagnostic labels. Since it could not be directly used for supervised learning, SimCLR and BYOL were applied to learn visual representations without manual annotations. The extracted embeddings were then analyzed in the latent space and clustered using K-Means and Agglomerative Clustering, resulting in four combinations: SimCLR + K-Means, SimCLR + Agglomerative Clustering, BYOL + K-Means, and BYOL + Agglomerative Clustering.

The resulting clusters were used to generate provisional pseudo-labels. However, these labels were not considered final diagnostic annotations. Therefore, expert validation was included to interpret coherent clusters, remove inconsistent samples, and improve dataset reliability.

The second phase used a labeled gastrointestinal parasite dataset containing eleven classes. It was introduced to validate and strengthen the first phase by applying the same methodological logic under controlled supervision scenarios. The experiments included a 0% labeled scenario, where labels were hidden, semi-supervised scenarios using 5%, 10%, and 50% of the labels, and a fully supervised baseline using 100%.

Overall, the methodology moves from unlabeled data organization to controlled classification experiments with increasing label availability. Its final aim is to support an AI-powered digital biobank capable of storing, organizing, analyzing, and reusing gastrointestinal parasite image data for diagnostic support, research, and epidemiological monitoring.

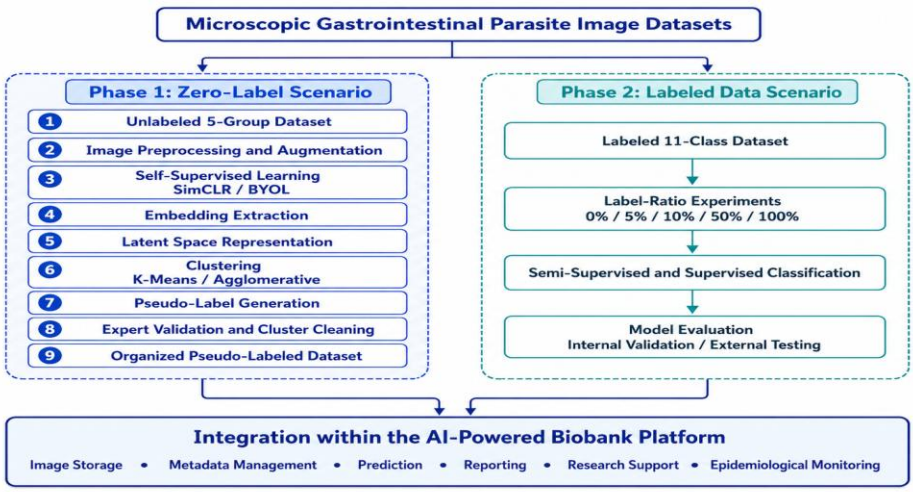


Figure 3.1: General Workflow of the Proposed Methodology

The first experimental phase was designed to address the lack of reliable diagnostic labels in the initial microscopic gastrointestinal parasite image dataset. Since the available images were organized into five non-validated groups, they could not be directly used to train a supervised classifier. Therefore, this phase adopted a zero-label strategy based on self-supervised learning.

In this phase, SimCLR and BYOL were used to learn meaningful visual representations from unlabeled microscopic images. The learned embeddings were then analyzed in the latent space and grouped using K-Means and Agglomerative Clustering. The resulting clusters were used to generate provisional pseudo-labels, which were later prepared for expert validation in order to improve their reliability.

Overall, this phase aimed to move from unlabeled image organization toward more reliable pseudo-labeled data that could support later classification experiments and expert-guided annotation.

3.3.1 Motivation and Problem Definition

The main motivation of this phase was the absence of expert-validated ground-truth labels in the initial dataset. Although the images were arranged into five groups, these groups were not considered confirmed diagnostic classes. Using such uncertain labels for supervised learning could introduce label noise and reduce the reliability of the trained model.

For this reason, self-supervised learning was selected as an appropriate strategy, because it allows the model to learn visual features directly from unlabeled images without depending on manual annotations. The problem addressed in this phase can therefore be defined as follows: given a collection of microscopic parasite images without reliable diagnostic labels, the objective is to learn visual embeddings that capture meaningful morphological information and organize the images into coherent visual groups.

This step is particularly important in gastrointestinal parasite image analysis, where expert annotation is difficult, time-consuming, and affected by visual similarity between parasite eggs, variation in image quality, and subtle morphological differences. Thus, learning representations without labels represents a necessary step toward building a scalable AI-assisted biobank system

3.3.2 Unlabeled Five-Group Dataset Description

The dataset used in the first experimental phase consisted of microscopic gastrointestinal parasite images provided as five separate folders: group0, group1, group2, group3, and group4. Although the images were arranged into five groups, these folders were treated only as organizational groups and not as expert-validated diagnostic classes. Therefore, the dataset was considered unlabeled for the purpose of model training.

The dataset contained a total of **8,695** microscopic images distributed across the five groups. Group3 contained the highest number of images with 2,297 samples, followed by group0 with 1,982 images, group1 with 1,867 images, group4 with 1,540 images, and group2 with 1,009 images.

	Group	Number of images
0	group0	1982
1	group1	1867
2	group2	1009
3	group3	2297
4	group4	1540
5	Total	8695

Table 3.1:Distribution of images in the unlabeled five-group dataset

3.3.3 Image Preprocessing and Augmentation

Before feature learning, the microscopic images were prepared to ensure compatibility with the deep learning pipeline. The images were converted into a standard RGB format, resized to a fixed input size, transformed into tensors, and normalized before being passed to the ResNet50-based encoder.

During the self-supervised learning stage, data augmentation was used to generate different views of the same image. This step is important for methods such as BYOL and SimCLR, because it helps the model learn stable visual representations from unlabeled data. The applied transformations aimed to simulate common variations in microscopic imaging, such as changes in orientation, scale, brightness, contrast, and image appearance.

The purpose of preprocessing and augmentation was to make the images suitable for training and to encourage the model to focus on meaningful morphological features rather than irrelevant visual variations. These learned features were later used for embedding extraction, latent-space analysis, clustering, and pseudo-label generation.

3.3.4 Supervised Representation Learning

In the first experimental phase, self-supervised representation learning was used to extract meaningful visual features from unlabeled microscopic parasite images. The aim was not to perform direct diagnosis, but to generate robust embeddings that could support the following methodological steps:

- Latent-space analysis**:** to examine how images are represented after feature extraction.
- Dimensionality reduction**:** to reduce embedding complexity using PCA.
- Clustering**:** to organize similar images using K-Means and Agglomerative Clustering.
- Visualization**:** to analyze the distribution of embeddings using t-SNE.
- Pseudo-label generation**:** to assign provisional labels based on cluster membership.
- Expert-guided review**:** to validate coherent clusters and remove inconsistent samples.

Two self-supervised learning methods were selected:

- SimCLR**:** used as a contrastive learning approach.
- BYOL**:** used as a non-contrastive learning approach.

Using both methods allowed a focused comparison between two major self-supervised learning strategies under zero-label conditions. Both methods were implemented using a ResNet50-based feature extraction pipeline, and the extracted embeddings were used as the basis for the next stages of the methodology.

3.3.5 Embedding Extraction and Latent Space Projection

After self-supervised feature learning, the trained encoders were used to extract embedding vectors from the microscopic parasite images. An embedding is a numerical representation of an image that summarizes its most important visual features. Instead of working directly with raw pixels, each image was transformed into a compact feature vector that represents morphological information such as shape, texture, boundaries, and visual similarity.

For the BYOL pipeline, the online ResNet50 encoder was used after training to extract embeddings from all images. During this stage, random augmentations were not applied, in order to obtain stable representations for each image. For the SimCLR pipeline, the previously extracted ResNet50 embeddings were used for the downstream analysis.

The extracted embeddings were then used to project the dataset into a latent representation space. In this space, visually similar images are expected to be located closer to each other, while visually different images are expected to be farther apart. This latent-space representation made it possible to analyze the internal organization of the dataset and prepare the embeddings for clustering. Since the original embedding vectors were high-dimensional, dimensionality reduction was later applied using PCA before clustering. This step reduced the complexity of the feature space while preserving the most relevant information. The resulting latent representations were used as input for K-Means and Agglomerative Clustering in order to organize the unlabeled images into visually coherent groups.

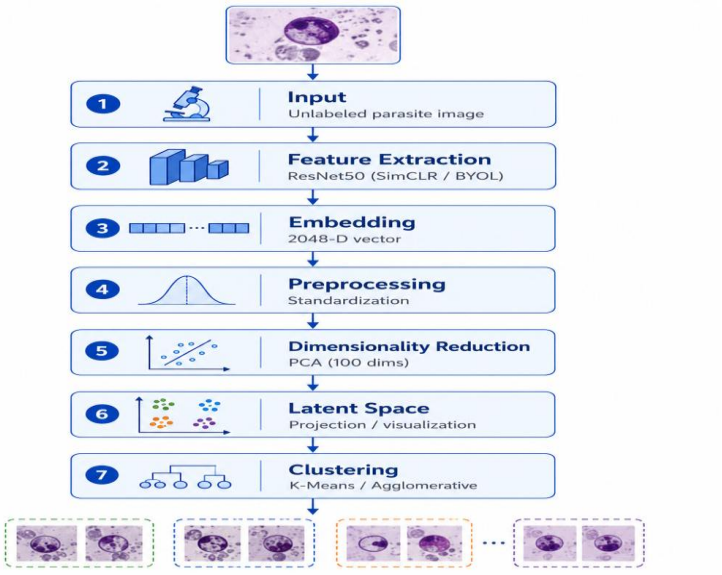


Figure 3.2: Embedding extraction and latent-space projection process

3.3.6 Clustering Strategy

After embedding extraction, clustering was applied to group microscopic parasite images according to visual similarity within the learned latent space. Since the dataset did not contain expert-validated labels, clustering was used as an exploratory step to identify hidden visual structures, support pseudo-label generation, and prepare representative images for expert review. Clustering is commonly used to identify homogeneous groups and compare different grouping strategies in applied research [47]. Before clustering, the original 2048-dimensional embeddings were reduced to 100 dimensions using Principal Component Analysis (PCA), in order to reduce computational complexity while preserving the main information contained in the feature space. In the SimCLR experiment, PCA preserved 87.69% of the original variance. Two clustering methods were then applied: K-Means and Agglomerative Clustering. K-Means was used to form compact groups around cluster centroids and to facilitate the selection of representative images, while Agglomerative Clustering was used to explore more flexible hierarchical structures in the latent space.

The clustering quality was evaluated using three internal validation metrics: Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, which are commonly used to assess cluster separation, compactness, and overall grouping quality [48]. These metrics were used to compare the different clustering strategies applied to SimCLR and BYOL embeddings before selecting the most suitable approach for representative image selection and expert review.



Figure 3.3: SimCLR-ResNet50 clustering results summary

3.3.6.1 SimCLR with K-Means

Before applying the clustering methods, microscopic images were first processed using a SimCLR-based representation learning pipeline. The model architecture was based on a ResNet50 backbone pretrained on ImageNet. Each input image was resized to 224 × 224 pixels and transformed into two augmented views using random cropping, flipping, rotation, color jittering, and Gaussian blur. These two views were passed through the same shared ResNet50 network to extract 2048-dimensional feature vectors. During contrastive training, these features were passed through a projection head composed of two linear layers, reducing the representation from 2048 to 512 and then to 128 dimensions. The model was optimized using the NT-Xent contrastive loss. After training, the projection head was removed, and the 2048-dimensional embeddings extracted from the ResNet50 backbone were used for PCA reduction and clustering using K-Means and Agglomerative Clustering.

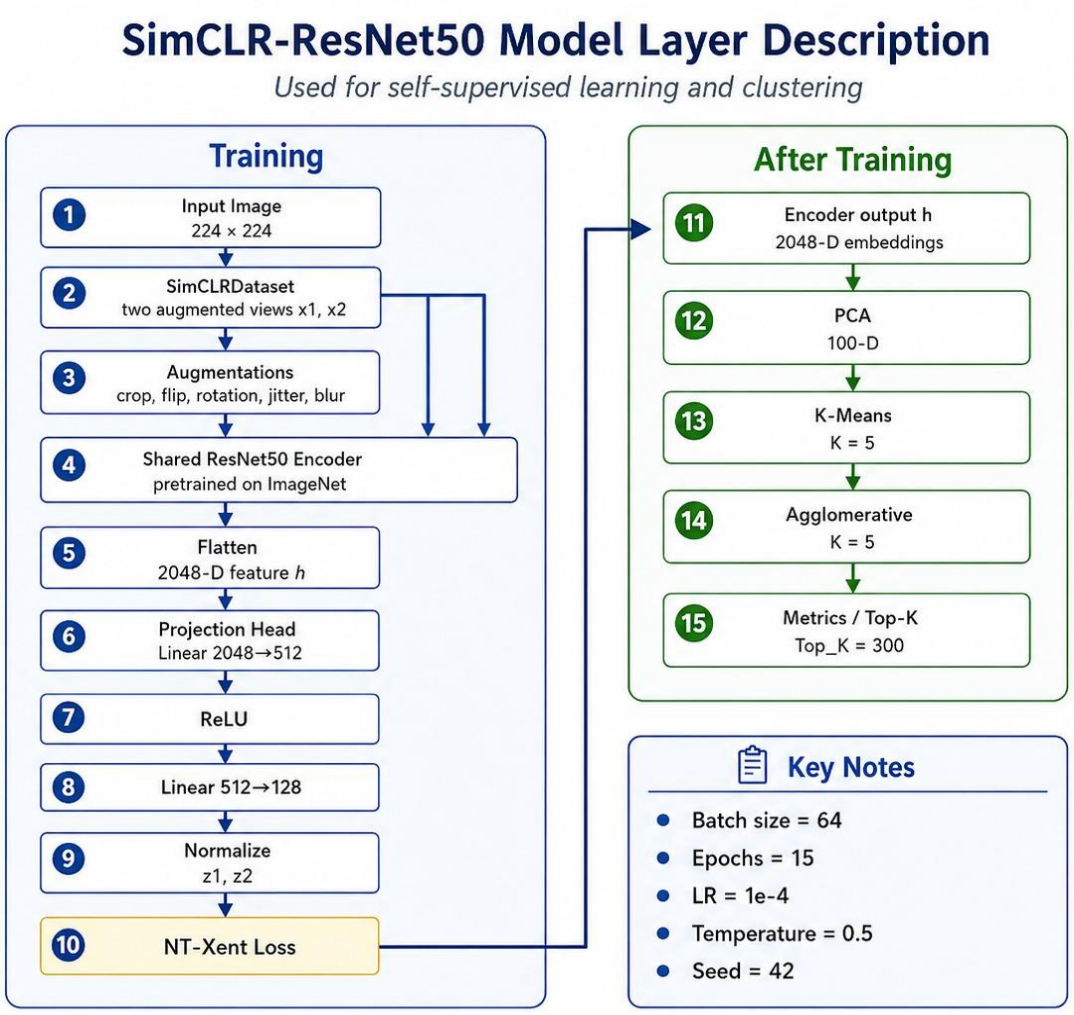


Figure 3.4: SimCLR-ResNet50 architecture used for self-supervised representation learning

This section presents the implementation of K-Means clustering on the PCA-reduced embeddings generated by the SimCLR model. In this experiment, K-Means was applied with K = 5 clusters, corresponding to the initial five-group structure of the dataset. The purpose of this step was to examine whether the SimCLR-based representations could form compact and visually coherent groups in the latent space. Since K-Means assigns samples according to their proximity to cluster centroids, it was considered suitable for identifying representative visual patterns within the unlabeled microscopic parasite images. The K-Means experiment achieved:

- **Silhouette Score:** 0.0952
- **Davies-Bouldin Score:** 2.7118
- **Calinski-Harabasz Score:** 510.035

These results indicate that K-Means was able to produce meaningful visual groupings from the SimCLR embeddings. While some overlap was still observed between clusters, especially in the t-SNE visualization, this can be explained by the high visual similarity between microscopic parasite images. Despite this overlap, several dense and structured regions were visible, suggesting that the SimCLR model captured useful visual patterns.

Based on these clustering results, K-Means was used to select 300 representative images from each cluster, resulting in a total of 1,500 images for expert review.

3.3.6.2 SimCLR with Agglomerative Clustering

This section presents the implementation of Agglomerative Clustering was applied to the same PCA-reduced SimCLR embeddings using **K = 5** clusters. Unlike K-Means, this method does not depend on fixed cluster centers. Instead, it builds clusters gradually by merging similar samples or groups. The Agglomerative experiment achieved:

- **Silhouette Score:** 0.1190
- **Davies-Bouldin Score:** 2.9345
- **Calinski-Harabasz Score:** 395.924

The higher Silhouette Score indicates that Agglomerative Clustering achieved better separation between clusters. However, its higher Davies-Bouldin Score and lower Calinski-Harabasz Score suggest that its clusters were less compact than those produced by K-Means.

In the t-SNE visualization, Agglomerative Clustering shows more flexible and natural group structures. This means that it was useful for understanding the organization of the learned latent space, especially when the visual groups were not perfectly spherical.

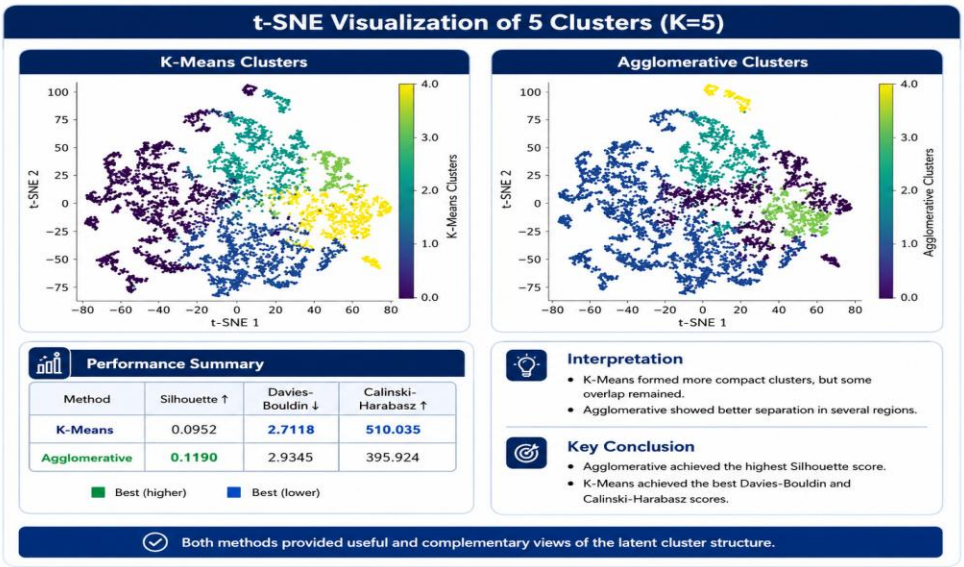


Figure 3.5: t-SNE visualization of SimCLR clusters using K-Means and Agglomerative Clustering

3.3.6.3 BYOL with K-Means

Before applying the clustering methods, microscopic images were first processed using a BYOL-based self-supervised representation learning pipeline. The model architecture was based on a ResNet50 backbone pretrained on ImageNet. Each input image was resized to 224 × 224 pixels and transformed into two augmented views using random cropping, horizontal and vertical flipping, rotation, color jittering, grayscale transformation, Gaussian blur, and normalization. Unlike SimCLR, BYOL uses two networks: an online network and a target

network. The online network contains a ResNet50 encoder, a projection head, and a predictor, while the target network contains a ResNet50 encoder and a projection head.

The target network was updated using an exponential moving average of the online network parameters. During training, the model learned by minimizing the regression loss between the online predictions and the target projections.

After training, only the online ResNet50 encoder was used to extract 2048-dimensional embeddings from the microscopic images.

These embeddings were then standardized, reduced using PCA to 100 dimensions, and clustered using K-Means and Agglomerative Clustering.

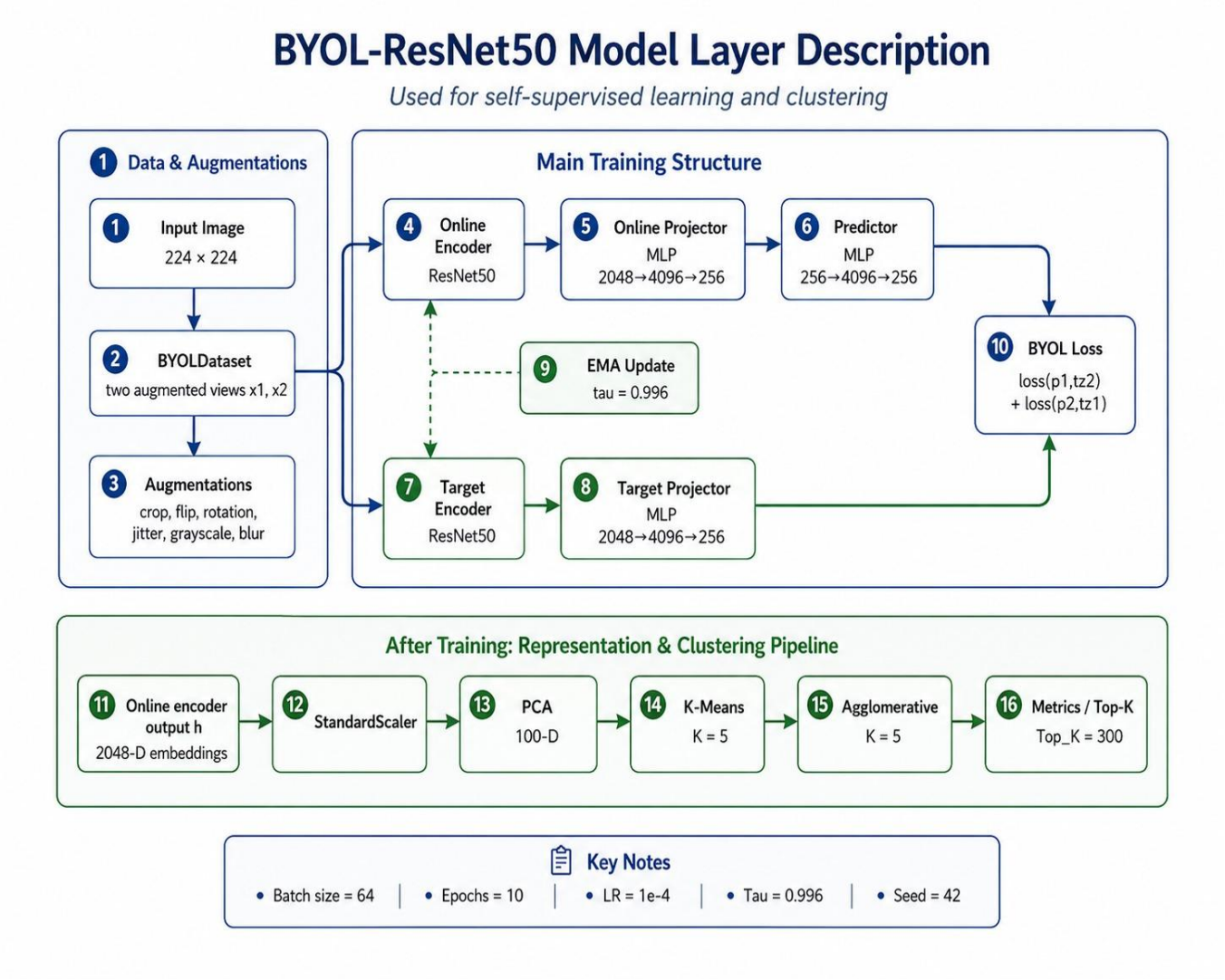


Figure 3.6: BYOL-ResNet50 architecture used for self-supervised representation learning

K-Means was applied to the PCA-reduced BYOL-ResNet50 embeddings using K = 5 clusters. The BYOL encoder produced 2048-dimensional embeddings, which were reduced to 100 principal components before

clustering. In this experiment, PCA preserved 63.80% of the original variance, indicating that the reduced representation still retained a relevant part of the learned visual information.

K-Means was used because it groups samples around cluster centers, making it suitable for forming compact visual groups and selecting representative images for expert review.

Recent studies emphasize that clustering can be useful for identifying homogeneous groups and comparing different grouping strategies in applied research (Gao et al., 2023). The BYOL with K-Means experiment produced the following results:

- **Silhouette Score:** 0.1050
- **Davies-Bouldin Score:** 2.7391
- **Calinski-Harabasz Score:** 488.694

These results indicate that K-Means produced relatively compact and coherent clusters. Compared with Agglomerative Clustering, K-Means achieved a higher Silhouette Score, a lower Davies-Bouldin Score, and a higher Calinski-Harabasz Score. This suggests that, for BYOL embeddings, K-Means provided better cluster separation and compactness.

In addition, K-Means was useful for preparing the expert-review dataset. The 300 images closest to each cluster center were selected, resulting in 1,500 representative images. These images were then used for expert inspection and cluster interpretation.

3.3.6.4 BYOL with Agglomerative Clustering

Following the previous BYOL with K-Means experiment, the same PCA-reduced BYOL embeddings were used to evaluate a second clustering strategy: Agglomerative Clustering**.** This experiment was conducted with **K = 5 clusters**, in order to compare how a hierarchical clustering method performs on the same learned representations.

Unlike K-Means, Agglomerative Clustering does not rely on fixed cluster centers. Instead, it follows a hierarchical strategy by progressively merging visually similar samples. This makes it useful for exploring more flexible structures in the latent space and for comparing whether BYOL embeddings are better organized using centroid-based or hierarchical clustering.

The BYOL with Agglomerative Clustering experiment produced the following results:

- **Silhouette Score:** 0.0806
- **Davies-Bouldin Score:** 3.0220
- **Calinski-Harabasz Score:** 397.172

Compared with K-Means, these results indicate weaker separation and lower compactness. The lower Silhouette Score suggests that the clusters were less clearly separated, while the higher Davies-Bouldin Score indicates more overlap between clusters. The lower Calinski-Harabasz Score also suggests that the internal cluster structure was less compact than the one obtained with K-Means.

The t-SNE visualization showed that Agglomerative Clustering captured a different organization of the learned BYOL embeddings. Some clusters appeared to follow more flexible and natural structures, while others remained

partially overlapping. This confirms that Agglomerative Clustering can provide complementary insight into the latent space, even when K-Means gives stronger quantitative results.

Overall, the BYOL clustering results show that K-Means was more effective for this embedding space. It produced better scores across the three internal metrics and was more suitable for selecting representative images near cluster centers. Agglomerative Clustering, however, remained useful for observing the hierarchical and non-spherical organization of the learned representations.

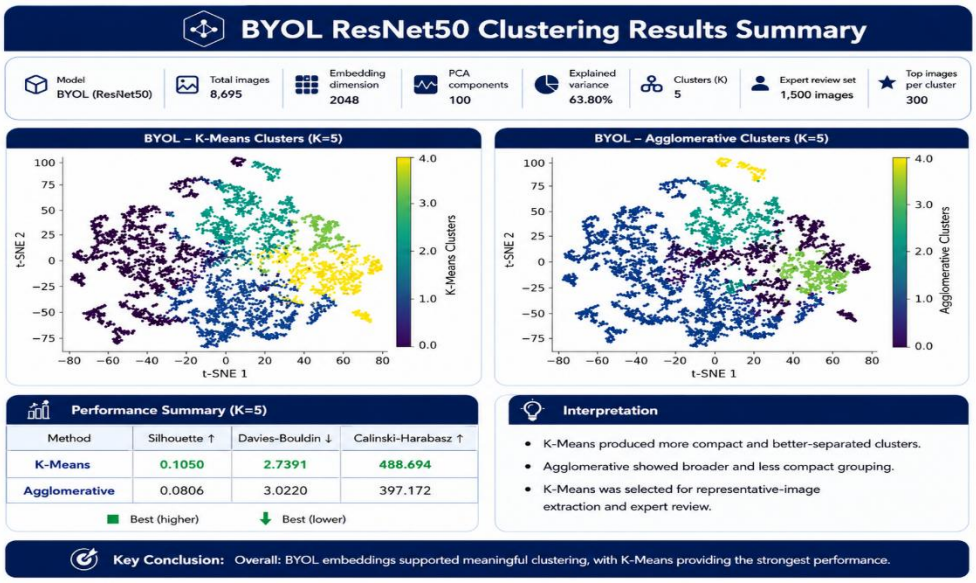


Figure 3.7: BYOL-ResNet50 clustering results and t-SNE visualization

3.3.6.5 Overall Comparison between SimCLR and BYOL

After evaluating the four clustering combinations, an overall comparison was conducted between SimCLR and BYOL in order to identify which self-supervised representation learning method produced more useful embeddings for the unlabeled parasite image dataset.

The comparison was based on PCA explained variance, clustering metrics, and t-SNE visualization. SimCLR preserved a higher amount of variance after PCA, with 87.69% compared to 63.80% for BYOL. This suggests that the SimCLR embeddings retained richer information after dimensionality reduction.

From the clustering perspective, SimCLR achieved the best overall results. Its best Silhouette Score was 0.1190, obtained with Agglomerative Clustering, while BYOL achieved a best Silhouette Score of 0.1050 using K-Means. SimCLR also obtained the best Davies-Bouldin Score of 2.7118 and the best Calinski-Harabasz Score of 510.035, both indicating better cluster compactness and separation compared with BYOL.

The t-SNE visual comparison also supported this conclusion. SimCLR showed a more organized latent space with clearer cluster structures, while BYOL produced useful but more overlapping representations. This indicates that both models were able to learn visual information from unlabeled parasite images, but SimCLR produced more discriminative and transferable embeddings for clustering.

Overall, SimCLR was considered the best-performing self-supervised method in this phase. It provided richer representations, better clustering metrics, and clearer latent-space organization. BYOL remained useful as an alternative non-contrastive method, but its embeddings were less structured compared with SimCLR in this experiment.

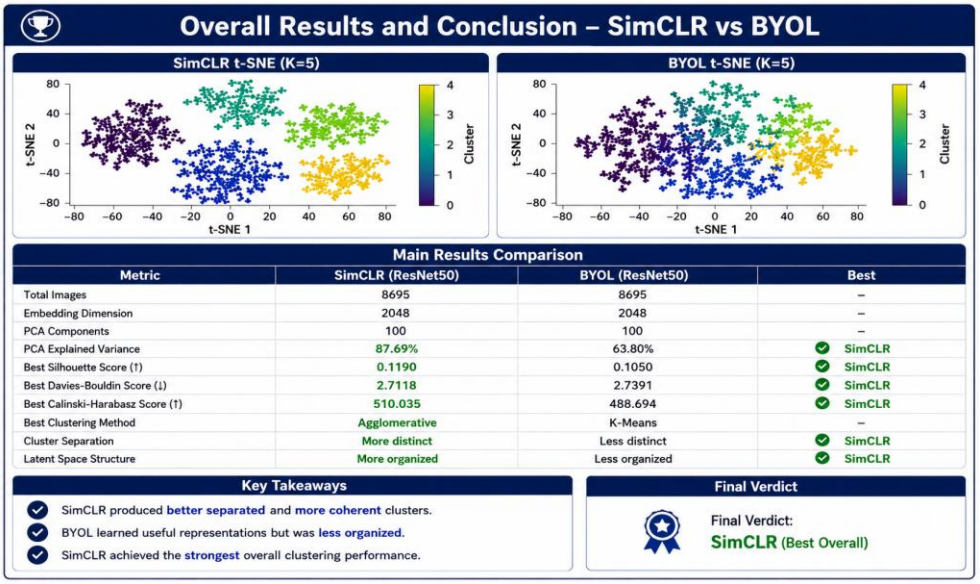


Figure 3.8: Overall comparison between SimCLR and BYOL clustering results

3.3.7 Latent Space Visualization

Latent space visualization was used to qualitatively inspect the organization of the learned embeddings after self-supervised representation learning and clustering. Since the extracted embeddings were high-dimensional, t-SNE was used to project them into a two-dimensional space and make the cluster structure easier to interpret.

The objective of this visualization was not to provide final diagnostic evidence, but to support the analysis of the learned representation space. It helped observe whether visually similar microscopic images were located close to each other and whether the generated clusters showed meaningful organization.

The t-SNE visualizations presented in the previous comparative figures showed that the learned embeddings were not randomly distributed. Some regions appeared more structured and compact, especially for the SimCLR-based representations. However, partial overlap between clusters was also observed, which is expected in microscopic parasite images due to morphological similarity and visual ambiguity between samples. Overall, latent space visualization supported the clustering analysis by showing that the self-supervised models were able to learn meaningful visual structures from unlabeled images. These visual observations helped justify the next step, where clusters were used to generate provisional pseudo-labels for expert-guided review.

3.3.8 Pseudo-Label Generation from Clusters

After clustering the learned embeddings, the cluster assignments were used to generate provisional pseudo-labels. Since the initial five-group dataset did not contain reliable expert-validated diagnostic labels, these pseudo-labels were not treated as final medical annotations, but as temporary labels derived from the learned latent-space structure.

In this phase, the SimCLR + K-Means pipeline was selected for pseudo-label preparation because it produced better-organized embeddings and provided cluster centroids. For each image, the assigned K-Means cluster was used as its pseudo-label, from cluster 0 to cluster 4. This created an initial structured organization of the unlabeled dataset.

To support expert review, the 300 images nearest to each K-Means centroid were selected, resulting in 1,500 representative images. This step aimed to reduce the manual annotation burden by allowing the expert to inspect the most representative samples from each cluster, assess their visual coherence, and support the transformation of provisional pseudo-labels into more meaningful annotations.

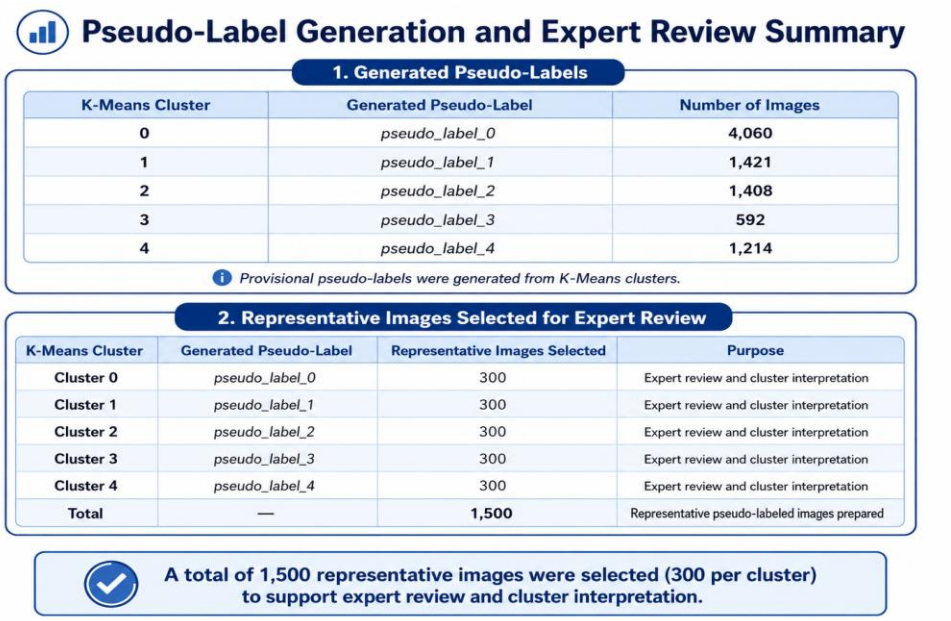


Figure 3.9: K-Means-Based Pseudo-Label Generation and Expert Review Summary

After pseudo-label generation, an expert-based review step was prepared to improve the reliability of the automatically organized dataset. Since the pseudo-labels were derived from clustering rather than direct medical diagnosis, they required expert interpretation before being considered meaningful annotations.

The representative images selected in the previous step were prepared for review by Dr. Mohammed Mebarek Bia, from the Department of Parasitology and Medical Research Institute, Chungbuk National University, South Korea. The expert review aimed to assess cluster coherence, support possible class naming, and identify visually inconsistent samples that should be removed or marked for further inspection.

This step created a bridge between automatic clustering and medically meaningful annotation, making the validation process more structured and efficient.

3.4 Validation Across Five Groups and Eleven Parasite Classes

The second experimental phase was introduced to validate the zero-label strategy used in the first phase. In Phase 1, the initial dataset was provided in five groups but did not contain reliable expert-validated diagnostic labels. Therefore, although self-supervised learning and clustering were used to organize the images and generate provisional pseudo-labels, it was not possible to fully verify the quality of the learned representations using supervised evaluation.

For this reason, a labeled gastrointestinal parasite dataset containing eleven classes was introduced as a validation resource. The purpose was not to replace the initial five-group dataset, but to test the same methodological logic under controlled conditions where true labels were available. By hiding the labels during training, the eleven-class dataset was first used to simulate a zero-label scenario similar to the first phase. The hidden labels were then used only for evaluation, allowing the study to measure whether the proposed approach could learn meaningful visual representations without direct supervision.

After this zero-label simulation, labeled data were progressively introduced using 5%, 10%, and 50% of the available annotations. Finally, a 100% labeled setting was used as a fully supervised baseline. This progressive design made it possible to evaluate the transition from zero-label learning to limited-label semi-supervised learning and fully supervised classification, while also showing how model performance changes as more expert annotations become available.

3.4.1 Eleven-Class Dataset Description

The second dataset used in this phase was a labeled gastrointestinal parasite image dataset containing eleven diagnostic classes. Unlike the initial five-group dataset, this dataset provided ground-truth labels for each image, which made it suitable for validating the proposed zero-label strategy under controlled conditions.

The dataset contained a total of 11,000 microscopic images distributed equally across eleven parasite classes. Each class included 1,000 images, which means that the dataset was balanced. This balanced distribution was important because it reduced the risk of class imbalance during training and evaluation, and allowed fair comparison between different supervision scenarios.

Class Distribution

Index	Label	Number of Images
0	Ascaris lumbricoides	1000
1	Capillaria philippinensis	1000
2	Enterobius vermicularis	1000
3	Fasciolopsis buski	1000
4	Hookworm egg	1000
5	Hymenolepis diminuta	1000
6	Hymenolepis nana	1000
7	Opisthorchis viverrine	1000
8	Paragonimus spp	1000
9	Taenia spp. egg	1000
10	Trichuris trichiura	1000
11	**Total**	**11000**
	Number of detected classes:	**11**
	Total labeled images:	**11000**

Table 3.2: Class distribution of the labeled eleven-class parasite

3.4.2 Experimental Design with Different Label Ratios

The experimental design of Phase 2 was based on controlling the amount of label information available during training. Since the eleven-class gastrointestinal parasite dataset contained ground-truth labels, different supervision scenarios could be simulated by hiding or progressively introducing labels.

The first scenario was the zero-label setting, where 0% of the labels were used during training. In this case, the dataset was treated as unlabeled, and the SimCLR-ResNet50 + PCA100 + K-Means pipeline was applied. The true labels were hidden during training and used only after clustering for evaluation.

After the zero-label scenario, limited-label experiments were planned using 5%, 10%, and 50% of the labeled data. These settings represented semi-supervised learning scenarios, where only a small portion of expert annotations was available. Finally, the 100% labeled setting was used as a fully supervised baseline.

This experimental design allowed the study to evaluate how the proposed approach behaves when no labels are available, and how performance changes when more labeled data are progressively introduced.

3.4.2.1 Zero-Label Scenario

Following the self-supervised representation learning and clustering strategy established in the previous phase, a zero-label experiment was conducted on the eleven-class gastrointestinal parasite dataset to validate the same methodological logic under controlled conditions.

- The eleven-class dataset was treated as unlabeled during training; ground-truth labels were hidden from the learning process and used only after clustering for evaluation.
- SimCLR with a ResNet50 encoder was used to extract 2048-dimensional embeddings from the microscopic images.
- The extracted embeddings were standardized and reduced to 100 principal components using PCA, which preserved 92.09% of the original variance.
- K-Means clustering was then applied with **K = 11**, corresponding to the eleven hidden parasite classes, and t-SNE was used to visualize the learned latent-space structure.
- The results showed that the zero-label pipeline could reveal meaningful visual structures, although some overlap between clusters remained. This supports the use of self-supervised learning and clustering as an exploratory pseudo-labeling strategy for the initial five-group dataset, while also motivating the later experiments using 5%, 10%, 50%, and 100% labeled data.

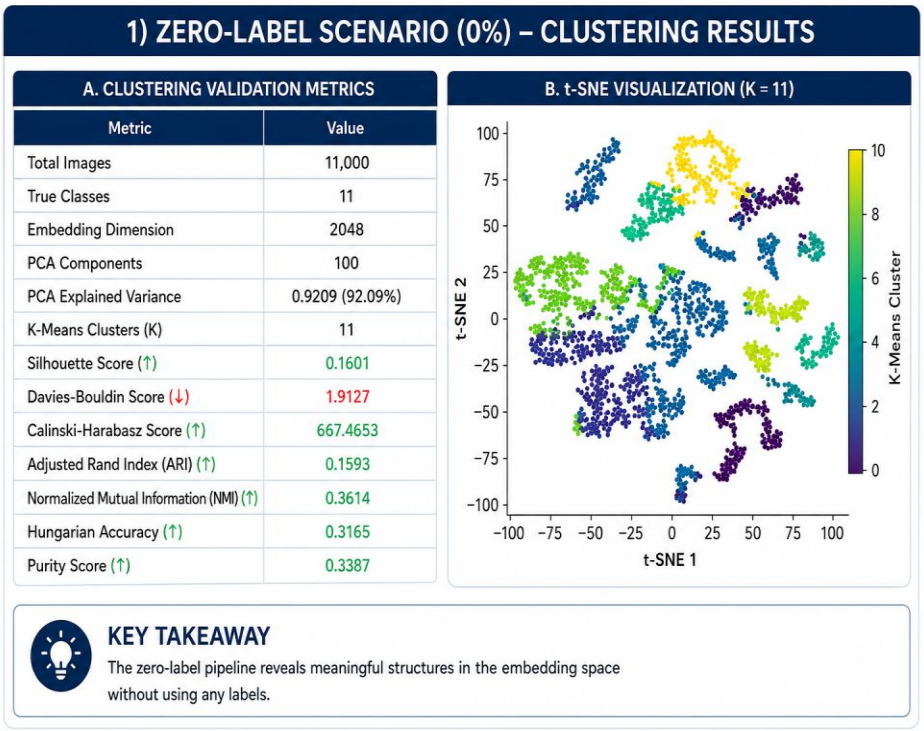


Figure 3.11: Zero-label SimCLR-ResNet50 clustering results using PCA100 and K-Means

3.4.2.2 Limited-Label Scenarios: 5%, 10%, and 50%

To ensure a fair comparison between the different limited-label scenarios, the same ResNet50 classifier architecture was used for all experiments. The network configuration remained unchanged across the 5%, 10%, and 50% settings, with the only difference being the number of labeled training samples available for supervised learning.

The internal dataset was divided into a training pool of 8,800 images and a validation set of 2,200 images. For each scenario, only a fraction of the training pool was used with labels: 440 images for the 5% scenario, 880 images for the 10% scenario, and 4,400 images for the 50% scenario.

- **Description of the ResNet50 Classifier Architecture**

The same ResNet50 classifier architecture was used for the three limited-label scenarios. The architecture remained unchanged across the 5%, 10%, and 50% experiments; only the number of labeled training images was varied.

The classifier was based on a ResNet50 model pretrained on ImageNet. The input microscopic images were first resized to 224 × 224 pixels and normalized using ImageNet normalization values. During training, data augmentation techniques were applied to improve the robustness of the model, including random horizontal flipping, vertical flipping, rotation, and color jittering.

After preprocessing, the image was passed through the ResNet50 backbone, which acts as a deep feature extractor. The first convolutional layers extract low-level visual features such as edges, textures, and simple patterns. These layers are followed by Batch Normalization and ReLU activation functions. Batch Normalization helps stabilize the training process, while the ReLU activation function introduces non-linearity and allows the model to learn more complex visual representations.

The main part of ResNet50 is composed of residual blocks. These blocks contain convolutional layers, Batch Normalization, and ReLU activations, together with skip connections. The skip connections allow information to pass directly from earlier layers to deeper layers, which helps reduce the vanishing gradient problem and makes it possible to train deeper networks effectively.

At the end of the network, the extracted feature maps are passed through a Global Average Pooling layer, which converts them into a compact feature representation. The original final fully connected layer of ResNet50 was then replaced with a new classification layer adapted to the 11 parasite egg classes. Finally, a Softmax function was used to generate class probabilities, and the class with the highest probability was selected as the predicted parasite class, together with its confidence score.

The architecture of the classifier can be summarized as follows:

- **Input:** microscopic parasite egg image.
- **Preprocessing:** resizing to 224 × 224 pixels and normalization using ImageNet values.
- **Data augmentation:** random horizontal flip, vertical flip, rotation, and color jittering during training.
- **Feature extraction:** ResNet50 backbone pretrained on ImageNet.
- **Main layers:** convolutional layers, Batch Normalization, ReLU activation, residual blocks, and skip connections.
- **Pooling:** Global Average Pooling to obtain a compact feature representation.
- Classifier adaptation: replacement of the original fully connected layer with a new layer adapted to 11 parasite egg classes.
- **Output:** predicted parasite class and confidence score.

The classifier was based on a ResNet50 model pretrained on ImageNet. Input microscopic images were resized to 224 × 224 pixels and normalized using ImageNet normalization values. During training, several data augmentation techniques were applied to improve model robustness, including random horizontal flipping, vertical flipping, rotation, and color jittering.

After preprocessing, each image was passed through the ResNet50 backbone, which acted as a deep feature extractor. The network consists of convolutional layers, Batch Normalization, ReLU activation functions, residual blocks, and skip connections. These components enable the extraction of both low-level visual features, such as edges and textures, and more complex high-level representations.

At the end of the network, the extracted feature maps were processed through a Global Average Pooling layer to obtain a compact 2048-dimensional feature vector. The original fully connected layer of ResNet50 was replaced with a new classification layer adapted to the 11 parasite egg classes. Finally, a Softmax function was applied to generate class probabilities, and the class with the highest probability was selected as the predicted parasite class along with its confidence score.

The model was trained for five epochs using a weighted Cross-Entropy loss function and the AdamW optimizer. The best-performing model was selected based on the highest Macro F1-score obtained on the internal validation set.

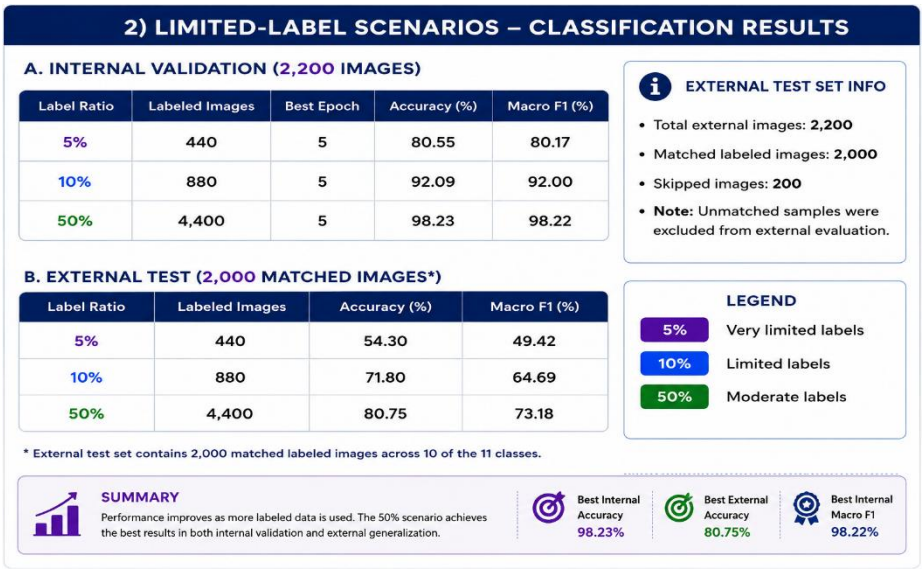


Figure 3.12: Performance results under limited-label scenarios

3.4.2.3 Fully Supervised Baseline: 100% Labels

The 100% labeled scenario was used as the fully supervised baseline. In this setting, all 8,800 images from the training pool were used with their ground-truth labels. This experiment represented the upper reference point for comparison with the zero-label and limited-label scenarios.

The ResNet50 classifier was trained for five epochs, and the best validation result was obtained at epoch 4. The fully supervised model achieved an internal validation accuracy of 99.05% and an internal macro F1-score of 99.04%. On the external test set, it achieved an accuracy of 82.60% and a macro F1-score of 75.52%.

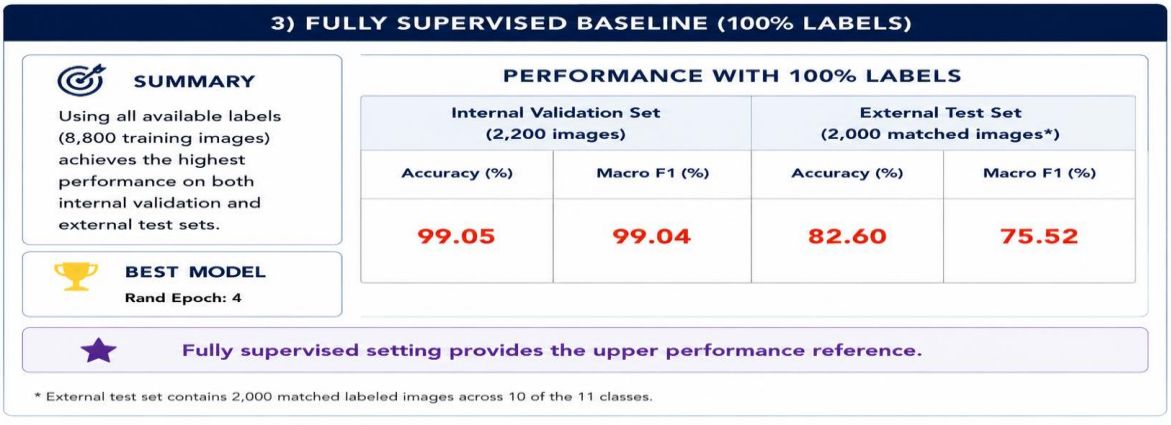


Figure 3.13: Fully supervised baseline performance using 100% labeled data

"The fully supervised baseline achieved the highest overall performance. This confirms that complete label availability allows the model to learn highly discriminative class-specific features. However, the improvement from 50% to 100% labels was smaller than the improvement observed from 5% to 10% labels. This suggests that a moderate amount of labeled data can already provide strong performance, while full supervision mainly refines the model further."

3.4.2.4 Comparative Analysis of Label Ratios

After conducting all the experiments adopted in this study, from the zero-label scenario to the limited-label and fully supervised scenarios, a comparative analysis was performed to examine the effect of label availability on model performance and generalization ability. The comparison between the different label ratios shows a progressive improvement in classification performance as the amount of labeled data increased. The most noticeable improvement was observed between the 5% and 10% scenarios, especially on the external test set, indicating that adding a relatively small number of expert-labeled images can significantly enhance the model's ability to learn discriminative parasite features. The 50% and 100% settings achieved the highest overall performance, confirming the positive effect of label availability on model accuracy and stability. However, the external results remained lower than the internal validation results across all scenarios, suggesting that the external dataset was more challenging due to possible differences in image quality, acquisition conditions, or data distribution. Overall, this comparison confirms the relevance of the progressive learning strategy adopted in this work, moving from zero-label representation learning to limited-label scenarios and finally to fully supervised classification.

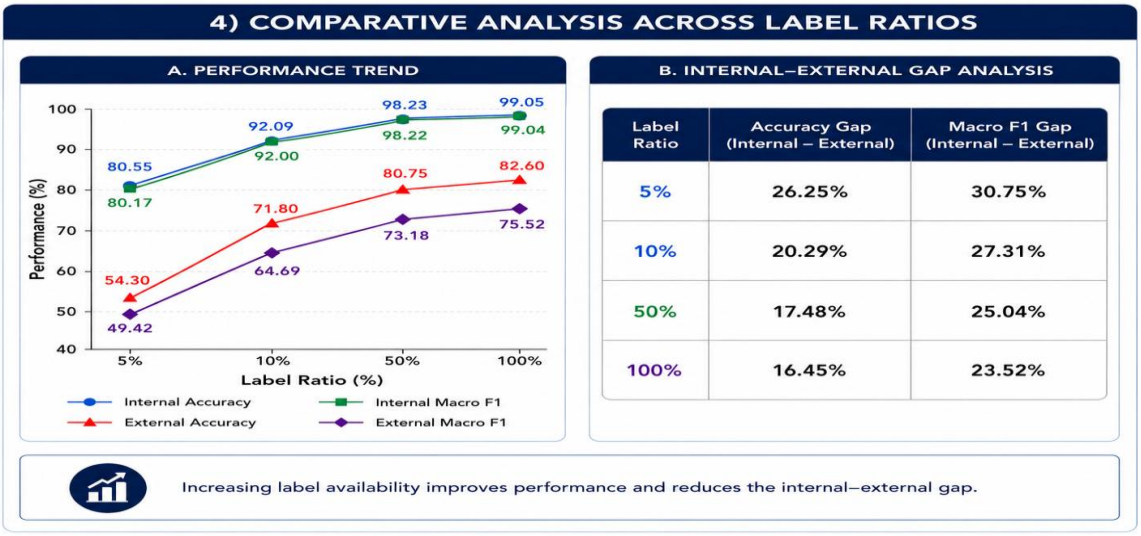


Figure 3.14: Performance trend and internal–external gap across label ratios

3.5 Evaluation Methodology

After defining the experimental scenarios and label-ratio settings, an evaluation methodology was established to measure the effectiveness of each stage of the proposed approach. The evaluation methodology was designed to assess both clustering quality and classification performance. For zero-label experiments, clustering metrics were used to evaluate how well the learned embeddings formed meaningful groups in the latent space. In the eleven-class dataset, labels were hidden during training and used only after clustering to compare the generated groups with the true diagnostic classes. For limited-label and fully supervised experiments, classification metrics were used to evaluate the ResNet50 classifier. Performance was measured on both an internal validation set and an external test set, in order to assess accuracy within the same data distribution and generalization to independent images.

3.5.1 Clustering Evaluation Metrics

Following the zero-label experiments, clustering metrics were used to evaluate whether the learned embeddings formed meaningful visual groups in the latent space. The internal quality of the clusters was assessed using Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, which measure cluster separation and compactness and are commonly used to compare clustering methods such as K-Means and hierarchical clustering [48].

When hidden true labels were available, additional agreement metrics were used only for evaluation: Adjusted Rand Index, Normalized Mutual Information, Hungarian Accuracy, and Purity Score. These metrics helped estimate how well the generated clusters corresponded to the real parasite classes.

3.5.2 Classification Evaluation Metrics

After evaluating the clustering stage, classification metrics were used for the limited-label and fully supervised experiments. A ResNet50 classifier was trained using different label ratios: 5%, 10%, 50%, and 100%.

The main metrics were accuracy and macro F1-score, while precision, recall, and F1-score were reported per class to analyze class-level performance. These metrics are commonly used for supervised machine learning evaluation, including multi-class classification tasks.

3.5.3 Internal Validation

To evaluate performance under the same data distribution, internal validation was performed on the labeled eleven-class gastrointestinal parasite dataset. The dataset was divided into 8,800 training images and 2,200 validation images.

For the limited-label scenarios, only part of the training set was labeled: 440 images for 5%, 880 images for 10%, and 4,400 images for 50%. The 100% scenario used all 8,800 labeled training images as the fully supervised baseline. The best model was selected according to the validation macro F1-score.

3.5.4 External Testing

After internal validation, external testing was used to assess model generalization on an independent dataset. The external set contained 2,200 images, but only 2,000 matched labeled samples were used for evaluation, while 200 images were skipped due to unmatched labels.

The external evaluation covered 10 of the 11 parasite classes, since no matched samples were available for Ascaris lumbricoides. Accuracy and macro F1-score were used again, and the external results were compared with internal validation results to measure the generalization gap caused by differences in image quality, acquisition conditions, or data distribution.

3.6 Return to the Initial Five-Group Dataset for Expert-Based Cluster Validation

After evaluating the zero-label strategy using the labeled eleven-class dataset, the methodology returned to the initial five-group dataset in order to apply the selected SimCLR + K-Means pipeline to the original unlabeled data. This step aimed to generate cluster-based groupings from the real target dataset and prepare representative microscopic images for expert-based interpretation.

The return to the five-group dataset was essential because this dataset represented the main annotation problem addressed in this study. Unlike the eleven-class dataset, which was used to assess the behavior of the proposed

zero-label strategy under known class conditions, the five-group dataset did not provide expert-validated diagnostic labels. Therefore, expert review was required to evaluate the biological relevance of the generated clusters, identify ambiguous or mixed-species images, and support the refinement of provisional pseudo-labels into more meaningful annotations.

3.6.1 Experimental Context

To evaluate the biological relevance of the unsupervised clustering results, an expert validation step was performed on a selected subset of microscopy images in collaboration with:

- Dr BIA Mohammed Mebarek affiliation: Parasite Research Center, Parasite Resource Bank Department of Parasitology, College of Medicine, Chungbuk National University
- Dr. BARAKAELI Abdieli Ndossi, affiliation: head of parasitology department on Tanzania Wildlife Research Institute.
- Prof. SEONGJUN Choe, affiliation: Parasite Research Center, Parasite Resource Bank Department of Parasitology, College of Medicine,Chungbuk National University

First, the model automatically grouped the images into five clusters without using any species labels during training. After obtaining the clusters, not all images within each cluster were used for validation. Instead, the 300 images closest to the center of each cluster were selected, in order to represent the most visually characteristic images of that cluster. These selected images were then reviewed by the experts to verify whether the clusters corresponded to biologically meaningful parasite-related patterns.

Based on this selection, the expert validation subset consisted of 1,500 images, with 300 representative images for each cluster. These images included different gastrointestinal parasites, namely Ascaris lumbricoides, Trichuris trichiura, hookworms, and Schistosoma mansoni, in addition to egg-free images. The images were collected within the AI4NTD consortium using the P3 prototype slide scanner from Kato-Katz thick smear preparations.

After selecting the representative images, a parasitology expert manually reviewed them and recorded the presence or absence of parasite eggs in each image. This step made it possible to assess whether the clusters produced by the model reflected meaningful biological organization or were mainly influenced by other visual factors, such as background, staining, or acquisition conditions.

3.6.2 Cluster-Level Findings

Cluster 0 was mainly enriched with Ascaris lumbricoides. Expert review showed that 233 images, corresponding to 77.7%, contained Ascaris eggs. However, this cluster was not completely pure, since 100 images, corresponding to 33.3%, also contained Trichuris trichiura, 38 images, corresponding to 12.7%, contained hookworms, and 31 images, corresponding to 10.3%, were empty. This indicates that the model captured some visual features associated with Ascaris eggs, but the cluster still included mixed and empty images.

Cluster 1 appeared to be mainly influenced by image background rather than parasite morphology. The selected images in this cluster shared a similar blue-green background, probably related to staining, illumination, or scanning conditions. Although Ascaris was present in 182 images, corresponding to 60.7%, hookworms were also frequent, appearing in 85 images, corresponding to 28.3%. Therefore, this cluster was visually coherent but biologically heterogeneous.

Cluster 2 showed clear Ascaris dominance with almost complete hookworm exclusion. Ascaris lumbricoides was found in 225 images, corresponding to 75.0%, while hookworms appeared in only one image, corresponding to

0.3%. This suggests that the model learned a visual separation between hookworm morphology and Ascaris-like patterns.

Cluster 3 represented the strongest biological result. It had the highest Ascaris presence, with 241 images, corresponding to 80.3%, and the highest pure Ascaris rate, with 219 images, corresponding to 73.0%. Most importantly, no hookworm eggs were detected in this cluster, corresponding to 0.0%. This complete exclusion supports the idea that the unsupervised model captured meaningful morphological differences between thick-walled Ascaris eggs and thin-walled hookworm eggs.

Cluster 4 was the most mixed cluster. It contained 84 hookworm images, corresponding to 28.0%, and 76 Trichuris trichiura images, corresponding to 25.3%. This makes it the main cluster receiving hookworm and Trichuris images that were almost absent from Clusters 2 and 3. Ascaris remained present in 186 images, corresponding to 62.0%, but this cluster showed greater species diversity than Clusters 2 and 3.

3.6.3 Conclusion of Cluster-Level Findings

The following table summarizes the expert validation results across the five clusters:

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Ascaris presence	77.7%	60.7%	75.0%	80.3%	62.0%
Ascaris only (pure)	42.7%	47.3%	67.3%	**73.0%**	45.3%
Trichuris presence	33.3%	11.7%	11.3%	11.0%	25.3%
Hookworm presence	12.7%	28.3%	**0.3%**	**0.0%**	28.0%
Schistosoma presence	3.0%	6.3%	9.0%	9.7%	0.3%
Multi-species images	35.0%	15.7%	7.7%	7.3%	21.7%
Empty images	10.3%	10.3%	12.0%	6.3%	6.3%

Table 3.3: Cross-cluster comparison. Yellow cells highlight the most notable findings: hookworm near-exclusion in Clusters 2–3, and highest Ascaris purity in Cluster 3.

3.6.4 General Interpretation

The expert validation revealed three main clustering behaviors. First, Clusters 0, 2, and 3 were enriched with Ascaris lumbricoides, especially Cluster 3, which showed the highest biological coherence. Second, Cluster 1 was mainly driven by a non-biological visual factor, namely the blue-green background. Third, Cluster 4 grouped a larger number of hookworm and Trichuris trichiura images, making it a mixed-species cluster.

The most important result was the hookworm exclusion pattern in Clusters 2 and 3. Hookworms appeared in only 0.3% of images in Cluster 2 and were completely absent from Cluster 3. This indicates that the unsupervised model learned a meaningful morphological boundary separating hookworm eggs from Ascaris-dominant images.

However, no cluster achieved complete species purity. In addition, no dedicated cluster was clearly formed for Trichuris trichiura or Schistosoma mansoni. The presence of a background-driven cluster also confirms that the

model can sometimes be influenced by technical factors such as color, illumination, or staining conditions, rather than relying only on biological egg morphology.

3.6.5 Conclusion of Expert-Based Cluster Validation

The expert validation confirms that the unsupervised clustering model produced partially meaningful biological groupings. The strongest evidence was observed in the separation of hookworm images from Ascaris-dominant clusters, especially in Cluster 3. However, the presence of mixed-species images, empty images, and a background-driven cluster highlights the limitations of relying on unsupervised clustering alone.

Therefore, unsupervised clustering can be considered an important exploratory tool for understanding the internal structure of microscopy images and identifying some biological patterns. Nevertheless, it is not sufficient on its own for reliable diagnosis or final parasite species classification. These findings support the need for more accurate subsequent stages, such as self-supervised learning or supervised classification, in order to improve performance and move toward a more reliable diagnostic system.

3.7 Conceptual Integration with the AI-Powered Biobank Framework

In the proposed AI-powered biobank framework, the eleven-class parasite classification model was selected for integration because it represents a broader and more clinically useful diagnostic setting than the initial five-group dataset. The five-group dataset was mainly used to study the annotation problem and test zero-label organization, while the eleven-class dataset provided a larger and more structured classification experiment. Therefore, the final AI component of the framework is intended to rely on the eleven-class model in order to support more detailed parasite identification, future report generation, and expert-assisted diagnostic review.

3.8 Conclusion

This chapter presented the methodology adopted for gastrointestinal parasite image analysis within the proposed AI-powered biobank framework. The work started from an initial five-group dataset with no reliable expert-validated labels, which motivated the use of self-supervised representation learning and clustering to organize unlabeled microscopic images. Several pipelines were explored, and SimCLR combined with K-Means was selected as the most suitable approach for generating provisional pseudo-labels and selecting representative samples for expert review. The methodology was then validated using an eleven-class labeled gastrointestinal parasite dataset under 0%, 5%, 10%, 50%, and 100% label settings, showing that performance improved progressively as more labeled data became available. Finally, the approach returned to the initial five-group dataset to prepare clustered images for expert interpretation and future dataset refinement, while the eleven-class model was identified as the main candidate for future integration into the proposed digital biobank platform.

This methodology highlights the importance of combining unsupervised exploration with supervised validation in medical image analysis. It also shows that self-supervised learning can be useful when expert annotations are limited, expensive, or not immediately available. The use of clustering helped structure the initial data and guide expert review instead of relying on random image selection. In addition, the progressive label-ratio experiments demonstrated how model performance evolves when more annotated data are introduced. Overall, this approach provides a practical foundation for building more reliable AI-assisted diagnostic tools within a future biobank system.

Chapter 4:

Design and Implementation of the AI-Powered Biobank Platform

4.1 Introduction

This chapter presents the implementation of the proposed AI-powered gastrointestinal parasite biobank platform. After defining the methodological approach and the artificial intelligence experiments in the previous chapter, this chapter focuses on the practical development of the system. The proposed platform was designed to support the digital management of gastrointestinal parasite samples, AI-based diagnosis, result storage, report generation, community interaction, and epidemiological visualization. The system combines a web platform, a database, an artificial intelligence service, and a mobile application in order to provide a complete digital environment for laboratories, researchers, and health-related institutions.

The chapter describes the development environment, the general system architecture, the database design, the main web interfaces, the mobile application screens, and the integration of the AI diagnosis service. It also explains how the implemented modules interact with each other to support the main objective of the project, which is the creation of an AI-powered biobank infrastructure dedicated to gastrointestinal parasites.

4.2 Development Environment and Technology Choices

The implementation of the proposed platform required the selection of several technologies capable of supporting web development, database management, artificial intelligence integration, and mobile access. Since the proposed system is not limited to a simple website, but represents an AI-powered digital biobank platform, the development environment had to be modular, secure, and flexible.

The main technologies used in this project include:

- PHP and Laravel for the web platform.
- MySQL for database management.
- Python and FastAPI for the AI diagnosis service.
- PyTorch for deep learning model development.
- Flutter for the mobile application.
- Blade and Tailwind CSS for the web user interface.

4.2.1 PHP and Laravel Framework

PHP was selected as one of the main backend technologies because it is widely used for server-side web development. It allows the creation of dynamic web applications, the processing of user requests, the management of sessions, and the interaction with databases. PHP has historically been associated with web development because it was designed to generate dynamic web pages and simplify communication between the user interface and the server.

However, developing a large web application using raw PHP can become difficult to maintain, especially when the project includes several modules such as authentication, user roles, database operations, forms, security rules, sample management, AI diagnosis, reports, and subscriptions. For this reason, the Laravel framework was chosen.

Laravel is a modern PHP framework that follows the Model–View–Controller architectural pattern. This architecture helps divide the application into three main layers: models for data management, views for user interfaces, and controllers for application logic. This structure was suitable for the proposed platform because the project includes several connected modules, such as users, organizations, samples, diagnosis results, reports, payment requests, community interaction, and activity logs.

Laravel was also useful because it provides many built-in features, including:

- Routing for managing application pages and API endpoints.
- Authentication for secure user access.
- Middleware for access control and verification.
- Validation for form data.
- Database migrations for organizing database structure.
- Eloquent ORM for interacting with MySQL tables.
- Blade templates for building dynamic web interfaces.

These features helped accelerate the development process and made the platform more organized, secure, and easier to maintain.

4.2.2 MySQL Database Management System

MySQL was used as the relational database management system of the platform. It was selected because the data handled by the system is structured and strongly connected. For example, each organization can have several users, each user can add several samples, and each sample can have one or more AI diagnosis results.

The use of a relational database was important because the platform requires clear relationships between different entities. These entities include users, organizations, samples, diagnosis results, reports, subscriptions, payment requests, community posts, comments, and activity logs.

In this project, MySQL stores the main information required by the platform, such as:

- User accounts and roles.
- Organization information.
- Sample records and image paths.
- AI diagnosis results.
- Generated reports.
- Subscription status.
- Payment requests.
- Community posts and comments.
- Activity logs.

MySQL was also suitable because it integrates easily with Laravel through migrations and Eloquent ORM. This simplified the creation, update, and management of the database tables. In addition, the database provides the information required to generate the parasite spread map using sample locations and diagnosis results.

4.2.3 Python and FastAPI for the AI Diagnosis Service

The artificial intelligence component was implemented separately from the Laravel platform using Python and FastAPI. Python was selected because it is widely used in artificial intelligence, deep learning, and medical image analysis. It provides powerful libraries for image processing, model training, and prediction.

FastAPI was used to expose the trained deep learning model as an independent diagnosis service. This service receives a microscopic image from the platform, processes it using the trained model, and returns the predicted parasite class with a confidence score.

The AI service performs the following tasks:

- Receives the microscopic image.
- Preprocesses the image.
- Sends it to the trained model.
- Generates the predicted parasite class.
- Returns the confidence score and model information.
- Sends the result back to the Laravel platform.

Separating the AI service from the web platform made the system more modular. This means that the AI model can be improved, replaced, or updated later without modifying the whole Laravel application. After receiving the prediction, the Laravel platform stores the diagnosis result in the MySQL database and displays it to the user through the web interface or the mobile application.

4.2.4 Flutter for Mobile Application Development

Flutter was used for the development of the mobile application. The objective of the mobile part was to make the platform accessible from smartphones, especially for users who need to add samples, launch AI analysis, check diagnosis results, or consult reports outside the desktop environment.

Flutter was selected because it allows the creation of modern and responsive mobile interfaces using a single codebase. This makes mobile development faster and more flexible. In this project, the mobile application communicates with the Laravel backend through API requests, allowing both the web platform and the mobile application to use the same database and services.

The mobile application includes several main screens:

- Login screen.
- Home dashboard.
- Sample submission screen.
- AI analysis action.
- Results list.
- Result details.
- Reports screen.
- Community screen.

This mobile access makes the system more practical for laboratories, researchers, and future field use.
This combination of technologies made it possible to develop a functional prototype that connects a web platform, a relational database, an AI diagnosis service, and a mobile application. The selected tools were suitable for the objectives of the project because they support modular development, structured data management, artificial intelligence integration, and mobile accessibility.

4.3 System Architecture

After presenting the development environment and the technologies used to implement the platform, this section describes the general architecture of the proposed system and the role of each component.

The proposed platform is based on a modular architecture composed of four main components: the web platform, the mobile application, the Laravel backend, and the AI diagnosis service. Users can access the system through

the web interface or the mobile application. Both interfaces communicate with the Laravel backend, which manages authentication, platform logic, database operations, and API requests.

The backend is connected to a MySQL database used to store users, organizations, samples, diagnosis results, reports, subscriptions, community data, and activity logs. It is also connected to a FastAPI service responsible for AI-based parasite diagnosis. This service receives microscopic images, processes them using the trained deep learning model, and returns the predicted parasite class with a confidence score.

This architecture separates the user interface, backend, database, and AI components, making the system easier to maintain and extend.

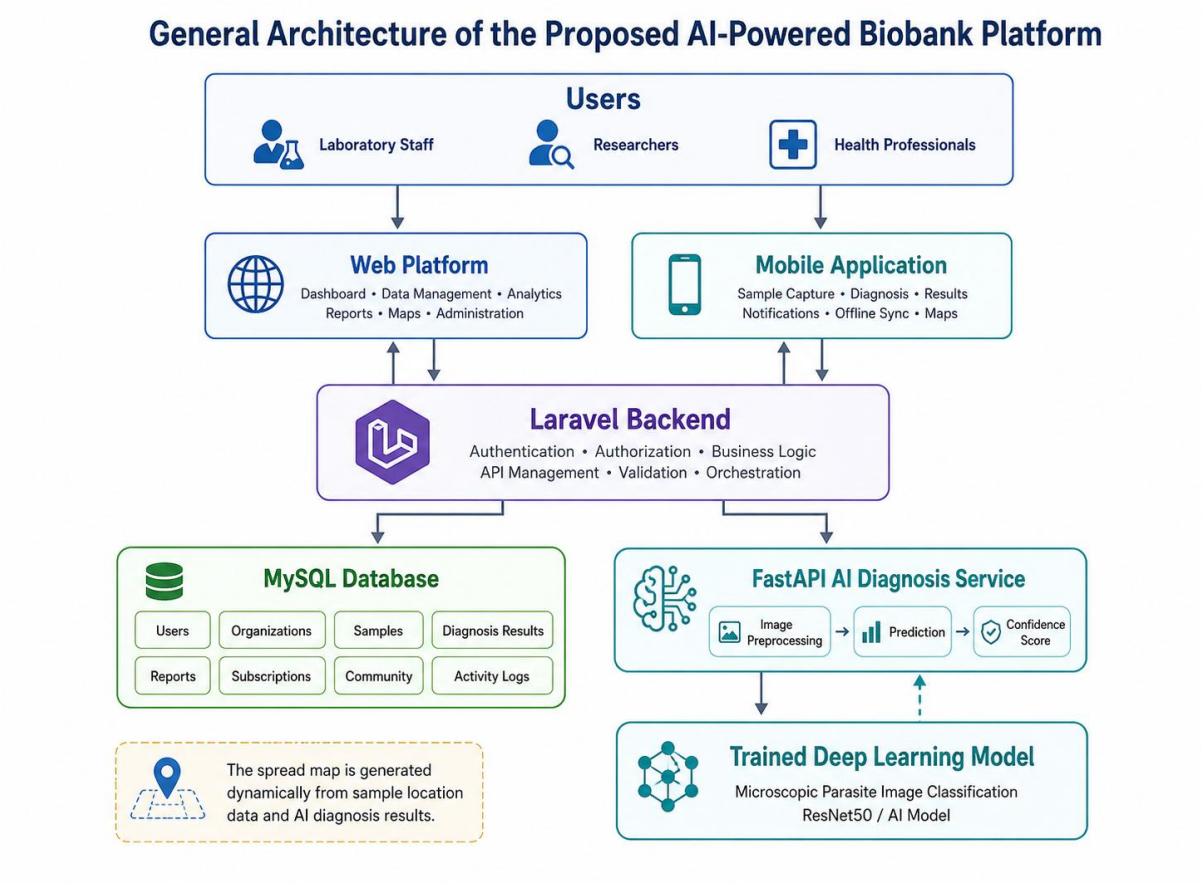


Figure 4.1: General Architecture of the Proposed AI-Powered Biobank Platform

After defining the general system architecture, the database was designed to organize and store the main information used by the platform. The database design represents an important part of the software engineering process, because it defines how the different system entities are structured and related before their implementation.

At the conceptual level, a final conceptual data model was created to describe the main entities of the platform and their relationships. This model includes Organization, User, Sample, Diagnosis_Result, Report, Subscription, Payment_Request, Community_Post, Community_Comment, and Activity_Log. These entities represent the main functional modules implemented in the system.

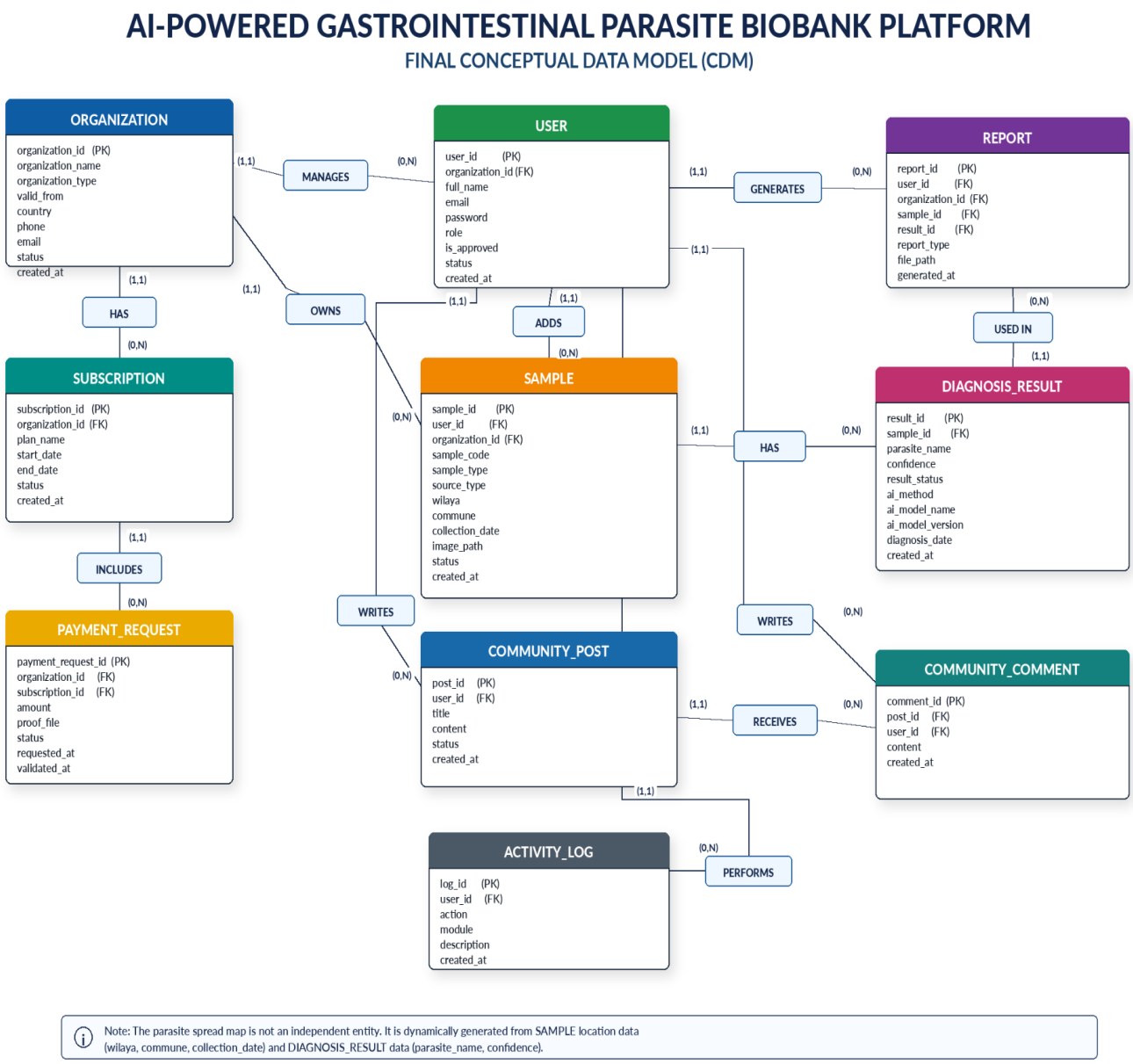


Figure 4.2: Final Conceptual Data Model of the AI-Powered Gastrointestinal Parasite Biobank Platform

At the logical and implementation level, the conceptual model was translated into relational tables using MySQL and Laravel migrations. Each table contains the necessary attributes, primary keys, and foreign keys required to maintain relationships between the different entities. For example, each user belongs to one organization, each sample is linked to both a user and an organization, and each diagnosis result is linked to a specific sample.

The database stores all essential platform data, including user accounts, organization information, sample records, microscopic image paths, AI diagnosis results, generated reports, subscriptions, payment requests, community interactions, and activity logs. This structure allows the platform to manage data in a clear, organized, and consistent way. The parasite spread map was not implemented as an independent table. Instead, it is generated dynamically from the existing database information, mainly sample location data and AI diagnosis results. This makes the spread map an analytical module based on stored data rather than a separate database entity.

In summary, the database implementation translates the conceptual design into a functional relational structure that supports sample management, AI diagnosis history, reporting, subscription control, community interaction, and epidemiological visualization.

4.5 Web Platform Implementation

This section presents the main interfaces implemented in the web platform. The objective is to show how the functional modules of the proposed biobank system were translated into practical user interfaces. The selected screenshots illustrate the most important parts of the platform, including the public home page, authentication, dashboard, sample management, AI diagnosis results, reports, community interaction, subscription management, and the epidemiological spread map.

4.5.1 Public Home Page

Before accessing the protected modules of the platform, users are first directed to the public home page. This interface presents the main purpose of the proposed biobank platform and summarizes its principal services, including sample management, AI-assisted diagnosis, reports, spread map, community interaction, and learning resources. It also provides access to login and registration. During registration, organization administrators can create an account request for their organization, while regular users must select an already registered organization in order to join the platform.

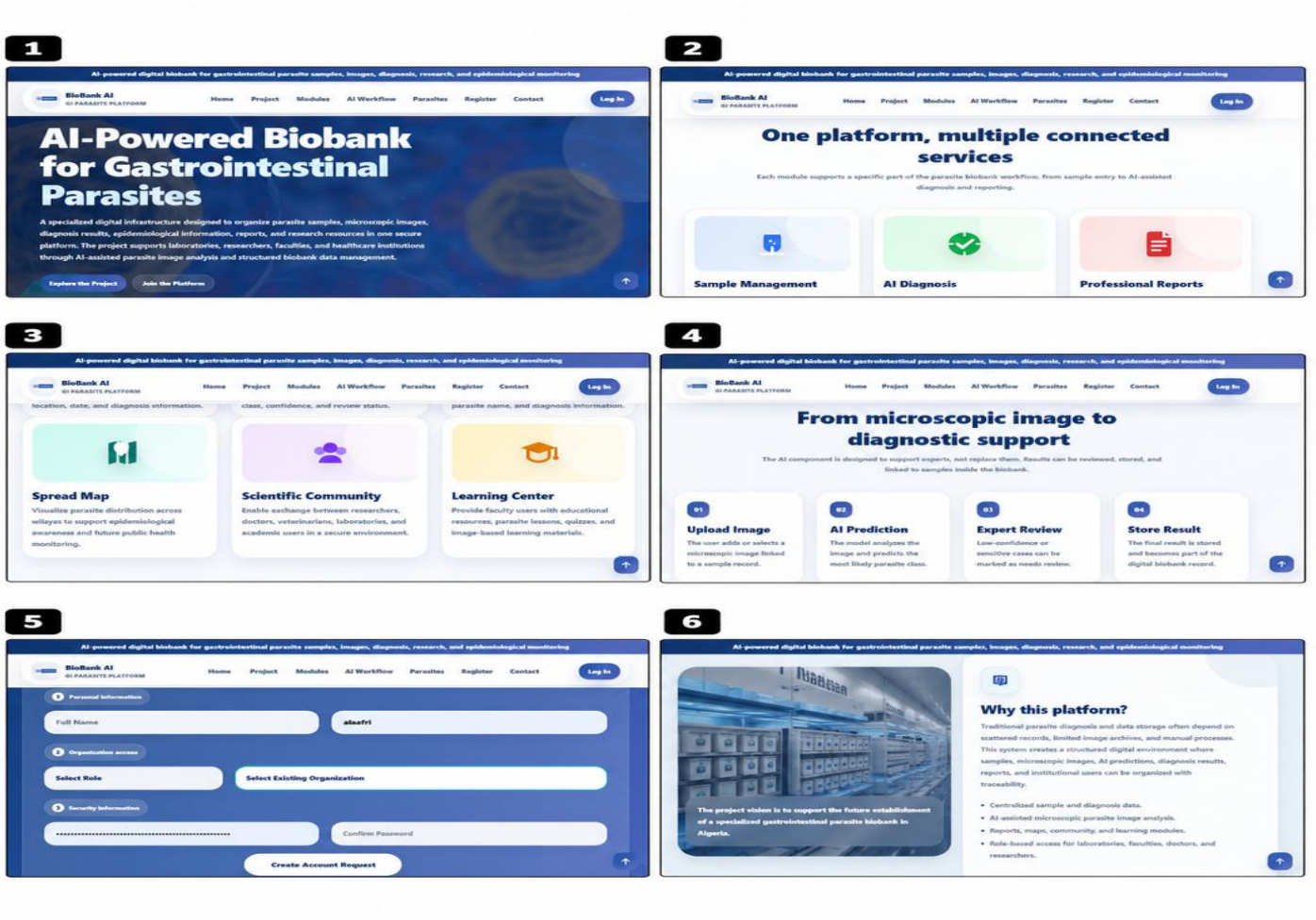


Figure 4.3: Public Home Page and Registration Access of the Web Platform

The login interface allows registered users to authenticate before accessing the protected modules of the platform. After entering their email and password, users can access the workspace only if their account has been approved by the organization administrator. This validation step ensures that only authorized users belonging to a registered organization can manage samples, view AI diagnosis results, generate reports, and access internal platform services.

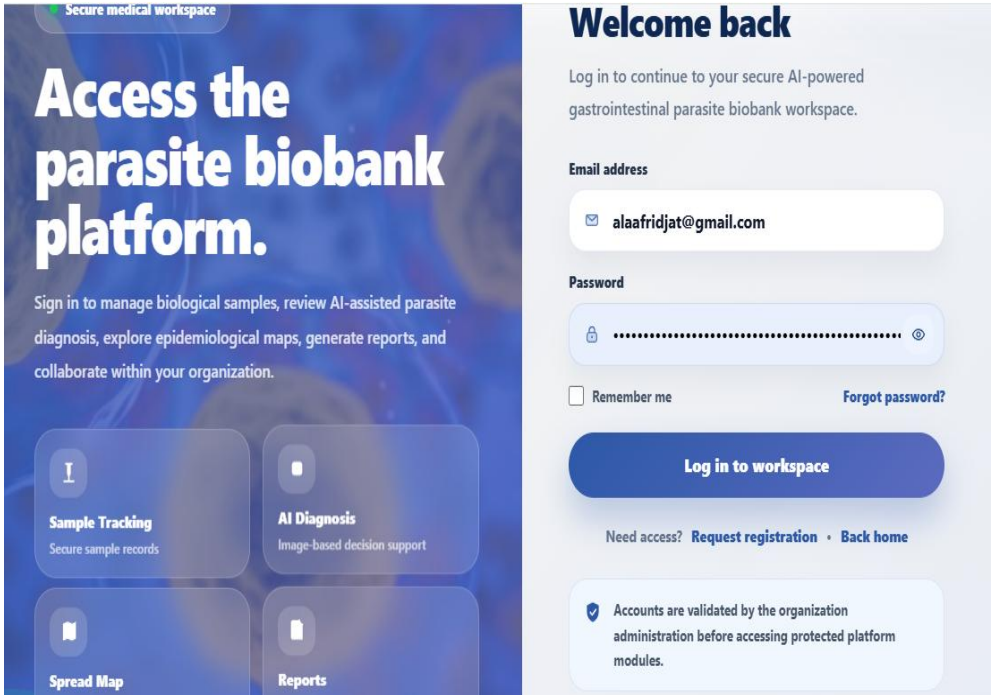


Figure 4.4: Web Platform Login and Authentication Interface

4.5.3 Dashboard Interface

After authentication, users are redirected to a dashboard adapted to their role within the platform. The organization administrator dashboard provides extended management functions, including organization overview, member management, join request approval or rejection, notifications, and recent activity monitoring. Through this dashboard, the organization administrator can supervise the internal structure of the institution, follow the status of registered members, and ensure that only authorized users are accepted within the organization. This contributes to better control of access, improved coordination between members, and more efficient management of platform activities.

In contrast, the regular user dashboard focuses on the operational modules available to the user, such as sample management, diagnosis results, reports, community, spread map, and learning center. This dashboard is designed to provide quick access to the main functions needed in daily use, especially for users involved in sample registration, diagnostic follow-up, report consultation, and knowledge sharing. The interface also helps users navigate easily between the different modules without accessing administrative functions that are outside their responsibilities.

This role-based dashboard design ensures that each user accesses only the functionalities corresponding to their responsibilities and permissions. It also improves data security by limiting sensitive management operations to authorized administrators only.

Moreover, separating administrative tasks from operational tasks makes the platform more organized, user-friendly, and suitable for different categories of users, including organization administrators, researchers, laboratory members, and healthcare professionals.

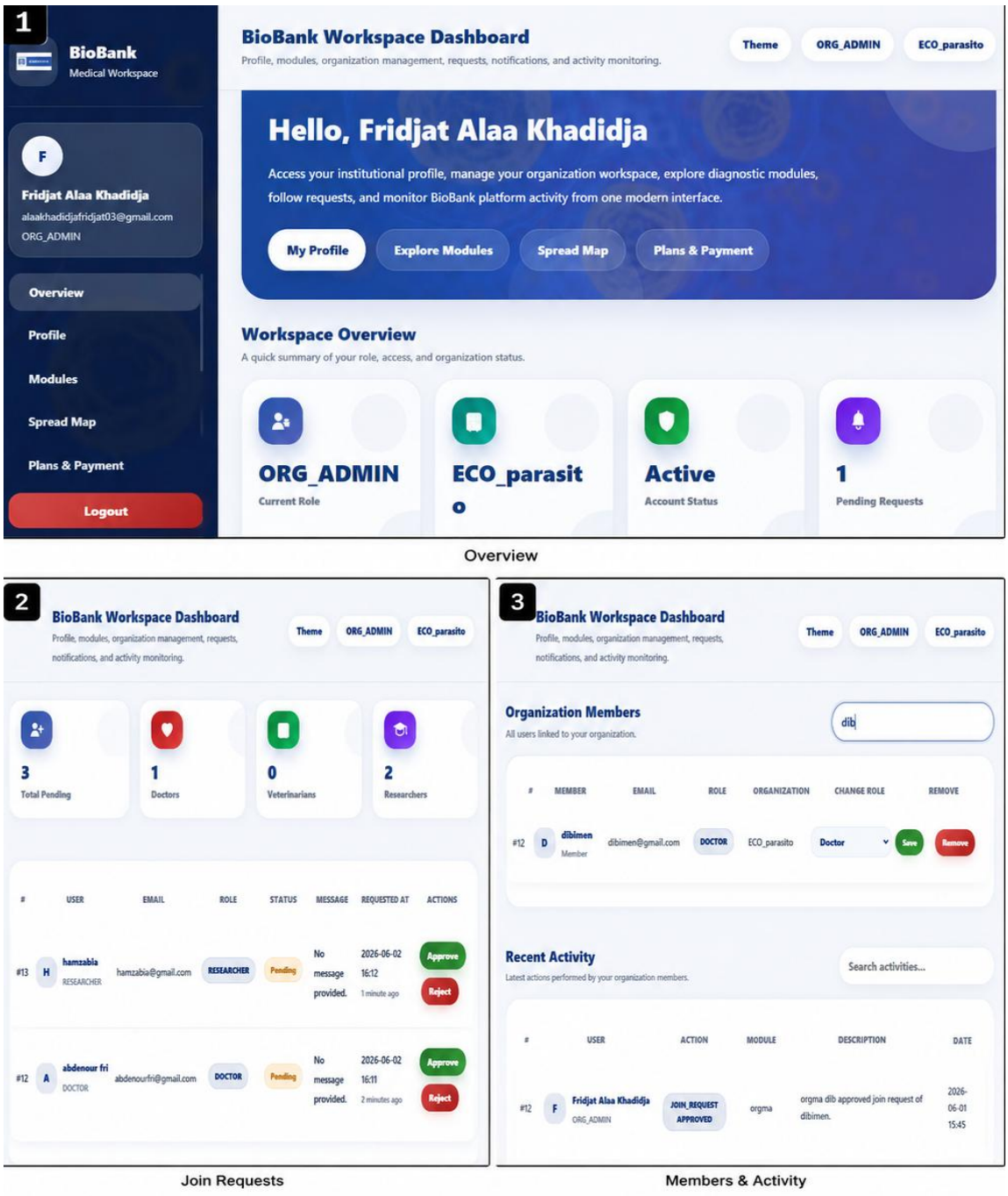


Figure 4.5: Organization Administrator Dashboard Interfaces

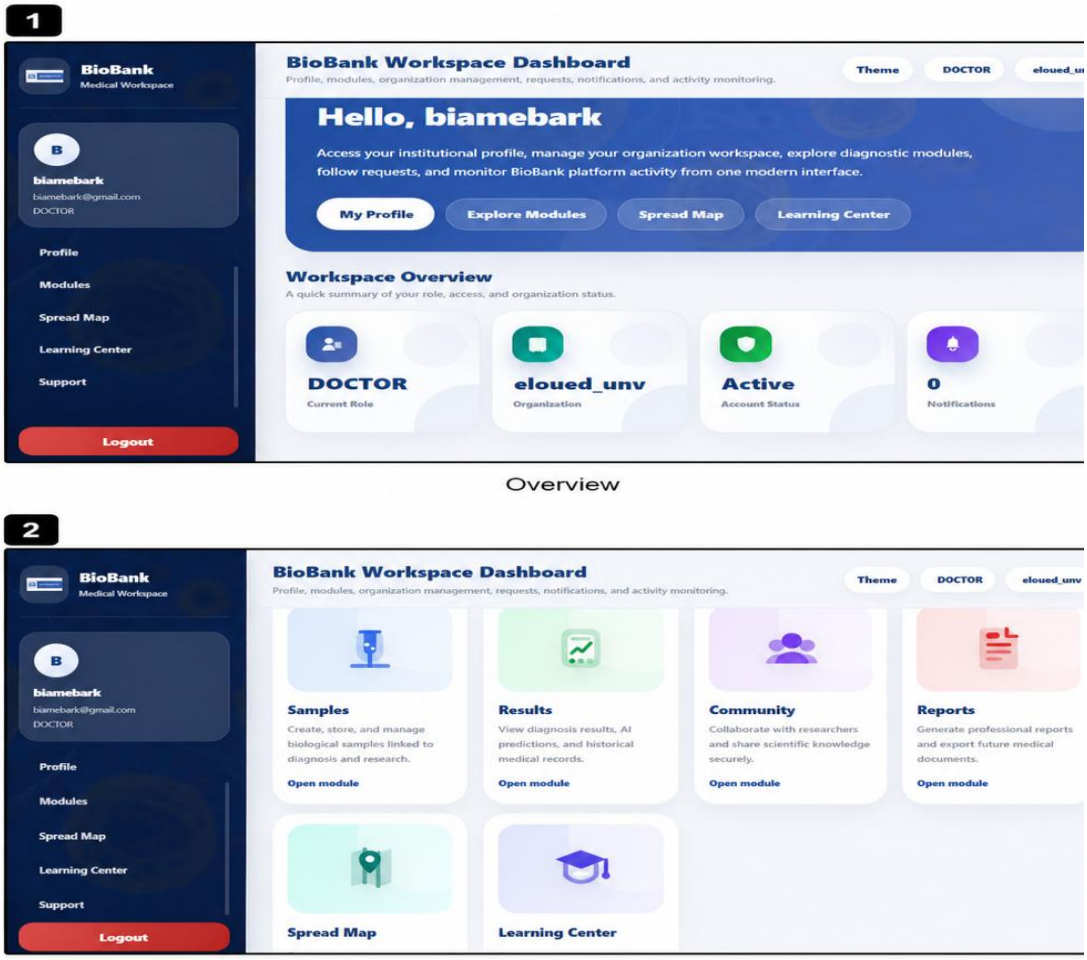


Figure 4.6: Regular User Dashboard Interfaces

4.5.4 Sample Management Interface

The sample management interface allows authorized users to create a new biological sample record by uploading a microscopic image and entering the related sample information. The form includes the sample type, source type, subject information, age, gender, geographical location, collection date, and additional notes. These fields help describe the sample in a clear and organized way, making it easier to identify, search, and manage within the platform.

This interface plays an important role in ensuring sample traceability from the moment the sample is registered. Each uploaded microscopic image is linked to its corresponding sample record, which reduces the risk of data confusion and helps maintain consistency between the biological sample, the image, and the associated medical information. The use of structured fields also facilitates data classification and improves the reliability of the stored information.

Furthermore, the sample management module supports the future use of the collected data for AI-assisted diagnosis, report generation, and epidemiological analysis. Once a sample is registered, it can be linked to diagnosis results generated either manually by specialists or automatically through the AI model. This makes the

sample record a central element in the platform workflow, connecting medical data, microscopic image analysis, and decision support.

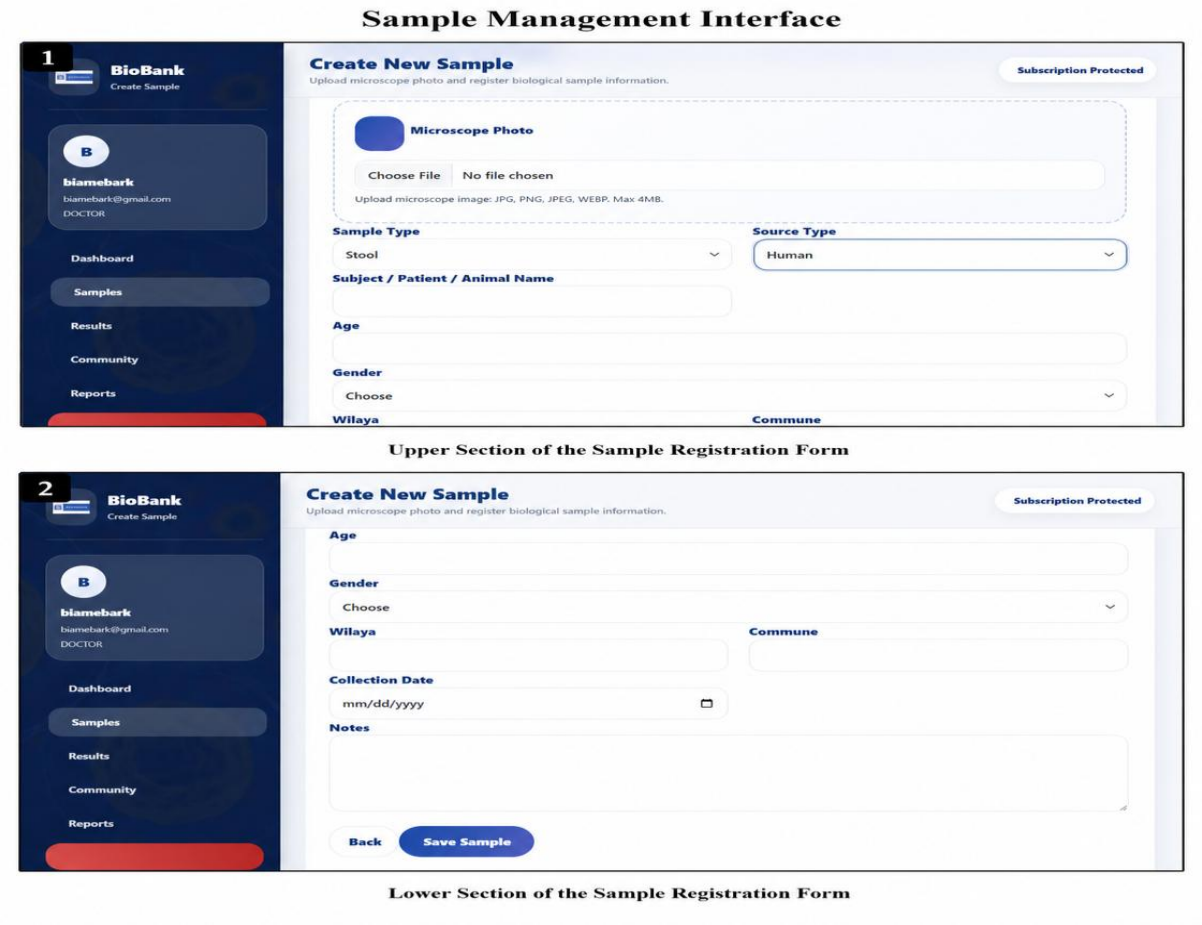


Figure 4.7: Sample Management Interface and Sample Registration Form

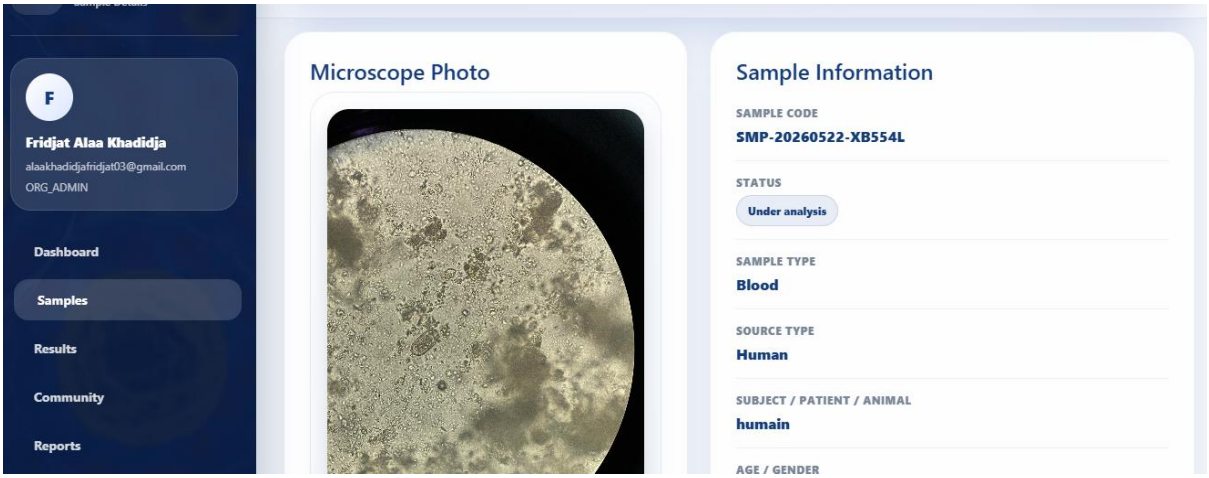


Figure 4.8: Sample Details and Microscope Image Interface

4.5.5 AI Diagnosis Results Interface

This interface manages the diagnostic results linked to registered samples. It includes a results workspace, a diagnosis creation form, and a detailed result view showing the predicted parasite, confidence score, AI model name, result status, and diagnosis notes.

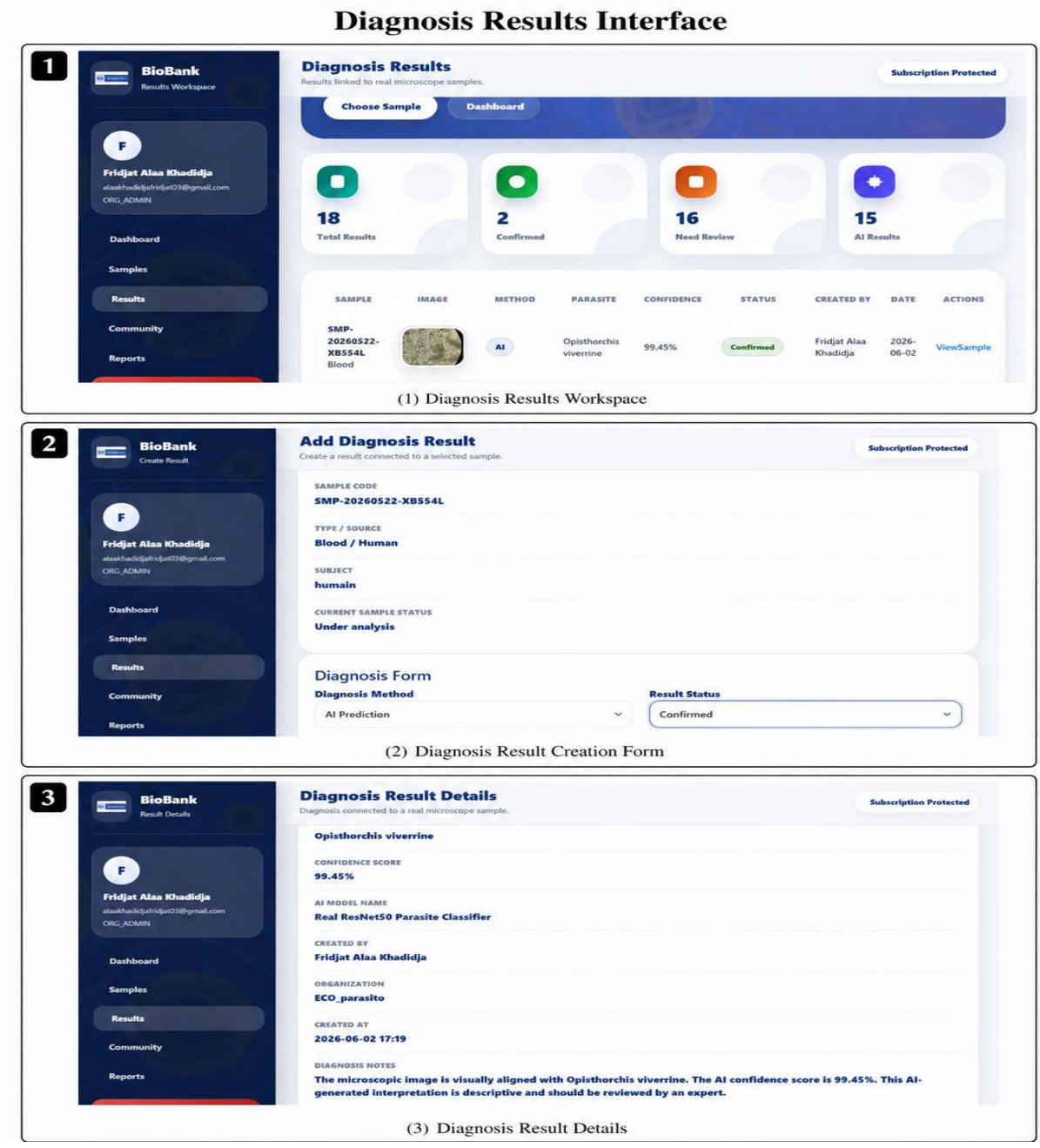


Figure 4.9: Diagnosis Results Interface

4.5.6 Reports Interface

After presenting the sample management interface, the reports interface represents the final step in using the registered sample data. It allows users to generate PDF diagnosis reports from saved sample results. Each report

summarizes the sample information, parasite diagnosis, confidence score, status, and microscope image in a structured document.

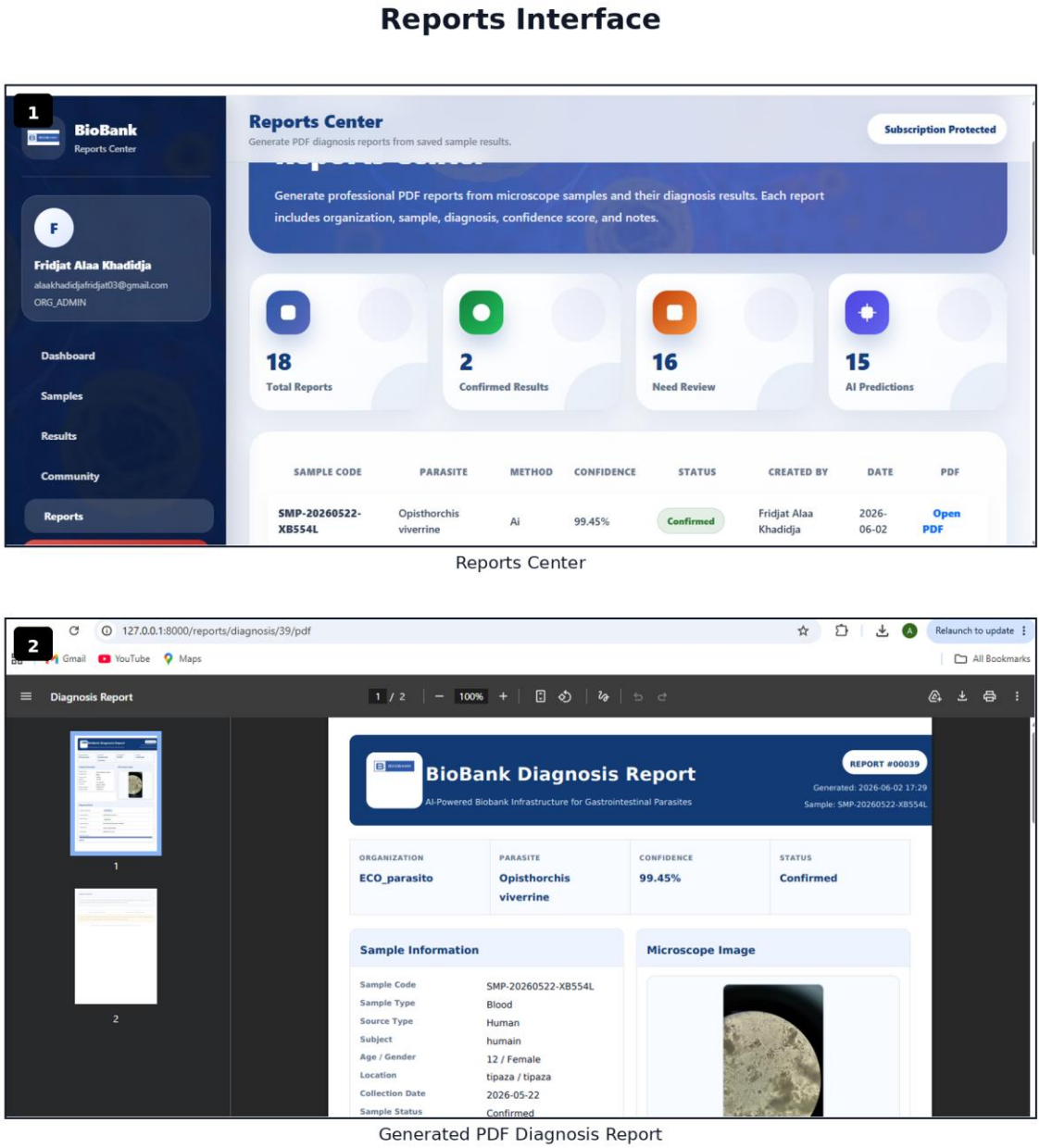


Figure 4.10: Reports Center and Generated Diagnosis PDF Report Interface

4.5.7 Epidemiological Spread Map Interface

The epidemiological spread map interface visualizes parasite-positive cases according to the geographical location of registered samples and AI diagnosis results. It provides filters by parasite type and date, as well as summary statistics such as total samples, positive cases, and covered wilayas. The interactive map allows users to identify parasite distribution across different regions and display detailed information for each wilaya, including the number of samples, diagnosed cases, positive cases, and detected parasite types.

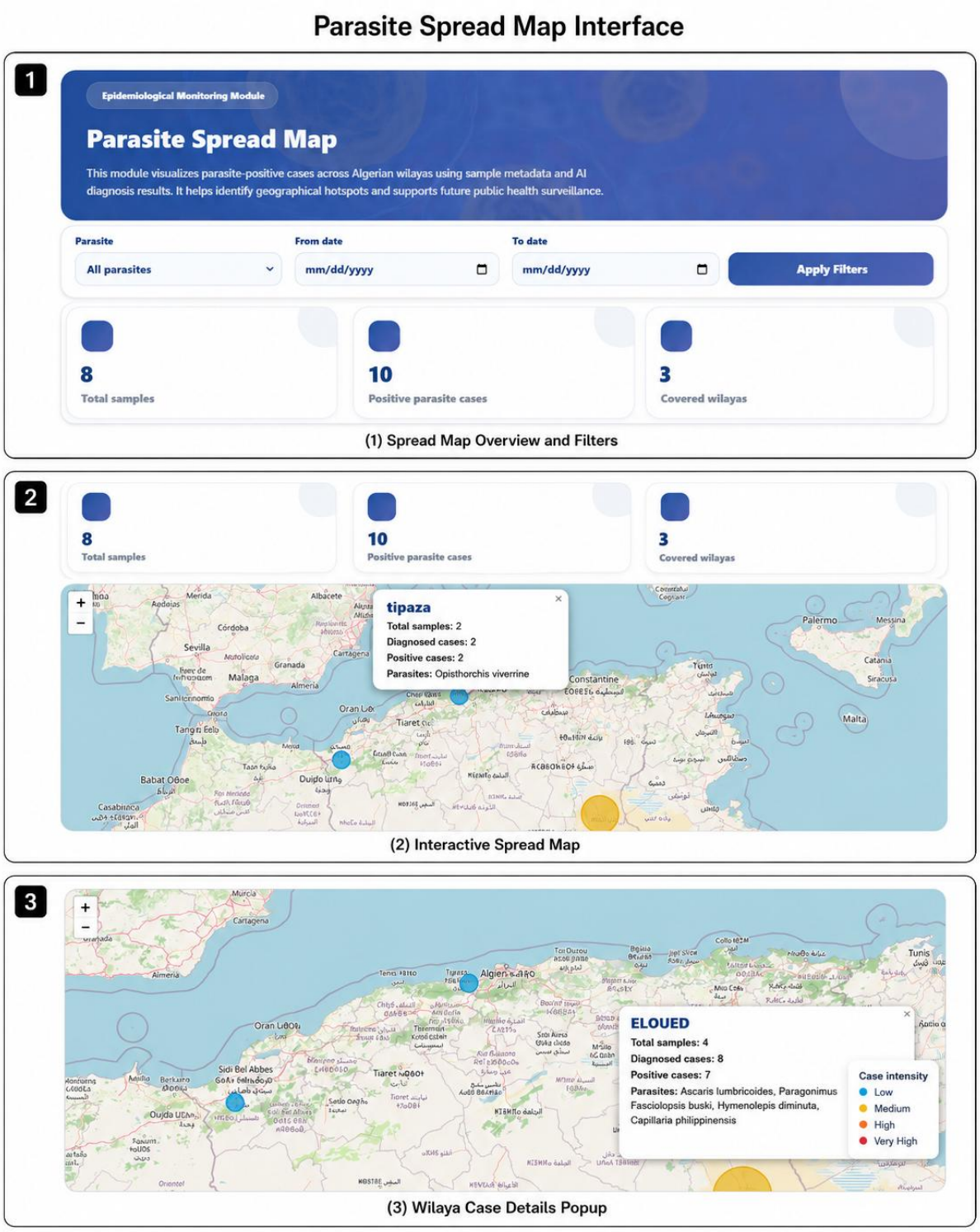


Figure 4.11: Epidemiological Parasite Spread Map Interface

4.5.8 Community Interface

The research community module represents an important feature of the platform because it allows users from the same digital biobank environment to share scientific posts, case discussions, observations, and questions. This module supports collaboration between researchers, doctors, laboratories, and organization members. It also helps

transform the platform from a simple sample management system into an interactive scientific workspace where knowledge, experience, and diagnostic observations can be exchanged securely.

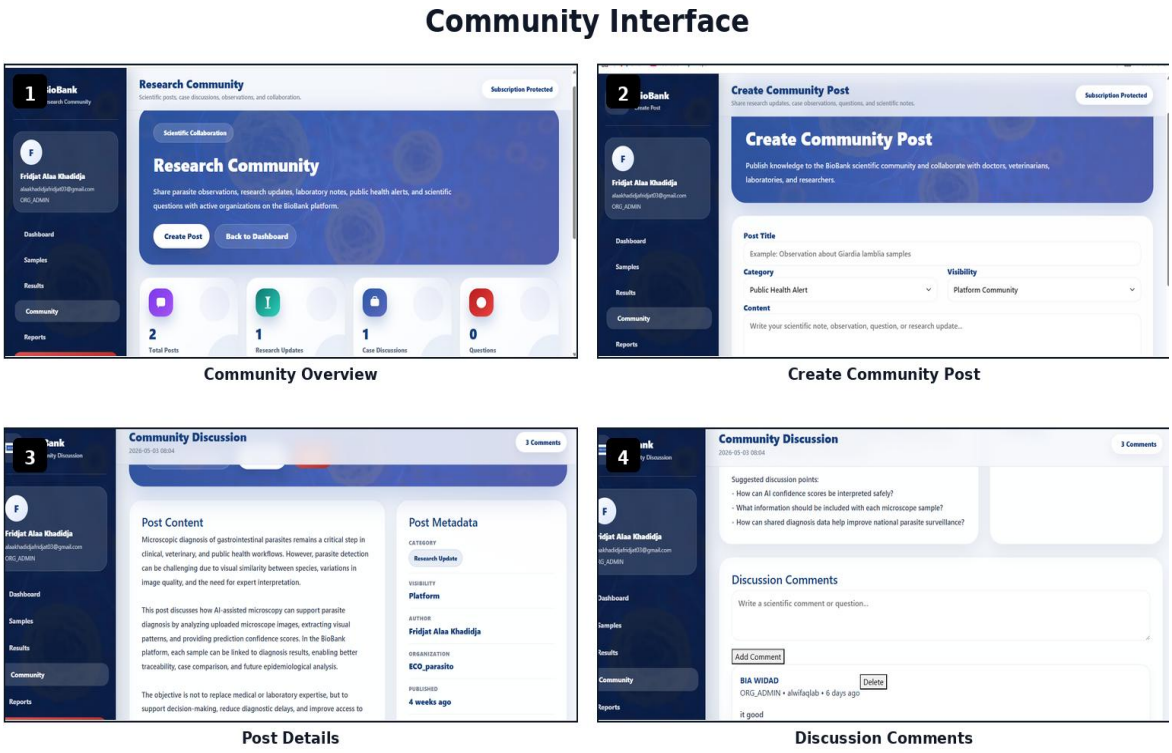


Figure 4.12: Research Community Interface: Overview, Post Creation, Discussion Details, and Comments

4.5.9 Learning Center Interface

The learning center interface was added as an educational support module within the platform. It provides users with access to learning resources related to gastrointestinal parasites, microscopy, diagnosis concepts, and platform usage. This module also includes quizzes that help users test their knowledge and review their answers. Therefore, the platform does not only manage samples and diagnosis results, but also supports training, awareness, and continuous learning.

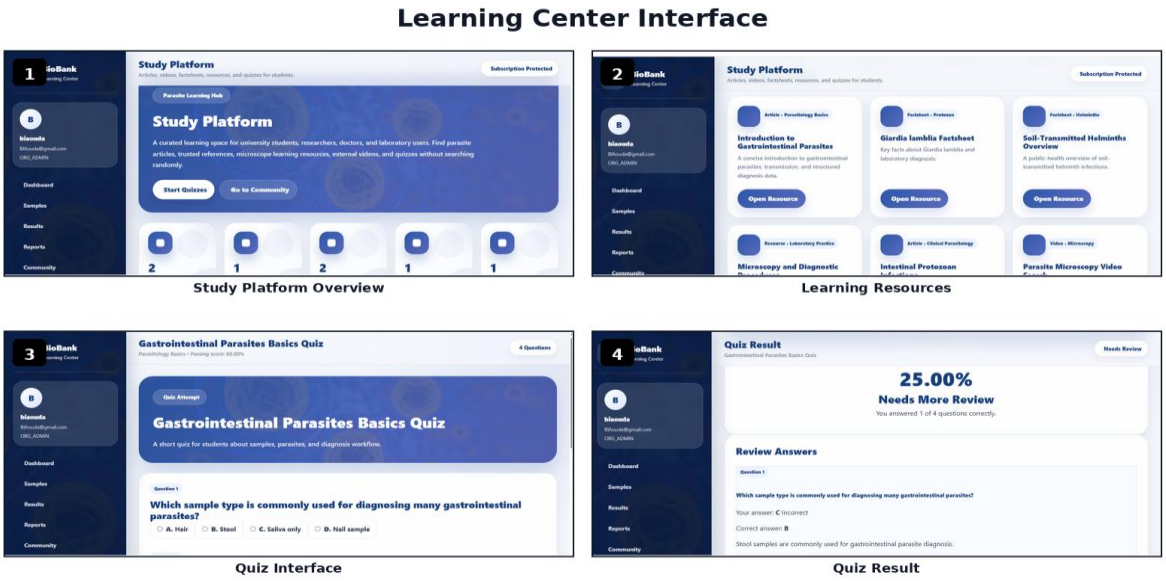


Figure 4.13: Learning Center Interface: Overview, Learning Resources, Quiz Interface, and Quiz Result

4.6 AI Diagnosis Service Integration

The artificial intelligence diagnosis component was integrated into the platform as an independent service using FastAPI. FastAPI is a Python web framework used to build APIs that can receive requests, process data, and return structured responses. It was selected because the AI models were developed in Python, and most deep learning tools used for image processing and prediction are also based on the Python ecosystem.

In this project, FastAPI acts as a bridge between the Laravel web platform and the trained deep learning models. Instead of placing the AI code directly inside Laravel, the diagnosis models were separated into an external service. This makes the system more modular and easier to maintain. The Laravel platform is responsible for user management, sample management, database operations, and interface display, while FastAPI is responsible for the AI analysis process.

This separation was important because the AI models were trained and executed using Python libraries, while the web platform was developed with Laravel and PHP. It also allows the AI service to be updated or replaced without modifying the whole Laravel project.

4.6.1 Autoencoder-Based Parasite / No-Parasite Screening

Before applying the parasite species classifier, an autoencoder was integrated as a preliminary screening model. Its role is to verify whether the uploaded image is parasite-like or No-Parasite.

The autoencoder was trained only on parasite images. It was not trained using bacteria, fungi, blood cells, or other non-parasite images. These non-parasite images were used only during testing in order to evaluate whether the model could reject images that do not belong to the parasite distribution.

The model is based on convolutional layers. The input image is resized to 64 × 64 RGB. The encoder compresses the image into a latent representation, and the decoder reconstructs the image from this representation. The decision is based on the reconstruction error. If the image is similar to parasite images, the reconstruction error is low. If the image is different from parasite images, the reconstruction error becomes higher.

The threshold was selected using the 99th percentile of reconstruction errors calculated on parasite images. The obtained threshold was 0.003563.

The decision rule was:

- Reconstruction error ≤ 0.003563: parasite-like image
- Reconstruction error > 0.003563: No-Parasite image

During sample creation, the uploaded image is first sent from Laravel to the FastAPI screening endpoint. If the image is detected as No-Parasite, the platform displays a warning message and the sample is not saved. If the image is considered parasite-like, the sample is accepted and can later be analyzed by the parasite classification model. This preliminary screening step improves the reliability of the platform by preventing irrelevant images from entering the diagnosis workflow. It also reduces unnecessary classification attempts, since the species classifier is applied only to images that are likely to contain parasite structures. By separating the screening task from the classification task, the system becomes more robust and better organized.

This design also helps protect the quality of the stored data, especially because the platform is intended to support future biobank development. Therefore, the autoencoder acts as a first control layer before sample registration and AI-assisted diagnosis. Its integration strengthens the overall workflow by improving image validation, data consistency, and diagnostic reliability.

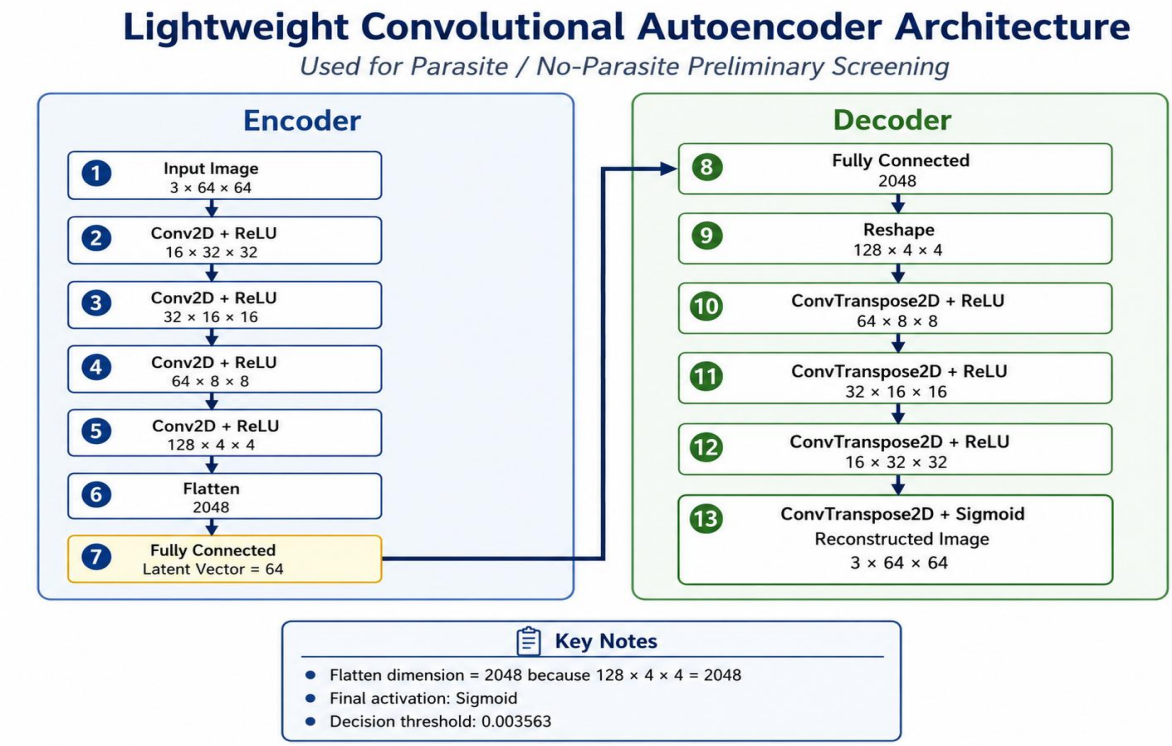


Figure 4.14 Architecture of the convolutional autoencoder used for Parasite / No-Parasite screening

4.6.2 ResNet50 11-Class Parasite Classification

After the autoencoder screening step, parasite-like images can be passed to the 11-class classification model. This model is based on ResNet50 and is used to identify the most probable parasite species among the predefined parasite classes.

The Laravel platform sends the accepted image to the FastAPI prediction endpoint. FastAPI preprocesses the image, applies the trained ResNet50 model, and returns the predicted parasite class with a confidence score. The returned result is then stored in the MySQL database as an AI diagnosis result.

The output of the classification model includes the predicted parasite species, the confidence score, the model name, and a descriptive report. This result can be displayed in the diagnosis results interface, included in PDF reports, and used in the epidemiological spread map.

4.6.3 Final AI Workflow in the Platform

The final AI workflow is composed of two main stages. First, the autoencoder performs Parasite / No-Parasite screening. If the image is detected as No-Parasite, the platform stops the process and does not save the sample. Second, if the image is considered parasite-like, it can be passed to the ResNet50 classifier for parasite species prediction.

This two-stage workflow improves the reliability of the system because the 11-class classifier is applied only to images that are likely to belong to the parasite distribution. Therefore, the AI service combines image verification, parasite screening, species classification, result storage, report generation, and epidemiological visualization in a coherent workflow.

4.7 Mobile Application Implementation

In addition to the web platform, a mobile application was developed to provide easier access to the main services of the proposed digital biobank. The objective of this application is to allow users to interact with the platform directly from a smartphone, especially in situations where using a computer is not practical. Through the mobile application, users can manage samples, consult AI-assisted diagnosis results, generate reports, access the research community, and use the assistant.

The mobile application was developed using Flutter and communicates with the Laravel backend through API requests. This allows the mobile application and the web platform to share the same database, authentication system, and core services. As a result, the information added or consulted through the mobile application remains synchronized with the web platform.

The application provides access to the main modules of the platform, including:

- User authentication through the login screen.
- Access to the main dashboard.
- Sample management and sample creation.
- Consultation of AI diagnosis results.
- Report generation and visualization.
- Access to the community and assistant services.

This mobile integration also improves the continuity of the workflow between the web platform and field use. For example, a user can register a sample or consult a diagnosis result directly from a mobile device, while the same data remains available on the web platform. This is particularly useful in laboratory and healthcare environments where users may need quick access to sample information, diagnostic outputs, or reports without returning to a desktop computer. Therefore, the mobile application supports both flexibility and accessibility while keeping the system connected to the same centralized database.

The first group of interfaces presents the main mobile workflow. It includes the login screen, the home dashboard, the core modules, the sample list, the sample creation form, and the AI prediction result. These interfaces show how the user can access the platform, navigate between modules, add a biological sample, upload a microscopic image, and consult the AI-assisted diagnosis result.

Overall, the mobile application represents an important extension of the web platform. It improves accessibility, supports field and laboratory use, and makes the digital biobank more practical for researchers, healthcare professionals, and laboratory members.

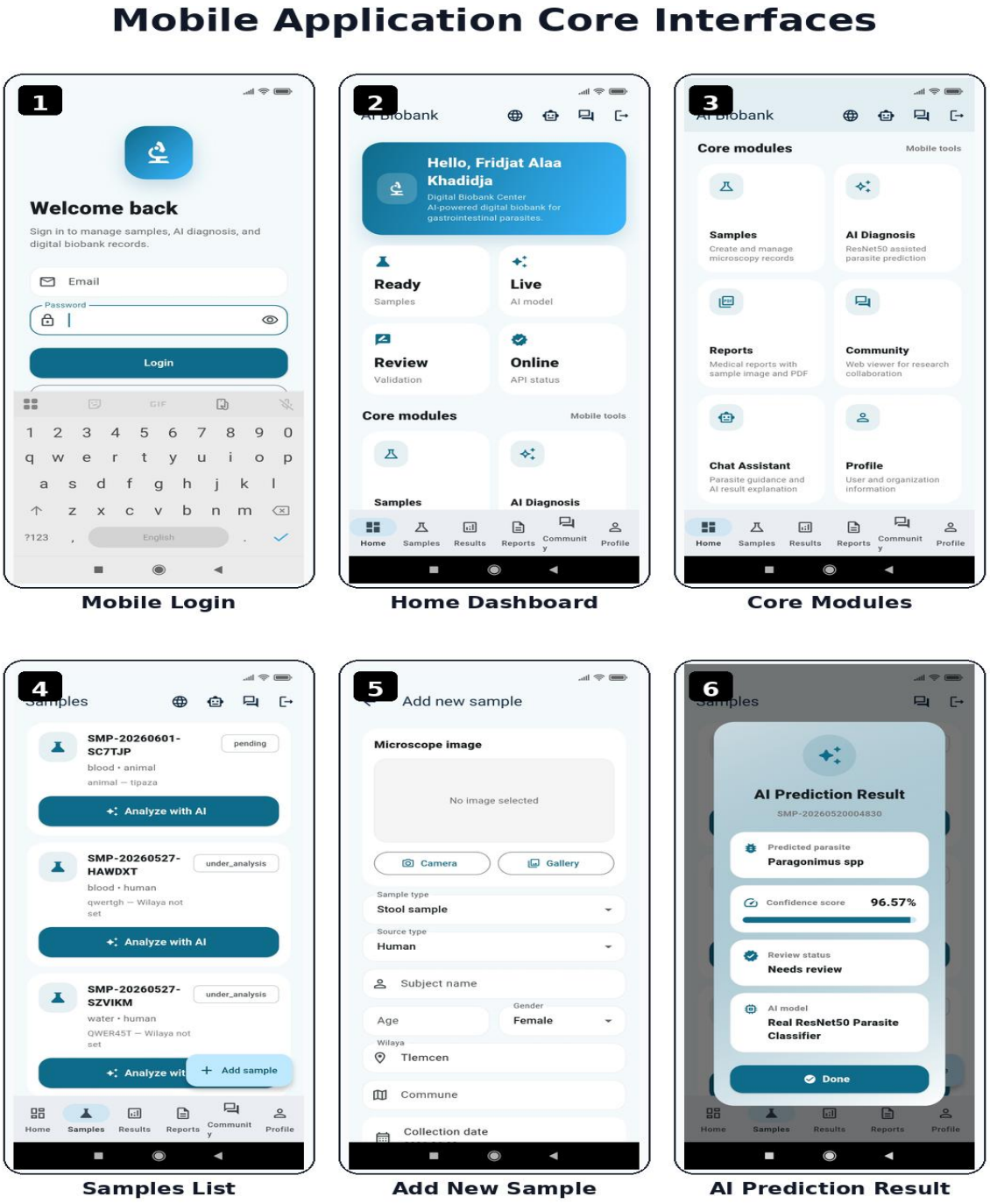


Figure 4.15: Mobile Application Core Interfaces

The second group of interfaces shows the result and report modules. The results screen displays the stored AI diagnosis outputs, including the predicted parasite, confidence score, model name, and review status. The report interface presents a structured medical AI report and allows the result to be exported or saved as a PDF document.

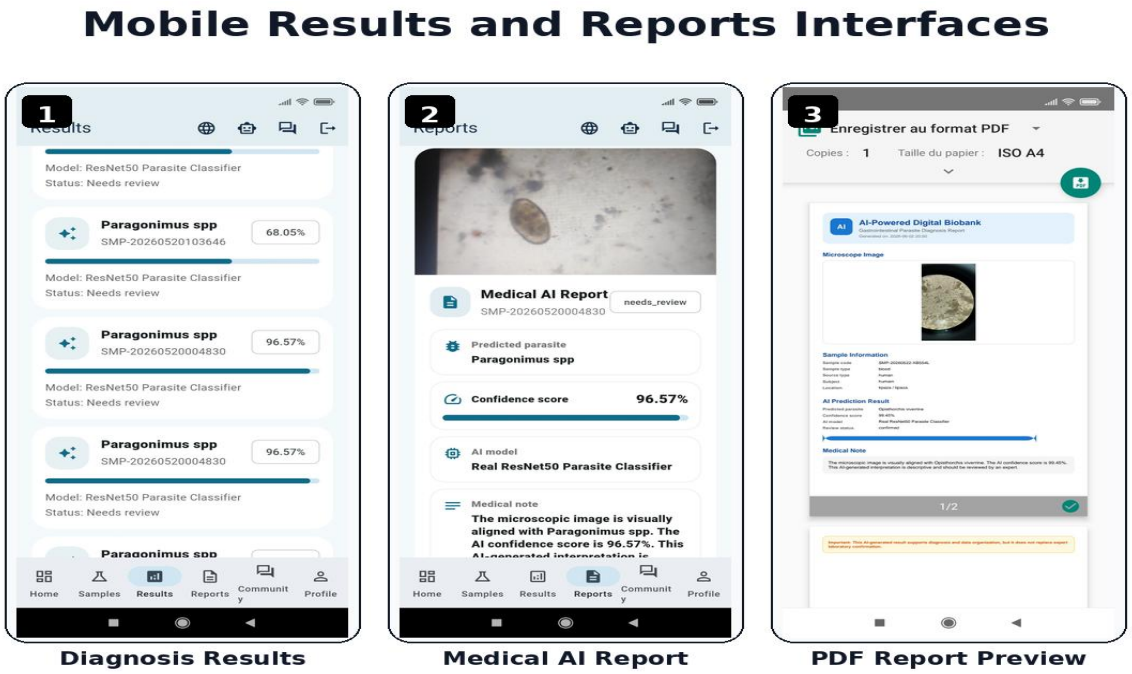


Figure 4.16: Mobile Results and Reports Interfaces

The third group of interfaces presents the community and assistant features. The research community allows users to consult scientific discussions and shared observations. The chat assistant provides guidance about parasites, AI diagnosis, confidence scores, prevention, symptoms, and the biobank workflow. A multilingual view was also included to show that the mobile application can support different interface languages.

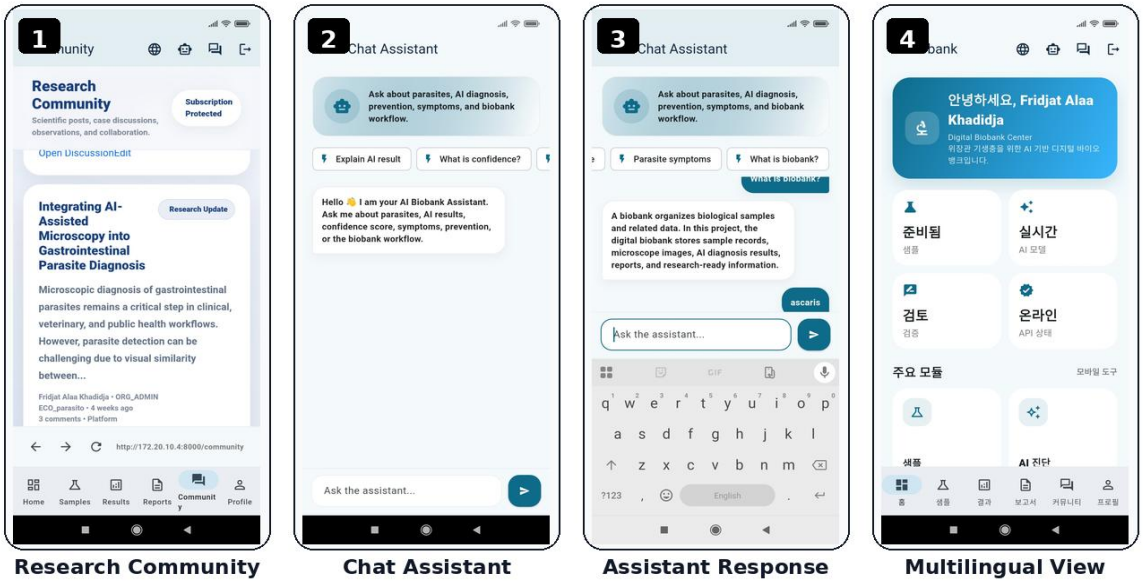


Figure 4.17: Mobile Community and Assistant Interfaces

4.8 Conclusion

This chapter presented the implementation of the proposed AI-powered digital biobank platform for gastrointestinal parasite samples. It started by describing the development environment and the main technologies used, including Laravel, MySQL, FastAPI, Python, PyTorch, and Flutter. The general system architecture was then presented, showing the interaction between the web platform, mobile application, database, and AI diagnosis service.

The database design and implementation were also described through the conceptual data model and its translation into relational tables. After that, the chapter presented the main web platform interfaces, including the public home page, authentication, dashboards, sample management, diagnosis results, reports, epidemiological spread map, research community, and learning center.

The chapter also explained the integration of the AI diagnosis service using FastAPI, which connects the Laravel platform with the trained deep learning model. This integration allows microscopic images to be analyzed and the predicted parasite class with confidence score to be stored and reused in results, reports, and epidemiological visualization. In addition, the mobile application was presented as a complementary tool that provides smartphone access to the main services of the platform.

Finally, the chapter discussed the main security and access control mechanisms, including authentication, user approval, role-based access, organization control, and subscription protection. Overall, this chapter demonstrated how the proposed methodology was translated into a functional prototype combining sample management, AI-assisted diagnosis, structured data storage, reporting, collaboration, learning, and mobile accessibility.

Chapter 5:

Entrepreneurial and Business Study of the AI-Powered Biobank Project

5.1 Introduction

This chapter presents the entrepreneurial approach of the AI-powered biobank project and its potential development from a scientific idea into an innovative and applicable startup initiative. The project responds to the need for better organization of biological samples, microscopic images, and diagnostic data related to gastrointestinal parasites.

The proposed solution starts with an intelligent digital platform that supports data management, AI-assisted image analysis, and diagnostic follow-up. This platform represents a first step toward the future establishment of a real biobank specialized in gastrointestinal parasites. This chapter also discusses the main entrepreneurial aspects of the project, including the problem, value proposition, beneficiaries, business model, costs, and legal and ethical considerations.

5.2 General Presentation of the AI Biobank Project

The AI-powered biobank project aims to create an organized digital platform specialized in gastrointestinal parasites. It allows sample registration, storage of medical data and microscopic images, follow-up of diagnostic results, and access to information for physicians, researchers, laboratories, and healthcare institutions.

This platform represents the first step toward establishing a real biobank in the future, combining biological sample preservation, medical data management, and scientific research support. The integration of artificial intelligence helps analyze microscopic images and assist specialists in diagnosis, while respecting legal and ethical requirements.

5.3 Project Objectives

The main objective of this project is to design and develop an AI-powered digital biobank platform dedicated to gastrointestinal parasitic diseases, as a first step toward the future establishment of a real biobank infrastructure. The project aims to improve the organization of biological sample data, support microscopic image analysis, facilitate AI-assisted diagnosis, and contribute to scientific research and public health monitoring.

5.3.1 Short-term objectives

1. Develop a self-funded digital platform for the registration, organization, and management of data related to gastrointestinal parasite samples.
2. Integrate an artificial intelligence model capable of assisting in the analysis of microscopic images and supporting the preliminary identification of parasite species.
3. Create a structured database linking sample information, microscopic images, diagnostic results, and user roles such as physicians, laboratories, researchers, and administrators.
4. Test the platform through an initial prototype in order to demonstrate its technical feasibility and practical usefulness in the management of parasitic disease data.

5.3.2 Medium-term objectives

1. Improve the platform by adding advanced services such as digital reports, case comparison, dashboard visualization, and epidemiological data organization.

2. Extend collaboration with medical laboratories, university hospitals, and research institutions to enrich the database and validate the system in real healthcare contexts.
3. Provide training and support for healthcare professionals and researchers on the use of AI-assisted diagnosis and digital sample management.
4. Use the digital platform as an organizational and technological foundation for preparing the creation of a real biobank specialized in gastrointestinal parasites.

5.3.3 Long-term objectives

1. Establish a real biobank infrastructure for the collection, preservation, and management of biological samples related to gastrointestinal parasitic diseases.
2. Contribute to improving diagnosis, scientific research, and epidemiological monitoring through the combination of biological resources, digital data, and artificial intelligence.
3. Develop strategic partnerships with healthcare institutions, public health authorities, universities, and research centers to ensure the sustainability and expansion of the project.
4. Position the project as a pioneering national initiative in AI-powered biobanking for parasitic diseases, with potential future extension to other infectious diseases.

5.4 Problem Identification and Market Need

Gastrointestinal parasites remain an important health problem, especially in areas with limited diagnostic resources. Their diagnosis still depends mainly on microscopic examination, which requires expertise and may lead to delays or interpretation errors.

In addition, data, microscopic images, and biological samples related to these parasites are often scattered between laboratories and healthcare institutions, without a unified system for storage, tracking, sharing, or research use. The absence of a specialized biobank also limits the availability of organized resources for physicians and researchers.

Therefore, there is a need for an intelligent digital solution that organizes parasite-related data and samples, supports diagnosis through artificial intelligence, and facilitates their use in healthcare and scientific research.

5.5 Entrepreneurial Opportunity Derived from the Identified Need

Based on the identified needs, the project represents a real entrepreneurial opportunity for developing a biobank specialized in gastrointestinal parasites. This opportunity is justified by the lack of specialized infrastructure in this field and the need for organized resources that support diagnosis, research, and epidemiological studies.

The project starts with an intelligent digital platform for managing sample data, microscopic images, and diagnostic follow-up. However, this platform is only the first step toward the creation of a real biobank that combines biological sample preservation, medical data management, and support for healthcare and research institutions.

Thus, the project is not limited to a digital solution, but represents an innovative initiative that combines biology, digital technology, and artificial intelligence to answer a real medical and scientific need.

5.6 Proposed Solution: Product and Services

The proposed solution can be better understood through the following workflow, which summarizes the main steps of the AI-powered biobank system, from sample collection to AI-assisted diagnosis and expert validation.



Figure 5.1: AI-Powered Biobank Workflow for Gastrointestinal Parasites

This figure illustrates the proposed workflow of the AI-powered biobank for gastrointestinal parasites. It shows how biological samples are collected, linked to digital data and microscopic images, analyzed using artificial intelligence, and finally validated by specialists before generating diagnostic reports.

5.7 Value Proposition

The value proposition of the AI-powered biobank project lies in offering an integrated solution that combines biological sample management, medical data organization, microscopic image analysis, and artificial intelligence to improve the diagnosis and study of gastrointestinal parasites.

The project provides a specialized infrastructure for linking samples, clinical data, microscopic images, and diagnostic results within a secure and organized system. Through AI-assisted analysis, it supports specialists in identifying parasite species and generating preliminary results, while maintaining expert validation.

This solution serves physicians, laboratories, researchers, and healthcare institutions by improving result traceability, supporting scientific research, enabling epidemiological studies, and contributing to the development of a future real biobank. Therefore, the project creates added value by combining digital technology, artificial intelligence, and biological resource preservation in a scalable and sustainable framework.

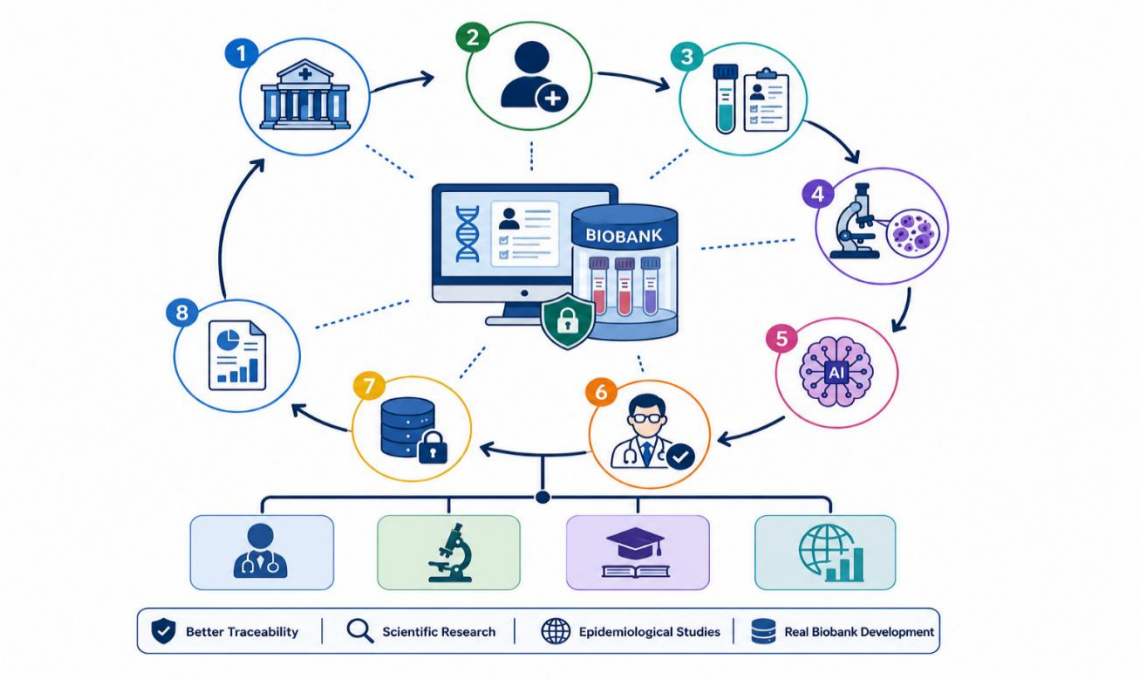


Figure 5.2: Proposition of the AI-Powered Biobank Project

5.8 Target Market and Beneficiaries

The AI-powered biobank project targets several actors in the healthcare and research sectors because it combines biological sample preservation, medical data management, and artificial intelligence-assisted diagnosis. The main beneficiaries are those involved in the diagnosis, study, and monitoring of gastrointestinal parasitic diseases.

- **Physicians and diagnostic specialists:** They can benefit from AI-assisted microscopic image analysis, which may help accelerate the diagnostic process and support more accurate medical decision-making.
- **Medical laboratories and analysis centers:** These institutions can use the platform to organize sample data, store microscopic images, follow diagnostic results, and generate clear and traceable digital reports.
- **Researchers, universities, and research centers:** The project provides structured biological and digital resources that can support scientific studies, epidemiological research, and the development of more accurate artificial intelligence models.
- **Veterinarians and veterinary laboratories:** They can also benefit from the platform when dealing with parasite-related samples of animal origin or zoonotic parasitic diseases. This can support the monitoring of parasites that may circulate between animals and humans and contribute to better collaboration between medical and veterinary sectors.
- **Healthcare institutions and hospitals:** They can benefit from improved management of medical information, better follow-up of cases, and access to organized diagnostic data within a secure digital environment.
- **Public health authorities:** Organized data generated through the project can support epidemiological monitoring, analysis of parasite distribution, and evidence-based public health decision-making.

- **Patients indirectly:** Although patients are not the direct users of the biobank, they can benefit indirectly through faster diagnosis, early detection, and improved quality of healthcare services.

Thus, the project addresses a broad healthcare and research ecosystem. This diversity of beneficiaries strengthens its potential for implementation, expansion, and long-term impact.

5.9 Competitive Advantage and Innovative Aspects

- Integrated system combining sample management, medical data, microscopic images, and AI-assisted diagnosis.
- Specialized focus on gastrointestinal parasitic diseases.
- Structured database linking samples, images, results, and clinical information.
- Support for both medical diagnosis and scientific research.
- Progressive development from a digital platform to a real biobank.
- AI used as a decision-support tool, not a replacement for specialists.
- Expert validation to ensure reliability of diagnostic results.
- Potential support for epidemiological monitoring and public health decision-making.
- Possibility of collaboration with hospitals, laboratories, universities, and research institutions.
- Innovative value through the combination of digital health, artificial intelligence, and biological resource management.

5.10 Strategic Market Analysis

The strategic market analysis aims to identify the main market segments targeted by the AI-powered biobank project and to evaluate the competitive environment in which the project may evolve. Since the project combines digital health, artificial intelligence, biological sample management, and future biobank infrastructure, it addresses a specific healthcare and research need related to gastrointestinal parasitic diseases.

5.10.1 Market Segmentation

The market of the AI-powered biobank project can be divided into several main segments according to their needs and level of involvement.

The priority segment includes medical laboratories, parasitology units, and university hospitals, as they are directly involved in microscopic diagnosis, sample handling, and result validation.

The research segment includes universities, research centers, and academic laboratories that may use the organized database and biological resources for scientific studies and AI model development.

The institutional segment includes hospitals and healthcare institutions that can benefit from sample follow-up, organized diagnostic data, and improved information management.

The public health segment includes health authorities that may use aggregated data for epidemiological monitoring and decision-making.

Finally, the future extension segment includes institutions interested in biological sample preservation, research collaboration, and the development of a real biobank infrastructure.

This segmentation shows that the project can first target laboratories and university hospitals, then gradually expand toward research institutions, healthcare organizations, and public health authorities.

5.10.2 Market Competition Analysis

The competition analysis identifies the existing and potential actors that may offer similar or alternative services. Since the project combines digital health, artificial intelligence, biological sample management, and future biobank development, competitors can be divided into direct and indirect competitors.

5.10.2.1 Direct Competitors

Direct competitors may include medical laboratories, hospital parasitology units, research laboratories, and biological sample storage structures. These actors already have experience in microscopic diagnosis, sample handling, and medical reporting.

However, most of them focus mainly on conventional diagnosis or internal sample storage. They generally do not provide an integrated system that combines sample preservation, digital data management, microscopic image analysis, and AI-assisted diagnostic support.

5.10.2.2 Indirect Competitors

Indirect competitors include digital health platforms, laboratory information systems, medical databases, and AI-based diagnostic tools. These solutions may support data storage or image analysis, but they are usually not specialized in gastrointestinal parasites or connected to a real biobank strategy.

Traditional microscopic diagnosis can also be considered an indirect competitor, since it remains widely used. However, it depends strongly on expert interpretation and does not offer integrated sample traceability, data sharing, epidemiological monitoring, or AI model development.

Some hospitals, public health institutions, and university laboratories may also develop internal systems. Nevertheless, they can also become strategic partners for sample collection, expert validation, and future biobank development.

5.10.2.3 Identified Strengths and Weaknesses of Competitors Strengths

Existing competitors possess several positive characteristics that have enabled them to establish a strong presence in the healthcare and laboratory sectors. Their experience in parasitological diagnosis, qualified personnel, direct access to patients and samples, and existing laboratory infrastructure represent valuable assets. These advantages contribute significantly to the delivery of diagnostic services and demonstrate the important role that healthcare institutions and laboratories play in managing parasitic diseases, including:

- Experience in conventional parasitological diagnosis.
- Availability of qualified healthcare and laboratory specialists.
- Direct access to patients, samples, and clinical data.
- Institutional credibility in the healthcare sector.
- Existing laboratory infrastructure.

However, despite these strengths, competitors also present several weaknesses that limit their ability to fully exploit medical and biological data for diagnosis, research, and long-term data management., such as:

- Fragmented data storage between laboratories and healthcare institutions.
- Limited sample traceability and research reuse.
- Absence of a structured biobank-oriented system.

- Dependence on manual microscopic diagnosis.
- Limited scalability for building large AI training datasets.

Overall, the competitive advantage of the proposed project lies in its integrated approach. Unlike conventional laboratories or general digital platforms, it connects sample management, medical data organization, microscopic image analysis, AI-assisted diagnosis, and future biological sample preservation within one structured system.

5.11 Production and Organization Plan

The production and organization plan explains how the AI-powered biobank project will operate in practice. Since the project provides digital and scientific services, the production process refers to the management of samples, microscopic images, diagnostic results, and medical data within the platform.

5.11.1 Operational Workflow

The operational workflow of the platform is organized through successive steps to ensure sample traceability, structured data management, and result validation. First, healthcare institutions, laboratories, research centers, or university hospitals register on the platform. Then, user accounts are created according to defined roles, such as platform administrator, organization administrator, physician, researcher, or laboratory member. After that, authorized users enter sample information, including the sample type, collection date, geographical origin, and related medical data. A microscopic image is then uploaded and linked to the corresponding sample record. The AI model analyzes the image and provides a preliminary result, including the predicted parasite species and confidence score. This result is reviewed by a physician or specialist, who validates, corrects, or completes it when necessary. Once validated, the result is stored and linked to the sample, image, and medical information. Finally, the platform generates digital reports and organizes the collected data for research, epidemiological monitoring, and public health use.

The workflow can be summarized as follows:



Figure 5.3: Operational Workflow of the AI-Powered Biobank Platform

5.11.2 Organization of Roles

The organization of roles ensures secure access, clear responsibilities, and reliable platform management.

- **Platform Administrator:** Manages the general platform, users, institutions, subscriptions, and system settings.
- **Organization Administrator:** Manages the account of a specific institution and supervises its internal users and activities.
- **Physician / Diagnostic Specialist:** Reviews and validates AI-assisted diagnostic results.
- **Laboratory Staff:** Registers samples, uploads microscopic images, and enters related data.
- **Researcher:** Uses authorized and anonymized data for scientific studies, epidemiological analysis, and AI model improvement.
- **AI System:** Analyzes microscopic images and provides preliminary diagnostic suggestions as a decision-support tool.

5.11.3 Future Real Biobank Organization

In the long term, the digital platform will serve as the foundation for a real biobank specialized in gastrointestinal parasites. This future phase will include the physical preservation of biological samples under scientific, ethical, and legal conditions.

The future biobank organization will include:

- **Reception of biological samples:** Samples will be received from partner laboratories, hospitals, or research institutions.
- **Sample coding and identification:** Each sample will receive a unique code to ensure traceability and confidentiality.
- **Scientific preservation of samples:** Samples will be stored under appropriate biological preservation conditions.
- **Linking samples with digital data:** Each sample will be connected to its microscopic images, diagnostic results, collection data, and authorized clinical information.
- **Ethical and legal compliance:** The storage, use, and sharing of samples and data will respect consent, confidentiality, and applicable regulations.

Thus, the project begins with a digital platform for data and diagnostic workflow management, then gradually evolves toward a real biobank infrastructure supporting healthcare, research, and epidemiological monitoring.

5.12 Legal, Ethical, and Regulatory Framework

The legal, ethical, and regulatory framework of the AI-powered biobank is essential to ensure the secure and responsible management of medical data, microscopic images, diagnostic results, and future biological samples. The project must protect patient confidentiality through secure authentication, role-based access, anonymization, and controlled data sharing. In future phases, the collection and use of biological samples should be based on informed consent, clear governance, traceability, and compliance with applicable regulations. Moreover, the AI system should remain a decision-support tool, with final diagnosis validated by qualified specialists, in order to guarantee ethical use, scientific reliability, and long-term trust in the biobank infrastructure.

5.13 Financial Plan

The financial plan presents the main financial requirements needed to launch the AI-powered biobank project as a startup activity. It includes the estimated initial investment, possible funding sources, projected revenues, expected expenses, working capital management strategies, and the cash flow plan for the first year.

5.13.1 Costs and Expenses

The launch of the AI-powered biobank startup requires both variable and fixed costs. These expenses are related to the deployment of the digital platform, artificial intelligence services, data storage, communication activities, administrative procedures, and the preparation of future biobank-oriented services.

5.13.1.1 Initial Investment Costs

The following table summarizes the estimated initial costs required to launch the project during the first operational year.

Type of Expense	Expense Category	Description	Estimated Amount (DA)
Variable Cost	Office Rent	Monthly rent for a small startup office: 35,000 × 12 months	420,000
Variable Cost	Staff Salaries	Salaries for a small core team: IT developer, AI engineer, biomedical/biobank coordinator, : (3 full-time × 50,000 × 12)+(2 part-time × 35,000 × 12)=2,640,000 DA	2,640,000
Variable Cost	Hosting and Cloud Services	Server, database hosting, AI API deployment, and secure storage for microscopic images and reports	300,000
Variable Cost	AI Computing and Software Tools	Cloud computing, AI testing tools, development environments, and model improvement	250,000
Variable Cost	Field Visits and Validation	Transport and meetings with laboratories, hospitals, and research institutions for testing and validation	300,000
Variable Cost	Operational Expenses	Internet, electricity, office supplies, communication, and basic daily expenses: 25,000 × 12 months	300,000
Variable Cost	Marketing and Communication	Posters, social media promotion, institutional presentations, and participation in scientific/professional events : 30,000 × 12 months	360,000
Fixed Cost	Office Furniture and Equipment	Desks, chairs, shelves, and basic office setup	100,000
Fixed Cost	IT Equipment	Laptops, printer, external storage, accessories, and testing devices	450,000

Fixed Cost	Administrative Fees	Legal creation, official documents, stamps, bank account setup, and administrative procedures	80,000
Fixed Cost	Visual Identity	Logo, brochures, presentation templates, and branding materials	50,000
Fixed Cost	Cybersecurity and Backup Setup	SSL certificate, backup configuration, security tools, and basic data protection setup	120,000
Fixed Cost	Maintenance	Equipment and workspace maintenance during the first year	80,000

Table 5.1: Estimated Initial Investment for the AI-Powered Biobank Project

"Total estimated initial investment: 5,450,000 DA"

This amount represents an estimated launch budget for transforming the AI-powered biobank platform into a startup activity. The values may vary depending on hosting choices, equipment quality, staff size, and partnership conditions.

5.13.1.2 Human Resources and Salary Plan

The launch of the AI-powered biobank startup requires a small multidisciplinary team capable of managing the technical, artificial intelligence, biomedical, administrative, and medical validation aspects of the project. Since the project combines a digital platform, AI-assisted diagnosis, medical data management, and future biological sample preservation, the team must include both full-time and part-time profiles.

During the first operational year, the team is organized as three full-time employees and two part-time collaborators. This structure allows the startup to cover the essential needs of the project while controlling salary expenses during the launch phase.

Position	Work Type	Main Role	Number	Estimated Monthly Salary (DA)	Estimated Annual Cost (DA)
IT Developer	Full-time	Maintains and improves the web platform, database, dashboards, user modules, and technical features	1	50,000	600,000
AI Engineer / Data Scientist	Full-time	Improves the AI model, microscopic image analysis pipeline, FastAPI service, and model performance	1	50,000	600,000
Biomedical / Biobank Coordinator	Full-time	Coordinates sample-related data, diagnostic workflow, future biobank preparation, and communication with laboratories	1	50,000	600,000
Administrative and Marketing Assistant	Part-time	Manages administrative tasks, communication, customer contact, documentation, promotion, and institutional follow-up	1	35,000	420,000
Medical / Parasitology Specialist	Part-time	Reviews AI-assisted diagnostic results, supports expert validation, and provides medical and scientific supervision	1	35,000	420,000

Table 5.2: Estimated Staffing Plan and Annual Salary Costs for the AI-Powered Biobank Project

This team represents a realistic human resource structure for the first year of activity. The three full-time profiles ensure the continuous technical and biomedical operation of the platform, while the two part-time profiles provide administrative support and medical validation without creating excessive salary pressure during the startup launch phase.

As the project evolves toward the establishment of a real biobank infrastructure, additional staff may be progressively recruited, such as laboratory technicians, data managers, biosafety officers, quality control supervisors, and full-time medical or parasitology specialists. These additional positions will depend on the growth of the project, the number of partner institutions, and the volume of biological samples and digital data managed by the biobank.

5.13.2 Funding Sources

The initial investment required to launch the project may be secured through a combination of complementary funding sources:

- **Startup support programs under Law 1275:** Partial financing may be requested through university incubators, innovation support structures, or national entrepreneurship programs dedicated to startup-oriented graduation projects.
- **Bank loans or public-private partnerships:** External funding may be considered through financial institutions, development-oriented public programs, or partnerships with healthcare and research institutions.
- **Personal and family contribution:** A part of the total investment may be covered through personal and family funds, especially during the first phase of platform deployment and administrative preparation.
- **Early revenues:** Income generated from first subscriptions, pilot diagnostic services, training sessions, or collaboration agreements may contribute to financing early operations.
- **Research and institutional partnerships:** Partnerships with universities, laboratories, hospitals, or public health institutions may support the validation and future development of the platform and the real biobank infrastructure.

To clarify the financing structure of the project, the following table presents a possible distribution of funding sources required to cover the updated estimated initial investment.

Funding Source	Estimated Amount (DA)	Percentage	Explantation
Personal and family contribution	1,150,000	21.1%	Initial contribution used to support the first expenses related to project preparation, administrative procedures, communication, and early deployment activities.
Startup support programs under Law 1275	1,700,000	31.2%	Possible support from university incubators, innovation programs, or national entrepreneurship mechanisms dedicated to startup-oriented graduation projects.
Bank loan or partnership support	1,800,000	33.0%	External financing that may be obtained through a bank loan, public-private partnership, or institutional collaboration with healthcare or research organizations.

Early revenues and prepaid services	800,000	14.7%	Initial income that may come from early subscriptions, prepaid AI-assisted diagnosis packages, training sessions, or research collaboration services.
Total Financing	**5,450,000**	**100%**	This amount covers the updated estimated initial investment required to launch the AI-powered biobank startup.

Table 5.3: Proposed Financing Plan for the AI-Powered Biobank Project

This financing structure shows that the project does not depend on a single source of funding. Instead, it combines personal and family contribution, startup support mechanisms, potential external financing, and early revenues. This diversified approach reduces financial risk and makes the launch of the project more realistic and progressive.

5.13.3 Revenue

The projected revenue of the AI-powered biobank startup is structured around several service categories related to digital health, AI-assisted diagnosis, training, and research collaboration.

The main revenue sources are:

- **Institutional subscription plans:** annual subscriptions for laboratories, clinics, university hospitals, research centers, and healthcare institutions using the platform.
- **AI-assisted diagnosis services:** paid microscopic image analysis and diagnostic support using the integrated artificial intelligence model.
- **Digital report generation:** structured diagnostic reports linked to sample data, microscopic images, and AI-assisted results.
- **Training courses:** paid training sessions for healthcare professionals, laboratory staff, students, and researchers on AI-assisted diagnosis and digital sample management.
- **Research collaboration services:** collaboration with universities and research institutions using authorized and anonymized data for scientific studies and AI model development.

To estimate the financial potential of the first year, two projection scenarios are proposed: an optimistic scenario and a pessimistic scenario.

Service Type	Price per Service (DA)	Estimated Quantity Optimistic	Total Revenue Optimistic (DA)	Estimated Quantity Pessimistic	Total Revenue Pessimistic (DA)
Institutional Subscription	250,000 / year	12 institutions	3,000,000	6 institutions	1,500,000
AI-Assisted Diagnosis Package	20,000 / package	70 packages	1,400,000	35 packages	700,000
Digital Report Service	5,000 / report package	100 packages	500,000	40 packages	200,000
Training Sessions	100,000 / session	8 sessions	800,000	4 sessions	400,000

Research Collaboration	300,000 / project	4 projects	1,200,000	1 project	300,000

Scenario	Total Projected Revenue (DA)
Optimistic Scenario	6,900,000
Pessimistic Scenario	3,100,000
Average Estimated Annual Revenue	5,000,000

Table 5.4: Estimated Annual Revenue Forecast for the AI-Powered Biobank Project

5.13.4 Projected Income Statement

The projected income statement summarizes the estimated annual revenue, annual expenses, and expected financial result during the first year of activity.

Item	Amount (DA)
Estimated Annual Revenue	5,000,000
Estimated Annual Operating Expenses	4,750,000
Projected Net Result	+250,000

Table 5.5: Estimated First-Year Financial Result of the AI-Powered Biobank Project

The projected result remains positive in the average scenario, although the profit margin becomes more limited due to the integration of two additional part-time profiles. This remains realistic for the first year of activity, as the startup is still in its launch phase and must invest in human resources, platform deployment, customer acquisition, and institutional validation.

5.13.5 Strategies for Reducing the Working Capital Requirement (WCR / BFR)

- Request partial prepayment before starting subscriptions or services.
- Encourage annual or semi-annual subscription plans.
- Offer prepaid AI-assisted diagnostic packages.
- Reduce delayed payments by setting clear payment deadlines.
- Limit physical stock during the first digital phase.
- Avoid unnecessary laboratory equipment before launching the real biobank.
- Start with a small and flexible team to control salary costs.
- Use part-time profiles, when possible, to reduce fixed expenses.
- Optimize hosting, storage, and maintenance costs.
- Use digital reports instead of paper-based documentation when possible.
- Build partnerships with laboratories, hospitals, and universities.
- Share some costs through collaboration and institutional support.
- Use partner laboratories for sample collection and expert validation.
- Develop the real biobank progressively after validating the digital platform.
- Reinvest early revenues in platform improvement and future biobank preparation.

- Maintain stable cash flow during the first year of activity.

5.13.6 Cash Flow Plan – Year 1

As part of the launch of the AI-powered biobank startup, a 12-month cash flow plan is proposed to track monthly financial movements. This plan includes expected revenues, expenses, investments, financing, and total monthly cash flow.

Month	Activity Level	Monthly Analysis	Operating Activities (DA)	Investments (DA)	Financing (DA)	Total Monthly Cash Flow (DA)
January	Average	Administrative preparation, platform organization, and initial project launch	100,000	-200,000	300,000	+200,000
February	Average	First contacts with laboratories, hospitals, and research institutions	120,000	0	0	+120,000
March	High	Platform demonstration and first pilot testing activities	250,000	0	0	+250,000
April	High	First subscription negotiations and AI diagnosis demonstrations	300,000	-100,000	0	+200,000
May	Medium	Preparation of training materials and technical documentation	180,000	-50,000	0	+130,000
June	High	Training sessions and diagnostic package testing	400,000	0	0	+400,000
July	Medium	Maintenance, user support, and improvement of platform modules	250,000	-80,000	0	+170,000
August	Low	Reduced activity due to holiday period, focus on technical updates	100,000	0	0	+100,000
September	High	New academic and institutional collaborations	500,000	-100,000	0	+400,000
October	High	Expansion of subscriptions and report generation services	600,000	0	0	+600,000
November	High	Research collaboration and AI model improvement	450,000	-120,000	0	+330,000
December	High	Annual evaluation and preparation for next development phase	700,000	0	0	+700,000

Table 5.6: Estimated Monthly Cash Flow Forecast for the AI-Powered Biobank Project

"Estimated annual cash flow balance: +3,600,000 DA"

This cash flow plan indicates that the project can maintain positive cash flow if early revenues are progressively generated through subscriptions, diagnostic services, training, and research collaborations.

5.13.7 Financial Sustainability

The financial sustainability of the AI-powered biobank project depends on combining scientific value with economic viability. In the short term, the digital platform can generate revenue through subscriptions, AI-assisted diagnosis, digital reports, and training sessions, which can support hosting, maintenance, security, and platform improvement.

In the long term, these revenues can help prepare the transition toward a real biobank specialized in gastrointestinal parasites, requiring investment in biological sample preservation, laboratory equipment, biosafety, and governance.

The addition of part-time profiles may reduce the first-year net result, but it strengthens the credibility of the project by ensuring administrative support and medical validation, which is important for building trust with laboratories and research partners.

5.14 Prototype of the Proposed Solution

Since the detailed technical implementation and platform interfaces were already presented in Chapter 4, this section focuses on the prototype from an entrepreneurial perspective. The objective is not to repeat the technical description of each interface, but to show how the developed prototype can represent a concrete starting point for a future startup and for the progressive creation of a real AI-powered biobank specialized in gastrointestinal parasitic diseases.

5.14.1 Prototype Overview

The proposed prototype is an AI-powered digital biobank platform designed to manage data, microscopic images, diagnostic results, and user activities related to gastrointestinal parasite samples. It represents the first functional version of the project and demonstrates the feasibility of the proposed solution.

The prototype includes a web platform, a mobile application, a database, and an AI module connected through an API. These components support the management of institutions, users, samples, AI-assisted diagnosis, reports, and collaboration spaces. From an entrepreneurial perspective, it can be considered a Minimum Viable Product (MVP), as it provides the essential functions needed to test the service with laboratories, physicians, researchers, and healthcare institutions.

5.14.2. Main Functional Modules of the Prototype

The prototype is composed of several functional modules that contribute to the operational and commercial value of the project.

Module	Main Function	Entrepreneurial Value
User and organization management	Allows the creation and management of users, roles, and healthcare or research institutions	Enables the platform to serve multiple laboratories, hospitals, and research organizations

Sample management	Allows the registration of parasite-related samples with associated information such as type, date, geographical origin, and status	Improves sample traceability and creates a structured digital workflow
Microscopic image upload	Allows users to upload microscopic images linked to registered samples	Provides the basic input for AI-assisted diagnosis and digital archiving
AI-assisted diagnosis	Analyzes microscopic images and provides a preliminary parasite classification with a confidence score	Creates a value-added diagnostic support service
Expert validation	Allows specialists to review, validate, or correct AI-generated results	Ensures medical reliability and maintains the role of human expertise
Results management	Stores diagnostic results and links them to samples, images, and users	Enables follow-up, consultation, and reuse of diagnostic data
Digital reports	Generates structured reports related to samples and diagnostic results	Provides a professional service for laboratories and healthcare institutions
Community module	Allows communication and collaboration between researchers and professionals	Supports scientific collaboration and knowledge sharing
Administration and subscription management	Supports platform supervision, institutional access, and service plans	Provides a basis for the business model and future revenue generation
Mobile application	Gives users mobile access to key services such as samples, results, reports, and community features	Improves accessibility and supports field or laboratory use

Table 5.7: Prototype Modules, Main Functions, and Entrepreneurial ValueTable

This modular structure shows that the prototype is not limited to a simple diagnostic tool. It is designed as an integrated digital ecosystem capable of supporting diagnosis, data organization, research collaboration, and future biobank development.

5.14.3 Integrated Prototype Components and Future Development

The developed prototype combines several integrated components that demonstrate the feasibility of the proposed AI-powered biobank solution and its potential evolution toward a future real biobank. The prototype includes a web platform for managing users, organizations, samples, diagnostic results, reports, and administrative activities; a mobile application that improves accessibility to platform services; an AI-assisted diagnosis module for microscopic image analysis and preliminary parasite identification; and a structured database that ensures traceability and efficient organization of medical and biological information. Together, these components support the complete workflow of sample registration, image analysis, expert validation, result storage, and report generation. In addition, the prototype establishes the digital and organizational foundations required for the future implementation of a real biobank dedicated to gastrointestinal parasites. Although the current system demonstrates the technical feasibility of the concept, future improvements remain necessary, including collaboration with partner laboratories, dataset expansion, cybersecurity enhancement, legal and ethical compliance, biological sample preservation integration, and continuous AI model improvement.

5.14.4. Materials and Cost Estimation

This section presents the main materials and resources required for the AI-powered biobank project. Since the project is developed progressively, the required materials are divided into two main categories: materials needed for the digital platform phase, and materials required for the future establishment of a real biobank infrastructure specialized in gastrointestinal parasites.

5.14.4.1. Materials and Equipment for the Digital Platform Phase

The first phase of the project is mainly based on a digital infrastructure. It requires computing equipment, hosting services, secure data storage, and technical tools to ensure the development, deployment, and maintenance of the AI-powered biobank platform.

Material / Equipment	Purpose	Estimated Cost (DA)
Laptops / Development computers	Development of the web platform, mobile application, database, and AI model testing	300,000 – 450,000
External storage devices	Backup of microscopic images, datasets, reports, and project files	30,000 – 80,000
Web hosting / VPS server	Deployment of the web platform, database, and AI API	80,000 – 150,000 / year
Cloud or GPU computing resources	AI model training, testing, and improvement	100,000 – 250,000 / year
Domain name and SSL certificate	Secure access to the platform and official online presence	15,000 – 40,000 / year
Backup and cybersecurity tools	Protection of medical data, user accounts, and platform files	80,000 – 150,000
Printer and office accessories	Printing documents, reports, administrative files, and communication materials	50,000 – 120,000
Internet and communication tools	Platform testing, data transfer, online meetings, and institutional communication	50,000 – 100,000 / year

Table 5.8: Digital Platform Materials, Purpose, and Estimated Costs

"Estimated cost for the digital platform materials: 705,000 – 1,340,000 DA"

These materials support the deployment and operation of the digital prototype. They are essential for managing users, samples, microscopic images, AI-assisted diagnosis, digital reports, and secure data storage.

5.14.4.2. Materials and Resources for the Future Real Biobank

The long-term objective of the project is to evolve from a digital platform toward a real biobank infrastructure. This future phase will require biological sample preservation equipment, laboratory consumables, coding systems, biosafety materials, and ethical documentation.

Material / Resource	Purpose	Estimated Cost (DA)

Laboratory space preparation	Preparation of a dedicated area for sample reception, preparation, and storage	500,000 – 1,000,000
Refrigerators and freezers	Preservation of biological samples under controlled conditions	700,000 – 2,500,000
Advanced freezer system	Long-term preservation of selected biological samples when needed	1,500,000 – 4,000,000
Temperature monitoring system	Continuous monitoring of storage conditions and sample safety	150,000 – 500,000
UPS / backup power system	Protection of storage equipment in case of power interruption	100,000 – 400,000
Sample coding system	Barcode or QR code printer, scanner, and labels for sample traceability	80,000 – 250,000
Sample storage boxes and racks	Organized storage of biological samples	100,000 – 300,000
Sterile containers and tubes	Collection and preservation of stool samples or parasite-related biological material	150,000 – 500,000
Microscopic slides and cover slips	Preparation and archiving of microscopic observations	50,000 – 150,000
Gloves, masks, lab coats, and disinfectants	Protection of staff and safe handling of biological samples	100,000 – 300,000
Biosafety equipment	Safe manipulation, waste management, and infection risk reduction	300,000 – 1,000,000
Microscope with camera system	Acquisition and documentation of microscopic images for diagnosis and AI training	300,000 – 900,000
Legal and ethical documentation	Consent forms, sample use agreements, data protection policies, and governance documents	50,000 – 200,000

Table 5.9: Future Real Biobank Materials, Purpose, and Estimated Costs

"Estimated cost for future real biobank materials: 4,080,000 – 12,000,000 DA"

This cost estimation concerns the future physical biobank phase and is not required for the initial digital prototype. It represents the resources needed to move from a digital biobank-oriented platform toward a real biological sample preservation infrastructure.

5.14.4.3. Biological and Digital Resources Needed to Fill the Real Biobank

In addition to equipment, the real biobank will require biological and digital resources collected through partnerships with laboratories, hospitals, and research institutions.

Resource	Description	Purpose

Biological samples	Stool samples and parasite-positive samples collected under ethical and legal conditions	To build the physical biological collection
Reference microscopic slides	Slides containing identified parasite eggs, cysts, or larvae	To support diagnosis, training, and quality control
Microscopic Image datasets	Images of gastrointestinal parasites and negative samples	To improve AI models and support research
Associated medical data	Sample type, collection date, geographical origin, diagnostic result, and authorized clinical information	To link each sample with structured digital information
Expert-validated results	Diagnostic results reviewed and validated by specialists	To ensure reliability of the database and AI training data
Consent and governance records	Ethical approvals, consent forms, and sample use permissions	To ensure responsible and legal use of samples and data

Table 5.10: Key Resources, Description, and Purpose for the Future Real Biobank

These resources will allow the future biobank to connect physical biological samples with digital records, microscopic images, diagnostic results, and research data. This connection is essential for traceability, scientific reliability, and long-term use of the biobank.

5.14.4.4. Summary of Estimated Material Costs

Category	**Estimated Cost (DA)**
Digital platform materials and equipment	705,000 – 1,340,000
Future real biobank materials and equipment	4,080,000 – 12,000,000
Total estimated material cost	**4,785,000 – 13,340,000 DA**

Table 5.11: Estimated Material Costs for the Digital Platform and Future Real Biobank

The estimated costs presented in this section are indicative and may vary according to market prices, equipment quality, supplier availability, storage capacity, and partnership conditions. The digital platform phase can be launched with limited resources, while the real biobank phase requires higher investment and institutional collaboration.

5.15. Business Model Canvas

The Business Model Canvas summarizes the entrepreneurial model of the AI-powered biobank project. It presents the main elements required to transform the developed prototype into a sustainable startup activity, while preparing the future establishment of a real biobank specialized in gastrointestinal parasitic diseases.

Business Model Canvas Block	**Description for the AI Biobank Project**

Customer Segments	The project targets medical laboratories, parasitology units, university hospitals, healthcare institutions, research centers, universities, and public health authorities. In the long term, it may also serve institutions interested in biological sample preservation, epidemiological studies, and infectious disease research.
Value Proposition	The project provides an integrated AI-powered biobank-oriented platform that combines sample data management, microscopic image storage, AI-assisted diagnosis, digital reports, expert validation, and research support. It improves sample traceability, supports faster diagnostic decision-making, organizes parasite-related data, and represents a first step toward the creation of a real biobank specialized in gastrointestinal parasites.
Channels	The project can reach users through the web platform, mobile application, institutional presentations, university and hospital collaborations, scientific events, professional networks, social media communication, and partnerships with laboratories and healthcare institutions.
Customer Relationships	The relationship with customers is based on continuous technical support, user training, institutional follow-up, platform updates, service customization, and collaboration with healthcare and research professionals. Trust, confidentiality, and expert validation are essential elements for maintaining long-term relationships.
Revenue Streams	The project can generate revenue through institutional subscription plans, AI-assisted diagnosis services, digital report generation, paid training courses, research collaboration, and, in the future, controlled biobank services related to biological sample preservation and authorized data access.
Key Resources	The main resources include the web platform, mobile application, artificial intelligence diagnostic model, database system, microscopic image datasets, secure storage infrastructure, technical team, biomedical and medical expertise, partner laboratories, and future biological sample preservation infrastructure.
Key Activities	The key activities include platform development and maintenance, user and institution management, sample data organization, microscopic image analysis, AI model improvement, expert validation, report generation, data security management, training, research support, and preparation for the future real biobank infrastructure.
Key Partners	Potential partners include medical laboratories, university hospitals, universities, research centers, physicians and parasitology specialists, public health institutions, university incubators, technology providers, cloud and hosting services, and future partners for biological sample storage and laboratory equipment.
Cost Structure	The main costs include platform hosting, cloud services, secure data storage, cybersecurity, software maintenance, AI model improvement, staff salaries, communication and marketing, training materials, administrative costs, and future investment in laboratory equipment, sample preservation systems, biosafety tools, and biobank governance procedures.

Table 5.12: Business Model Canvas of the AI-Powered Biobank Project

The Business Model Canvas shows that the AI-powered biobank project is based on a progressive business model. In the first phase, value is created through the digital platform, AI-assisted diagnosis, digital reports, training services, and research collaboration. In the long term, this model can evolve toward a real biobank infrastructure, offering biological sample preservation and structured scientific resources. This confirms that the project has both healthcare value and entrepreneurial potential.

5.16 Conclusion

This chapter presented the entrepreneurial dimension of the AI-powered biobank project and its potential evolution from a scientific prototype into a startup initiative. The project proposes an intelligent digital platform

for managing samples, microscopic images, diagnostic results, and AI-assisted diagnosis of gastrointestinal parasites.

This platform represents a first practical step toward the future establishment of a real biobank. By combining digital health, artificial intelligence, biological resource management, and research support, the project offers an innovative and scalable solution that can contribute to improving diagnosis, supporting scientific research, and strengthening healthcare infrastructure.

Bibliography

[1] World Health Organization, “Soil-transmitted helminth infections,” World Health Organization, Jan. 18, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections. [Accessed: Jun. 15, 2026].

[2] P. J. Hotez, P. J. Brindley, J. M. Bethony, C. H. King, E. J. Pearce, and J. Jacobson, “Helminth infections: the great neglected tropical diseases,” Journal of Clinical Investigation, vol. 118, no. 4, pp. 1311–1321, Apr. 2008, doi: 10.1172/JCI34261.

[3] E. Christaki, “Classification of Parasitic Diseases,” in The Surgical Management of Parasitic Diseases, G. Tsoulfas, J. J. Hoballah, G. C. Velmahos, and Y.-H. Ho, Eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 23–45, doi: 10.1007/978-3-030-47948-0_2.

[4] R. C. A. Thompson and A. Smith, “Zoonotic enteric protozoa,” Veterinary Parasitology, vol. 182, no. 1, pp. 70–78, 2011, doi: 10.1016/j.vetpar.2011.07.016.

[5] K. Hajissa, M. A. Islam, A. M. Sanyang, and Z. Mohamed, “Prevalence of intestinal protozoan parasites among school children in Africa: A systematic review and meta-analysis,” PLOS Neglected Tropical Diseases, vol. 16, no. 2, p. e0009971, Feb. 2022, doi: 10.1371/journal.pntd.0009971.

[6] W. Checkley, A. C. White Jr., D. Jaganath, M. J. Arrowood, R. M. Chalmers, X.-M. Chen, et al., “A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for Cryptosporidium,” The Lancet Infectious Diseases, vol. 15, no. 1, pp. 85–94, Jan. 2015, doi: 10.1016/S1473-3099(14)70772-8.

[7] L. O. B. Andersen and C. R. Stensvold, “Blastocystis in health and disease: Are we moving from a clinical to a public health perspective?,” Journal of Clinical Microbiology, vol. 54, no. 3, pp. 524–528, Mar. 2016, doi: 10.1128/JCM.02520-15.

[8] GBD 2017 DALYs and HALE Collaborators, “Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017,” The Lancet, vol. 392, no. 10159, pp. 1859–1922, Nov. 2018, doi: 10.1016/S0140-6736(18)32335-3.

[9] W.-M. Sohn and J.-Y. Chai, “Practical guide for the diagnosis of helminth ova in stools,” Annals of Clinical Microbiology, vol. 27, no. 2, pp. 49–67, Jun. 2024, doi: 10.5145/ACM.2024.27.2.3.

[10] R. M. Maizels and H. J. McSorley, “Regulation of the host immune system by helminth parasites,” Journal of Allergy and Clinical Immunology, vol. 138, no. 3, pp. 666–675, Sep. 2016, doi: 10.1016/j.jaci.2016.07.007.

[11] J. Bethony, S. Brooker, M. Albonico, S. M. Geiger, A. Loukas, D. Diemert, and P. J. Hotez, “Soil-transmitted helminth infections: Ascariasis, trichuriasis, and hookworm,” The Lancet, vol. 367, no. 9521, pp. 1521–1532, May 2006, doi: 10.1016/S0140-6736(06)68653-4.

[12] M. R. Rivero, C. De Angelo, P. Nuñez, M. Salas, C. E. Motta, A. Chiaretta, O. D. Salomón, and S. Liang, “Environmental and socio-demographic individual, family and neighborhood factors associated with children intestinal parasitoses at Iguazú, in the subtropical northern border of Argentina,” PLOS Neglected Tropical Diseases, vol. 11, no. 11, p. e0006098, Nov. 2017, doi: 10.1371/journal.pntd.0006098.

[13] R. L. Pullan and S. J. Brooker, “The global limits and population at risk of soil-transmitted helminth infections in 2010,” Parasites & Vectors, vol. 5, article no. 81, Apr. 2012, doi: 10.1186/1756-3305-5-81.

[14] R. J. Stoltzfus, M. L. Dreyfuss, H. M. Chwaya, and M. Albonico, “Hookworm control as a strategy to prevent iron deficiency,” Nutrition Reviews, vol. 55, no. 6, pp. 223–232, Jun. 1997, doi: 10.1111/j.1753-4887. 1997.tb01609. x.

[15] M. C. H. Jukes et al., “Heavy schistosomiasis associated with poor short-term memory and slower reaction times in Tanzanian schoolchildren: Partnership for Child Development,” Tropical Medicine & International Health, vol. 7, no. 2, pp. 104–117, Feb. 2002, doi: 10.1046/j.1365-3156.2002.00843.x.

[16] D. Stark, J. L. N. Barratt, S. van Hal, D. Marriott, J. Harkness, and J. T. Ellis, “Clinical significance of enteric protozoa in the immunosuppressed human population,” Clinical Microbiology Reviews, vol. 22, no. 4, pp. 634–650, Oct. 2009, doi: 10.1128/CMR.00017-09.

[17] J. Ndibazza et al., “Effects of deworming during pregnancy on maternal and perinatal outcomes in Entebbe, Uganda: A randomized controlled trial,” Clinical Infectious Diseases, vol. 50, no. 4, pp. 531–540, Feb. 2010, doi: 10.1086/649924.

[18] A. L. Braseth, D. E. Elliott, and M. N. Ince, “Parasitic infections of the gastrointestinal tract and liver,” Gastroenterology Clinics of North America, vol. 50, no. 2, pp. 361–381, Jun. 2021, doi: 10.1016/j.gtc.2021.02.011.

[19] World Health Organization, Helminth Control in School-Age Children: A Guide for Managers of Control Programmes, 2nd ed. Geneva, Switzerland: World Health Organization, 2011.

[20] N. Argy et al., “Selecting a multiplex PCR panel for accurate molecular diagnosis of intestinal protists: A comparative study of Allplex® (Seegene®), G-DiaParaTrio (Diagenode®), and RIDA®GENE (R-Biopharm®) assays and microscopic examination,” Parasite, vol. 29, article no. 5, 2022, doi: 10.1051/parasite/2022003.

[21] I. H. McHardy, M. Wu, R. Shimizu-Cohen, M. R. Couturier, and R. M. Humphries, “Detection of intestinal protozoa in the clinical laboratory,” Journal of Clinical Microbiology, vol. 52, no. 3, pp. 712–720, Mar. 2014, doi: 10.1128/JCM.02877-13.

[22] M. Tanyuksel and W. A. Petri Jr., “Laboratory diagnosis of amebiasis,” Clinical Microbiology Reviews, vol. 16, no. 4, pp. 713–729, Oct. 2003, doi: 10.1128/CMR.16.4.713-729.2003.

[23] R. Augustine et al., “Loop-mediated isothermal amplification (LAMP): A rapid, sensitive, specific, and cost-effective point-of-care test for coronaviruses in the context of COVID-19 pandemic,” Biology, vol. 9, no. 8, article no. 182, Jul. 2020, doi: 10.3390/biology9080182.

[24] EFSA Panel on Biological Hazards (BIOHAZ) et al., “Public health risks associated with food-borne parasites,” EFSA Journal, vol. 16, no. 12, article no. 5495, Dec. 2018, doi: 10.2903/j.efsa.2018.5495.

[25] G. Bonizzi et al., “Standard operating procedures for biobank in oncology,” Frontiers in Molecular Biosciences, vol. 9, article no. 967310, Aug. 2022, doi: 10.3389/fmolb.2022.967310.

[26] Organisation for Economic Co-operation and Development, “OECD Guidelines on Human Biobanks and Genetic Research Databases,” OECD, 2009. [Online]. Available: https://www.oecd.org/health/biotech/44054609.pdf. [Accessed: Jun. 15, 2026].

[27] International Society for Biological and Environmental Repositories, “ISBER Best Practices: Recommendations for Repositories, Fourth Edition,” ISBER, 2018. [Online]. Available: https://www.isber.org/page/BPR. [Accessed: Jun. 15, 2026].

[28] C. Bycroft et al., “The UK Biobank resource with deep phenotyping and genomic data,” Nature, vol. 562, no. 7726, pp. 203–209, Oct. 2018, doi: 10.1038/s41586-018-0579-z.

[29] The Malaria Genomic Epidemiology Network, “A global network for investigating the genomic epidemiology of malaria,” Nature, vol. 456, no. 7223, pp. 732–737, Dec. 2008, doi: 10.1038/nature07632.

[30] P. H. J. Riegman, M. M. Morente, F. Betsou, P. de Blasio, and P. Geary, “Biobanking for better healthcare,” Molecular Oncology, vol. 2, no. 3, pp. 213–222, 2008, doi: 10.1016/j.molonc.2008.07.004.

[31] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. R. Thoma, “Image analysis and machine learning for detecting malaria,” Translational Research, vol. 194, pp. 36–55, Apr. 2018, doi: 10.1016/j.trsl.2017.12.004.

[32] Association of Public Health Laboratories, Laboratory Information Systems Project Management: A Guidebook for International Implementations. Silver Spring, MD, USA: APHL, 2019. [Online]. Available: https://aphl.org/docs/default-source/technical/GH-2019May-LIS-Guidebook-web.pdf. [Accessed: Jun. 15, 2026].

[33] N. Rieke et al., “The future of digital health with federated learning,” npj Digital Medicine, vol. 3, article no. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[34] F. Cabitza and J.-D. Zeitoun, “The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence,” Annals of Translational Medicine, vol. 7, no. 8, article no. 161, Apr. 2019, doi: 10.21037/atm.2019.04.07.

[35] N. O’Mahony, S. Campbell, L. Krpalkova, A. Carvalho, J. Walsh, and D. Riordan, “Representation learning for fine-grained change detection,” Sensors, vol. 21, no. 13, article no. 4486, Jun. 2021, doi: 10.3390/s21134486.

[36] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in Computer Vision – ECCV 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer International Publishing, 2018, pp. 139–156, doi: 10.1007/978-3-030-01264-9_9.

[37] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.

[38] J.-B. Grill et al., “Bootstrap your own latent: A new approach to self-supervised learning,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 21271–21284.

[39] X. Chen and K. He, “Exploring simple Siamese representation learning,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 15745–15753, doi: 10.1109/CVPR46437.2021.01549.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[41] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[43] H.-C. Shin et al., “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. 37th International Conference on Machine Learning (ICML), 2020, pp. 1597–1607.

[45] J. E. van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” Machine Learning, vol. 109, no. 2, pp. 373–440, Feb. 2020, doi: 10.1007/s10994-019-05855-6.

[46] Scikit-learn developers, “Metrics and scoring: Quantifying the quality of predictions,” Scikit-learn documentation. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed: Jun. 15, 2026].

[47] C. X. Gao et al., “An overview of clustering methods with guidelines for application in mental health research,” Psychiatry Research, vol. 327, article no. 115265, Sep. 2023, doi: 10.1016/j.psychres.2023.115265.

[48] M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, “Quantitative comparison of machine learning clustering methods for tuberculosis data analysis,” Engineering Proceedings, vol. 60, no. 1, article no. 20, 2024, doi: 10.3390/engproc2024060020.