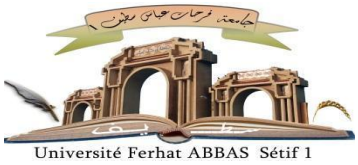
الجمهورية الجزائرية الديمقراطية الشعبية

Democratic and Popular Algerian Republic

Ministry of Higher Education and Scientific Research



FERHAT ABBAS UNIVERSITY - SETIF1

FACULTY OF SCIENCES

THESIS

Presented for the Department of Computer Science

For the graduation of

PhD

Domain: Mathematics and Computer Science

Field: Computer Science **Option: Data Science**

By : Aya MESSAI

THEME

Automatic medical decision for diagnosis of infectious diseases based on artificial intelligence approaches

Thesis defended in 17/05/2025 in front of the jury members:

SLIMANI Yacine	**MCA**	**Faculty of Technology, UFAS 1**	**President**
DRIF Ahlem	**MCA**	**Faculty of Sciences, UFAS 1**	**Supervisor**
OUYAHIA Amel	**Professor**	**Faculty of Medecine, UFAS 1**	**Co-supervisor**
ABDOUN Meriem	**Professor**	**Faculty of Medecine, UFAS 1**	**Examiner**
KARA-MOHAMED Chafia	**Professor**	**Faculty of Sciences, UFAS 1**	**Examiner**
LOUNAS Razika	**MCA**	**Dep of Computer Science, M'Hamed Bougara University, Boumerdès**	**Examiner**
RAIS Mounira	**MCA**	**Faculty of Medecine, UFAS 1**	**Invited**
KADERALI Lars	**Professor**	**Bioinformatics Institute, Greifswald University, Germany**	**Invited**

Abstract

Infectious diseases present complex diagnostic challenges due to the overlapping clinical manifestations caused by diverse pathogens. Meningitis, in particular, remains a significant global health concern due to its high morbidity and mortality, especially when diagnosis and treatment are delayed. Traditional diagnostic methods often involve invasive procedures and extensive laboratory testing, which can be time-consuming and resource-intensive. This Ph.D. research investigates the integration of artificial intelligence (AI) into the diagnostic process, aiming to enhance accuracy, speed, and interpretability through the use of explainable AI (XAI) techniques.

The first phase of this study examines cerebrospinal fluid (CSF) biomarker variations across different age groups—children, adults, and the elderly—within various types of meningitis. By analyzing these patterns, we aim to improve the understanding of diagnostic and clinical variations and their implications for treatment strategies. This analysis establishes a foundational understanding of how biomarkers behave in different populations and infection contexts.

Our next contribution focuses on diagnosing multiple meningitis types using ensemble models and SHapley Additive exPlanations (SHAP) to interpret feature importance. Using data from Setif Hospital (Algeria) and Brazil's SINAN database, we validated our findings across diverse populations. Extreme Gradient Boosting achieved strong performance (accuracy: 0.90, AUROC: 0.94, F1-score: 0.98). SHAP revealed distinct biomarker profiles such as elevated neutrophils in meningococcal, high lymphocytes in tuberculous, and neutrophil dominance in H. influenzae meningitis, along with clinically relevant diagnostic patterns. These results highlight the model's ability to distinguish bacterial, viral, and pathogen-specific meningitis, increasing trust in AI-driven diagnostics.

Our third contribution develops specialized models for meningococcal meningitis, emphasizing local explainability for precise diagnosis. We tested several models on 934 cases, with gradient boosting performing best (accuracy: 0.88, AUROC: 0.93, F1-score: 0.87). Using XAI tools like ELI5 and LIME, we provided local explanations that highlighted key diagnostic factors, including Neisseria meningitidis presence, CSF WBC count, patient age, and neutrophil levels. These insights support clinical trust by aligning model predictions with medical reasoning.

To enhance AI transparency, we introduced a novel explainable approach that integrates medical expertise into the interpretation of black-box models. Using concept vector analysis, we assessed the contribution of symptoms and biomarkers in identifying pneumococcal meningitis. Our deep learning model showed strong performance (accuracy: 92.23%, F1-score: 92.98%, AUROC: 92.36%) and remained robust in real-world validation, correctly identifying most cases with high agreement (Cohen's Kappa: 0.75). Bio-TCAV revealed clinical signs (0.92), medical history (0.79), and CSF aspect (0.88) as key influences on predictions, while biomarkers had a moderate effect (0.56). Tests like PCR, culture, LATEX, and bacterioscopy were most influential (TCAV = 1) aligning with their critical role in real-world meningitis diagnosis. Welch's t-test confirmed that these differences in TCAV scores were statistically significant.

Keywords: Black box model, Infectious Diseases, Meningitis diagnosis, Artificial Intelligence (AI) in Clinical Diagnostics, Explainable AI (XAI), Interpretable Diagno-

sis, Trustworthy AI in Healthcare.

Résumé

Les maladies infectieuses posent des défis diagnostiques complexes en raison des manifestations cliniques qui se chevauchent et de la diversité des agents pathogènes. La méningite, en particulier, demeure un problème de santé publique majeur à l'échelle mondiale en raison de sa forte morbidité et mortalité, notamment lorsque le diagnostic et le traitement sont retardés. Les méthodes de diagnostic traditionnelles reposent souvent sur des procédures invasives et des analyses biologiques approfondies, ce qui peut être long et coûteux. Cette recherche doctorale explore l'intégration de l'intelligence artificielle (IA) dans le processus diagnostique, avec pour objectif d'améliorer la precision, la rapidité et l'interprétabilité grâce aux techniques d'Explainable AI (XAI).

La première phase de l'étude s'intéresse aux variations des biomarqueurs dans le liquide céphalorachidien (LCR) selon les tranches d'âge — enfants, adultes et personnes âgées — pour différents types de méningite. En analysant ces profils, nous visons à mieux comprendre les variations diagnostiques et cliniques et leurs implications pour les stratégies de traitement. Cette analyse pose les bases d'une compréhension approfondie du comportement des biomarqueurs en fonction des populations et des contextes infectieux.

Notre contribution suivante porte sur le diagnostic de plusieurs types de méningite à l'aide de modèles ensemblistes et de SHapley Additive exPlanations (SHAP) pour interpréter l'importance des variables. À partir de données issues de l'hôpital de Sétif (Algérie) et de la base SINAN (Brésil), nous avons validé nos résultats sur des populations diversifiées. Le modèle Extreme Gradient Boosting a obtenu de très bonnes performances (accuracy : 0.90, AUROC : 0.94, F1-score : 0.98). L'analyse SHAP a mis en évidence des profils distincts de biomarqueurs : neutrophiles élevés dans la méningite à méningocoque, lymphocytes élevés dans la méningite tuberculeuse, et prédominance des neutrophiles dans la méningite à H. influenzae, ainsi que des motifs diagnostiques cliniquement pertinents. Ces résultats soulignent la capacité du modèle à différencier les méningites bactériennes, virales et spécifiques à un agent pathogène, renforçant la confiance dans les diagnostics pilotés par l'IA.

Notre troisième contribution développe des modèles spécialisés pour la méningite à méningocoque, mettant l'accent sur l'explicabilité locale pour un diagnostic plus précis. Plusieurs modèles ont été testés sur un ensemble de 934 cas, le modèle gradient boosting offrant les meilleurs résultats (accuracy : 0.88, AUROC : 0.93, F1-score : 0.87). En utilisant des outils XAI tels que ELI5 et LIME, nous avons fourni des explications locales révélant les facteurs clés du diagnostic, comme la présence de Neisseria meningitidis, le taux de globules blancs dans le LCR, l'âge du patient, et les niveaux de neutrophiles dans le LCR. Ces éléments renforcent la confiance clinique en alignant les prédictions de l'IA avec le raisonnement médical.

Pour améliorer la transparence de l'IA, nous avons proposé une approche explicable innovante intégrant l'expertise médicale dans l'interprétation des modèles boîte noire. À l'aide de l'analyse des vecteurs de concepts, nous avons évalué la contribution des symptômes et biomarqueurs dans l'identification de la méningite à pneumo-

coque. Notre modèle de deep learning a montré d'excellentes performances (accuracy : 92.23%, F1-score : 92.98%, AUROC : 92.36%) et a conservé sa robustesse lors de la validation en conditions réelles, identifiant correctement la majorité des cas avec une forte concordance (Cohen's Kappa : 0.75). L'approche Bio-TCAV a révélé que les signes cliniques (0.92), les antécédents médicaux (0.79) et l'aspect du LCR (0.88) influençaient fortement les prédictions, tandis que les biomarqueurs avaient un impact plus modéré (0.56). Les tests comme le PCR, les cultures, le LATEX et la bactérioscopie se sont avérés les plus influents (TCAV = 1), en accord avec leur rôle critique dans le diagnostic réel de la méningite. Un test t de Welch a confirmé que ces différences de scores TCAV étaient statistiquement significatives.

Mots-clés: Modele boîte noire, Maladies infectieuses, Diagnostic de la méningite, IA appliquée au diagnostic clinique, IA transparente (XAI), Diagnostics intelligibles, fiabilité de l'IA en milieu clinique.

ملخص

تُشكّل الأمراض المعدية تحديات تشخيصية معقدة بسبب تداخل الأعراض السريرية الناتجة عن تنوّع العوامل الممرضة. وتُعدّ التهاب السحايا بشكل خاص مشكلة صحية عالمية خطيرة بسبب ارتفاع معدلات المراضة والوفيات، خاصة عند تأخر التشخيص والعلاج. غالبًا ما تعتمد طرق التشخيص التقليدية على إجراءات جراحية وتحاليل مخبرية معقّدة، مما يجعلها تستغرق وقتًا طويلاً وتتطلب موارد كبيرة. تهدف هذه الأطروحة إلى دراسة دمج الذكاء الاصطناعي (AI) في عملية التشخيص، مع التركيز على تحسين accuracy والسرعة وقابلية التفسير باستخدام تقنيات Explainable AI (XAI) .

في المرحلة الأولى من هذا البحث، قمنا بدراسة اختلافات المؤشرات الحيوية في السائل الدماغي الشوكي (CSF) عبر الفئات العمرية المختلفة — الأطفال، البالغين، وكبار السن — لأنواع متعددة من التهاب السحايا. ومن خلال تحليل هذه الأنماط، نسعى لفهم الفروقات التشخيصية والسريرية بشكل أفضل وتحديد آثارها على خطط العلاج.

تركز مساهمتنا الثانية على تشخيص أنواع متعددة من التهاب السحايا باستخدام نماذج learning ensemble وتقنية SHapley Additive exPlanations (SHAP) لتفسير أهمية كل ميزة. استُخدمت بيانات من مستشفى سطيف (الجزائر) وقاعدة SINAN (البرازيل). حقق نموذج Boosting Extreme Gradient أداءً قويًا (accuracy: 90.0, AUROC: 94.0, F1-score: 98.0). كشفت SHAP عن أنماط مميزة للمؤشرات الحيوية مثل ارتفاع العدلات في حالات المكورات السحائية، وارتفاع الخلايا اللمفاوية في التهاب السحايا السلي، وهيمنة العدلات في حالات H. influenzae.

في المرحلة الثالثة، طورنا نماذج لتشخيص التهاب السحايا بالمكورات السحائية، مع التركيز على التفسير المحلي (local explainability) للحصول على دقة أعلى. اختُبرت عدة نماذج على 934 حالة، وكان أداء boosting gradient هو الأفضل (accuracy: 88.0, AUROC: 93.0, F1-score: 87.0). باستخدام أدوات XAI مثل ELI5 وLIME، وفرنا تفسيرات محلية لتنبؤات النموذج.

ولتطوير شفافية الذكاء الاصطناعي، اقترحنا منهجًا تفسيريًا جديدًا يدمج خبرة الأطباء في تفسير قرارات نماذج black-box. باستخدام تحليل concept، vectors قمنا بتقييم مساهمة الأعراض والمؤشرات الحيوية في

التعرف على حالات التهاب السحايا بالمكورات الرئوية. أظهر نموذج learning deep أداءً ممتازًا (accuracy: 92.23%, F1-score: 92.98%, AUROC: 92.36%)، وأثبت فعاليته على بيانات حقيقية (Kappa Cohen's: 75.0). أظهر Bio-TCAV أن العلامات السريرية (92.0)، والتاريخ الطبي (79.0)، ومظهر السائل الشوكي (88.0) كانت العوامل الأكثر تأثيرًا، بينما كانت مساهمة المؤشرات الحيوية متوسطة (56.0). وكانت اختبارات مثل PCR، culture، LATEX والبكتريوسكوبية من الأكثر تأثيرًا (TCAV = 1)، بما يتماشى مع دورها في التشخيص الطبي. وأكد اختبار t-test Welch's أن الفروقات في درجات TCAV كانت ذات دلالة إحصائية.

الكلمات المفتاحية: نموذج الصندوق الأسود، الأمراض المعدية، تشخيص التهاب السحايا، الذكاء الاصطناعي في التشخيص السريري، الذكاء الاصطناعي القابل للتفسير، التشخيصات القابلة للتفسير، الذكاء الاصطناعي الجدير بالثقة في الرعاية الصحية.

Dedications

*****

I dedicate this work to my lovely family especially to my father and mother. To my sister Rayene, to my brother Rabeh, to my brother in-law and to my grand mother.

Acknowledgements

*****

I am one of those who firmly believe that there is no strength or power except from God. My deepest gratitude goes first to Almighty God for granting me the will, health, and patience to persevere through all these years of study.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, **Dr. Drif Ahlem**, for her invaluable advice, continuous support, and patience throughout my PhD journey. Her immense knowledge and vast experience have been a constant source of encouragement during my research. Her positivity made this process both enjoyable and inspiring.

I am equally grateful to my co-supervisor, **Prof. Ouyahia Amel**, whose guidance helped me make the decision to pursue my research in this field in the future. Her unwavering support throughout these years has been invaluable, and her contributions have played a significant role in bringing this thesis to completion.

Also, I would like to show appreciation for the support and contributions provided by **Dr. Rais Mounira** and **Doctor Guechi Meriem**, as well as the infectious disease service at **CHU Sétif**, whose collaboration greatly benefited this work. Their involvement was instrumental in completing this research. I am deeply thankful to **Prof. Lars Kaderali** and **Prof. Hocine Cherifi** for their fruitful discussions and valuable advice.

I would also like to extend my sincere thanks to all **jury members** and the president of the jury, **Dr. Slimani Yacine** for accepting the task of evaluating and discussing this thesis. Their insightful feedback and valuable suggestions will undoubtedly contribute to improving this work and guiding further research.

I would also like to extend my deepest appreciation to my parents, siblings and brother in-law. Their immense understanding, love, and encouragement over the past four years have made this accomplishment possible. I am especially grateful to my grandmother, who always reminded us that there is nothing greater than the pursuit of knowledge.

To my dear friends, thank you for your companionship and humor, without you, I might have graduated two months earlier! but life would have been far less entertaining! A special thanks to my supportive cousins, Rahma and Nora, and my best friend, Lina, for always being there to remind me that food, laughter, and occasional procrastination are essential ingredients of success.

Contents

*****

List of Figures

*****

List of Tables

*****

Introduction

*****

The global threat of infectious diseases, intensified by climate change, urbanization, and frequent international travel, highlights the urgent need for innovative healthcare solutions. Despite significant achievements such as the eradication of smallpox and near-elimination of polio, new diseases continue to spread quickly and claim millions of lives each year. Health professionals in infectious disease, microbiology, public health, and front-line care must address current health threats while preparing for emerging challenges. The ease of global travel has facilitated the rapid spread of infections across borders, as seen in the SARS, H1N1 flu, and COVID-19 outbreaks [1].

In this context, artificial intelligence (AI) has transformative potential, supporting key objectives of precision medicine [2, 3] such as early detection, targeted interventions, and personalized treatments for diverse populations. The capacity of AI to analyze vast health data can significantly improve healthcare decision-making, from identifying infection risk factors to recommending timely interventions and reducing misdiagnoses or unnecessary procedures. Eric Topol's concept of "Deep Medicine" [4] illustrates how AI can integrate complex patient data, improving clinical understanding and outcomes. However, for AI to be effective in managing infectious diseases, it must be designed to be ethical and interpretable. Transparent, accountable, and human-centered AI models enable healthcare providers to confidently use these tools and communicate clear ethical explanations to patients, an essential aspect in today's interconnected world [5].

Additionally, integrating AI into healthcare brings ethical considerations to the forefront, particularly around transparency and accountability in automated diagnostics. Since AI influences life-critical decisions, its predictions must be understandable and justifiable to clinicians and patients. This research aims to improve diagnostic accuracy and foster trust in AI tools by providing interpretable models that help medical teams make informed, ethical decisions in managing meningitis.

By exploring state-of-the-art Explainable AI (XAI) methods, this study builds an AI model that delivers clear, case-specific explanations, making AI-driven diagnoses more practical and trustworthy for clinicians. Through detailed analysis of cerebrospinal fluid (CSF) data, lab results, and patient symptoms, this work offers a pathway to more interpretable and actionable AI diagnostics—especially valuable in complex meningitis cases where accurate differential

diagnosis is challenging.

In medical diagnostics, especially concerning infectious diseases like meningitis, several significant issues, and research objectives arise due to the complexity of diagnosing diseases with overlapping symptoms. Below, we outline the main challenges and the specific research goals intended to address these issues.

Main issues:

1. Diagnostic complexity:
   Meningitis and other infectious diseases often present with non-specific symptoms such as fever, headache, and neck stiffness, which can overlap with other conditions. This makes differential diagnosis challenging, especially in cases involving multiple pathogens.
   Traditional diagnostic methods, such as cultures, PCR, and imaging, can be time-consuming and may not always be available in resource-limited settings.

2. Limited accuracy and speed of conventional methods:
   Current diagnostic approaches often require significant time for laboratory results (such as blood culture or cerebrospinal fluid analysis), leading to delays in treatment. This can increase the risk of complications, especially in conditions like meningitis, where timely intervention is crucial.

3. Transparency and trust in AI models:
   As Artificial Intelligence technologies, particularly machine learning models, are increasingly applied to clinical diagnostics, one of the major challenges is ensuring that these models are transparent and interpretable. Many AI systems, such as deep learning models, are often described as ”black boxes,” making it difficult for clinicians to trust or fully understand the rationale behind AI-driven recommendations. Following the recommendation of the World Health Organization (WHO), healthcare professionals and patients need to understand how AI systems make decisions. Promoting transparency in AI algorithms and ensuring that developers and providers are accountable for their decisions is essential to building trust in AI systems.

Research Objectives:

1. Enhancing diagnostic accuracy using AI:
   The primary goal of this research is to explore how AI can improve diagnostic accuracy in identifying meningitis and other infectious diseases. AI models can potentially identify diseases accurately and faster by analyzing complex patterns in clinical signs, laboratory data, and medical histories.

2. AI for rapid diagnosis:
   One key objective is to reduce diagnostic time using AI. By developing algorithms that can quickly analyze patient data, the research aims to accelerate decision-making processes and reduce the time it takes to diagnose, thereby improving patient outcomes.

3. Development of interpretable and transparent AI models:
   Ensuring that AI models are interpretable is critical for their acceptance and use in clinical settings. The research will focus on explainable AI approaches, such as SHAP (SHapley Additive explanations), which provide insights into how different features (clinical signs, laboratory results, etc.) contribute to the AI model's predictions. This can help clinicians understand and trust AI-assisted decisions.

4. Investigating biomarker patterns for accurate diagnosis:
   A key objective is to examine the role of biomarkers (e.g., cerebrospinal fluid analysis) in distinguishing between different types of meningitis and other infections. The research will focus on identifying specific biomarkers and their patterns, which can serve as valuable indicators for AI models in improving diagnostic precision.

5. Cross-population validation:
   The research also aims to validate the AI models across diverse patient populations, including different age groups and regions, to ensure that the models are generalizable and can be applied universally. This will include validating the model using datasets from different geographical locations and healthcare systems to confirm its robustness.

6. AI in low-resource settings:
   One of the research objectives is to explore how AI-based diagnostic tools can be adapted for use in resource-limited settings where access to advanced diagnostic technologies may be restricted.

Thesis outline

The research work consists of several chapters, namely:

Chapter 1: In this chapter, we introduce Explainable Artificial Intelligence (XAI) by exploring its various facets and providing a comprehensive definition. We present a taxonomy of XAI methods, categorizing them based on key dimensions such as scope, stage, and model specificity. In addition, we discuss the applications of XAI in the context of infectious diseases, emphasizing how these methods improve diagnostic accuracy and trustworthiness in healthcare. Furthermore, we examine several applications of XAI in various medical settings, illustrating how these approaches improve the transparency, reliability, and practical utility of AI systems in clinical practice. These contributions highlight the pivotal role of XAI in bridging the gap between complex machine-learning models and their adoption in real-world medical applications.

Chapter 2: This chapter comprehensively reviews meningitis and other infectious diseases, highlighting their diagnostic challenges. It begins with an introduction to infectious causes and explores the temporal trends of these diseases. The chapter delves into central nervous system infections, particularly meningitis. It presents an in-depth analysis of cerebrospinal fluid laboratory test results across various types of meningitis and age categories. Additionally, it examines the role of XAI in the healthcare domain. This chapter constitutes a background of our contributions to chapters 4, 5, and 6. Our academic paper related to this topic is:

1. MESSAI, A., DRIF, A., OUYAHIA, A., RAIS, M., GUECHI, M., KADERALI, L., & CHERIFI, H. "A comprehensive investigation into the ranges of laboratory tests present in cerebrospinal fluid across various types of meningitis within different age categories." BATNA JOURNAL OF MEDICAL SCIENCES 2024, VOL. 11, NO. 3, 347–356 DOI: 10.48087/BJMSoa.2024.11310

Chapter 3: This chapter presents our main contribution, which introduces a new approach to make AI models more understandable when diagnosing meningitis. Unlike traditional "black-box" AI models, which offer results without clear explanations, this chapter outlines a method that enhances the transparency of AI systems, making them easier for doctors to trust and use in clinical decision-making. We started the process by acquiring domain knowledge to define the rules for this research study, establishing the etiological diagnoses for Meningococcaemia, Meningococcal Meningitis, Tuberculous Meningitis, Aseptic Meningitis, Haemophilus influenzae Meningitis, and Pneumococcal Meningitis. The following is the preprocessing step of the SINAN dataset, comprising 6,729 patients aged over 18 years, obtained from the Brazilian Government's Health Information System on Notifiable Diseases, which was used for training the models, while additional data was collected from Setif Hospital in Algeria for further testing. Tree-based ensemble methods were then applied to assess the model's performance. Finally, an XAI-agnostic explainability approach was applied using the SHapley Additive exPlanations (SHAP) method to determine the contribution of each feature to the model's predictions. The academic publication of this work is:

2. Messai, A., Drif, A., Ouyahia, A., Guechi, M., Rais, M., Kaderali, L., & Cherifi, H. (2024). Toward XAI agnostic explainability to assess differential diagnosis for Meningitis diseases. Machine Learning: Science and Technology, 5(2), 025052, DOI: https://doi.org/10.1088/2632-2153/ad4a1f

Chapter 4: This chapter uses local explainability to improve how AI diagnoses meningococcal meningitis. Local explainability means providing case-by-case insights into how an AI model makes each specific prediction. Doctors can see why the AI model thinks a patient might have meningococcal meningitis based on their unique symptoms and test results. We addressed relevant research questions, such as: - How can AI provide more tailored, understandable explanations for individual diagnoses of meningococcal meningitis? - What

specific features does the AI consider most important for each case? and - How can these detailed explanations improve doctors' trust in AI tools and support their decision-making?. We train and test various Machine learning (ML) models, including logistic regression, K-nearest neighbors, support vector machine, decision tree, gradient boosting, AdaBoost, random forest, and LightGBM classifier, on a dataset comprising 460 cases of Meningococcal meningitis and 474 cases of other types of meningitis. Subsequently, we employ XAI tools (ELI5 and LIME) to elucidate the importance of features in the ML models. Our contribution has been published in:

3- Messai, A., Drif, A., Ouyahia, A., Guechi, M., Rais, M., Kaderali, L., & Cherifi, H. (2024, August). Transparent AI Models for Meningococcal Meningitis Diagnosis: Evaluating Interpretability and Performance Metrics. In 2024 IEEE 12th International Conference on Intelligent Systems (IS) Varna, Bulgaria, 2024, pp. 1-8. doi: 10.1109/IS61756.2024.10705255 , URL=https://ieeexplore.ieee.org/document/10705255

Chapter 5: In this chapter, we developed a Deep Neural Network (DNN) for diagnosing pneumococcal meningitis, integrating key medical concepts such as clinical signs, biomarkers, medical history, and laboratory test results. The model achieved strong performance metrics, including an accuracy of 92.23%, precision of 94.64%, recall of 91.38%, and an F1-score of 92.98%, demonstrating its high diagnostic accuracy. When tested on real-world clinical data from Algeria, the model correctly identified all pneumococcal meningitis cases and most cases of other types of meningitis, with a Cohen's Kappa score of 0.75, indicating substantial agreement between the model's predictions and actual clinical diagnoses. This result highlights the robustness and reliability of the model. To improve the interpretability of the model, we applied Bio-TCAV, a technique that links model predictions to clinically relevant concepts, offering valuable insight into how factors like clinical symptoms and CSF analysis influence the diagnosis. Including real-world medical concepts that align with clinical practice, is crucial for building trust in AI-driven diagnostics. Medical professionals rely on well-established indicators like clinical signs and laboratory results to make informed decisions. By aligning our model's decision-making process with these trusted medical concepts, we improve its credibility and transparency. This trust is vital for widespread clinical adoption, as healthcare professionals need to understand the reasoning behind model predictions to incorporate them into their practice effectively. Our contribution has been submitted in IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS (J-BHI).

In the context of disseminating our research and fostering collaboration, our work was showcased in two notable scientific events. First, it was presented at a scientific day held at the University of Bouira under the theme *Developing Research Skills Using Artificial Intelligence Techniques*. This event provided a platform for researchers and students to explore the integration of AI technologies in advancing scientific inquiry. Notably, our

contribution was recognized with the award for the best poster, underscoring the significance and quality of our work.

Additionally, our research was featured in the *Medical Physics Study Day*, organized by the Department of Physics at the University of Ferhat Abbas, Setif 1. These events focused on the intersection of medical physics, artificial intelligence, and advanced instrumentation, offering a unique opportunity to discuss and demonstrate the application of Explainable AI (XAI) in these fields. Presenting our work in such interdisciplinary and cutting-edge forums underscores its relevance and potential impact in advancing AI methodologies and their medical applications.

Finally, we conclude the thesis and introduce possible future directions.

Chapter 1

Explainable AI background: Fundamental theories and literature review.

*****

1.1 Introduction

Machine learning models are widely applied across various fields, each with distinct methodologies and objectives. Data analysts and scientists use these models for clustering and data visualization, while economists rely on them for predictive tasks, such as forecasting market trends or stock prices. In the medical field, machine learning plays a crucial role in diagnosing and analyzing medical data. In many of these applications, understanding how a model functions and why it produces specific results is essential. However, as machine learning models become increasingly complex, they often cross a threshold where the relationship between inputs and predictions is no longer easily interpretable. Such models are commonly referred to as ”black box” models. This chapter will explore key concepts and definitions related to the need for explainability in AI, focusing on methods to make machine learning models more transparent and understandable.

1.2 Explainable Artificial Intelligence (XAI)

Explainable AI (XAI) is a field of research focused on making AI system outputs more comprehensible to humans. The term 'explainable AI' was first introduced by Van Lent et al. in 2004 to describe the capability of their system to elucidate the actions of AI-driven entities in simulation games [6]. However, the need for AI explainability dates back to the mid-1970s, when researchers initially studied ways to clarify decisions in expert systems. Despite the early interest, progress in explainability research slowed as machine learning advanced rapidly, shifting the focus toward optimizing predictive accuracy over interpretability. As a result, recent AI research has often prioritized models and algorithms for their predictive power, leaving the ability to explain their decision-making processes as a secondary concern.

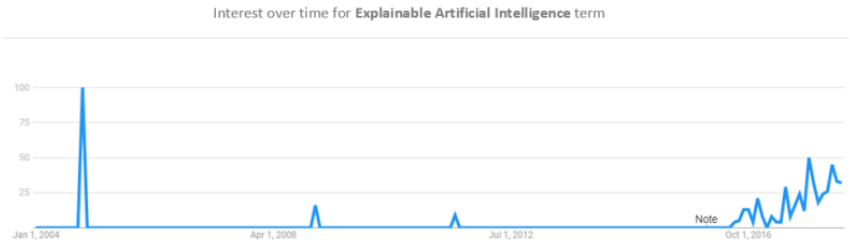


Figure 1.1: Analysis of Explainable Artificial Intelligence's research interest using Google Trends

The field of Explainable AI has recently regained significant attention from researchers and industry professionals. As illustrated in Figure 1.1, there has been a notable resurgence in research interest surrounding the term "XAI" [7]. This renewed focus is driven by the rapid advancements in artificial intelligence, particularly through sophisticated machine learning techniques and deep neural networks (DNNs). While these techniques often achieve high-performance levels, they are frequently criticized for their lack of interpretability. Notably, methods with the highest predictive accuracy, such as DNNs, are typically the most opaque. In contrast, more interpretable approaches like decision trees generally offer lower predictive performance. This trade-off between accuracy and explainability underscores the critical need for research in XAI, aiming to bridge the gap between performance and interpretability in machine learning models [8], Figure 1.2 illustrates the trade-off between explainability and learning performance across various machine learning models.

1.2.1 Making AI understandable to end users

Humans struggle to process large volumes of data, whereas computational programs excel in this domain, constructing complex models that often defy human understanding. Figure 1.3 illustrates this concept. The term "interpretability methods" refers to techniques designed to

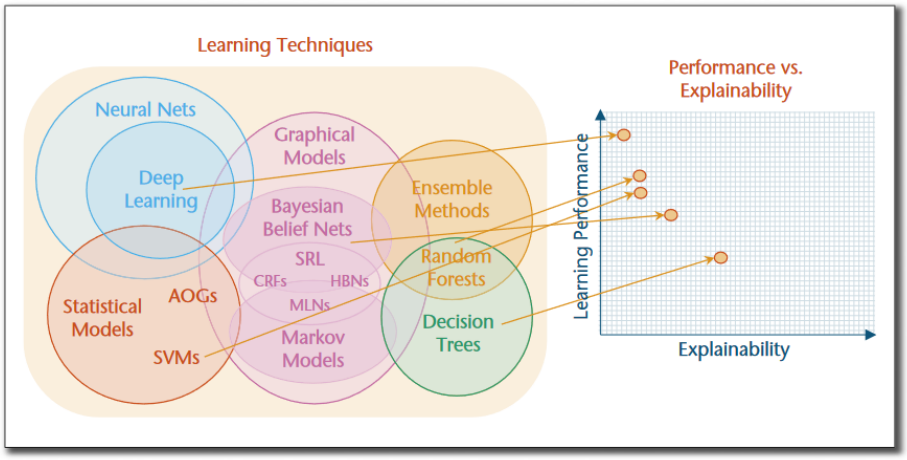


Figure 1.2: The trade-off between explainability and learning performance in different learning models. [9]

bridge the gap between opaque (”black box model”) machine learning models and human cognition. These methods operate at multiple levels, translating raw data into increasingly abstract representations that are meaningful and understandable to humans.

Interpretability methods serve as a critical link between the predictions and decisions made by machine learning models and human understanding, fostering trust in these systems. However, establishing a universal standard for the type and depth of explanations required in AI is inherently challenging. This difficulty arises from the diversity of machine learning models and the end users' wide-ranging backgrounds, knowledge, and experiences. Therefore, effective explanations must be adaptable to the specific context and tailored to the needs of the target audience [10].

In many cases, creating a model involves a human domain expert who provides domain knowledge to guide the model's development. In return, the model generates predictions and explanations. This model type is called a ”human-in-the-loop model,” as it is designed to integrate human input throughout its operation.

As illustrated in Figure 1.4, humans are actively involved in every stage of developing a model that delivers predictions and explanations. Building such a model requires meeting two key criteria: first, the machine must be able to interpret the information provided by humans (e.g., data, labels, or domain knowledge), though this does not imply that the machine processes or represents the information in the same way as humans. Second, humans must understand the machine's predictions, explanations, and decisions; this ensures that the system is transparent, trustworthy, and actionable for human users. Providing knowledge to the machine involves organizing and structuring data meaningfully, transforming it from appearing unstructured to being more interpretable and useful [11]. This principle serves as a foundational example of a human-in-the-loop model.

Two principles have been proposed to assess the interaction between human knowledge

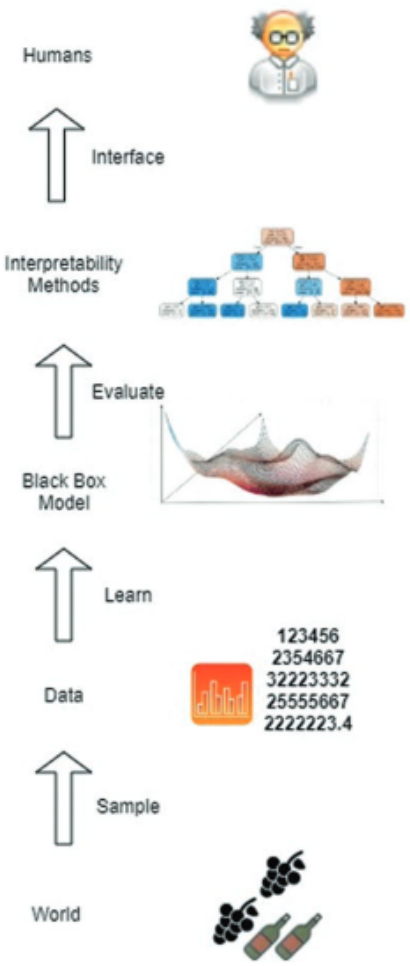


Figure 1.3: Bridging the real world and humans with Machine Learning [10].

and machine learning models. The first principle states that providing domain knowledge, such as expert insights or structured data, improves a model's performance, indicating that the machine can understand the knowledge. However, the reverse is not necessarily true, if performance does not improve, it does not imply that the knowledge is not understandable. This principle highlights that improved performance is sufficient evidence of understanding but is not a mandatory condition. The second principle focuses on machine-generated explanations, suggesting that it must be understandable if an explanation can be refuted or proven wrong. This is because the ability to critique or challenge an explanation inherently requires understanding. These principles emphasize the importance of mutual intelligibility between humans and machines in human-in-the-loop models.

1.2.2 Where is XAI crucial?

As AI grows more sophisticated and complex, it becomes increasingly challenging for humans to understand and trace the steps by which algorithms reach their results. This entire computational process is often called a "black box," an essentially opaque and difficult-to-

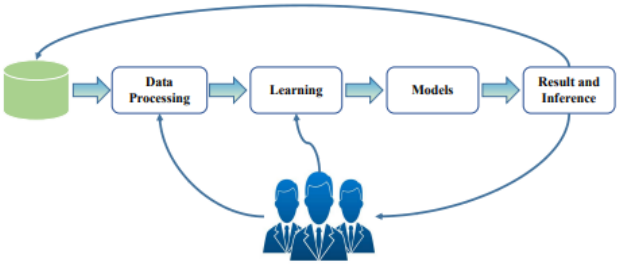


Figure 1.4: The role of humans in guiding the model's development. [11]

interpret model. Such black box models are constructed directly from data, meaning that even the engineers or data scientists behind the algorithms may struggle to explain precisely how they function or why they generate certain outcomes. Understanding how an AI system produces specific outputs offers numerous benefits, especially in critical domains [12].
in some cases, it wouldn't be fair to be satisfied with predictions without answering the "why"; that is, the understanding of how a certain model got a prediction or why it led us to some results, so getting a high accuracy or high confidence prediction is not enough. Consider an AI system employed for predictive tasks, such as forecasting market trends or stock prices, or models applied in the medical field for analyzing patient data or predicting disease outcomes. Alternatively, envision an autonomous vehicle that drives unpredictably, causing a fatal accident even under normal road conditions. Understanding models' behavior and decision-making processes is essential in such critical cases.

AI has emerged as a transformative tool in the medical field, offering numerous benefits in diagnosis, treatment planning, and patient care [13]. One of the most prominent applications of AI in medicine is in medical imaging. AI-powered computer vision algorithms are increasingly used to analyze medical images, such as X-rays, MRIs, and CT scans [14]. These algorithms can assist in detecting and diagnosing conditions such as tumors, fractures, or other abnormalities with high accuracy, often at earlier stages than human clinicians might detect [15]. For example, AI-based systems have been shown high performance in tumor recognition, supporting oncologists in identifying malignant growths and enabling timely intervention.

However, adopting AI in medical practice raises significant trust, accountability, and transparency challenges, where explainable AI becomes crucial [16]. As AI systems make increasingly complex decisions, clinicians and patients must understand the rationale behind these decisions. For example, in medical imaging, while AI can accurately identify anomalies like tumors [17], understanding the specific features that led to the detection is vital. Clinicians must be able to interpret the model's reasoning to ensure that it aligns with medical knowledge and the patient's context. Without transparency, even the most accurate AI-driven predictions might be met with skepticism, particularly when decisions could impact patient health.

Another significant application of AI in healthcare is using robotics [18]. AI-driven robotics, particularly in the form of reinforcement learning, is being leveraged for surgical purposes. These systems learn and adapt to complex tasks through trial and error, allowing robotic systems to assist surgeons in performing highly precise and minimally invasive procedures [13]. While these systems may outperform human surgeons in some aspects, their lack of interpretability can raise concerns, particularly in critical situations where understanding the machine's decision-making process is necessary to ensure patient safety.

Additionally, AI is being applied to diagnosis by analyzing large datasets of patient information. Machine learning algorithms can recognize patterns in patient records, laboratory results, genetic data, and clinical notes, helping doctors make more accurate diagnoses and recommend personalized treatment plans [13]. In these scenarios, XAI methods are needed to ensure that the model's reasoning is clear and its conclusions are consistent with clinical best practices. The integration of explainable AI (XAI) into medical systems is crucial for fostering trust between healthcare providers and patients. By offering transparency, traceability, and interpretability, XAI ensures that AI complements human expertise rather than replacing it. This approach not only enhances confidence in AI applications but also promotes their responsible use, making them more reliable and effective in supporting healthcare decisions while prioritizing patient well-being.

1.2.3 What is "Easily Interpretable"?

Defining interpretability in machine learning is inherently challenging due to its subjective and context-dependent nature. Miller et al. [19] offers a widely appreciated non-mathematical definition: *"Interpretability is the degree to which a human can understand the cause of a decision."* Another perspective defines interpretability as *"the degree to which a human can consistently predict the model's result."* These definitions emphasize that interpretability centers on a model's ability to provide insights into how and why decisions or predictions are made.

A model is considered more interpretable if its decisions are easier for a human to comprehend compared to another model. For example, a linear regression model, which provides clear relationships between features and outcomes through coefficients, is often more interpretable than a deep neural network with millions of parameters [20, 21]. However, even simple models can become complex when involving large numbers of features, underscoring that interpretability is not solely tied to a model's complexity.

Interpretable machine learning involves extracting meaningful insights from models about the relationships in the data or those learned by the model. For example, decision trees are inherently interpretable, as their structure mirrors human reasoning, enabling users to trace the decision-making process step by step. Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) further enhance the interpretability

of complex models, such as neural networks or gradient boosting machines, by attributing feature importance to specific predictions.

1.2.4 Performance and interpretability trade-off

The trade-off between performance and interpretability in machine learning is often confusing, especially regarding the role of explainable AI. Implementing XAI does not inherently reduce accuracy, the trade-off occurs when choosing models. Simpler models like decision trees are inherently interpretable but may lack predictive power, while complex models like deep neural networks (DNNs) offer high accuracy but are harder to explain. The goal is not to sacrifice performance but to enhance both explainability and accuracy, shifting the performance-explainability curve upward [10, 21] (See Figure 1.5).

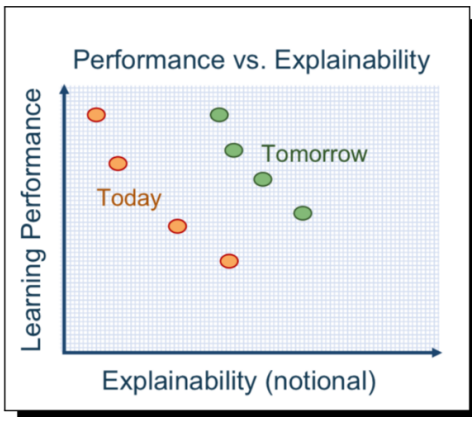


Figure 1.5: Accuracy explainability trade off myth [10]

In high-stakes applications like healthcare, interpretability is critical for understanding model predictions, such as diagnoses or treatment recommendations, especially when lives are at risk. For instance, DNNs can provide highly accurate predictions for patient outcomes but require explainability to ensure trust and adoption in clinical practice. In contrast, low-risk environments prioritize accuracy over interpretability, like movie recommendation systems. The challenge lies in balancing completeness and interpretability, as interpretable models often lack comprehensive explanations, while complex models are harder for humans to understand [20, 21]. XAI aims to address this by bridging the gap and ensuring reliable predictions and actionable insights.

1.2.5 Interpretability metrics

It is challenging to define interpretability in purely mathematical terms. Interpretability refers to the extent to which a human can comprehend the reasoning behind a decision made by a model [19]. It can also be understood as the degree to which a human can consistently predict the model's outputs [22]. In machine learning, interpretability is generally defined

as the ability to explain or present the model's decision-making process in terms that are understandable to humans [23]. Thus, interpretability lacks a universal quantitative metric and is inherently subjective, depending on an individual's understanding and context.

In their work, [23] proposed three approaches to evaluate interpretability:

1. **Application-level evaluation (real task):** This approach involves testing experts' explanations in real-world applications. For instance, in the case of fracture detection software that uses machine learning to identify fractures in X-rays, radiologists would directly evaluate the software's predictions and compare them with their own clinical decisions. This method assesses how well the explanations align with expert knowledge and whether the system's reasoning is understandable and useful in practice.

2. **Human-level evaluation (simple task):** In this approach, the explanations are tested by non-experts or regular users, allowing for a larger pool of testers. For example, different explanations would be presented to users, who would select the most understandable or convincing. This method evaluates the clarity and comprehensibility of explanations from the general public's perspective without requiring specialized knowledge.

3. **Function-level evaluation (proxy task):** This evaluation method does not involve human testers. It is typically used when the model class has already been assessed in a human-level evaluation. For instance, if it is known that end users understand decision trees, the proxy for explanation quality could be the tree's depth. Shorter trees would score better in explainability as they are simpler and easier to interpret. However, this evaluation should also ensure that the model's predictive performance does not significantly degrade when the tree is pruned to improve explainability. The goal is to balance simplicity with performance.

1.2.6 Explainability vs. Inerpretability

While often used interchangeably, interpretability and explainability are distinct concepts in machine learning, each addressing different aspects of understanding a model. Although no universally accepted quantitative definitions exist, the distinction between these terms is critical for clarifying their roles in explainable AI [10].

Interpretability refers to the ability to understand the internal mechanics of a model and how inputs are transformed into outputs without necessarily explaining why a specific decision was made. It focuses on making the model's behavior more transparent, often by simplifying its structure or visualizing its processes.

Explainability, in contrast, extends beyond transparency by addressing "why" a model produces certain predictions. It encompasses reasoning about how changes in input features influence the outputs, including counterfactual scenarios, such as "What would happen if

feature X were different?" or "How would the prediction change if value Y were removed?" Explainability thus aims to provide a more comprehensive understanding of the model's decision-making process.

In a medical context, a diagnostic model predicting heart disease is considered. Interpretability would involve understanding that the model prioritizes factors like cholesterol levels, age, and smoking history. This information helps medical professionals understand which attributes influence the model's predictions. Explainability, on the other hand, focuses on specific cases. For instance, if the model predicts a high risk of heart disease for a particular patient, explainability might reveal that this decision was primarily driven by the patient's elevated cholesterol level and smoking history. This case-specific rationale offers actionable insights for the individual scenario.

Interpretability is inherently broader, facilitating understanding a model's overall behavior, while explainability provides localized insights into individual predictions. Together, they form complementary tools for improving trust, usability, and adoption of machine learning models, particularly in high-stakes domains like healthcare [10].

1.2.7 Model transparency: White Box vs. Black Box

This section introduces the concepts of white-box and black-box models, focusing on their differing degrees of transparency and interpretability. The discussion will explore how these models impact understanding and trust in machine learning systems.

- **White Box Models (Transparent):** White-box models, including decision trees, rule-based systems, linear regression, and naive Bayes, are characterized by their transparency and interpretability. These models are easily understood because their decision-making processes are straightforward and resemble human reasoning [24]. For example, decision trees break down decisions into simple if-then rules, allowing practitioners to trace the logic behind any given output. As illustrated in Figure 1.6, the decision-making process is transparent, and the rules leading to a particular classification can be identified. While models like decision trees can be more interpretable, they are often pruned for simplicity, which may affect the accuracy of their decision-making process. Other white-box models, such as generalized linear models (GLMs), logistic regression, k-nearest neighbors (K-NN), and rule-based systems, provide insights into how features contribute to predictions. For instance, GLMs and logistic regression offer coefficients that quantify feature contributions, enhancing interpretability, while K-NN relies on feature similarity, making its predictions intuitively understandable [20, 25].

- **Black Box models (Opaque Models):** The term 'black box' refers to machine learning models that are inherently difficult to interpret or explain, especially from a mathematical or practical perspective [24]. Examples include Support Vector Machines (SVMs),

Deep Neural Networks (DNNs), ensemble methods like random forests and gradient boosting, kernel methods, and unsupervised k-means clustering and Gaussian mixture models. DNNs, in particular, epitomize black-box models due to their intricate architectures and vast number of parameters. Similarly, ensemble methods combine multiple models to improve accuracy but often obscure the individual contributions of each component. While these models are highly effective and can approximate complex functions, their internal structures provide little to no insight into how specific predictions are made, presenting challenges for interpretability and transparency in critical applications.

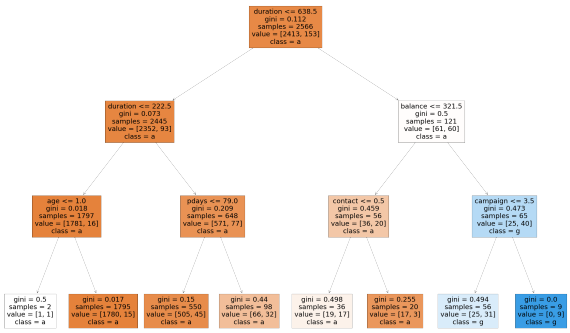


Figure 1.6: Decision tree constraint to three levels deep for classification of diabetes disease [20]

1.3 Explainable Artificial Intelligence (XAI): taxonomy and methods

1.3.1 Introduction

Explainable AI methods have proliferated significantly recently, driven by the increasing complexity of machine learning models and the demand for transparency in AI decision-making. Figure 1.7 presents a taxonomy of XAI methods, organized by their approach and characteristics, to provide a structured overview of this rapidly evolving field.

Given the diversity and continuous growth of XAI techniques, this taxonomy is not exhaustive and will likely expand as new methods are developed. For this thesis, we have focused on three key dimensions of classification: scope (global versus local explanations), stage (ante-hoc versus post-hoc methods), and model specificity (agnostic versus specific methods).

1.3.2 Ante-Hoc vs. Post-Hoc Interpretability

- **Ante-Hoc:** Ante hoc interpretability, also known as intrinsic interpretability, focuses on models designed with simplicity and transparency, often referred to as ”glass-box” or

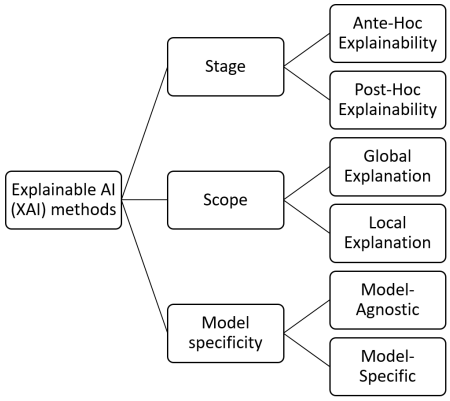


Figure 1.7: Explainable AI methods taxonomy.

"white box" methods. These approaches inherently provide explanations by maintaining a straightforward structure, such as linear relationships or rule-based logic, which directly links features to predictions [26]. Techniques like linear and logistic regression, decision trees, rule-based learners, and fuzzy systems prioritize interpretability. However, these models are generally less effective in capturing complex relationships in data compared to black-box approaches. Despite their transparency, they often face a trade-off between accuracy and interpretability, highlighting the need for post-hoc explainability methods in scenarios requiring both trust and precision [27].

- **Post-Hoc:** Post-hoc interpretability techniques address the black-box nature of complex models by offering insights into their decision-making without altering the model structure. These methods are particularly advantageous because they work across diverse algorithms, provide flexibility in internal representations, and allow multiple types of explanations for the same model. However, a trade-off exists between fidelity (how well the explanation represents the model) and comprehensibility (how easily humans can understand it). Post-hoc methods can be categorized into three main types:

  1. **Visual Explanations:** Visual explanations use graphical representations to relate features to predictions in a human-readable manner. One common example is the Partial Dependence Plot (PDP).

     - **Partial Dependence Plots (PDP)** are global methods for visualizing feature effects in predictive models. They show how input features influence model output by marginalizing over all other features. For a model $f(x_1, x_2)$, the PDP function for x_1 is defined as:

$$\text{PDP}(x_1) = \mathbb{E}_{x_2}[f(x_1, x_2)] = \int p(x_2) f(x_1, x_2) dx_2 \tag{1.1}$$

       where p (x2) represents the marginal distribution of feature x2, and approxi-

mated using n samples:

$$\text{PDP}(x_1) = \frac{1}{n}\sum_{j=1}^{n} f(x_1, x_{2,j}) \tag{1.2}$$

PDP plots can extend to multiple features. For a subset C of features, the PDP function becomes:

$$\text{PDP}(C) = \frac{1}{n}\sum_{j=1}^{n} f(C, \bar{C}_j), \tag{1.3}$$

where $\bar{C}_j$represents the values of the complementary set of features (not included in C) for the j-th sample, and C denotes the subset of features of interest. Two-dimensional plots are used for pairs of features, while higher-dimensional plots are harder to interpret.
PDP plots are efficient and offer insight into how features affect predictions, including interactions between features. However, they assume feature independence, leading to bias when features are correlated, and they may smooth out heterogeneous effects in the data [20].

2. **Feature Importance Methods:** Feature importance methods quantify the impact of features on predictions, either globally or for specific instances. Examples include:

   - **Permutation Feature Importance (PFI):** This method evaluates feature significance by measuring the change in model performance when a feature's values are randomly shuffled. While effective, PFI can lead to unrealistic data instances when features are permuted. Computing PFI on test data is recommended to avoid overfitting, but this can increase variance. Cross-validation can reduce variance but limits the use of the full dataset. PFI also accounts for feature interactions, which can dilute the importance of correlated features, leading to underestimation of their contributions [20].

$$PFI(x_j) = \mathbb{E}[M(f(X), y)] - \mathbb{E}[M(f(X_{\text{permuted},j}), y)] \tag{1.4}$$

   where M is a performance metric such as accuracy or error, and $\mathbb{E}[M(f(X), y)]$ is the expected performance metric of the model f, with original data X and true labels y, and $\mathbb{E}[M(f(X_{\text{permuted},j}), y)]$ is the expected performance metric of the model f with feature x_j permuted (i.e., its values shuffled), and true labels y.

- **Shapley Additive Explanations (SHAP):** Shapley values, derived from cooperative game theory, are used to fairly distribute the importance of features in a model's predictions. Each feature is treated as a "player" in a game, and its importance is determined based on its average contribution across all possible subsets of features (coalitions). For a model f with d features $X = \{x_1, x_2, \ldots, x_d\}$, the Shapley value for feature x_j is given by:

$$\phi_j(f) = \sum_{S \subseteq X \setminus \{x_j\}} \frac{|S|!(d - |S| - 1)!}{d!} \big[f(S \cup \{x_j\}) - f(S) \big] \tag{1.5}$$

where S is a subset of features excluding x_j, $f(S)$ is the model's prediction using only features in S, and $f(S \cup \{x_j\})$ includes x_j. The Shapley value represents the average marginal contribution of a feature across all coalitions. Shapley values ensure a fair and interpretable distribution of importance, adhering to key properties:

  * **Local accuracy**: The sum of feature contributions equals the model's prediction for instance.
  * **Missingness**: Features not affecting the prediction have a Shapley value of zero.
  * **Consistency**: If a feature's contribution increases in a model, its Shapley value will not decrease.

  However, computing Shapley values is resource-intensive due to the exponential number of possible feature coalitions. To address this, SHAP (Shapley Additive Explanations) offers an efficient approximation by estimating Shapley values using a background dataset. SHAP relies on the conditional expectation of the model's predictions:

$$f(X) = \mathbb{E}[f(X)|S], \tag{1.6}$$

  where S is a subset of features, the expectation is taken over all other features. SHAP provides additive and visually interpretable feature importance for individual predictions, making it suitable for large-scale models [21, 28]. Despite its advantages, Shapley values and SHAP can struggle with strongly correlated features, and their computational cost remains a challenge for high-dimensional datasets [20].

- **Local Surrogate Models:** Local surrogate models are designed to explain the behavior of black-box machine learning models at the level of individual predictions rather than the entire model. These models prioritize *local*

fidelity, ensuring that they closely approximate how the black-box model behaves near a specific data instance. A prominent example of this approach is **Local Interpretable Model-Agnostic Explanations (LIME)**, introduced in 2016 [29], which provides interpretable explanations for predictions by approximating the local behavior of a black-box model using a simpler, interpretable model. LIME operates through the following key steps:

(a) *Interpretable Representations*: LIME maps the data instance of interest into a human-understandable representation, such as binary vectors representing the presence or absence of features.

(b) *Neighborhood Approximation*: It approximates the behavior of the black-box model in the local neighborhood of the instance with an interpretable model (e.g., linear regression, logistic regression, or decision trees).

(c) *Data Perturbation and Weighting*: The data instance is perturbed to generate new samples. These samples are weighted using a kernel function that considers the proximity of each sample to the original instance.

(d) *Fitting the Surrogate Model*: The interpretable model is then trained on the perturbed samples, with weights applied to ensure it focuses on the local behavior of the black-box model.

Formally, LIME seeks to minimize the following objectives:

$$\zeta(x) = \arg\min_{h \in H} L(f, h, \pi_x(z)) + \Omega(h) \tag{1.7}$$

Where f is the Black-box model, x is the instance of interest, h represents an interpretable explanation model chosen from a class H, $\pi_x(z)$ represents a Kernel function that weights the samples z based on their distance to x, L represents the loss function quantifying how poorly h approximates f in the locality of x, and $\Omega(h)$ represents the complexity penalty ensuring the interpretability of h.

A critical aspect of LIME is the mapping to interpretable representations. For instance, binary vectors can indicate the presence or absence of specific features, values, or data patches. While this transformation facilitates human interpretability, it may introduce limitations such as loss of information or reduced representation power, particularly for high-dimensional data.

LIME offers several advantages, including flexibility, simplicity, and the ability to generate easy explanations for humans to understand. However, it has limitations. For highly non-linear models, LIME may fail to capture the full complexity of the black-box model's behavior around the instance of interest. Additionally, random perturbations used to sample the local neighborhood

may not adequately capture the diversity of the surrounding data, particularly in high-dimensional spaces, where defining meaningful distances between data points becomes difficult. Defining the kernel weighting function and ensuring it works effectively across dimensions requires careful consideration, and similar data instances may lead to significantly different explanations, which reduces interpretability. Moreover, studies [30] have shown that perturbation-based methods like LIME and SHAP can be exploited to generate misleading explanations for biased models.

3. **Example-Based Explanations** Example-based explanations are a valuable approach to understanding machine learning models by selecting specific instances from the dataset to illustrate model behavior or the underlying data distribution. Unlike other interpretability methods that summarize features, such as feature importance or partial dependence plots, example-based methods provide insight by directly referencing representative examples. These methods are typically model-agnostic, meaning they can be applied to any machine learning model, and are particularly effective for structured data, like images or text, where humans can easily interpret individual examples. For example, an image can be viewed directly, or a text can be read to understand its context. However, applying these methods to tabular data is more challenging due to the unstructured nature of such data and the potentially large number of features. Listing all feature values of a single instance often lacks meaning unless the data can be presented in a summarized and interpretable manner. Despite this challenge, example-based explanations remain an effective tool for helping users construct a mental model of how the machine learning system operates and the data on which it was trained. They can be particularly useful for understanding complex data distributions and offering relatable and intuitive explanations. For instance, a physician recalling a past case with similar symptoms to diagnose a current patient exemplifies how example-based reasoning aids decision-making, illustrating the power of this method in both human reasoning and machine learning interpretability. Molnar [21] cited several prominent Example-based techniques:

   - **Counterfactual explanations:** This type of explanation shows how an instance must change to produce a significantly different prediction. For example, in a loan approval system, a counterfactual explanation might identify that increasing an applicant's income by a specific amount would result in loan approval. These explanations help us understand the boundaries of the model's decision-making by highlighting the minimal changes required to alter the prediction. They are particularly useful for explaining individual predictions and for users seeking actionable insights.

- **Adversarial examples:** Adversarial examples are intentionally crafted counterfactuals designed to fool machine learning models. Unlike counterfactual explanations, their primary focus is not on interpretability but on exposing weaknesses in the model. For example, an adversarial example might involve subtle changes to an image that causes a model to misclassify it, even though the changes are imperceptible to humans. While adversarial examples are primarily used in the context of model robustness, they also reveal vulnerabilities in the model's decision-making process.
- **Prototypes and criticisms:** Prototypes are representative examples that summarize the dataset by capturing the central patterns or characteristics of the data. Criticisms, in contrast, are instances that are poorly represented by these prototypes and highlight outliers or regions of the dataset that the prototypes fail to capture. Together, prototypes and criticisms provide a balanced perspective on the dataset, offering insights into its structure and diversity. For example, in a medical dataset, a prototype might represent a typical case of a disease, while a criticism might highlight a rare or atypical presentation.
- **Influential instances:** Influential instances are the training data points that impact the model's parameters or predictions most. Identifying such instances can help debug the model, uncover biases in the training data, and understand the relationship between the data and the model's behavior. For example, suppose a specific data point disproportionately influences the prediction for a test instance. In that case, understanding this relationship can reveal whether the model is over-reliant on certain patterns in the data.
- **k-Nearest Neighbors (k-NN):** The k-NN algorithm is an interpretable machine learning model that relies entirely on examples. It makes predictions based on the k most similar instances in the dataset, using a defined distance metric. The simplicity of k-NN lies in its reliance on actual data points, making it intuitive for users to understand how a prediction is derived. For instance, in a classification problem, the label of a test instance is determined by the majority label of its nearest neighbors, providing a clear rationale for the decision.

1.3.3 Global vs. Local Explainability

- **Global methods:** Global explainability methods focus on interpreting the overall behavior of a model, providing insights into how features influence predictions across all cases. For instance, in a model predicting tuberculosis (TB), a global explanation could highlight how features such as persistent cough, abnormal chest X-ray findings, and

positive sputum test results collectively contribute to predictions across the dataset. By examining the importance of these features, clinicians can assess whether the model aligns with established diagnostic guidelines. Common techniques for global explainability include feature importance measures (e.g., permutation feature importance) and Partial Dependence Plots (PDPs), which illustrate how changes in a specific feature influence the model's predictions. Additionally, SHAP (SHapley Additive exPlanations) provides global insights by aggregating Shapley values for each feature across all predictions. For example, a SHAP summary plot could reveal that positive sputum test results have the highest overall influence on TB risk predictions. In contrast, abnormal chest X-ray findings have a more variable impact depending on other clinical features. Using these methods, clinicians comprehensively understand how features interact and contribute to the model's overall behavior, enhancing trust and ensuring alignment with diagnostic expectations. Another method, Testing with Concept Activation Vectors (TCAV), offers a unique approach by quantifying the influence of high-level, human-interpretable concepts on model predictions [31]. Unlike feature-based methods, TCAV evaluates how concepts affect predictions. By providing insights into how models leverage domain-specific concepts, TCAV facilitates trustworthiness and ensures the model's reasoning aligns with clinical understanding.

- **Local Methods:** Local methods explain individual predictions by identifying the specific factors that influenced the model's decision for a single patient. For example, in the case of a patient being assessed for COVID-19 severity, a local explanation might reveal that the model's prediction was driven primarily by the patient's oxygen saturation level, elevated D-dimer, and the presence of ground-glass opacities on a CT scan. Local methods such as SHAP values or counterfactual explanations can pinpoint the contribution of these features, providing clinicians with a clear rationale for the model's recommendation, such as prioritizing the patient for intensive care. Popular local interpretability methods include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). In this case, SHAP values could quantify how much each feature—like oxygen saturation—contributed to the severity score. At the same time, LIME might approximate the model's behavior for the patient by generating a simpler, interpretable local surrogate model. These methods provide clinicians with actionable insights into why a specific decision was made, fostering trust and facilitating personalized care.

1.3.4 Model-Agnostic vs. Model-Specific Methods

In explainable machine learning, methods can broadly be classified into two categories based on their applicability: model-agnostic and model-specific.

- **Model-Agnostic explanation methods:** Model-agnostic methods are designed to be applicable across a wide range of machine learning models, regardless of their underlying architecture. These methods do not rely on the internal specifics of a model and can be used to explain the predictions of any model, from decision trees to deep neural networks. They focus on the model's behavior as a whole and generate explanations that humans can interpret, regardless of the model's internal structure. Popular examples include feature importance measures, surrogate models, and explanation techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) [28, 29]. The primary advantage of model-agnostic methods is their versatility, as they can be applied to any machine-learning model without needing model-specific modifications. However, model-agnostic techniques often make certain trade-offs regarding accuracy or fidelity to the model. Since they work independently of the model's internals, they may not fully capture the intricate relationships between inputs and predictions, especially in complex models like deep neural networks. Consequently, the explanations provided by these methods might not always reflect the true underlying mechanisms of the model's decision-making process.

- **Model-Specific explanation methods:** Model-specific explanation methods are designed for particular model types, leveraging their architecture to provide detailed and context-aware insights. For instance, in convolutional neural networks (CNNs), they may visualize neuron activations, while in recurrent neural networks (RNNs), attention mechanisms can highlight influential input sequences [32, 33]. These methods offer precise explanations by utilizing the model's internal characteristics, but their limitation lies in being specific to certain models, making them less generalizable across different frameworks.

The choice between model-agnostic and model-specific methods depends on the application's needs. Model-agnostic methods are suited for flexible scenarios involving multiple or undefined models. In contrast, model-specific methods are ideal for cases requiring high-fidelity explanations, offering more accurate insights tailored to a particular model's architecture.

1.4 Properties of explanation

The rapid development of interpretable and explainable machine learning techniques in recent years has led to a growing need for formalizing the evaluation of these explanations. While significant progress has been made in defining the quality of explanations, challenges persist regarding the formalism, the assessment, and the effective measurement of these explanations. To evaluate the quality of an explanation, it is essential to establish its core attributes and

map various explanation techniques to these properties [34]. In the context of explainable machine learning, several key properties have been identified as essential for the effectiveness of explanations. These properties include clarity, parsimony, broadness, completeness, and soundness [35].

- **Clarity**: Refers to the ability of an explanation to provide a single, unambiguous reason for a decision or outcome. It ensures that the explanation is easily understandable and transparent.
- **Parsimony**: Implies that the explanation should be simple and compact, avoiding unnecessary complexity while conveying the necessary information.
- **Broadness**: Indicates that the explanation should be generalizable, meaning that the same explanation can be applied to a wide range of observations and instances within the system.
- **Completeness**: Refers to the extent to which the explanation provides sufficient information to fully understand the decision-making process and to compute the outcome for a given input.
- **Soundness**: This is an indicator of the correctness and truthfulness of the explanation, ensuring that the reasoning is accurate and reflects the true factors contributing to the decision.

When assessed collectively, these properties provide a comprehensive framework for understanding the effectiveness of explanations in machine learning systems [34].

1.5 Categories of explanation

In addition to the identification of key properties, explanations in machine learning can be classified into several categories, each serving distinct purposes and addressing different aspects of the decision-making process. Based on the work of the Information Commissioner's Office (ICO) and Webb et al., the following six categories of explanations have been proposed [20, 36, 37]:

1. **Rationale explanation**: This category addresses the "why" behind the decision made by the machine learning system. A rationale explanation explains why a particular decision was made, thereby allowing users to understand the justification behind the outcome. This type of explanation enables users to critique or validate the decision by identifying potential flaws or confirming the correctness of the reasoning.

2. **Responsibility explanation**: Responsibility explanations answer the "who" in the decision-making process. This category clarifies the roles and contributions of different stakeholders in the data processing, modeling, and decision-making steps. It ensures accountability by making the participants and their responsibilities traceable throughout the ML pipeline.

3. **Data explanation**: Data is fundamental to machine learning systems, as it drives the training, validation, and testing processes. Data explanations focus on the types of data used to train the model, the sources of the data, and how data preprocessing impacts the decisions made by the model. This explanation helps users understand the input factors influencing the model's behavior by providing insights into the data used.

4. **Fairness explanation**: Fairness explanations ensure that the machine learning model operates without bias or discrimination. These explanations focus on identifying and addressing disparities in the treatment of different groups or individuals within the system's decision-making process. They are critical for building trust in ML systems, as fairness is central to ethical decision-making and regulatory compliance.

5. **Safety and performance explanation**: This category of explanations emphasizes the model's decisions' reliability, robustness, and accuracy. Safety and performance explanations assess the model's performance in terms of its ability to produce consistent, reliable outcomes and to handle unexpected or extreme conditions without compromising its effectiveness. These explanations are crucial for understanding the model's operational limits and ensuring that it can perform under diverse scenarios.

6. **Impact explanation**: Impact explanations describe the broader consequences of the machine learning system's decisions. These explanations focus on the effects of the system's outputs on individuals, organizations, and society. By articulating the potential consequences, impact explanations help stakeholders understand the implications of deploying such systems.

Zhou et al. [34] further extend the classification by differentiating between *ethical* and *technical* explanations. Responsibility and fairness explanations are categorized as ethical explanations, focusing on accountability, equity, and justice. On the other hand, rationale, data, safety, and performance explanations are more directly related to the technical aspects of the model's explainability. The impact explanation, while connected to the use of the system, bridges both ethical and technical dimensions by evaluating the societal implications of the system's outcomes [20].

Through these categories, it is possible to map various explanation techniques to the properties of explainable systems, thus providing a structured framework for designing and evaluating machine learning explanations.

1.6 XAI model for infectious diseases diagnosis

The systematic literature review was presented in two focus groups. The first part focuses on the current state-of-the-art Clinical Decision Support Systems (CDSS) for Meningitis, and the second part focuses on current explainability approaches for various healthcare domains.

1.6.1 Advancements in clinical decision support systems for diagnosing Meningitis

Clinical Decision Support Systems (CDSS) are intelligent systems that assist medical professionals in facilitating decision-making at different stages of the diagnosis and treatment of diseases using specific recommendations [38].

CDSSs are classified into knowledge-based or non-knowledge-based, with the latter leveraging machine learning and artificial intelligence or statistical pattern recognition. However, non-knowledge-based systems face challenges, such as understanding AI's logic (black boxes) and obtaining high-quality data due to fragmentation, inconsistent formats, and privacy concerns [39]. Several studies have been proven effective in diagnosing different pathologies.

Regarding Meningitis diagnosis, the work presented by D'Angelo et al. [40] aims to improve the discrimination between bacterial and viral Meningitis etiologies through machine learning-based methodologies. Two cases were considered: one in which both blood and cerebrospinal parameters were taken into account and another in which only blood data were used. The results showed that a combination of clinical parameters is necessary to properly distinguish between the two Meningitis etiologies. The study used four classifiers: Naive Bayes, Multilayer perceptron (MLP), Decision tree-J48, and genetic programming (GP). The GP classifier achieved the best performance. It obtained 100% sensitivity in detecting bacterial Meningitis in nine out of ten folds.

Zaccari et al. [41] focused on developing a quantitative measure to help healthcare professionals decide whether or not patients need to undergo a CSF exam to diagnose Meningitis. Their approach involves using machine learning techniques to analyze data from blood and urine exams and patient chief complaint reports to identify patterns that could indicate the presence of Meningitis. The study used seven classifiers: Adaptive Boosting (AdaBoost), Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest and Support Vector Machines (SVM). Their analysis found that the Decision Tree model performed best, with an accuracy of 96.18%, 100% sensitivity, and 92.36% specificity. Although the ML model cannot fully substitute for the CSF exam, it can help doctors make more informed decisions about whether or not to recommend it.

Authors in [42] aimed to develop a system to classify subjects with Meningitis using a feedforward Artificial Neural Network (ANN). They employed two learning algorithms to develop the ANN: The Levenberg-Marquardt training algorithm, suitable for pattern recognition and particle swarm optimization (PSO) to adjust the decision threshold. The goal

was to achieve better performance by optimizing the decision threshold using a database that included several parameters such as temperature, CSF/blood glucose ratio, proteins, CSF leukocytes, glucose, lactates, erythrocyte sedimentation rate (ESR), and C-reactive protein (CRP).

This study [43] aimed to identify the best classification model to assist in diagnosing Meningitis. The researchers examined the performance of seven classification techniques applied to nine clinical symptoms of a patient, as well as their age, sex, and geographical location. They found that all models could predict Meningitis even before the completion of laboratory tests, indicating the possibility of a non-invasive and early diagnosis. The best classification technique was determined to be the J48 decision tree.

The researchers expanded their research on Meningitis [44] by creating a computerized decision support system (CDSS) that can help doctors identify the illness. They developed three decision models: DM1 determines if a patient has Meningitis based on observable symptoms, DM2 predicts the probability of meningococcal Meningitis using the same symptoms, and DM3 explores the disease's cause using chemical and cytological test data. The decision models achieved a high classification accuracy of 94.3% for Meningococcal Disease Meningitis. Evaluation of the system with real patient data showed that diagnosing Meningitis based solely on observable symptoms is challenging, but the CDSS correctly diagnosed 88% of Meningitis cases from the database.

The same researchers [45] explored data-driven techniques to differentiate between viral and bacterial Meningitis using a dataset of 26,228 patients and 19 attributes. They experimented with various sampling, feature selection, and classification models, finding that combining ensemble methods with decision trees achieved the best performance. The best classifiers had precision, recall, and f-measure of 89% and an AUC value of 95%. Their results suggest that this approach outperforms previous work using only decision trees.

Mentis et al. [46] studied the differential diagnosis of bacterial and viral meningitis using ML algorithms: multiple logistic regression (MLR), Random Forest (RF), and naïve-Bayes (NB). They analyzed patients of different age groups (0–14 years and >14 years) using both culture and molecular (PCR) methods. Various biomarkers, including CSF neutrophils, CSF lymphocytes, neutrophil-to-lymphocyte ratio (NLR), blood albumin, blood C-reactive protein (CRP), glucose, blood soluble urokinase-type plasminogen activator receptor (suPAR), and CSF lymphocytes-to-blood CRP ratio (LCR) were used. MLR and RF showed the best performance, indicating over 95% accuracy for viral meningitis and 78% for bacterial meningitis. The work in [47] used machine learning algorithms, including Logistic Regression (LR), K Nearest Neighbors (KNN), and Random Forest, to diagnose bacterial meningitis. These models achieved high accuracy rates: RF (90.6%), LR (90.3%), and KNN (90.1%). The study identified low education levels and the presence of red blood cells in the CSF as key predictors of patient mortality, suggesting the possibility of intracranial hemorrhage.

In [48], researchers employed Bayes Server to construct the predictive models. These models demonstrated high accuracy (99.99%) and sensitivity (97.12%) for meningococcal meningitis and its serogroup types (Serogroup type A, Serogroup type B, Serogroup type C, and *Neisseria meningitidis*) with 95.42% sensitivity.

1.6.2 Models Explainibility

Despite the promising results of previous studies in accurately predicting diagnosis, the black-box nature of these models poses a challenge for their adoption in clinical settings [49], as it can be challenging to comprehend the reasoning behind the model's predictions. This transparency is essential as it involves acknowledging AI usage and understanding how AI arrives at its conclusions or classifications [50]. Ensuring transparency in AI usage is essential, and it involves both the acknowledgment of AI usage and the understanding of how AI arrives at its conclusions [50]. Applying rigorous controls and testing from the medical field to AI deployment in healthcare reinforces this transparency, providing clear explanations of AI decision-making to ensure safety, accountability, and responsible use in medical settings [51].

Choi et al. [52] conducted a study on meningitis and encephalitis classification in patients hospitalized within the initial 24 hours. Various machine learning models were applied, including XGBoost, Random Forest, Light Gradient Boosting Machine, K-Nearest Neighbour, Gaussian Naive Bayes, and TabNet. An ensemble model (80% XGBoost, 20% TabNet) achieved the highest performance, with accuracy, precision, recall, and F1 score of 0.89 and AUROC of 0.91. Classifiers were applied to baseline characteristics, medical history, vital signs, and diagnostic results (CT, CXR, EEG). Laboratory findings from CSF, blood, and urine were also considered. Model-agnostic techniques (PIMP, LIME, SHAP) provided explainability. AI models slightly outperformed human clinicians due to the absence of certain factors considered in actual clinical practice. However, the researchers highlighted that AI still performed very well, suggesting it could be helpful for neurologists in making quick treatment decisions.

Yang et al. [53] conducted an insightful retrospective study on febrile infants aged ≤ 60 days, using a deep neural network to develop a predictive model of invasive bacterial infection (IBI). The model's performance was then compared to that of the IBI score. The SHapley Additive Explanations (SHAP) technique explained the model's different-level predictions. Five influential predictive variables (absolute neutrophil count, body temperature, heart rate, age, and C-reactive protein) were identified using SHapley Additive exPlanations. The study developed an explainable deep learning model that performs better than previous scoring systems and provides insight into how it arrives at its predictions through individual features and cases.

Sial et al. [54] introduced a non-invasive screening method for infant meningitis using

artificial intelligence and high-resolution ultrasound imaging. The study aim to address the limitations of lumbar punctures (LP), which are invasive, often yield negative results, and are not feasible in low-resource settings. The dataset comprised 2194 ultrasound images from 30 infants suspected of meningitis, collected across three Spanish University Hospitals. The authors developed a three-stage deep learning framework, with Stage 1 focusing on quality control and artifact removal, Stage 2 employing a deep learning model to classify images based on white blood cell count, and Stage 3 incorporating XAI techniques, such as GradCAM, to enhance model interpretability. The model achieved 96% accuracy in quality control, 93% precision, 92% accuracy in image-level meningitis detection, and 94% patient-level accuracy. With a single misclassification, it demonstrated 100% sensitivity and 90% specificity in identifying meningitis cases.

This study [55] aimed to predict the severity of COVID-19 by using Machine Learning and Deep Learning algorithms that consider various clinical markers, vital signs, and critical factors. The researchers evaluated five data-balancing techniques and twelve classifiers to find the most effective method. They discovered that Random Forest trained on Borderline SMOTE balanced data was the best-performing method, achieving an 83% recall rate in predicting COVID-19 severity. To better understand the models, the team deployed Explainable Artificial Intelligence tools such as Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), ELI5, Qlattice, Anchor, and Feature Importance to determine the importance of critical features in predicting COVID-19 severity. Their findings showed that respiratory rate, blood pressure, lactate, and calcium values were the primary contributors to the increase in severity of a COVID-19 patient. Ultimately, this architecture aims to serve as an explainable decision-support triaging system for medical professionals in countries lacking advanced medical technology and infrastructure to reduce COVID-19 fatalities.

Latifa et al. [56] investigate the impact of cytokines on the severity of SARS-CoV-2 infection. Plasma levels of 48 cytokines were measured in 87 participants from the COVID-19 study. Five models (Random Forest, XGBoost, Bagging Classifier, Decision Tree, and Gradient Boosting Classifier) were trained on synthetic data, with the Gradient Boosting Classifier showing superior performance. The interpretations of the Gradient Boosting model by Shapley additive explanation (SHAP) and the LIME (Local Interpretable Model-agnostic Explanations) provide detailed insights into the cytokine dataset. The results revealed significant variations in cytokine levels among COVID-19-infected patients, with VEGF-A, MIP-1b, and IL-17A showing elevated levels in severe cases. At the same time, M-CSF, IL-27, IL-9, IL-12p40, RANTES, and TNF were associated with non-severe cases and healthy individuals. These findings suggest the involvement of these cytokines in disease promotion and offer new possibilities for prevention and treatment.

In contrast, Mercaldo et al. [57] adopted a different approach by utilizing medical images

to detect coronavirus disease. They introduced a deep learning method that categorized computed tomography (CT) images into healthy patients, patients with pulmonary disease, and patients affected by Coronavirus 19. To provide explainability in their model, they employed the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm, which automatically highlighted the symptomatic areas of infection within CT images. This technique enhanced the diagnostic process by offering visual insights into the regions contributing to disease detection. Integrating Grad-CAM with deep learning improved the efficiency and accuracy of disease detection, providing valuable information for medical professionals. Shi et al. [58] employed machine learning techniques to diagnose tuberculous Meningitis, offering a potential solution to enhance diagnostic accuracy.
Several works have explored the application of the Testing with Concept Activation Vectors (TCAV) explainability approach in diverse healthcare settings, Kaveri et al [59] utilized TCAV to infer image concepts used by CNNs in detecting glaucoma. The study compared TCAV results with eye fixations of clinicians to validate AI predictions in ophthalmology. The researchers developed CNN architectures that demonstrated robust detection of glaucoma in optical coherence tomography (OCT) images. They found that the TCAV/eye-fixation comparison suggested important sub-images consistent with the areas of interest for clinicians. Mincu et al. [60] extends the TCAV method to time series data in electronic health records (EHRs), enabling sequential predictions and providing human-understandable explanations for adverse outcomes in clinical settings. The researchers applied TCAV to recurrent neural networks (RNNs) used for modeling adverse outcomes in EHRs, demonstrating increased performance compared to other approaches. This extension allowed for a better understanding of high-level concepts through the network's gradients. The approach was evaluated on an open EHR benchmark from the intensive care unit and synthetic data to isolate individual effects. Janik et al. [61] applied an extended version of TCAV, known as Discovering and Testing with Concept Activation Vectors (D-TCAV), to cardiac MRI analysis. The aim was to enhance the diagnosis of cardiac diseases by extracting important features from MRI data. The method provided score-based values of qualitative concepts and key performance metrics. D-TCAV offered a user-independent approach and reduced pre-processing time for clinicians, making it a valuable tool for clinical applications. The study found that the D-TCAV method provided meaningful and clinically relevant explanations for cardiac disease classification.

Building on previous works, our current work determines which signs or indicators have the highest predictive value by analyzing the Meningitis disease, providing valuable insights for accurate disease diagnosis. Our method derives the underlying factors influencing the IA-based algorithms' outcomes, allowing clinicians to trust and effectively incorporate AI-based recommendations into their clinical practice of Meningitis.

Table 1.1: Summary of studies on Meningitis diagnosis using AI techniques and the application of explainable AI in infectious disease diagnosis

Authors and Year	Objective	Dataset	Methods	Results	Limitations
D'Angelo et al. (2019) [40]	Discrimination between bacterial and viral Meningitis etiologies	Blood and CSF parameters from patients	Machine learning (Naive Bayes, MLP, Decision tree-J48, Genetic Programming)	Genetic Programming: 100% sensitivity in 9/10 folds	Limited to two etiologies (bacterial and viral)
Zaccari et al. (2019) [41]	Support meningitis diagnosis, minimizing invasive procedures	Blood, urine tests, and patient complaints from a children's hospital in Sao Paulo, Brazil.	Machine learning classifiers (ML algorithms (AdaBoost, Decision Tree, Gradient Boosting, KNN, Logistic Regression, Random Forest, SVM))	Decision Tree achieved 96.18% testing accuracy	Excludes non-structured patient complaints, assumes perfect CSF accuracy, binary diagnosis (the patient either has the disease or not), ignores time-dependency of meningitis.
Šeho, L., et al. (2022) [42]	Classify subjects with Meningitis	1,000 subjects (800 diseased, 200 healthy) with features including temperature, CSF/blood glucose ratio, proteins, CSF leukocytes, glucose, lactates, ESR, and CRP	Feedforward ANN with Levenberg-Marquardt training and Particle Swarm Optimization (PSO) for threshold optimization	96.69% accuracy for diagnosis	Limited to binary classification (diseased vs. healthy)
Lelis et al. (2017) [43]	Develop a statistical classifier to diagnose meningococcal meningitis early and non-invasively	22,602 records from Bahia, Brazil. Features: 9 symptoms, age, sex, and area of residence.	7 classifiers: J48, C4.5 (ID3) Decision Trees (DTs), SVM, ADTree, Random Forest (RF), Naive Bayes (NB), Bayesian Network (BN)	J48 classifier: Precision = 0.942, ROC area ¿ 0.95	Specific to meningococcal meningitis diagnosis
Lelis et al. (2020) [44]	Develop a Clinical Decision Support System (CDSS) for early Meningitis diagnosis	26,228 records from Brazil, data from clinical symptoms and test results (chemical/cytological data)	TThree tree-based decision models (DM1, DM2, DM3) and expert knowledge techniques	94.3% classification accuracy for the class Meningococcal Disease Meningitis	Classification limited to four meningitis categories (viral, bacterial, meningococcal, and other)
Guzman et al. (2022) [45]	Differentiate between viral and bacterial Meningitis using ensemble methods	26,228 patients, 19 attributes, including information related to the person, observable symptoms, and laboratory test results	27 classification models (19 ensemble methods, decision trees, and combinations of both)	bagging and NBTrees achieved 89% precision, recall, F-measure, and 95% AUC	Limited to binary classification (Bacterial vs. viral)
Mentis et al. (2021) [46]	Different between bacterial vs. viral Meningitis	Data from patients with meningitis in two age groups (0–14 and ¿14 years), with predictors including CSF neutrophils, CSF lymphocytes, NLR, blood albumin, CRP, glucose, suPAR, and LCR	Multiple Logistic Regression (MLR), Random Forest (RF), Naïve-Bayes (NB)	MLR and RF showed over 95% accuracy for viral Meningitis and 78% for bacterial Meningitis	Limited to binary classification (Bacterial vs. viral)
Pinheiro et al. (2022) [47]	Diagnose bacterial Meningitis using ML algorithms	Patient data from 2019 to 2021 from Brazil including demographics, symptoms, CSF lab results, and medical history	Machine learning algorithms (Logistic Regression, KNN, Random Forest)	Random Forest achieved 90.6% accuracy, with low education levels and the presence of red blood cells in the CSF as key predictors for mortality identified	Limited to binary classification
Alile et al. (2020) [48]	Construct predictive models for meningococcal Meningitis and serogroup types	Data on Meningococcal Meningitis serogroup types (A, B, C, Neisseria meningitidis)	Bayesian Belief Network implemented via Bayes Serve	99.99% accuracy, 97.12% sensitivity for meningococcal Meningitis, 95.42% sensitivity for serogroup types A, B, C, and and Neisseria Meningitidis	Focuses on meningococcal Meningitis, may not generalize to other types of Meningitis
Choi et al. (2023) [52]	Develop and validate an AI model for early aetiological determination of meningitis and encephalitis	Retrospective data (patients ¿18 years old) from two South Korean centers, training/testing (n=283), external validation (n=220), includes clinical variables within 24 hours of admission: vital signs (blood pressure, heart rate, respiratory rate, body temperature), diagnostic results (brain CT, chest X-ray, EEG), and laboratory findings from CSF, blood, and urine samples	Eight AI models (XGBoost, Random Forest, LightGBM, Category Boosting, KNN, Naive Bayes, TabNet), ensemble models (80% XGBoost, 20% TabNet), model-agnostic techniques (PIMP, LIME, SHAP)	Ensemble model (XGBoost + TabNet) achieved accuracy of 0.89, AUROC of 0.91. ML models slightly outperformed human clinicians	Classification limited to four categories (autoimmunity, bacteria, virus, and tuberculosis), and Small dataset

Authors & Year	Objective	Dataset	Methods	Results	Limitations
Yang et al. (2023) [53]	Develop and validate an explainable deep learning model to predict invasive bacterial infection (IBI) in febrile infants ≤60 days old	Retrospective data of febrile infants (n=1847) presenting to a Taiwanese emergency department (2011–2019), including demographics, vital signs, and laboratory findings such as blood counts, CRP, and urinary tests	Deep neural network, SHAP (Shapley Additive Explanations) for explainability	Deep learning model outperformed traditional scoring systems in specificity (54%), positive predictive value (5%), and AUROC (0.87). Identified key predictive variables: absolute neutrophil count, body temperature, heart rate, age, and C-reactive protein	Limited to binary classification (IBI vs. non-IBI)
Sial, Hassan, et al. (2024) [54]	Non-invasive screening of infant meningitis using AI	30 infants from Spanish hospitals, 2194 ultrasound images	Deep learning, GradCAM, statistical analysis	96% accuracy in quality control, 93% precision and 92% accuracy in image-level detection, and 94% patient-level accuracy	Small sample size
Khanna et al. (2023) [55]	Predict COVID-19 severity using ML/DL algorithms with explainable insights into key features	Open-source Kaggle data from Sírio-Libanês Hospital, includes 1925 COVID-19 patients with 231 parameters (demographics, grouped diseases, blood parameters, vital signs)	12 classifiers: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, XGBoost, ExtraTrees, AdaBoost, LightGBM, CatBoost, and 1-D Convolutional Neural NetworkRandom, Explainable AI Tools: SHAP, LIME, ELI5, Qlattice, Anchor, and Feature Importance	Random Forest with Borderline SMOTE data achieved 83% recall in predicting COVID-19 severity	May not apply to other diseases beyond COVID-19
Latifa et al. (2023) [56]	Investigate the impact of cytokines on SARS-CoV-2 infection severity and provide explainability for the model	Plasma levels of 48 cytokines measured in the blood of COVID-19 patients	Random Forest, XGBoost, Bagging Classifier, Decision Tree, Gradient Boosting Classifier and SHAP, LIME for model explainability	Gradient Boosting Classifier showed superior performance, Elevated VEGF-A, MIP-1β, IL-17 indicate severity and cytokine storm, Higher RANTES, TNF are linked to no infection, and IL-27, IL-9, IL-12p40, MCP-3 are associated with mild disease	Small sample size, findings may require validation in larger populations, limited to COVID-19
Mercaldo et al. (2023) [57]	Develop a deep learning method for detecting coronavirus disease (COVID-19) using CT scans medical images	Consists of medical images from CT scans: 20 COVID-19 patients, 9 patients with other pulmonary diseases, and 16 healthy patients	Deep learning model to classify CT images into three categories and Gradient-weighted Class Activation Mapping (Grad-CAM) used to highlight symptomatic areas in CT scans for better interpretability	The method achieved high accuracy (0.95) in classifying CT images and could identify areas symptomatic of COVID-19 infection within 8.9 seconds	Small sample size, findings may require validation in larger populations, limited to COVID-19

1.7 Conclusion

Explainable AI (XAI) is an essential and innovative approach that enhances human understanding of the results generated by AI systems. In critical domains, such as healthcare, XAI plays a vital role in interpreting the outputs of AI and machine learning models. Ethical explanations play a pivotal role in addressing issues of accountability, fairness, and equity, ensuring that AI systems operate justly and responsibly within societal frameworks. On the other hand, technical explanations provide insights into the model's underlying rationale, data handling, and performance metrics, enabling a deeper understanding of the system's functionality and reliability. By integrating these ethical and technical dimensions, XAI not only builds trust and transparency but also evaluates the broader societal implications of AI applications, fostering accountability and aligning these technologies with human values. This dual approach ensures that AI systems are both robust and socially responsible, paving the way for their ethical deployment across critical sectors.

In healthcare, where decisions can be life-changing, using XAI is especially important. XAI makes AI predictions and recommendations easier to understand, which helps build trust between doctors, patients, and healthcare teams. It also supports better decision-making and ensures AI tools are used responsibly in diagnosing, planning treatments, and caring for patients. This clarity helps people trust and use AI more, while also protecting patients, making XAI a key part of future healthcare advancements.

Chapter 2

Comprehensive review of infectious diseases

*****

2.1 Introduction

Infectious diseases continue to pose major public health challenges worldwide, impacting individuals, communities, and healthcare systems alike. Among these illnesses, meningitis, a life-threatening inflammation of the meninges, the protective layers surrounding the brain and spinal cord, stands out due to its rapid onset, high risk of severe complications, and considerable mortality if not rapidly diagnosed and treated. Meningitis can arise from a variety of pathogens, including bacteria, viruses, fungi, and parasites, each presenting distinct clinical and diagnostic complexities. The evolving understanding of infectious diseases like meningitis has emphasized the need for rapid, precise diagnostic tools that are essential for improving patient outcomes and guiding effective treatment strategies.

This chapter provides an in-depth review of infectious diseases, beginning with an overview of various infectious causes, their reservoirs, transmission, biological characteristics, and how they are quantified across different populations. It then explores the temporal patterns of infectious diseases, highlighting their evolving nature and the factors that contribute to these trends. The chapter further investigates central nervous system infections, with a focused discussion on meningitis, covering both viral and bacterial causes, their epidemiology, clinical presentation, and diagnostic challenges. Special attention is given to

differentiating bacterial from viral meningitis and the diagnostic tests used for accurate identification. Additionally, the chapter includes a comprehensive analysis of laboratory tests in cerebrospinal fluid across different age categories in the context of meningitis. Finally, we discuss the integration of artificial intelligence in healthcare, emphasizing its role in enhancing diagnostic accuracy and the importance of explainability in medical AI systems.

2.2 Infectious causes

Healthcare professionals often classify infectious diseases by the primary clinical symptoms or by the organ systems predominantly affected. For instance, Table 2.1 illustrates an example of such a clinical classification. Microbiologists, on the other hand, categorize diseases based on the properties of the responsible pathogen, as shown in Table 2.2. From an epidemiologic standpoint, classification typically focuses on transmission patterns or the natural habitat of the organism. Infectious diseases can be grouped into five categories based on their transmission method, as outlined in Table 2.3. Additionally, epidemiologists may classify diseases by the reservoir of the pathogen, whether it is primarily associated with humans, animals, soil, or water. Table 2.4 provides examples of infectious diseases organized according to their reservoir.

Table 2.1: Clinical classification of infections [62]

Classification	**Infection type**
Diarrheal diseases	Secretory (e.g., cholera, traveler's diarrhea) Invasive (e.g., dysentery)
Respiratory diseases	Upper respiratory (e.g., sinusitis, pharyngitis) Lower respiratory (e.g., pneumonia, bronchitis)
Central nervous system infections	Meningitis: Bacterial (e.g., pneumococcal, meningococcal) vs. Aseptic (often viral) Encephalitis (e.g., viral encephalitis) Abscess (e.g., brain abscess)
Cardiovascular infections	Endocarditis (infection of heart valves) Myocarditis (infection or inflammation of heart muscle) Vasculitis (inflammation of blood vessels)
Sepsis	Disseminated infection (e.g., indicated in latex tests)

When a new infectious disease emerges, the microorganism's characteristics are often poorly understood, and the complete clinical profile may remain unclear. For example, it was initially unknown that *Borrelia burgdorferi*, the bacteria causing Lyme disease, led not only to *erythema chronica migrans* (ECM) skin lesions but also to arthritis, cardiovascular issues, and neurological symptoms such as Bell's palsy and encephalitis. The complete spectrum of symptoms caused by *B. burgdorferi* continues to be investigated. Even with limited microbiological data, knowing the agent's reservoir and transmission method can

Table 2.2: Microbiologic classification of infectious diseases [62]

Classification	**Organism type**
Bacterial	Gram-negative Gram-positive
Viral	DNA virus RNA virus Enveloped vs. non-enveloped viruses
Fungal	Disseminated (biphasic) Localized
Parasitic	Protozoa Helminths Trematodes Cestodes
Prion	Protein

Table 2.3: Transmission-based classification of infectious agents [62]

Transmission type	**Characteristics**
Contact	Requires direct or indirect contact (indirect: infected fomite, blood, or body fluid; direct: skin or sexual contact)
Food- or Water-borne	Ingestion of contaminated food (outbreaks may be large and dispersed, depending on distribution of food)
Airborne	Inhalation of contaminated air
Vector-borne	Dependent on biology of the vector (mosquito, tick, snail, etc.), as well as the infectivity of the organism
Perinatal	Similar to contact infection, however, the contact may occur in utero during pregnancy or at the time of delivery

help establish strategies to prevent its spread. John Snow's work during the 1853 cholera outbreak in London demonstrated that contaminated water was a source, paving the way for control measures even before *Vibrio cholerae* was identified by Robert Koch in 1884. Similarly, Budd's findings in 1858 linked human carriers to outbreaks of typhoid fever long before *Salmonella typhi* was isolated in 1880 by Eberth. Walter Reed's studies in 1901 showed that yellow fever could be transmitted by *Aedes aegypti* mosquitoes, while the virus was only isolated in 1928 by Stokes and his team. Later, the pneumonia outbreak at the American Legion convention in 1976 was traced to airborne transmission from a hotel air-conditioning system, suggesting prevention strategies by avoiding exposure, although *Legionella pneumophila* was only identified in 1978 by the federal Centers for Disease Control and Prevention (CDC) researchers McDade and Sheppard. Identifying a disease's reservoir is often crucial to developing effective control measures. Before Snow's demonstration that contaminated water was the source of cholera in London, the prevailing miasma theory mistakenly attributed the disease to foul air exposure. Snow's analysis, showing higher cholera rates among individuals using water from a particular company, led to effective outbreak

Table 2.4: Reservoir-Based classification of infectious agents [62]

Reservoir	**Some typical organisms**
Human	*Treponema pallidum, Neisseria gonorrhoeae*, HIV, hepatitis B and C virus, *Shigella, S. typhi*
Animals (zoonoses)	Rabies, *Yersinia pestis, Leptospira*, nontyphoid *Salmonella, Brucella*
Soil	*Histoplasma capsulatum* (and other systemic fungi), *Clostridium tetani, Clostridium botulinum*
Water	*Legionella, Pseudomonas aeruginosa, Mycobacterium marinum*

control by shutting down the implicated water source, providing compelling evidence for the waterborne transmission route [62].

2.2.1 Biologic Characteristics of the organism

In understanding infectious diseases, the biological characteristics of the causative organisms are crucial, as they influence both the pathogenicity and the spread of infections [63]. These characteristics encompass structural features, genetic makeup, and reproductive methods that distinguish different pathogens and dictate how they interact with hosts, evade immune responses, and cause disease. Examining these properties provides insight into the mechanisms of infection and guides the development of targeted treatments and preventive measures. In [62], the authors provide an in-depth discussion on various biologic characteristics of pathogens:

- **Infectivity:** refers to an agent's capacity to establish an infection in a susceptible host. This is typically measured by the minimum quantity of infectious particles necessary to initiate an infection. Infectivity can also be assessed by the secondary attack rate for diseases transmitted between individuals, representing the proportion of susceptible people who develop the infection following exposure.

- **Pathogenicity:** describes the capacity of a microbial agent to cause disease. Certain illnesses, including rabies, smallpox, measles, chickenpox, and rhinovirus colds, demonstrate high pathogenicity. In contrast, conditions like Poliomyelitis and arbovirus infections, transmitted by mosquitoes, exhibit lower levels of pathogenicity. In other terms, infectivity measures how well a pathogen can invade and reproduce in a host. In contrast, pathogenicity measures the extent to which that pathogen can cause disease and the severity of the resulting illness.

- **Virulence:** Virulence and pathogenicity are often confused, but they refer to different aspects of infectious agents. Pathogenicity describes an organism's ability to cause disease, whereas virulence measures how severe that disease is once it occurs. For

example, smallpox and the common cold can cause illness, but smallpox is much more virulent because it can lead to severe complications or death. We can categorize infectious agents by their ability to spread (infectivity), cause disease (pathogenicity), and the severity of that disease (virulence). It's also essential to understand that these characteristics can change over time; diseases like syphilis and streptococcal infections, once very virulent, have become much less so.

- **Immunogenicity:** Immunogenicity refers to an organism's ability to trigger an immune response that protects against future infections by the same or similar pathogens. For example, infections from measles, Poliomyelitis, hepatitis B, and rubella usually lead to long-lasting immunity. In contrast, some organisms, like *Neisseria gonorrhoeae* and *Plasmodium falciparum*, provoke weaker immune responses, resulting in frequent reinfections. Research into the antigens responsible for protective immunity has been instrumental in developing effective vaccines. However, some immune responses can be harmful. For instance, certain group strains *A streptococci* can cause conditions like glomerulonephritis due to antibodies mistakenly targeting the body's tissues. Moreover, some infections can produce antibodies indicating past or current infections but not confer immunity. These are binding antibodies found in diseases like hepatitis C, HIV, and HSV-2, where individuals may have antibodies but remain susceptible to reinfection.

- **Inapparent infections:** refers to an infection where an organism can be detected through culture, polymerase chain reaction (PCR) testing, or specific immune responses, yet the individual shows no symptoms. The prevalence of such asymptomatic infections can provide insight into the organism's pathogenicity. Inapparent infections are common with many pathogens and can significantly contribute to the spread of epidemics. For instance, most Poliomyelitis cases are asymptomatic, and many individuals can carry *Neisseria meningitidis* in their nasopharynx without showing signs of illness, especially during outbreaks. Identifying and treating carriers of bacteria like *meningococci* or *Staphylococcus aureus* is crucial for controlling transmission, as asymptomatic individuals can be key transmitters. In the U.S., individuals who test positive for *Mycobacterium tuberculosis* but remain asymptomatic are often treated to prevent the development of active tuberculosis and reduce the risk of spreading the infection to others. In contrast, some diseases, such as measles, varicella, smallpox, and hantavirus infections, usually present with symptoms. Understanding the ratio of symptomatic to asymptomatic infections is vital for managing disease transmission and developing effective control strategies during epidemics.

- **The carrier state** Asymptomatic carriers, or healthy individuals who carry infectious agents without showing symptoms, are essential to understanding disease spread. A

classic example is "Typhoid Mary," an Irish cook in early 1900s New York City. Although healthy, she spread typhoid fever to over 250 people while working in various homes and hospitals. She unknowingly carried Salmonella typhi in her gallbladder [62, 64], shedding the bacteria in her stool, which contaminated the food she prepared. Her case highlighted the risk asymptomatic carriers pose, particularly for diseases spread through person-to-person contact. Asymptomatic carriers play a role in spreading many infections. They can harbor bacteria or viruses in various body parts, such as the respiratory or genital tracts, stool, or blood.

- **Transfusion-transmitted infection:** Infection transmission through blood transfusion has become a growing concern over the past 20 years. While hepatitis B virus (HBV) transmission was recognized early, the introduction of blood donor screening for HBsAg in 1973 significantly lowered the risk. However, even after this screening, cases of post-transfusion hepatitis continued, pointing to other causes. When hepatitis C virus (HCV) was identified, screening for it began in 1990, further reducing risk. The spread of HIV/AIDS through transfusions among hemophiliacs and other recipients underscored the importance of screening. Today, blood donors are carefully questioned about potential infection risks and screened for multiple pathogens. Pooled plasma products are also treated to inactivate viruses, including heat treatments. Despite these precautions, new infectious agents with the potential to be transmitted by blood transfusion are still being discovered.

2.2.2 Quantification of infectious diseases

Epidemiologists employ various metrics to assess disease occurrence, aiming to estimate either the overall disease burden in a population or the rate of new cases (incidence) [62, 65].

- **Prevalence:** Epidemiologists calculate prevalence to measure the proportion of a population affected by a disease. This is determined by dividing the number of infected individuals by the total population. The numerator may represent those who are symptomatic, exhibit specific symptoms, or show microbiologic evidence of infection without symptoms, depending on the study focus. Each approach offers a valid view of prevalence as long as infection criteria are defined. The denominator may include the entire population or just those exposed to the disease. Measuring prevalence in the full population indicates the overall disease burden, while focusing on exposed individuals reveals prevalence within that group. Age-specific prevalence is used when exposures are common, while rates by exposure group are used for rare exposures.

- **Incidence:** This measure reflects the rate at which new cases arise in a population or the transmission rate of the infectious agent. Unlike prevalence, incidence always

includes a time factor, quantifying the number of cases occurring within a specific period, such as annually, monthly, or weekly.

2.3 Temporal patterns of infectious diseases

The incidence of many infectious diseases fluctuates over time. This variability can often be attributed to changes in exposure to the infectious agent, which may differ across seasons or between years.

- **Seasonal variation:** Vector-transmitted diseases, such as malaria, dengue, and St. Louis encephalitis (SLE), rely on contact with infected mosquito vectors for transmission [66]. In temperate climates, these diseases appear only during the warmer months when mosquitoes are active. For example, a study by Teshager Zerihun Nigussie et al. [67] on malaria transmission in northwest Ethiopia reveals seasonal trends influenced by climate and environmental factors. The researchers found that malaria transmission was seasonal, with a higher number of cases occurring from September to November, with notable spatial effects in areas near the Abay gorge and the Sudan border. The long-term trend of malaria incidence decreased between 2012 and 2018 but increased since 2019, with notable spatial effects in areas near the Abay gorge and the Sudan border. The study highlighted how climate and environmental predictors have nonlinear effects on malaria incidence, with spatial, temporal, and space-time interaction effects playing more significant roles in explaining transmission patterns. These findings emphasize the challenges in predicting malaria outbreaks, as they depend on various factors such as rainfall, temperature, and ecological features.

- **Annual variation:** Before effective vaccines were developed for common childhood infections like measles, mumps, rubella, and varicella, these diseases showed distinct, repetitive cycles. These cycles largely depended on each epidemic depleting the population of susceptible individuals, with new birth cohorts gradually restoring this group. For example, in urban areas of the United States, measles epidemics typically occurred every two years, with case numbers roughly doubling compared to surrounding years. However, with the widespread adoption of the measles vaccine, incidence has dramatically decreased, and this cyclical pattern has been disrupted. In contrast, pertussis (whooping cough) continued to show a persistent 3- to 4-year cycle in reported cases between 1967 and 1997, despite the availability and use of a vaccine. This ongoing pattern suggests that pertussis transmission still occurs frequently, possibly due to waning immunity from the whole-cell pertussis vaccine, transmission by older children and adults, and the steady replenishment of susceptible individuals. Most childhood infections also tend to peak in winter and early spring, likely due to higher transmission

rates when people spend more time indoors. Additionally, the low humidity of indoor air and the presence of other respiratory infections, which promote coughing and sneezing, are thought to play a role in promoting the transmission during the winter.

- **Herd immunity:** Before the epidemiological theories proposed by Kermack and McKendrick, as well as Reed and Frost from Johns Hopkins, the prevailing belief was that epidemics were primarily caused by variations in the infectivity of pathogens. However, the research conducted by these scientists revealed that epidemic patterns could be better understood by examining the proportion and distribution of susceptible individuals within a population. In diseases that spread from person to person, the immunity level within the population is crucial for determining the likelihood of an epidemic and the associated risk of infection for susceptible individuals. Transmission relies on interactions between infected and susceptible individuals, so if a sufficient number of individuals are immune, it becomes unlikely for a susceptible person to encounter an infected one. This phenomenon is known as herd immunity. While some susceptible individuals may still exist in the population, epidemics cannot sustain themselves because the daily interactions do not facilitate contact between contagious infected persons and those who remain susceptible. The specific level of immunity required to achieve herd immunity varies depending on the characteristics of the infectious disease. Diseases with higher transmissibility necessitate a greater level of population immunity than those that spread less easily. The dynamics of herd immunity and individual susceptibility are key epidemiological factors influencing the periodicity and long-term trends observed in diseases such as measles, rubella, varicella, and Poliomyelitis.

2.4 Central Nervous System infections

Infections of the central nervous system (CNS) are fortunately infrequent but can have severe consequences. The brain and spinal cord are encased within the rigid structures of the skull and spinal canal, which limits the space for inflammation and swelling. As a result, any edema can lead to significant tissue damage, including infarction, with the potential for permanent neurological impairments or fatality. The brain and spinal cord are surrounded by cerebrospinal fluid (CSF), which is produced by the choroid plexus in the cerebral ventricles and absorbed by the arachnoid villi, draining into the superior sagittal sinus. These neural structures are covered by three layers of protective membranes called the meninges. The innermost two layers, the pia mater and the arachnoid, together form the leptomeninges. The outermost layer, the dura mater, provides additional structural protection. The CSF circulates through the subarachnoid space between the pia mater and the arachnoid, serving as a cushioning fluid that helps protect the CNS from mechanical injury (Figure 2.1).

CNS infections are categorized based on their location. Infection of the brain tissue is

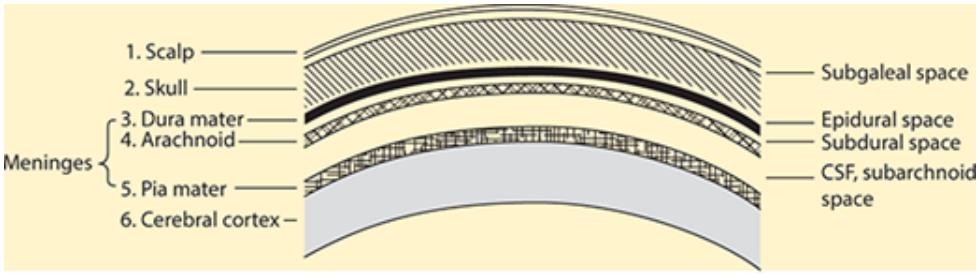


Figure 2.1: Representation of the subgaleal, epidural, subdural, and subarachnoid spaces within the central nervous system. [68]

known as encephalitis, while infection of the meninges is called meningitis. Abscesses can form in three main areas of the CNS: within the brain itself (brain abscess), between the dura and arachnoid layers (subdural abscess), or outside the dura (epidural abscess). The blood vessels of the brain and spinal cord have unique tight junctions, forming a blood-brain barrier that limits permeability, protecting the CNS from pathogens and toxins. However, this barrier also restricts the entry of immune defenses like immunoglobulins, complement, and many antibiotics. When pathogens cross the blood-brain barrier, the CNS's defense mechanisms are compromised, contributing to CNS infections' rapid and severe nature. Treating these infections requires drugs that can cross the barrier effectively, often at high doses (referred to as "meningeal doses") to ensure therapeutic levels within the CNS [68].

2.4.1 Meningitis

Meningitis is a significant health condition that needs appropriate treatment, and preventative strategies. It can be caused by viruses, bacteria, parasites, and fungus. Viral meningitis can be caused by enterovirus, parechovirus, herpes simplex virus, influenza virus, cytomegalovirus, varicella-zoster virus, mumps virus, measles virus, coronavirus, adenovirus, and human immunodeficiency virus. The most common bacteria isolated are Haemophilus influenzae, Streptococcus pneumoniae, Neisseria meningitidis, and Listeria monocytogenes . Aseptic meningitis can refer to inflammation of the meninges caused by other than pus-producing bacteria , such "atypical" bacteria (e.g., Borrelia burgdorferi, Leptospira spp., Mycobacteria, Brucella, Treponema pallidum), fungi (e.g., Cryptococcus neoformans Candida and Aspergillus species), parasites (e.g.Plasmodium falciparum, Naegleria fowlerie).

2.4.2 Viral Meningitis

2.4.2.1 Enteroviruses/parechoviruses:

Enteroviruses (EV) and parechoviruses (HPeV), are members of Picornavirus family. Numerous human infections, such as poliovirus, coxsackie A and B viruses, echoviruses, numbered

enteroviruses (such as EV-A71 and EV-D68), and rhinoviruses, are members of the genus Enterovirus. There are six species in the genus Parechovirus, ranging from Parechovirus A to Parechovirus F, and each species has a variety of genotypes. HPeV was categorized as Parechovirus A, which has 19 genotypes ranging from HPeV-1 to -19 and is only found in humans.

EV and HPeV, have a worldwide distribution and are seasonal in temperate regions, often peaking in summer and fall [69–72]. Transmission primarily occurs via the fecal-oral route and, less frequently, through respiratory droplets [71, 73]. Infants and young children are most vulnerable, with risk factors for severe illness, including the absence of oral lesions, seizures, and lethargy [74, 75]. In adults, EV and HPeV infections mostly present as aseptic meningitis with generally good outcomes [70]. Neonates are particularly susceptible, and severe infections can lead to meningoencephalitis or sepsis-like syndromes involving multiple organs, disseminated intravascular coagulation, seizures, focal neurological signs, and cardiovascular collapse may develop [71, 76]. A multiplex real time PCR (Fast Track Diagnostic FTD viral meningitis) assay has been used since 2015 for assessing CSF as virological diagnostic testing.

2.4.2.2 Herpes Viruses

Herpes viruses encompass several types, including varicella-zoster virus, human herpesviruses 6, 7, and 8, Epstein-Barr virus, cytomegalovirus, and herpes simplex virus (HSV) types 1 and 2 [69]. One of the main causes of meningitis and encephalitis is the herpes simplex virus (HSV). About 90% of HSV encephalitis is caused by HSV type 1 (HSV-1) and can affect people of any age [77]. Oral contact is the primary way that HSV-1 transmits and is known to result in oral sores. Usually, a virus from an earlier infection reactivates to cause meningitis and encephalitis. Herpetic meningoencephalitis commonly presents with symptoms such as headache, fever, stiff neck, hallucinations, and confusion [78]. Diagnosis typically involves polymerase chain reaction (PCR) testing of cerebrospinal fluid (CSF), and without timely and appropriate therapy, mortality rates can rise to 70% [79]. Varicella-zoster virus (VZV) is also linked to aseptic meningitis, sometimes presenting without typical skin lesions (a form known as zoster sine herpete) [80]. VZV may be underdiagnosed, as only 1.2% of aseptic meningitis cases receive CSF VZV PCR testing [70].

2.4.2.3 Arboviruses

Arboviruses (arthropod-borne viruses), transmitted by mosquitoes, ticks, or sandflies, include several virus families responsible for aseptic meningitis [81]. West Nile virus (WNV) is a virus that belongs to the flavivirus family and develops after an infected mosquito bites especially Culex species [82]. The virus circulates in a cycle between mosquitoes and birds.

WNV has been expanding geographically becoming the most cited arbovirus worldwide in regions that include the American, Europe, and Mediterranean coastal countries. In Algeria, cases of West Nile Virus (WNV) meningoencephalitis have been reported as early as 1994 in Tinerkouk, located in the southwestern Sahara region. Additionally, a neurotropic WNV infection was identified in the eastern coastal region during the fall of 2012 [83, 84]. West Nile fever (WNF) has been reported as a mild clinical but can progress to episodes of meningoencephalitis or even flaccid paralysis [85]

Other arboviruses causing aseptic meningitis in the USA include mosquito-borne *St. Louis encephalitis* and the California encephalitis group of viruses, along with tick-borne Powassan virus in northern central and eastern USA, and coltivirus in the mountainous and western regions of the USA and Canada [81]. Other less common arboviruses in the USA that can cause aseptic meningitis are the two mosquito-borne illnesses, *St. Louis encephalitis* and the California encephalitis group of viruses, and two tick-borne illnesses, Powassan virus in northern central and eastern USA and coltivirus (agent of Colorado tick fever) in the mountainous and western regions of the USA and Canada [81]. Toscana virus has emerged as one of the most common causes of meningitis or encephalitis during the summer in the Mediterranean countries [86]. It is transmitted by sandflies and is caused by a bunyavirus. In Europe, tick-borne encephalitis can be associated with a complex syndrome of meningoencephaloradiculitis (MER), which carries a relatively high risk of severe disease, including the need for intensive care and mechanical ventilation. Age, male sex, and preexisting diabetes mellitus were predictive of the more severe MER [87].

2.4.2.4 Other Viruses

Lymphocytic choriomeningitis virus (LCMV), though now rarely reported, can cause aseptic meningitis [69]. LCMV is primarily transmitted to humans via contact with rodents or their excreta [69, 88], posing a risk to laboratory workers, pet owners, and those in poor living conditions. No human-to-human transmission has been reported. Mumps, in unimmunized populations, is another cause of aseptic meningitis [69]. However, since the measles-mumps-rubella (MMR) vaccine was introduced, mumps-related meningitis has dropped to $< 1\%$ of all meningitis cases in the US and UK [89, 90]. Human immunodeficiency virus (HIV) can also lead to aseptic meningitis, often during seroconversion, presenting with mononucleosis-like symptoms [69]. Vaccine-preventable Japanese encephalitis continues to cause meningitis and encephalitis where vaccination is unavailable [91]. Dengue, chikungunya, and Zika viruses are emerging as causes of meningitis and encephalitis globally [92, 93]. Additionally, the Ebola epidemic in West Africa revealed that Ebola can lead to viral relapse with acute meningitis, treatable with experimental antiviral therapy and corticosteroids [94].

2.4.3 Bacterial Meningitis

Bacterial meningitis may present with a negative Gram stain [95]. Typical symptoms include fever, headache, neck stiffness, and altered mental status, though presentation can vary by age, health status, and bacterial pathogen [96]. CSF findings generally show >1000 WBC/mm^3 with neutrophil predominance, protein >100 mg/dl, and glucose <40 mg/dl, however, similar findings may occasionally appear in viral meningitis as well [97, 98]. Patients with infective endocarditis caused by *Staphylococcus aureus* or *Streptococcus pneumoniae* may develop meningitis [69], and infections near the meninges, such as epidural or subdural empyemas from sinusitis, otitis, or mastoiditis, can also lead to meningitis [69].

2.4.3.1 Epidemiology

According to the National Guidelines for the Management of Community-Acquired Bacterial Meningitis and Invasive Meningococcal Infections by the General Directorate for Prevention and Health Promotion in Algeria in 2022, the frequency of bacterial meningitis remains relatively low compared to clear-fluid meningitis, primarily in children under 5. It fluctuates from year to year, with a minimum rate of 11.2% in 2014 and a maximum rate of 29.8% in 2009 among all meningitis cases. The proportion of meningitis cases caused by undetermined pathogens constitutes a significant share of all reported cases. Annual incidences of meningococcal meningitis have declined, from 0.30 cases per 100,000 inhabitants in 2008 to 0.04 cases per 100,000 in 2020. However, a peak was recorded in 2013, with an incidence of 0.23 per 100,000 inhabitants. *Haemophilus influenzae type B* (Hib) and *Streptococcus pneumoniae (S. pneumonia)* are included in the Expanded Program on Immunization, which aims to prevent infections caused by these pathogens. In October 2008, vaccination against *Haemophilus influenzae type B* was introduced in Algeria. Within a very short time, the number of infections caused by Hib, including meningitis, decreased significantly. The introduction of pneumococcal vaccination in April 2016 also contributed to a decrease in the burden of pneumococcal infections.

The primary pathogens responsible for community-acquired bacterial meningitis in children and adults are defined in [68] as four major organisms:

- ***Streptococcus pneumoniae:*** *S. pneumoniae* is the leading cause of community-acquired meningitis in the United States. The infection typically begins in the ear, sinuses, or lungs before spreading to the bloodstream, where it reaches the meninges. *S. pneumoniae* is also the most common cause of recurrent meningitis in patients with a CSF leak following head trauma.

- ***Neisseria meningitidis:*** This organism can cause both isolated cases and epidemics. It typically infects the nasopharynx, causing a sore throat, and in individuals without

antimeningococcal antibodies, it can enter the bloodstream and infect the meninges. Crowded settings like college dormitories or military facilities increase the risk of *N. meningitidis* transmission, particularly in winter, when respiratory spread is more frequent.

- ***Listeria monocytogenes:*** *L. monocytogenes* primarily affects people with compromised cell-mediated immunity, such as pregnant women, newborns, immunosuppressed patients, HIV-positive individuals, and those over 60. This meningitis is typically contracted through ingesting contaminated food, as *L. monocytogenes* can grow in refrigerated conditions (4°C). Foods at risk include unpasteurized soft cheeses and improperly processed hot dogs and fish. Upon entry through the gastrointestinal tract, Listeria can invade the lining, enter the bloodstream, and infect the meninges.

- ***Haemophilus influenzae:*** Prior to widespread use of the Hib vaccine, *H. influenzae* was the most common cause of meningitis in children. However, cases of meningitis from this pathogen have become rare.

2.4.4 Differentiation between bacterial and viral Meningitis

Although viral infections are the most common causes of meningitis and encephalitis, most patients are admitted and receive empirical antibiotic treatment due to the challenges in differentiating bacterial from viral causes based solely on initial clinical presentation [90,95]. To aid in this distinction, various clinical models have been experimented with. For instance, in a study of 422 immunocompetent patients older than one month with acute bacterial (ABM) or viral meningitis, specific CSF markers—including a CSF glucose level below 34 mg/dl, a CSF-to-blood glucose ratio under 0.23, a CSF protein concentration exceeding 220 mg/dl, over 2000 leukocytes/mm^3 of CSF, and more than 1180 neutrophils/mm^3 of CSF—were identified as predictors of bacterial meningitis with high certainty (99% or greater) [99]. The Bacterial Meningitis score, derived and validated in 4896 patients, helps identify children with CSF pleocytosis at low risk for bacterial meningitis. Low-risk features include a negative CSF Gram stain, CSF absolute neutrophil count under 1000 cells/mm^3, CSF protein under 80 mg/dl, and peripheral absolute neutrophil count below 10,000 cells/mm^3 [100]. Notably, a positive Gram stain is among the most critical predictors in this scoring system. For adults, a recent study involving 960 patients developed a risk score for those with meningitis and a negative Gram stain, identifying a "zero risk" subgroup for any urgent treatable cause (e.g., bacterial meningitis, herpes simplex encephalitis, fungal encephalitis), with 100% sensitivity [95]. Despite the availability of these clinical models, empirical treatment for bacterial meningitis remains common [90]. Biomarkers, such as CSF lactate, can further support differentiation between bacterial and viral meningitis in patients without prior antimicrobial therapy [101, 102]. One meta-analysis includes 25 studies with 1692

patients (adults and children) [101]. The other, including 31 studies with 1885 patients [102], concluded that the diagnostic accuracy of CSF lactate is better than that of the CSF white blood cell count, glucose concentration, and protein level in the differentiation of bacterial from aseptic meningitis, sensitivities of 93% and 97% and specificities of 96% and 94%, respectively, were seen. Additionally, serum and CSF C-reactive protein (CRP) levels and serum procalcitonin concentrations have been shown to elevate bacterial meningitis, aiding in the distinction from viral cases. One study reported that serum CRP could differentiate Gram stain-negative bacteria from viral meningitis at admission with 96% sensitivity, 93% specificity, and a negative predictive value of 99% [103]. Serum procalcitonin, when above 0.2 ng/ml, achieved sensitivities and specificities up to 100% in bacterial meningitis diagnosis, though occasional false negatives have been documented [104, 105].

2.4.5 Clinical presentation

Acute bacterial meningitis is a rapidly progressing condition that can lead to rapid deterioration in patients before or shortly after hospital admission. Typical symptoms include fever, headache, neck stiffness, and altered mental status. Notably, two of these symptoms appear in 90–95% of cases, though all four occur together in only 30–40%, making the clinical presentation atypical for many patients [106]. Other frequent symptoms include nausea, vomiting, photophobia, and hypersensitivity to sound. While Kernig's and Brudzinski's signs may in some cases be observed. Early symptoms often suggest respiratory infections, such as earache, rhinorrhea, or cough for pneumococcal meningitis, and sore throat for meningococcal disease. A petechial rash commonly appears in meningococcal meningitis, potentially indicating severe sepsis and septic shock with multiorgan failure. In elderly patients, these typical symptoms may be absent, complicating diagnosis based solely on clinical findings [107]. Convulsions, as new-onset seizures, occur in 10–15% of cases (especially in children), and focal neurological deficits, typically cranial nerve palsies, are seen in around 5% of cases. Psychomotor agitation, signaling elevated intracranial pressure, can indicate a risk of rapid deterioration into coma or cerebral herniation. Signs of herniation include coma with dilated pupils, irregular breathing, increasing blood pressure alongside bradycardia, opisthotonus, or lack of responsiveness [96].

The defining characteristic of acute bacterial meningitis is a rapid progression of cerebral symptoms, typically over hours, prompting most patients to seek hospital care within 12–24 hours [106]. This progression contrasts with cerebral mass lesions, where symptoms evolve gradually over several days, and subarachnoid hemorrhage, where a sudden severe headache emerges momentarily in seconds. Viral meningitis, a common differential diagnosis, shares symptoms such as fever, headache, and neck stiffness but typically lacks altered mental status and generally has a longer symptom duration [86].

In infants, fever and altered mental status should raise suspicion for acute bacterial

meningitis, though symptoms may be less specific, irritability, lethargy, or weakness being common [108]. A bulging fontanelle may be observed, while neck stiffness is usually absent. Some infants exhibit seizures or temperature and skin color changes, indicating a circulatory impairment associated with severe sepsis and septic shock [96].

2.4.6 Diagnostic tests

For diagnosing acute bacterial meningitis (ABM), initial testing should include blood cultures, routine chemical and hematologic analyses, and arterial blood gas with lactate measurement upon hospital admission. Lumbar puncture and cerebrospinal fluid (CSF) analysis are essential for diagnosis, as they are the only methods capable of confirming or excluding ABM [109, 110]. In cases where the CSF is cloudy and the spinal opening pressure is elevated (>300 mmH_2O), a preliminary diagnosis can be made at the bedside. Within 1–2 hours of CSF analysis, markers such as leukocyte count (>500–1000 × 10^9/L with a predominance of polymorphonuclear cells), a low CSF/serum glucose ratio (<0.4), elevated lactate (>4–5 mmol/L), and protein levels (>1 g/L) can further support the diagnosis. Bacteria may be detected in CSF through microscopy or antigen tests, typically within a few hours. Definitive diagnosis is achieved through culture and/or polymerase chain reaction (PCR) testing on CSF or blood samples, with newer PCR methods often providing results within 24 hours [111, 112]. Culture also facilitates antibiotic susceptibility testing, allowing targeted treatment adjustments.

Prompt lumbar puncture is critical for early diagnosis and treatment, though the decision to proceed immediately or after a computerized tomography (CT) scan remains controversial. Some guidelines recommend pre-lumbar puncture CT if elevated intracranial pressure or cerebral mass lesions are suspected due to concerns about potential herniation following lumbar puncture [109,110]. However, evidence linking lumbar puncture directly to herniation is limited, and the natural progression of ABM or a brain lesion may independently lead to herniation [113, 114]. Studies show that CT is unreliable in predicting herniation risk in ABM cases and rarely provides helpful information in suspected bacterial meningitis cases [115, 116]. Given these findings, immediate empiric antibiotic treatment is recommended if lumbar puncture is delayed, as delays in lumbar puncture and neuroimaging are linked to higher mortality and unfavorable outcomes [117–119]. CT is generally advised before lumbar puncture only when a mass lesion is more likely than ABM, indicated by focal neurological deficits (other than cranial nerve palsy) or symptoms lasting over four days. Some guidelines also recommend CT in cases with altered mental status, recent seizures, immunocompromised status, or papilledema [109, 110]. However, these signs can also occur in ABM, making timely clinical evaluation challenging in the emergency setting. In adults, especially the elderly, ABM is often among several differential diagnoses considered in emergency departments. Early intensive care is frequently necessary, with some patients

requiring neuro-intensive care to manage intracranial pressure [120]. Starting antibiotics before CT and lumbar puncture carries risks, such as CSF sterilization leading to negative culture results, complicating further treatment choices and decisions regarding treatment duration [121, 122]. Although blood cultures can identify the causative organism, only 50–70% of ABM cases yield positive results [106, 117]. Delays in lumbar puncture also disrupt the diagnosis of conditions similar to ABM, such as viral or tuberculous meningitis, herpes simplex encephalitis, and various non-infectious neurologic disorders. This concern is especially relevant in adults, where differential diagnoses are more complex and symptoms less distinct than in children.

2.5 A comprehensive investigation into the ranges of laboratory tests present in cerebrospinal fluid across various types of meningitis within different age categories

Meningitis is an inflammation of the membranes surrounding the brain and spinal cord, often caused by bacterial, viral, fungal, or parasitic infections [123]. Among these, bacterial meningitis is particularly concerning due to its high mortality rate and the severity of its complications. According to the World Health Organization (WHO), bacterial meningitis has the highest fatality rate among the different types, estimated at approximately 10%. However, this varies depending on the causative agent, age, and demographics of the patients [124]. It is reported that roughly 20% of individuals diagnosed with bacterial meningitis experience severe complications, resulting in mortality for 1 in 6 cases and serious complications for 1 in 5 survivors [125]. Prompt and accurate diagnosis of meningitis is crucial for initiating appropriate treatment and improving survival rates. Tests on cerebrospinal fluid (CSF), such as culture, Gram stain, molecular analyses, and blood tests, are essential for identifying the causative agent and guiding prompt treatment [126]. Cerebrospinal fluid biomarkers play a significant role in diagnosing different types of meningitis, helping to differentiate between bacterial, viral, fungal, or other forms, each with distinct profiles [127]. Biomarker analysis assists clinicians in choosing appropriate antibiotics, antivirals, or antifungals tailored to the identified pathogen. Specific biomarkers also offer insights into infection severity and potential complications while monitoring treatment response over time. Meningitis outbreaks, such as the recent one observed in Niger, highlight the importance of a comprehensive understanding and accurate diagnosis of the various forms of meningitis. From 1 November 2022 to 27 January 2023, Niger reported 559 meningitis cases, including 111 confirmed cases and 18 deaths, compared to 231 cases during the same period in 2021-2022. Most laboratory-confirmed cases were caused by *Neisseria meningitidis*. The outbreak in Niger is severe and spreading rapidly, posing a high risk of transmission across West Africa, particularly to neighboring countries due to shared borders and concurrent outbreaks. The WHO has clas-

sified the outbreak in Niger as high-risk, indicating potential spread throughout Africa [128]. This study aims to comprehensively investigate the ranges of laboratory tests present in CSF across various types of meningitis over different populations. Different populations can display diverse responses in infection biomarkers, influenced by factors like age, genetics, health status, and environmental conditions. Studying these differences helps improve diagnostic accuracy tailored to specific demographics or geographical regions. By comparing CSF laboratory test results across diverse populations and types of meningitis, this research seeks to enhance our understanding of diagnostic variations and improve treatment strategies globally. Ultimately, these insights will contribute to more effective public health interventions and tailored medical responses in managing meningitis outbreaks.

2.5.1 Materials and methods

This study used data on patients diagnosed with meningitis cases reported to SINAN, the Information System on Notifiable Diseases of the Brazilian Government's Health Department [1], from 2003 to 2022. To assess age-related differences, the full cohort was divided into three major age groups: Children [0-12 years] with 6,408 cases; Adults [13-64 years] with 4,319 cases; and the Elderly [ages $\geq$ 65 years] with 499 cases. Table 2.5 presents the distribution of reported cases of meningitis is based on the affected individuals' age groups and the causative agents. Statistical analyses are conducted to uncover nuanced insights

Table 2.5: Distribution of reported cases of meningitis based on age groups (children, adults and elderly).

Type of causative agent	**Children (0-12 years)**	**Adults (13-64 years)**	**Elderly ($\geq$ 65 years)**
Meningococcaemia	47	31	2
Meningococcal meningitis	68	64	2
Tuberculous meningitis	20	121	7
Haemophilus influenzae meningitis	39	12	1
Pneumococcal meningitis	110	182	25
Aseptic meningitis	5195	2659	257
Meningitis by other bacteria	890	826	172
Meningitis due to other aetiology	39	424	33

within the dataset. This analysis aims to discern patterns, trends, and statistically significant differences in meningitis cases across the specified age groups. The proposed methodology for this study is illustrated in figure 2.2. A series of statistical analyses are conducted.

[1]The SINAN database can be accessed at https://datasus.saude.gov.br/transferencia-de-arquivos# [129].

Descriptive statistics provide a detailed summary of the distribution and central tendencies of various cerebrospinal fluid biomarkers across different types of meningitis. This allows for a comprehensive understanding of the variations in biomarker levels within each meningitis type. The Welch ANOVA test is performed to explore potential differences in biomarker levels across different meningitis sub-types. Subsequently, the Games-Howell. Post hoc test is employed to discern which specific types of meningitis exhibit significantly different levels of the examined biomarkers. This post hoc test is well-suited for datasets with unequal variances and sample sizes, providing robust comparisons between groups. The statistical

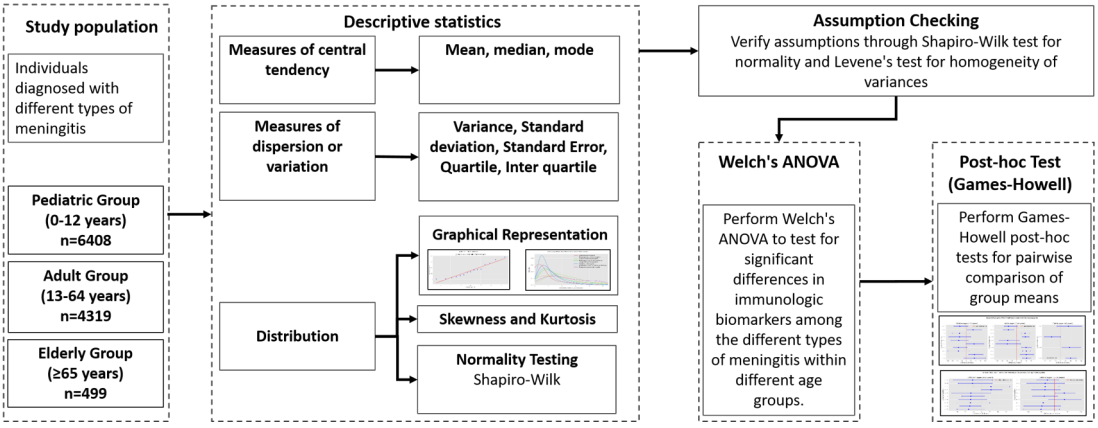


Figure 2.2: Flowchart illustrating the proposed methodology.

analyses were implemented using Python programming language, which uses scipy.stats module from the SciPy[2] library and the Pingouin[3] library. These libraries collectively provide a comprehensive and reliable toolkit for executing descriptive statistics, the Welch ANOVA test, and the Games-Howell post hoc test.

2.5.2 Results

The investigation delved into descriptive statistics for various types of meningitis, including Meningococcaemia, Meningococcal meningitis, Tuberculous meningitis, Haemophilus influenzae meningitis, Pneumococcal meningitis, Aseptic meningitis, Meningitis by other bacteria, and Meningitis due to other etiology. The analysis was stratified across distinct age groups: Children (ages 0-12 years), Adults (ages 13-64 years), and the Elderly (ages ≥ 65 years). Tables 2.7, 2.8, and 2.9 provide detailed overviews of the descriptive statistics for children, adults, and elderly populations, respectively. These tables offer insights into the central tendency and distribution of key variables, including Neutrophils, Lymphocytes, CSF WCC (Cerebrospinal Fluid White Blood Cell Count), CSF/serum Glucose ratio, and CSF/serum Protein ratio among different group ages. The table includes various statistical measures: median, mean ± standard deviation (SD), minimum (min), maximum (max), 25%

[2]https://scipy.org
[3]https://pingouin-stats.org/build/html/index.html

(First Quartile, Q1), 75% (Third Quartile, Q3), interquartile range (IQR), and 95% confidence intervals (CI). Figure 2.3 depicts box plots illustrating the distribution of cerebrospinal fluid biomarkers across different meningitis cases within three populations: children, adults, and elderly individuals. The plots reveal the variability in CSF WCC, neutrophil and lymphocyte levels, Glucose ratios, and Protein ratios among different causative agents, aiding in understanding the diagnostic profiles of meningitis Each plot helps illustrate statistical differences in these biomarkers depending on the causative agent type and age, offering insights into diagnostic profiles and potential disease severity. WCC varies significantly across different types of meningitis. Meningococcal meningitis and Haemophilus influenza meningitis show higher median values in both populations of children and adults, indicating more severe inflammatory responses compared to other types like Meningicoccaemia and aseptic meningitis, which shows lower WCC across all age groups. Neutrophil levels are notably higher in bacterial types of meningitis such as meningococcal, Haemophilus influenzae, pneumococcal meningitis, and meningitis by other bacteria, reflecting a typical acute bacterial response. Lymphocyte levels are more variable across different types. Higher lymphocyte counts in conditions like tuberculous meningitis and viral meningitis (as seen in cases labeled as ”aseptic meningitis”) align with their profiles as predominantly lymphocyte-driven responses. Children tend to show a wider range of lymphocyte counts, particularly in bacterial meningitis such as meningococcemia, meningococcal, and Pneumococcal meningitis, as well as Aseptic meningitis, which may indicate a more robust immune response in younger individuals. The Glucose ratio is significantly lower in bacterial meningitis, including meningococcal, tuberculous, Haemophilus influenzae meningitis, and pneumococcal meningitis, consistent with bacterial consumption of Glucose. Elevated Protein ratios are seen in bacteria, indicating a breach in blood-CSF barrier integrity typically associated with these infections [130]. Adults often show higher Protein ratios, possibly due to more severe disease manifestations or comorbidities that affect Protein levels in CSF.

Table 2.7: Descriptive statistics of CSF biomarkers in pediatric patients within different meningitis

Type of causative agent	Neutrophils								Lymphocytes								CSF WCC							
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI
Meningococcaemia	2.00	17.96 ± 27.17	0.00	88.00	0.00	25.00	25.00	9.98 - 25.94	25.00	39.30 ± 41.72	0.00	100.00	0.00	89.50	89.50	27.05 - 51.55	5.00	54.87 ± 163.64	0.00	840.00	2.00	14.50	12.50	6.83 - 102.92
Meningococcal meningitis	57.00	53.46 ± 33.69	0.00	100.00	24.50	85.25	60.75	45.30 - 61.61	13.50	27.76 ± 29.56	0.00	100.00	3.75	45.00	41.25	20.61 - 34.92	224.00	367.63 ± 386.77	0.00	1190.00	39.00	605.00	566.00	274.01 - 461.25
Tuberculous meningitis	15.50	23.25 ± 26.11	0.00	77.00	4.00	35.00	31.00	11.03 - 35.47	65.50	59.95 ± 32.51	0.00	97.00	47.75	84.25	36.50	44.74 - 75.16	88.00	130.95 ± 129.60	0.00	448.00	53.75	173.75	120.00	70.29 - 191.61
Haemophilus influenzae meningitis	71.00	60.05 ± 31.35	0.00	98.00	38.00	83.50	45.50	49.89 - 70.21	23.00	32.97 ± 30.59	0.00	97.00	10.00	48.50	38.50	23.06 - 42.89	320.00	406.08 ± 415.39	0.00	1184.00	59.00	662.50	603.50	271.42 - 540.73
Pneumococcal meningitis	69.00	58.53 ± 30.73	0.00	100.00	40.00	83.00	43.00	52.72 - 64.33	30.00	33.88 ± 30.11	0.00	100.00	7.00	52.00	45.00	28.19 - 39.57	180.50	310.22 ± 329.59	0.00	1250.00	47.00	476.75	429.75	247.93 - 372.50
Aseptic meningitis	19.00	27.02 ± 25.70	0.00	100.00	6.00	42.00	36.00	26.32 - 27.72	69.00	59.95 ± 31.51	0.00	100.00	34.00	87.00	53.00	59.09 - 60.81	76.00	150.01 ± 197.79	0.00	1259.00	30.00	180.00	150.00	144.63 - 155.39
Meningitis by other bacteria	68.00	60.16 ± 26.64	0.00	100.00	46.00	80.00	34.00	58.41 - 61.91	25.00	31.21 ± 25.21	0.00	100.00	12.00	45.00	33.00	29.55 - 32.87	195.00	322.18 ± 320.50	0.00	1234.00	65.25	491.50	426.25	301.09 - 343.26
Meningitis due to other aetiology	6.00	17.77 ± 25.60	0.00	97.00	3.00	19.00	16.00	9.47 - 26.07	82.00	66.77 ± 35.14	0.00	99.00	44.50	94.50	50.00	55.38 - 78.16	131.00	276.38 ± 336.27	3.00	1024.00	19.00	355.00	336.00	167.38 - 385.39
Type of causative agent	Glucose ratio CSF/serum								Protein ratio CSF/serum															
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI								
Meningococcaemia	0.65	0.62 ± 0.20	0.11	0.99	0.49	0.77	0.28	0.56 - 0.68	0.21	0.37 ± 0.46	0.10	2.81	0.18	0.31	0.14	0.24 - 0.51								
Meningococcal meningitis	0.27	0.36 ± 0.29	0.01	1.06	0.09	0.64	0.55	0.29 - 0.43	1.26	1.27 ± 0.91	0.02	3.61	0.41	1.90	1.49	1.05 - 1.50								
Tuberculous meningitis	0.29	0.32 ± 0.20	0.02	0.80	0.21	0.35	0.14	0.22 - 0.41	1.43	1.60 ± 0.96	0.31	3.35	0.85	2.46	1.61	1.15 - 2.05								
Haemophilus influenzae meningitis	0.18	0.23 ± 0.22	0.01	0.80	0.05	0.36	0.31	0.16 - 0.30	1.76	1.61 ± 0.89	0.01	3.19	1.01	2.29	1.29	1.32 - 1.90								
Pneumococcal meningitis	0.16	0.22 ± 0.22	0.00	0.81	0.04	0.37	0.33	0.18 - 0.27	1.61	1.65 ± 0.96	0.00	3.52	1.00	2.40	1.40	1.47 - 1.83								
Aseptic meningitis	0.57	0.57 ± 0.14	0.00	1.07	0.50	0.65	0.15	0.57 - 0.58	0.35	0.47 ± 0.42	0.00	3.63	0.25	0.54	0.29	0.46 - 0.49								
Meningitis by other bacteria	0.53	0.50 ± 0.21	0.00	1.07	0.39	0.63	0.24	0.48 - 0.51	0.61	0.91 ± 0.79	0.00	3.63	0.36	1.25	0.89	0.86 - 0.96								
Meningitis due to other aetiology	0.48	0.48 ± 0.23	0.05	0.94	0.37	0.58	0.20	0.41 - 0.55	0.63	0.87 ± 0.76	0.12	3.15	0.23	1.05	0.82	0.63 - 1.12								

Table 2.8: Descriptive statistics of CSF biomarkers in adult patients within different meningitis

Type of causative agent	Neutrophils								Lymphocytes								CSF WCC							
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI
Meningococcaemia	1.00	22.77 ± 33.50	0.00	95.00	0.00	57.00	57.00	10.48 - 35.06	21.00	37.16 ± 39.03	0.00	100.00	0.00	68.00	68.00	22.84 - 51.48	5.00	35.55 ± 89.01	0.00	400.00	1.00	14.50	13.50	2.90 - 68.20
Meningococcal meningitis	78.50	68.59 ± 28.44	0.00	100.00	60.00	90.00	30.00	61.49 - 75.70	16.00	24.75 ± 25.05	0.00	100.00	5.75	33.50	27.75	18.49 - 31.01	339.00	376.61 ± 362.81	0.00	1225.00	25.00	640.00	615.00	285.98 - 467.24
Tuberculous meningitis	20.00	29.94 ± 27.75	0.00	97.00	7.00	50.00	43.00	24.95 - 34.94	69.00	59.38 ± 31.87	0.00	100.00	38.00	87.00	49.00	53.64 - 65.12	150.00	204.07 ± 205.98	0.00	1200.00	60.00	277.00	217.00	166.99 - 241.14
Haemophilus influenzae meningitis	80.00	63.75 ± 32.33	12.00	99.00	35.50	88.50	53.00	43.21 - 84.29	17.00	30.08 ± 28.10	1.00	71.00	9.25	63.50	54.25	12.23 - 47.94	344.50	392.83 ± 340.20	15.00	880.00	26.50	681.25	654.75	176.68 - 608.98
Pneumococcal meningitis	80.00	70.45 ± 27.72	0.00	100.00	60.00	91.00	31.00	66.40 - 74.50	15.00	21.79 ± 23.37	0.00	100.00	5.00	28.00	23.00	18.37 - 25.21	285.00	422.79 ± 373.08	0.00	1260.00	110.00	677.50	567.50	368.22 - 477.36
Aseptic meningitis	10.00	18.31 ± 22.23	0.00	99.00	3.00	25.00	22.00	17.46 - 19.15	82.00	71.74 ± 28.88	0.00	100.00	61.00	93.00	32.00	70.64 - 72.84	106.00	195.84 ± 228.35	0.00	1215.00	35.00	277.00	242.00	187.15 - 204.52
Meningitis by other bacteria	66.00	60.42 ± 28.16	0.00	100.00	47.00	83.00	36.00	58.50 - 62.35	28.00	33.35 ± 26.69	0.00	100.00	12.00	45.00	33.00	31.53 - 35.17	234.50	362.68 ± 347.70	0.00	1255.00	64.25	563.75	499.50	338.94 - 386.43
Meningitis due to other aetiology	8.00	20.07 ± 25.67	0.00	100.00	2.00	30.00	28.00	17.62 - 22.52	76.50	59.77 ± 37.23	0.00	100.00	21.75	92.00	70.25	56.22 - 63.32	60.00	134.32 ± 188.20	0.00	1215.00	19.00	174.25	155.25	116.36 - 152.29
Type of causative agent	Glucose ratio CSF/serum								Protein ratio CSF/serum															
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI								
Meningococcaemia	0.61	0.53 ± 0.22	0.00	0.85	0.42	0.69	0.27	0.45 - 0.62	0.39	0.80 ± 0.91	0.00	3.60	0.24	0.93	0.69	0.47 - 1.13								
Meningococcal meningitis	0.28	0.30 ± 0.23	0.01	1.04	0.11	0.46	0.35	0.24 - 0.36	1.23	1.34 ± 0.96	0.00	3.62	0.49	2.02	1.53	1.10 - 1.58								
Tuberculous meningitis	0.28	0.34 ± 0.23	0.01	1.04	0.17	0.43	0.26	0.30 - 0.38	1.82	1.82 ± 0.88	0.15	3.63	1.17	2.60	1.43	1.66 - 1.98								
Haemophilus influenzae meningitis	0.28	0.43 ± 0.35	0.00	1.05	0.18	0.77	0.60	0.21 - 0.65	1.07	1.38 ± 0.97	0.33	3.41	0.73	1.92	1.20	0.77 - 2.00								
Pneumococcal meningitis	0.12	0.22 ± 0.23	0.00	1.03	0.04	0.37	0.33	0.19 - 0.26	1.55	1.57 ± 0.96	0.00	3.52	0.81	2.27	1.45	1.43 - 1.71								
Aseptic meningitis	0.56	0.56 ± 0.17	0.00	1.07	0.47	0.66	0.19	0.56 - 0.57	0.67	0.84 ± 0.61	0.00	3.64	0.42	1.06	0.64	0.81 - 0.86								
Meningitis by other bacteria	0.46	0.44 ± 0.24	0.00	1.06	0.25	0.61	0.36	0.43 - 0.46	1.07	1.27 ± 0.84	0.00	3.64	0.61	1.74	1.13	1.21 - 1.33								
Meningitis due to other aetiology	0.35	0.36 ± 0.19	0.00	1.04	0.23	0.48	0.25	0.34 - 0.38	0.89	1.05 ± 0.73	0.00	3.60	0.49	1.45	0.96	0.98 - 1.12								

Table 2.9: Descriptive statistics of CSF biomarkers in elderly patients within different meningitis

Type of causative agent	Neutrophils								Lymphocytes								CSF WCC							
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI
Pneumococcal meningitis	75.00	71.24 ± 29.80	0.00	100.00	60.00	94.00	34.00	58.94 - 83.54	14.00	18.32 ± 20.56	0.00	84.00	2.00	30.00	28.00	9.83 - 26.81	160.00	292.44 ± 333.74	6.00	1046.00	48.00	341.00	293.00	154.68 - 430.20
Aseptic meningitis	10.00	22.92 ± 26.38	0.00	97.00	3.00	35.00	32.00	19.68 - 26.16	77.00	65.83 ± 31.93	0.00	100.00	55.00	90.00	35.00	61.91 - 69.75	74.00	147.11 ± 187.14	1.00	1200.00	30.00	180.00	150.00	124.12 - 170.10
Meningitis by other bacteria	75.00	65.82 ± 26.42	0.00	97.00	52.75	87.25	34.50	61.84 - 69.80	20.00	28.53 ± 23.68	0.00	96.00	11.00	39.25	28.25	24.96 - 32.09	210.00	333.44 ± 329.62	0.00	1260.00	77.75	476.75	399.00	283.83 - 383.05
Meningitis due to other aetiology	16.00	27.30 ± 27.66	0.00	95.00	3.00	46.00	43.00	17.49 - 37.11	75.00	60.91 ± 31.77	0.00	100.00	32.00	87.00	55.00	49.64 - 72.18	97.00	224.88 ± 280.94	0.00	970.00	47.00	272.00	225.00	125.26 - 324.50
Type of causative agent	Glucose ratio CSF/serum								Protein ratio CSF/serum															
	median	mean ± sd	min	max	25%	75%	IQR	95% CI	median	mean ± sd	min	max	25%	75%	IQR	95% CI								
Pneumococcal meningitis	0.08	0.17 ± 0.25	0.00	0.94	0.01	0.14	0.13	0.06 - 0.27	1.54	1.56 ± 0.88	0.18	3.58	0.86	1.90	1.04	1.19 - 1.92								
Aseptic meningitis	0.60	0.60 ± 0.24	0.00	1.07	0.43	0.74	0.31	0.57 - 0.63	0.80	1.04 ± 0.80	0.00	3.63	0.47	1.35	0.88	0.95 - 1.14								
Meningitis by other bacteria	0.53	0.50 ± 0.29	0.01	1.06	0.26	0.73	0.47	0.46 - 0.54	1.36	1.47 ± 0.81	0.00	3.58	0.85	1.98	1.13	1.34 - 1.59								
Meningitis due to other aetiology	0.30	0.30 ± 0.19	0.00	0.68	0.18	0.44	0.26	0.23 - 0.36	0.83	1.01 ± 0.72	0.00	3.48	0.63	1.05	0.42	0.76 - 1.27								

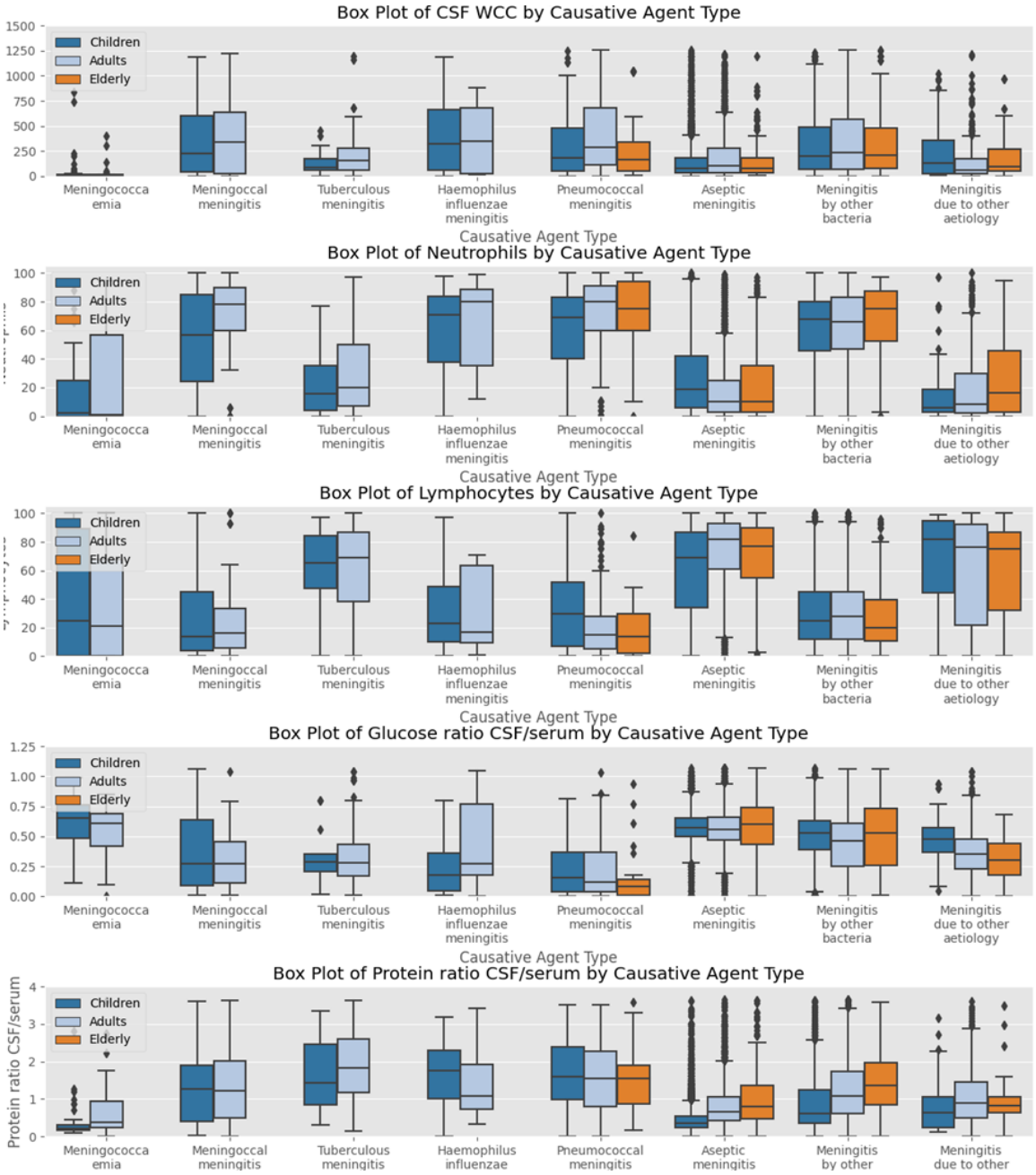


Figure 2.3: Box plots illustrating the distribution of cerebrospinal fluid biomarkers across different meningitis cases in three distinct age groups.

Normality test

Shapiro-Wilk test was employed to assess the normality of the laboratory biomarkers variables in our study. This statistical test evaluates whether the observed data significantly deviates from a normal distribution. The null hypothesis (H_0) states that the data follows a normal distribution, while the alternative hypothesis (H_1) asserts that the data does not follow a normal distribution. A low p-value (< 0.05) indicates a significant departure from normality, leading to the rejection of the null hypothesis. Additionally, skewness was computed for each

variable to measure the asymmetry of the distributions. Kurtosis was analyzed to determine the shape of the distribution, with positive kurtosis indicating heavier tails than a normal distribution and negative kurtosis suggesting lighter tails. The results of normality tests for various features across different types of meningitis, as detailed in Table 5, reveal significant deviations from normal distribution for most variables in pediatric patients. Specifically, neutrophil and lymphocyte levels and CSF WCC do not follow a normal distribution, as indicated by the rejection of the null hypothesis (H_0). In contrast, the Glucose ratio CSF/serum is normally distributed for Meningococcaemia, Haemophilus influenzae meningitis, and Meningitis due to other etiology, with the null hypothesis not rejected. The Protein ratio CSF/serum also follows a normal distribution for Tuberculous meningitis and Haemophilus influenzae meningitis. In the adult population, Neutrophils, Lymphocytes, Glucose ratio CSF/serum, CSF WCC, and Protein ratio across all causative agents reject the null hypothesis (H_0), indicating non-normal distribution. However, for Haemophilus influenzae meningitis, the Glucose ratio, CSF WCC, and Protein ratio data fail to reject the null hypothesis (H_0), suggesting a normal distribution. Yet, interpretation may be limited due to the low Shapiro-Wilk test power for small sample sizes [131]. Samples with a size of fewer than 7 were excluded from the study involving the elderly population. Table 2.10 indicates deviations from normality across various features for different causative agents, except for the Protein ratio in Pneumococcal meningitis and the Glucose ratio in Meningitis due to other aetiologies within the elderly population.

Table 2.10: Statistical features of various causative agents in meningitis by population, including sample size, kurtosis, skewness, and shapiro-wilk test results for normality. Each row represents a specific causative agent and population group, with columns detailing the associated statistical attributes.

Type of causative agent	Population	Sample Size	Feature	Kurtosis	Skewness	Shapiro-Wilk Test Statistic	P-Value
Meningococcaemia	Children	47	Glucose ratio CSF/serum	-0.33	-0.39	0.98	0.45
Tuberculous meningitis	Children	20	Protein ratio CSF/serum	-1.14	0.43	0.93	0.14
Haemophilus influenzae meningitis	Children	39	Protein ratio CSF/serum	-0.78	-0.15	0.96	0.26
	Adults	12	Glucose ratio CSF/serum	-1.23	0.46	0.90	0.16
			CSF WCC	-1.66	0.22	0.88	0.08
			Protein ratio CSF/serum	0.05	0.80	0.91	0.22
Pneumococcal meningitis	Elderly	25	Protein ratio CSF/serum	0.19	0.67	0.95	0.22
Meningitis due to other aetiology	Children	39	Glucose ratio CSF/serum	-0.12	0.21	0.96	0.19
	Elderly	33	Glucose ratio CSF/serum	-0.76	0.34	0.96	0.20

Analysis of variance

In consideration of the potential non-homogeneity of variances assumption (see Figure 2.3), Welch ANOVA (Analysis of Variance) was employed as an alternative due to its capability to generate reliable results in situations characterized by unequal variances across the studied groups [132]. This test explored potential differences in biomarker levels among various types of meningitis across different age groups. Table 2.11 presents a detailed summary of

the Welch ANOVA analysis results, including the statistical significance of differences in dependent variables across different types of meningitis considered in this study. Each row corresponds to a specific variable, displaying the F-statistic, p-value, and effect size (η_p^2). The significance is determined by whether the p-value is less than 0.05, suggesting a statistically significant difference between groups and supporting rejecting the null hypothesis (H_0: No difference among groups). If the means show significant differences (p-value > 0.05), this supports the alternative hypothesis (H_1: There are differences among groups). Significant differences were observed in various biomarkers (CSF WCC, Neutrophils, Lymphocytes, Glucose ratio CSF/serum, and Protein ratio CSF/serum) across different causative agent groups within children and adults. For an elderly population, significant differences were observed only in Neutrophil levels. This suggests that the impact of causative agents on these biomarkers may be different or negligible in the elderly compared to children and adults. The partial eta-squared (η_p^2) values indicate the proportion of variance explained by causative agent groups in biomarker levels [133]. Higher values suggest stronger effects, aiding in understanding the impact of causative agents across populations. Effect sizes for certain biomarkers vary across age groups. For example, Neutrophils exhibit higher effect sizes in adults and the elderly than in children. This indicates that the impact of causative agents on Neutrophil levels is more pronounced in older age groups.

Table 2.11: Welch's ANOVA results for different biomarkers across populations

Dependent Variable	**Population**	**F-statistic**[4]	**p-value**	η_p^2	**Significance**
CSF WCC	Children	44.73	0.00	0.008	The means of CSF WCC vary significantly across different types of meningitis in both populations.
	Adults	58.72	0.00	0.10	
	Elderly	0.00	1.00	0.11	Means of CSF WCC are likely equal between meningitis types in elderly population.
Neutrophils	Children	188.56	0.00	0.19	The means of Neutrophils levels vary significantly across different types of meningitis in all three populations.
	Adults	305.53	0.00	0.38	Elderly population shows a higher amount of variance (40%).
	Elderly	170.58	0.00	0.40	
Lymphocytes	Children	260.61	0.00	0.26	The means of Lymphocytes levels vary significantly across different types of meningitis in both populations.
	Adults	144.37	0.00	0.11	Pediatric patients explain a higher amount of variance (26%) compared to adults (11%).
	Elderly	0.00	1.00	0.31	Means of Lymphocyte levels are likely to equal between meningitis types in the elderly population.
Glucose ratio CSF/serum	Children	134.48	0.00	0.20	The means of Glucose ratio vary significantly across different types of meningitis in both populations.
	Adults	73.43	0.00	0.14	Pediatric patients explain a higher amount of variance (20%) compared to adults (14%).
	Elderly	0.00	1.00	0.18	Means of Glucose ratio are likely equal between meningitis types in the elderly population.
Protein ratio CSF/serum	Children	59.51	0.00	0.12	The means of Protein ratio vary significantly across different types of meningitis in both populations.
	Adults	78.29	0.00	0.18	Adult patients explain a slightly higher amount of variance (18%) than in pediatric population (12%).
	Elderly	0.00	1.00	0.09	Means of Protein ratio are likely to equal between meningitis types in the elderly population.

Post hoc test

Games-Howell post hoc test is employed to identify specific pairs of meningitis types that exhibited statistically significant differences in biomarker levels. This post hoc analysis is well-suited for data with unequal variances and sample sizes [134], facilitated a nuanced understanding of the variations in biomarker concentrations among different meningitis categories. Notably, the Games-Howell test revealed distinct patterns of significance, specifying which pairs of meningitis types demonstrated significant differences in biomarker levels. These findings contribute valuable insights into key variables' central tendencies and distributions and provide a better understanding of the heterogeneity within the studied meningitis subtypes across different age groups. This study involves comparing mean differences in different biomarker levels among all types of meningitis of this study. Multiple comparisons are made among means of groups with unequal variances and sample size; another group is compared against it. The zero-difference line indicates no difference between groups. Points to the right suggest the first group has higher means; points to the left indicate the second group has higher means. In the examination of white cell count parameters in the pediatric population (Figure 2.4a), Meningococcal meningitis (group 2) significantly differs from Aseptic meningitis (group 6), displaying a mean difference of 217.62 (95% CI: 147.68, 158.14). Similarly, Haemophilus influenzae meningitis (group 4) exhibits a substantial difference in white cell count from Aseptic meningitis (group 6), with a mean difference of 256.07 (95% CI: 146.45, 157.61). Significant mean differences are observed between Meningococcaemia (group 1) and Haemophilus influenza meningitis (group 4) -351.20 (95% CI: 137.86; 288.34). In our analysis, the confidence intervals very close to the zero reference line on the forest plot are not considered, as these values may suggest a potential lack of statistical significance or effect. In the analysis of the adult population, Pneumococcal meningitis (group 5) demonstrates a significant mean difference of 226.95 (95% CI: 201.7, 219.80) from Aseptic meningitis (group 6). Furthermore, Meningococcal meningitis (group 2) versus Meningitis due to other etiology (group 8) reveals a mean difference of 242.29 (95% CI: 145.02, 187.05), emphasizing notable distinctions. Pneumococcal meningitis (group 5) versus Meningitis due to other etiology (group 8) shows a substantial mean difference of 288.47 (95% CI: 198.55, 245.06).

Based on the results of the Games-Howell test (Figure 2.4b), significant mean differences and their associated confidence intervals suggest that neutrophil levels in Meningococcal meningitis (group 2) and Haemophilus influenzae meningitis (group 4) are statistically different from those in Meningitis due to other etiologies (group 8) in the children's population. In the adult population, distinct neutrophil responses were observed across various types of meningitis. Aseptic meningitis (group 6) showed a significant mean decrease of -42.12 in neutrophil levels compared to meningitis by other bacteria (group 7). Neutrophil levels in Haemophilus influenzae meningitis (group 4) differ significantly from both Meningitis due

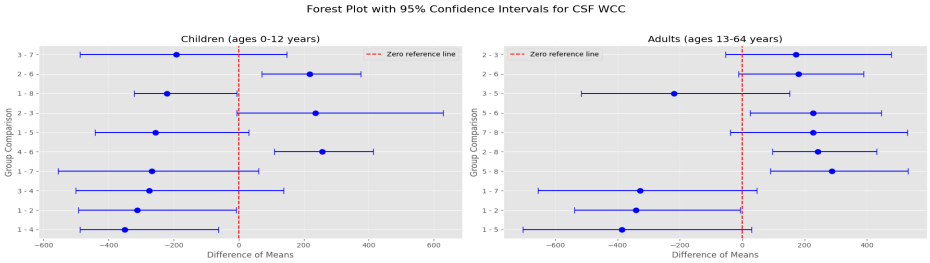
to other etiology (group 8) 43.68 (95% CI: 18.78, 23.71) and Aseptic meningitis (group 6) 45.44 (95% CI: 17.69, 19.31). Meningococcal meningitis (group 2) also shows significant differences in neutrophil levels compared to Meningitis due to other etiology (group 8) 48.52 (95% CI: 23.71, 29.27) and Aseptic meningitis (group 6) 50.29 (95% CI: 18.61, 20.39). Pneumococcal meningitis (group 5) demonstrates significant differences in neutrophil levels compared to Meningitis due to other etiology (group 8) 50.38 (95% CI: 32.39, 37.94) and Aseptic meningitis (group 6) 52.14 (95% CI: 20.72, 22.58). In the elderly population, the results indicate neutrophil responses when affected by Pneumococcal meningitis (group 5) compared to other etiologies (group 8), and Aseptic meningitis (group 6). The test results indicate significant mean differences in Lymphocytes levels only in adults (See figure 3(C)) between Tuberculous meningitis (group 3) and Pneumococcal meningitis (group 5) with a mean difference of 37.59 (95% CI: 33.18, 40.46).

The Glucose ratio in cerebrospinal fluid/serum was examined in pediatric and adult populations (Figure 2.4d). In children, the comparison between Meningococcaemia (group 1) and Pneumococcal meningitis (group 5) revealed a mean difference of 0.40 (95% CI: 0.30, 0.39). Similarly, the corresponding mean difference in adults was 0.31 (95% CI: 0.23, 0.30). These findings suggest notable differences in the Glucose ratio between the two specified meningitis types across age groups.

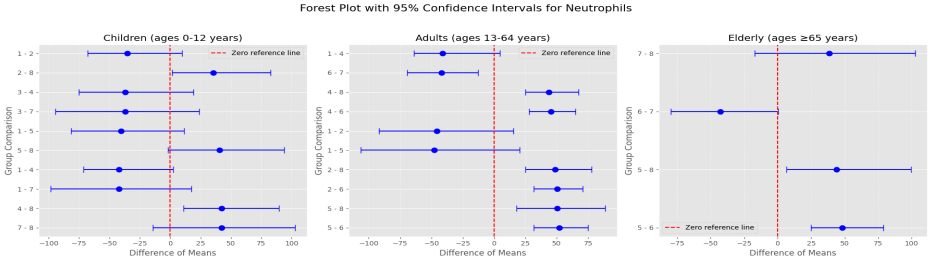
The analysis of Protein ratios in cerebrospinal fluid/serum among pediatric subjects revealed distinct patterns across diverse cases of meningitis (Figure 2.4e). Meningococcal meningitis (group 2), Tuberculous meningitis (group 3), Haemophilus influenza meningitis (group 4) and Pneumococcal meningitis (group 5), demonstrated a mean difference of 0.80 (95% CI: 0.47, 0.50), 1.13 (95% CI: 0.47, 0.49), 1.14 (95% CI: 0.47, 0.50) and 1.18 (95% CI: 0.49, 0.51) relative to Aseptic meningitis (group 6) respectively. Comparisons involving Meningoccaemia (group 1) unveiled inverse Protein ratios of -1.23 (95% CI: 0.55, 0.95) and -1.24 (95% CI: 0.75, 1.13) with Tuberculous meningitis (group 3) and Haemophilus influenzae meningitis (group 4), respectively. Negative values suggest a lower Protein ratio in Meningococcemia compared to the other types of meningitis being compared. In adults, the specific comparison of Tuberculous meningitis with Aseptic meningitis demonstrated a mean difference of 0.99 (95% CI: 0.85, 0.90), illustrating distinct Protein ratio patterns in this age group.

2.5.3 Discussion

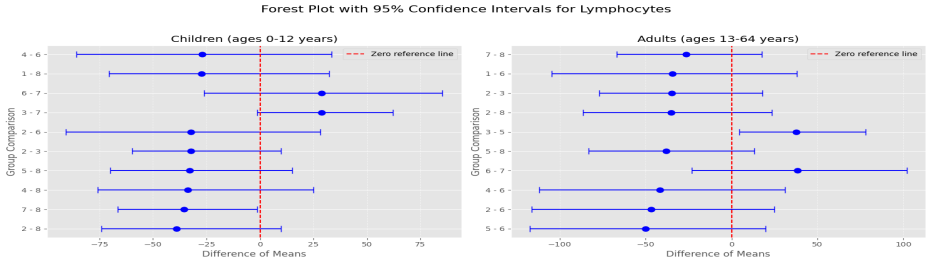
This study aims to compare the neutrophils and lymphocyte counts, CSF WCC, Glucose and Protein ratio (CSF/serum) among patients with different types of meningitis, stratified by age groups. The findings provide insights into the immune response profiles and biochemical characteristics associated with various types of meningitis. There was a marked increase in neutrophil levels across all age groups in cases of bacterial meningitis (including Meningococcal, Tuberculous, Haemophilus influenzae, Pneumococcal, and other bacterial



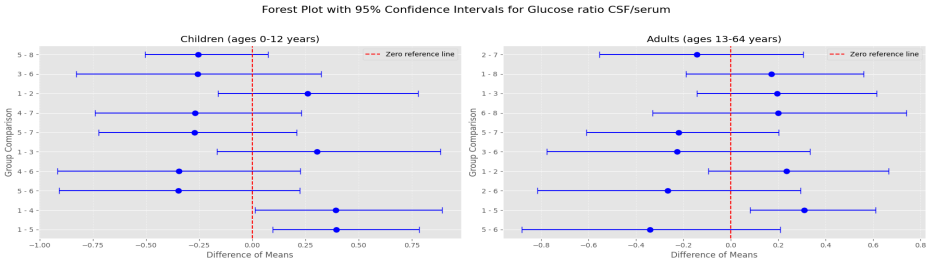
(a) Forest plot comparing CSF WCC across different age groups



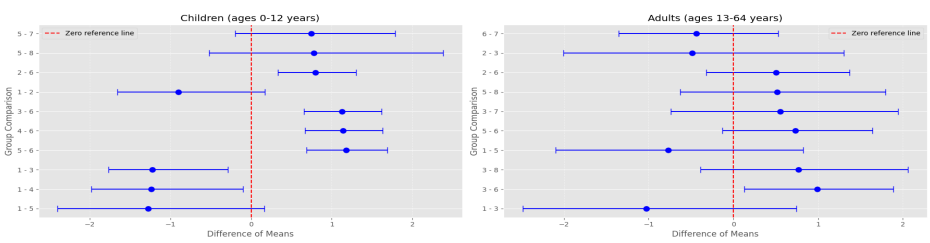
(b) Forest plot comparing Neutrophils level across different age groups



(c) Forest plot comparing lymphocyte level across different age groups



(d) Forest plot comparing Glucose ratio across different age groups



(e) Forest plot comparing Protein ratio across different age groups

Figure 2.4: Forest plot comparing different attributes across different age groups with 95% confidence interval

etiologies). This aligns with the established understanding that bacterial infections elicit a robust neutrophilic response. Neutrophil levels were generally higher in adults compared to children in conditions like Meningococcal meningitis. For instance, adults with Meningococcal meningitis had a median neutrophil level of 78.5% [60%, 90%], compared to 57% [24.5%, 85.25%] in children. This could reflect differences in immune system maturity and response mechanisms. Viral meningitis cases showed higher lymphocyte levels, consistent with viral infections typically inducing a lymphocytic response. The median lymphocyte level was notably higher among children 69% [34%, 87%], adults 82% [61%, 93%], and elderly 77% [55%, 90%] with aseptic meningitis compared to those with bacterial meningitis. While lymphocyte levels were generally higher in adults than in children for viral infections, the differences were less pronounced compared to bacterial infections. CSF WCC was significantly elevated in bacterial meningitis cases, particularly in adults. For instance, the median CSF WCC in children with Meningococcal meningitis was 224 cells/mm^3, compared to 339 cells/mm^3 in adults. This elevation is indicative of the intense inflammatory response to bacterial pathogens. CSF WCC was elevated in viral meningitis but generally lower than in bacterial cases. This reflects the typically less intense inflammatory response seen in viral infections. The CSF WCC in patients with meningococcemia tends to be relatively low in children and adults. For children with meningococcemia, the median CSF WCC was five cells/mm^3, with a mean ± SD of 54.87 ± 163.64 cells/mm^3. The range was from 0 to 840 cells/mm^3. Despite some extreme values, the central tendency (median) remained low. For adults aged 13-64, the median CSF WCC was also five cells/mm^3, with a mean ± SD of 35.55 ± 89.01 cells/mm^3. The range extended from 0 to 400 cells/mm^3. Meningococcemia is primarily a bloodstream infection. The bacteria can cause sepsis without necessarily leading to meningitis, where CSF WCC would be expected to rise significantly, resulting in a lower CSF WCC. This highlights the importance of considering other clinical factors and diagnostic tools when evaluating and managing patients with suspected meningococcemia. The Glucose ratio was notably lower in bacterial meningitis, including Meningococcal, Tuberculosis, *Haemophilus influenzae*, and Pneumococcal meningitis, with values often below 0.3, reflecting the consumption of Glucose by bacteria in the CSF. In viral meningitis, the Glucose ratio was higher, often above 0.5. The Protein ratio was higher in bacterial meningitis, including Meningococcal, Tuberculosis, *Haemophilus influenzae*, and Pneumococcal meningitis, reflecting increased permeability of the blood-brain barrier and Protein leakage into the CSF. Building on the scoring tool established by Pascal Chavanet et al. for pediatric and adult meningitis, where a CSF WCC >1700 cells/mm^3, CSF neutrophil percentage >80%, CSF Protein >2.3 g/L, and Glucose CSF/blood ratio <0.33 are identified as key indicators of bacterial etiology in adults, and a CSF WCC >1800 cells/mm^3, CSF neutrophil percentage >80%, CSF Protein >1.2 g/L, and Glucose CSF/blood ratio <0.3 in children, our findings align closely with these thresholds. We observed significantly elevated CSF WCC

and Protein ratios and low Glucose ratios in bacterial meningitis cases, thus reinforcing these diagnostic criteria [135]. Sérgio Monteiro de Almeida et al. compared the effectiveness of lactate and Glucose (GL) in CSF, as well as the CSF/blood GL ratio, in distinguishing between acute bacterial meningitis (BM) and viral meningitis (VM) with typical and atypical CSF characteristics. They observed that the median WCC was significantly higher in bacterial meningitis (560 × 10^6/L) than in viral meningitis (36 × 10^6/L). Additionally, the median neutrophil percentage was markedly higher in bacterial meningitis (83%) than in viral meningitis (10%). The Glucose ratio was also different, with a median of 0.30 in bacterial meningitis and 0.56 in viral meningitis [127]. Despite the valuable insights gained from our data, its specific regional focus may limit the generalizability of our findings to broader populations. Additionally, the interpretation of CSF parameters can be influenced by factors such as the timing of lumbar puncture, prior antibiotic therapy, and other infections. These factors highlight the need for careful consideration in diverse clinical settings and underscore the importance of future research to further validate and expand upon our findings. Furthermore, it is necessary to compare other biomarkers, such as CSF lactate, as it has demonstrated superior operational characteristics compared to CSF Glucose in differentiating between bacterial and viral meningitis [127].

2.6 AI in Healthcare

From an epistemological perspective, AI systems, particularly those based on machine learning, operate by detecting patterns and correlations within large datasets, unlike human reasoning, which often emphasizes causal relationships. While AI's strength lies in its ability to uncover complex patterns that surpass human capabilities, this also introduces the challenge of 'explicability' [136]. Explicability, or the capacity to explain how an AI system arrives at a particular conclusion, is often limited in AI, particularly with complex models like deep learning. This phenomenon is known as "epistemic opacity," meaning that even AI developers may not fully understand the step-by-step computational processes that lead to a specific result [137]. This lack of transparency leads to the so-called "black box" problem, where the AI's inner workings are unclear. This is especially concerning because trust in AI often depends on transparency and explainability [138]. A key question arises: how can people trust AI systems if they don't understand how these systems reach their decisions? Authors argue that for humans to confidently rely on AI, the processes behind AI decisions must be transparent and comprehensible [139]. This is particularly crucial in medicine, where informed consent is foundational. Physicians must provide patients with the best available information to allow them to agree to or decline treatments. This includes explaining why a specific medical action is recommended. However, the issue of explicability leads to further questions: What exactly needs to be understood about AI processes, and what kind of

explanations are required to ensure trust and informed consent? [13, 140].

2.6.1 Justifying decisions

Making clinical decisions, such as selecting a treatment option or reaching a medical diagnosis, requires a solid evidence base, as intuition or reliance solely on professional experience no longer suffices in the era of Evidence-Based Medicine (EBM). In a Medical AI (MAI) setting, physicians must be able to evaluate and understand the mechanisms driving an algorithm's outcome [141]. This is crucial for justifying decisions like choosing a particular medication or therapy. Durán and Jongsma refer to this as an 'epistemic warrant', which obligates physicians to base their choices on reliable knowledge [142]. Physicians must be able to assess whether a recommended treatment will truly benefit the patient. However, the complexity and opacity of algorithms may prevent them from fully understanding or verifying the underlying evidence, raising concerns about their ability to trust the outcomes of algorithmic decision-making.

The lack of transparency in algorithmic decisions conflicts with the ethical duty of clinicians to provide clear justifications for their medical decisions [143]. Like other experts, clinicians are expected to justify their actions with causal explanations, especially within Evidence-Based Medicine (EBM) [144]. When such explanations are absent, clinicians may rely on algorithmic outcomes they cannot evaluate, which could limit their ability to identify or prevent errors in these decisions. Machine learning algorithms detect patterns and make predictions based on statistical correlations, not causation, which can lead to unique errors. Without interpretability, clinicians may miss potential confounding errors or biases, as they cannot fully verify whether a causal inference is valid or if a confounding factor is at play [145]. For instance, overfitting, where an algorithm performs well during testing but fails in real-world applications, can lead to severe misdiagnoses [146]. Due to such risks, some argue that opaque, non-interpretable algorithms should be avoided in high-stakes fields like medicine, with only fully explainable algorithms permissible [147]. Some argue that patients have a right to understand algorithm-supported decisions, making explainability crucial for trust and legal standards [139, 146].

Some argue that in real-world medicine, not all clinical decisions can be explained strictly through science, as intuition and experience also play roles [143, 146]. With the demonstrated capabilities of MAI technologies, particularly deep learning in areas like tumor recognition, prioritizing explainability over accuracy may be impractical and could risk depriving patients of beneficial advancements.

To understand the problems caused by the lack of explainability in MAI, we first need to break down what 'opacity' means. Burrell [148] identifies different types of opacity. One type is intentional secrecy, where companies hide the workings of their algorithms to protect their interests and maintain a competitive edge. While this might make sense in

commercial areas like online shopping, it becomes an issue in healthcare. Sharing data, models, and algorithms in healthcare is key to making MAI technologies work effectively. Intentional secrecy could block progress toward a learning healthcare system (LHR) and might even be used to avoid regulations or manipulate users. Opacity can also come from the complexity of AI systems, where multiple teams build different components. A programmer working on one part may not know how others function. Similarly, making the algorithm available to healthcare professionals doesn't mean they can interpret it. Additionally, machine learning algorithms handle large amounts of data, which can make them too complex to fully understand. Deep learning algorithms, in particular, are used for tasks where linear logic doesn't work, making it hard to explain all the steps needed to produce a result. This raises the question of how much we can expect AI to be explainable [13].

2.6.2 Explainability

Explainability is essential in developing machine learning systems, beyond just ethical concerns, for several key reasons [149]. First, it helps justify the results, especially when AI produces new or unexpected findings. Second, understanding system behavior aids in preventing and identifying errors, an important part of debugging. Third, explainability is crucial for improving the system, as refinement requires ongoing insight into its mechanisms. Fourth, explainability supports scientific discovery, delivering only results isn't enough, understanding how those results are produced is necessary for building knowledge. Beisbart and Räz [140], however, argue that explainability itself is complex. On the one hand, we need to define what exactly requires explanation. On the other hand, understanding and explanation don't always align. There's a difference between explanatory understanding, which clarifies why something occurs, and objective understanding, which provides general knowledge of a domain based on theories or models without necessarily explaining the "why." Thus, explainability in medical AI (MAI) needs careful consideration to clarify what "explaining" truly requires and if it's essential for understanding. Explainability also involves interpretability, that is presenting a system's operations in simple, understandable terms and completeness or fully detailing how a system functions. Yet, achieving both can be challenging, a detailed mathematical description may satisfy completeness but lack interpretability for users [150].

Not all experts prioritize explainability, with some stressing the importance of accuracy in algorithmic decision-making. While some view the trade-off between accuracy and explainability as problematic [151], others argue that accuracy should take precedence [143]. Some even suggest forgoing explainability entirely, instead focusing on system reliability. As Durán and Formanek argue, reliability, which means ensuring consistent and verified results, is more critical than understanding every detail [138]. This view, known as computational reliabilism, implies that trust in algorithms should be based on proven performance rather than transparency. Physicians and patients can trust algorithmic decisions based on demonstrated

accuracy and validity rather than explainability alone. Ferrario et al. describe this trust as growing with each validated use [152], and only using machine learning systems for empirically proven tasks can further ensure autonomy and accountability [143].

Beyond ethical concerns, explainability may also be legally required, as laws on liability and individual rights may demand it [153]. Solutions combine technical fixes, social practices, and regulations to address ethical issues around AI opacity. A common approach is to treat explainability in machine learning as essential for healthcare AI, making it a core design goal. Explainable AI (XAI) aims to answer three basic questions: Why did the algorithm do this? Can I trust the results? How can I fix an error? [154].

2.7 Conclusion

This chapter offers a detailed examination of meningitis and its associated infectious causes. It analyzes trends and prevalence across different populations to underscore the evolving nature of infectious diseases and the necessity for adaptive, accurate diagnostic approaches. Furthermore, we highlighted the crucial role of analyzing CSF in enhancing the diagnosis and management of meningitis. We found significant variations in biomarker levels associated with different causative agents among children and adults. At the same time, the elderly displayed less variability in certain biomarkers, suggesting age-related distinctions in how meningitis manifests. Understanding these unique immune responses and biochemical profiles across age groups enables clinicians to enhance diagnostic accuracy and tailor treatment plans to better meet patient needs. Additionally, we defined several explainability aspects of AI models in medical diagnostics, emphasizing the need for transparent algorithms to support clinician decision-making and enhance patient trust. Explainability fosters trust in AI-driven tools by helping healthcare providers and patients understand the rationale behind diagnostic insights. This study explored XAI's ethical and practical implications, underlining its importance in maintaining patient autonomy and informed consent.

Chapter 3

Towards XAI agnostic explainability to assess differential diagnosis for Meningitis diseases

*****

3.1 Introduction

Meningitis diseases pose a significant threat to public health, and their rapid transmission can lead to widespread outbreaks if not managed effectively. Early detection and interventions are critical to reducing the morbidity and mortality associated with these diseases. Delayed diagnosis and treatment can result in severe complications such as brain damage, hearing loss, and even death [155]. According to the World Health Organization (WHO), approximately 1 in 10 people with bacterial Meningitis die, and 1 in 5 suffer from long-term severe complications [156]. Rapid and accurate diagnosis allows for promptly initiating appropriate treatment and isolation measures to prevent further transmission. Therefore, developing and implementing fast and accurate diagnostic tools are crucial for preventing the spread of Meningitis diseases and reducing associated morbidity and mortality. Although culture and smear microscopy are commonly used for meningitis diagnosis, low sensitivity limits their effectiveness. One can miss up to 30% of cases [157], especially when patients have received antibiotics before testing. Additionally, traditional diagnostic methods can be time-consuming, invasive, and costly, leading to delays in treatment and increased healthcare expenditures. This limitation highlights the pressing need for more reliable diagnostic methods to improve the early detection of Meningitis and enhance patient outcomes.

Machine Learning (ML) models have shown great potential in diagnosing various medical conditions, including infectious diseases [2]. A recent study implements and validates an artificial AI model for early aetiological determination of patients with encephalitis and Meningitis. Considering four categories (autoimmune, bacterial, tuberculosis, and viral), based on the initial 24-h data, it identifies essential factors among these aetiologies in the classification process [52]. Furthermore, Artificial intelligence (AI) played a significant role in the healthcare and medical field, particularly during the COVID-19 pandemic, assisting with vaccination and improving human decision-making. AI has been employed to analyze large datasets related to COVID-19, encompassing infection rates, transmission patterns, and demographic information. This analysis enables targeted interventions, diagnosis, and preventive measures to reduce human-to-human infectivity of COVID-19. The employment of AI has also shown great potential in advancing COVID-19 vaccine development by predicting potential epitopes with antigenic characteristics and detecting virus mutations through Deep Convolutional Neural Networks (CNN) [158, 159]. Similarly, applying deep learning approaches has yielded promising results in identifying Acute Lymphoblastic Leukemia (ALL). A novel approach leveraging ensemble CNN models was introduced to overcome the limitations of traditional methods, such as Peripheral Blood Smear (PBS) examination, which is laborious, time-consuming, and heavily reliant on specialists' expertise to detect ALL. These models have demonstrated promising results for feature extraction from images and the classification of B-ALL lymphoblast and normal cells [160]. Moreover, recent studies employing CNN models have shown promising outcomes in diagnosing different types of malaria using blood smear images comprising various strains, specifically Falciparum, Vivax, and Ovale, and samples from healthy individuals [161]. These models can analyze vast amounts of data, offering accurate predictions that can improve diagnostic accuracy and reduce the time required for diagnosis. However, the lack of interpretability in these models poses challenges in clinical practice. Indeed, clinicians struggle to understand and trust what they perceive as black boxes. This work deals with ML interpretability for diagnosing meningitis disease. Indeed, interpreting machine learning models for diagnosing Meningitis is critical and challenging, given the diverse clinical presentations and the importance of understanding the model's decisions in medical applications. Recognizing the need for effective resolution of medical problems, we are focusing on a clinical decision support system (CDSS) to aid medical teams in making automated decisions for diagnosing different types of meningitis cases, including Meningococcaemia, Meningococcal Meningitis, Tuberculous Meningitis, Aseptic Meningitis, Haemophilus influenzae meningitis, and Pneumococcal Meningitis.

In this chapter, we propose an innovative approach to improve meningitis diagnosis using SHapley Additive explanations (SHAP) model-agnostic techniques applied to the outcome predictions of an XGBoost classifier. Our study addresses a significant gap in the exist-

ing literature by examining the clinical presentations of meningitis across various classes, including Meningococcaemia, Meningococcal Meningitis, Tuberculous Meningitis, Aseptic Meningitis, Haemophilus influenzae meningitis, and Pneumococcal Meningitis. We employ SHAP values to identify influential factors in the model's predictions, highlighting essential biomarkers and attributes such as Neutrophils and lymphocyte levels, White Cell Count (WCC), Protein and Glucose ratios, and clinical signs. Notably, our analysis reveals variations in the impact of these features across different meningitis types, aligning with established clinical associations. These insights lay the groundwork for an automated meningitis diagnosis tool and suggest avenues for further investigation. Furthermore, our experiments, conducted on both the SINAN database and a real dataset from Setif's hospital in Algeria, demonstrate the efficacy of our methodology in balancing model accuracy and interpretability, offering a promising approach to enhance meningitis diagnosis and improve patient outcomes.

3.2 Method

AI-driven techniques, powered by machine learning algorithms and data analysis, have the potential to analyze vast amounts of patient data, recognize patterns, and identify subtle indicators that might escape human observation [162–165]. For this purpose, our proposed methodology to diagnose the Meningitis cases and discover the differential diagnosis is illustrated in Figure 3.1. In the first phase, the data processing techniques were applied, and synthetic data was generated using the SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous technique) to address the class imbalance issue. Secondly, multiple classifiers were trained using 70% of the overall dataset. We evaluate their performance using the remaining 30% for testing. Subsequently, the fitting performances of the models were compared, and the best-performing model was selected. In the third phase, we investigated the SHAP analyses (Shapley Additive exPlanations) to study and visualize the impact of different features across the various diagnosis cases. The explanation provides valuable insights into the contribution of each feature to the model's predictions and aids in understanding the importance of different factors in diagnosing Meningitis diseases. Finally, the interpretation phase was performed in collaboration with the medical experts to discover the AI-powered diagnostic indicators that can assist in early detection and swift identification of Meningitis cases. Furthermore, early detection often translates to timely intervention and treatment, potentially preventing disease progression and improving prognosis.

3.2.1 Data preparation: Study case

We conducted the experiments on 6729 notified Meningitis cases of individuals over 18, retrieved from SINAN, the Information System on Notifiable Diseases of the Brazilian Government's Health Department from 2003 to 2022. SINAN is a database that encompasses

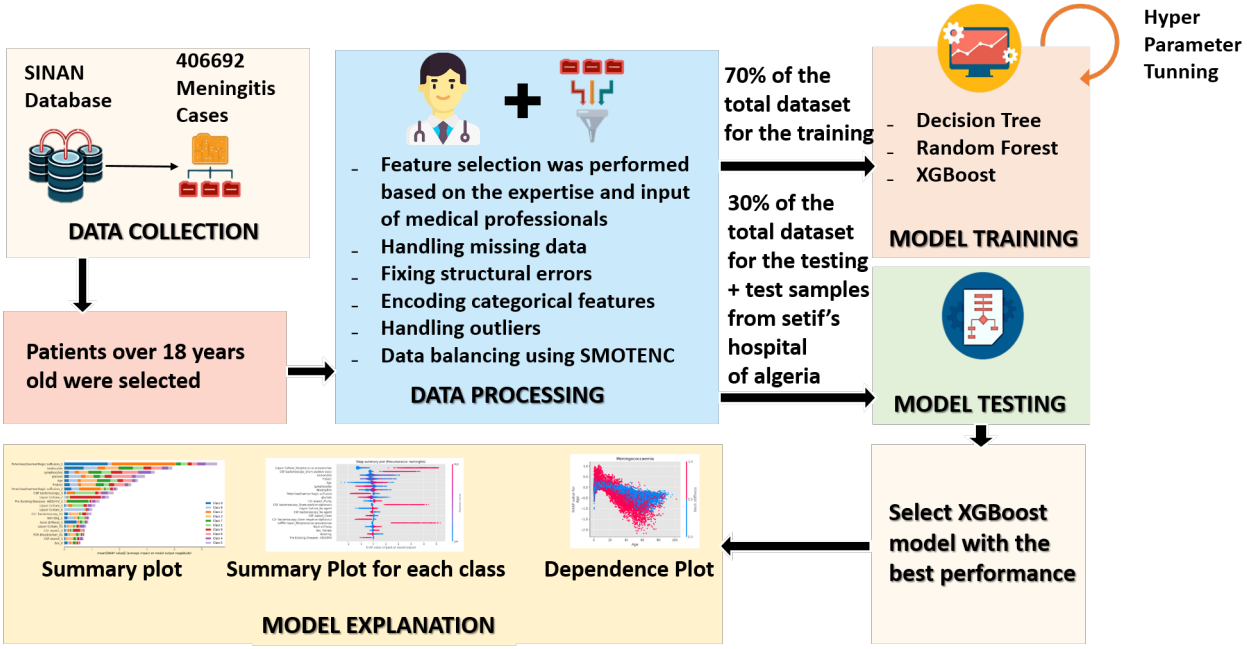


Figure 3.1: Illustration of the proposed Meningitis diagnostic workflow, based on an XAI-agnostic explainability framework. Preprocessing involves data augmentation via SMOTENC. Multiple machine learning classifiers are trained and evaluated, with SHAP analyses identifying feature relevance in diagnosis. The interpretation phase, supported by medical professionals, extracts meaningful explanations for the AI's decision-making, ensuring transparency without focusing on internal model mechanics.

compulsory disease notifications throughout Brazil. Notifications and investigations are stored in the SINAN NET database, with dedicated tables for each specific disease, including meningitis.

Initially, the dataset consists of 123 attributes. In our work, we focused on 34 specific attributes that are centered around clinical signs and biological examinations. These attributes were carefully chosen in consultation with the infectious disease experts at Setif Hospital in Algeria, as they are the most informative and relevant for Meningitis diagnosis. Table 3.1 and 3.2 summarizes the description and the possible values of the selected attributes from the dataset.

We collected a dataset from Setif's Hospital in Algeria, comprising cases notified as Meningitis. Notable disparities were observed compared to the dataset sourced from the Brazilian SINAN database. This dataset diversification aims to enhance our evaluation samples and offer a broader perspective on the performance and generalizability of our explainable AI model. Within Setif's hospital dataset, we identified cases suitable for testing our model. However, these cases only encompass two types of Meningitis, constituting a subset of the classes found in the training dataset of the SINAN database.

Table 3.1: Key prognostic factors in Meningitis diagnosis, including demographics, medical history, pre-existing illnesses, and clinical signs.

Diagnostic evaluation	**Attributes**	**Categories**	**Description**
Demographic information	Age	Numerical	Informs the patient's age
	Sex	Categorical M - Male F - Female	Informs the patient's sex
Medical History Pre-existing illnesses	Pre-Existing Diseases - AIDS/HIV	Categorical 1 - Yes 2 - No 9 - Ignored	Provide information about the presence of pre-existing diseases or illnesses in an individual's medical history
	Pre-Existing Illnesses - ARI		
	Pre-Existing Diseases - Tuberculosis		
	Pre-Existing Illnesses - Trauma		
	Pre-Existing Illnesses - Hospital Infection		
	Pre-Existing Illnesses - Other		
	Vaccination - Meningococcal C Conjugate		Indicate whether the person has received specific vaccinations to protect against certain diseases or infections
	Vaccination - BCG		
	Vaccination - Triple Viral		
	Vaccination - Hemophile – Tetravalent or Hib		
	Vaccination - Pneumococcus		
Clinical Signs/Symptoms	Headache	Categorical 1 - Yes 2 - No 9 - Ignored	Informs if this clinical manifestation has occurred in the patient
	Fever		
	Vomiting		
	Seizures		
	Neck stiffness		
	Petechiae / haemorrhagic suffusion		
	Kernig/Brudzinski		
	Coma		

Table 3.2: Key prognostic factors in meningitis diagnosis from biological test results

Diagnostic evaluation	**Attributes**	**Categories**	**Description**
Biological tests	Lymphocytes	Numerical	Chemocytological examination
	Neutrophils		
	CSF White Cell Count (WCC)		
	Protein ratio CSF/serum		
	Glucose ratio CSF/serum		
	CSF aspect	Categorical 1 - Clear 2 - Purulent 3 - Haemorrhagic 4 - Cloudy 5 - Xanthochromic 6 - Other 7 - Ignored	Appearance of the cerebrospinal fluid (CSF) obtained through the puncture
	PCR Blood/serum	Categorical 62 - Ignored 37 - Mumps 38 - Measles 39 - Herpes Simplex 40 - Chickenpox/Herpes Zoster 41 - Rubella 55 - Influenza 72 - Dengue 61 - Unrealized 75 - Not identified 70 - Adenovirus 56 - Echovirus 63 - Coxsackie Virus 59 - Other Enteroviruses 71 - West Nile Virus 73 - Other Arboviruses 74 - Other viruses 01 - Neisseria meningitidis 06 - Haemophilus influenzae 07 - Streptococcus pneumoniae 28 - Other bacteria 43 - Cryptococcus/Torula 42 - Other fungi 48 - Toxoplasma 52 - Other Parasites 51 - No agent	List of causative agents for meningitis identified through PCR testing

	CSF Bacterioscopy	Categorical 61 - Unrealized 62 - ignored 32 - Gram negative bacilli 31 - Gram positive bacilli 36 - Coccobacilli 34 - Gram negative cocci 33 - Gram positive cocci 35 - Gram negative diplobacilli 03 - Gram negative diplococci 08 - Gram positive diplococci 28 - Other Bacteria 51 - no agent	Etiological agent identified in the examination
	CSF Culture Blood/serum culture	Categorical 61 - Unrealized 62 - Ignored 01 - Neisseria meningitidis 06 - Haemophilus influenzae 07 - Streptococcus pneumoniae 28 - Other bacteria 51 - No agent	Etiological agent identified in the examination
	LATEX CSF LATEX Blood/serum	Categorical 61 - Unrealized 62 - Ignored 01 - Neisseria meningitidis 06 - Haemophilus influenzae 07 - Streptococcus pneumoniae 14 - Streptococci (sp, piogens, alpha, hemolytic, faecalis, agalactiae) 28 - Other bacteria 43 - Cryptococci 51 - No agent	Etiological agent identified in the examination
Diagnosis	Type of causative agent	0 - Meningococcaemia 1 - Meningococcal meningitis 2 - Tuberculous meningitis 3 - Meningitis by other bacteria 4 - Unspecified meningitis 5 - Aseptic meningitis 6 - Meningitis due to other aetiology 7 - Haemophilus influenzae meningitis 8 - Pneumococcal meningitis	Confirmed diagnosis

3.2.2 Data preprocessing

In this phase, a series of data preparation steps have been performed on the data set, including addressing duplicate data and missing values and handling outliers. Additionally, a crucial step involved balancing the data to mitigate any biases and enhance the overall learning of the model. These rigorous data preparation steps have been undertaken to minimize distortions and improve the accuracy of predictions:

- The data type transformation: We converted the data type on selected features to ensure compatibility with the model's requirements and enhance prediction accuracy.

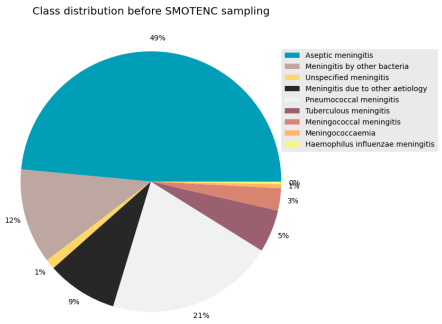
- The missing Data: To preserve the integrity of the analysis and consider the substantial volume of data available, all observations with missing values were omitted from the analysis.

- The structural errors: Fixing structural errors in the preprocessing step, including eliminating typos and inaccurate information, ensures consistent features. This step enhances accuracy, reduces biases, and improves subsequent analysis and modeling.

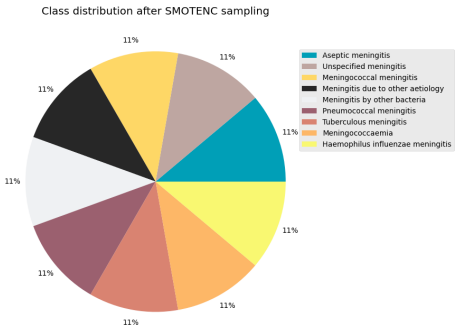
- The categorical features encoding: We identified several categorical attributes pre-encoded from the original data in the dataset. We selectively converted the necessary attributes into integer form for model compatibility and prediction generation, enabling their effective utilization within our analysis. It is worth noting that certain machine learning algorithms, such as decision trees and rule-induction methods (e.g., CART, C5.0, etc.), possess inherent capabilities to handle high-cardinality categorical attributes without the need for external preprocessing steps [166].

- The outliers: In medical research, outliers hold valuable insights into rare or atypical cases, contributing to a comprehensive analysis. These extreme values may or may not represent aspects of data intrinsic variability and may have a legitimate place in the dataset [167]. We applied the Interquartile Range (IQR) method to handle outliers specifically to numerical attributes, including laboratory biomarkers. The IQR method is a statistical technique employed to detect and manage outliers within the dataset. Using the IQR method, we systematically deal with outliers, leading to a more comprehensive data variability. Following this process, the dataset size was reduced to 5072 samples.

- The data balancing: we employed the SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) method to deal with the class imbalance. SMOTE-NC extends SMOTE to handle categorical and numerical features, generating synthetic samples. Choosing SMOTE-NC over SMOTE allows us to balance the dataset



(a) Original dataset class distribution before SMOTE-NC.



(b) Dataset class distribution after applying SMOTE-NC.

Figure 3.2: Comparison of class distribution before and after applying SMOTE-NC. Subfigure (a) shows the class distribution before using SMOTE-NC, revealing a significant imbalance, with 'aseptic Meningitis' as the majority class at 49%. Subfigure (b) demonstrates the post-SMOTE-NC class distribution, which effectively balances the classes, resulting in an equal number of instances for each class, totaling 2460 instances.

while maintaining feature integrity, which ensures fair representation and reduces bias towards the majority class [168]. Figure 3.2 shows the class distribution pre-application and post-application of SMOTE-NC.

3.2.3 Models investigation

In this section, we justify the choice of three tree-based decision models: the decision tree classifier, random forest classifier, and XGBoost classifier. Among the several attractive properties of tree-based methods is their ability to capture complex interactions between predictors [169]. Moreover, they are considered less prone to outliers, require no distributional assumptions or data transformations, and are intuitive [170].

Various metrics assess distinct characteristics of the classifier generated by the classification algorithm [171]. This study utilized diverse performance metrics to evaluate the developed classifiers. The chosen metrics encompass accuracy, recall, precision, F1-score, and area under the receiver operating characteristic curve (AUROC). The dataset was divided into two subsets to evaluate the classifiers' performance appropriately: 70% for training the models and 30% for testing their generalization capabilities. This division allowed us to effectively train the classifiers on a significant portion of the data while reserving a separate portion for independent assessment, serving as a benchmark to evaluate the classifier's predictive accuracy on unseen instances.

3.2.4 Model agnostic explainibility

The SHAP (Shapley Additive exPlanations) method is a model-agnostic interpretation technique used to interpret the results of a predictive model. It employs the Shapley value, a cooperative game theory concept, to quantify each feature's contribution to the model's prediction. The Shapley value formula ($\phi_j(val)$) considers all possible feature combinations and calculates the marginal contribution of each feature. It satisfies desirable properties like efficiency, symmetry, dummy, and additivity.

The formula for the Shapley value ($\phi_j(val)$) is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{j\}) - val(S)) \tag{3.1}$$

In the formula 3.1, S represents a subset of features used in the model, x is the vector of feature values for the instance being explained, and p is the total number of features. The function $val(x)$ represents the prediction for the feature values x, marginalized over the features not included in the set S. It is obtained by integrating the model's predictions over the ranges of the excluded features, weighted by their respective probabilities. The Shapley value for feature j is calculated by summing the contributions of all subsets S that do not contain feature j and weighing each contribution by the number of possible orderings of the features in the subset [21, 28].

The Shapley value satisfies several desirable properties, such as:

1. Completeness/Efficiency: The sum of the Shapley values for all features equals the difference between the model's prediction for the instance x and the average prediction for all possible instances:

$$\sum_{i=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X)) \tag{3.2}$$

2. Symmetry: If two features values j and k contribute equally to all possible coalitions, their Shapley values should be the same:

$$\begin{aligned} \text{if} \quad & val(S \cup \{j\}) = val(S \cup \{k\}) \\ \text{for all} \quad & S \subseteq \{1, \ldots, p\} \setminus \{j, k\} \\ \text{then} \quad & \phi_j(val) = \phi_k(val) \end{aligned} \tag{3.3}$$

3. Dummy: If a feature j has no impact on the model's prediction, regardless of its

inclusion in coalitions, its Shapley value should be zero:

$$\begin{aligned} \text{if} \quad & val(S \cup \{j\}) = val(S) \\ \text{for all} \quad & S \subseteq \{1, \ldots, p\} \\ \text{then} \quad & \phi_j(val) = 0 \end{aligned} \tag{3.4}$$

4. Linearity: When combining two models, represented by val and val' , the overall prediction should correspond to the sum of the contributions from each model.

$$\phi_j(val + val') = \phi_j(val) + \phi_j(val') \tag{3.5}$$

The linearity property is beneficial when using ensemble models like XGBoost with TreeSHAP. It enables the computation of the Shapley Value for a feature by averaging the contributions of each tree in the ensemble. This simplifies the calculation process and enhances the interpretation of individual feature contributions.
In our study, we employed the TreeSHAP method to interpret the results of the XGBoost model. We chose SHAP due to its model-agnostic nature, theoretical grounding, and ability to provide local and global interpretations. This approach improves the transparency, interpretability, and identification of potential issues or biases in the model.

3.3 Results

3.3.1 Model validation

The performances of the classifiers are provided in Table 3.3. It includes key metrics such as precision, recall, F1-score, and support for each Meningitis class within the dataset.
Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive, indicating the model's ability to avoid false positives. Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances out of all actual positive instances, indicating the model's ability to capture all relevant positives. The F1-score is the harmonic mean of precision and recall, providing a balanced evaluation of the classifier's performance by considering both metrics. Lastly, the support column in the classification report indicates the number of instances belonging to each class, providing context and understanding of the distribution and representation of classes in the dataset.
Table 3.3 provides valuable insights into the model's ability to classify different Meningitis classes accurately. Random Forest (RF) and XGBoost (XGB) achieved consistently high precision scores, ranging from 0.83 to 1.00 and 0.78 to 1.00, respectively, indicating their effectiveness in minimizing false positives. The Decision Tree (DT) classifier also performed

Table 3.3: Performance Metrics by Meningitis class

Meningitis cases	Precision			Recall			F1-Score			Support
	DT	RF	XGB	DT	RF	XGB	DT	RF	XGB	
Meningoccaemia	0.99	1.00	1.00	0.99	0.99	1.00	0.99	1.00	1.00	738
Meningoccal Meningitis	0.97	1.00	0.99	0.90	0.95	0.98	0.93	0.97	0.98	738
Tuberculous Meningitis	0.83	0.94	0.92	0.71	0.86	0.94	0.76	0.90	0.93	738
Meningitis by other bacteria	0.71	0.84	0.78	0.49	0.53	0.80	0.58	0.65	0.79	738
Unspecified Meningitis	0.50	0.88	0.79	0.42	0.42	0.73	0.45	0.57	0.76	738
Aseptic Meningitis	0.73	0.83	0.78	0.51	0.57	0.80	0.60	0.68	0.79	738
Meningitis due to other aetiology	0.83	0.96	0.92	0.64	0.77	0.90	0.72	0.85	0.91	738
Haemophilus influenzae Meningitis	0.91	0.95	0.94	0.94	0.98	0.99	0.92	0.96	0.97	738
Pneumococcal Meningitis	0.97	0.99	0.99	0.94	0.94	0.97	0.96	0.97	0.98	738

well, although slightly lower than RF and XGB, with precision scores ranging from 0.50 to 0.99.

Regarding recall, XGB demonstrated superior performance, consistently achieving scores above 0.73 for all classes, indicating its capability to capture the most positive instances. Decision Tree and Random Forest classifiers also exhibited lower recall scores, ranging from 0.42 to 0.99.

The F1 score combines precision and recall, evaluating the model's overall performance for each class. XGB achieved the highest F1 scores across all classes, ranging from 0.76 to 1.00, followed closely by RF with scores ranging from 0.57 to 1.00. DT showed slightly lower F1 scores, ranging from 0.45 to 0.99, indicating a trade-off between precision and recall.

The support metric ensures a fair evaluation and comparison of the classifiers' performance across classes. The findings demonstrate that XGBoost and Random Forest consistently perform strongly across all metrics. These classifiers are reliable for accurate Meningitis classification, providing high precision in identifying positive cases and effectively capturing the most positive instances. Their ensemble nature, robustness to overfitting, and ability to capture complex relationships contribute to their superior performance compared to the Decision Tree classifier.

Table 3.4 summarizes the results to provide an overview of the classifiers' outcomes. It can be concluded that the XGBoost model outperformed the other classifiers. With the highest accuracy, precision (Macro Avg), recall (Macro Avg), F1-Score (Macro Avg) scores, and One-vs-Rest AUROC (Area Under the Receiver Operating Characteristic), the XGBoost model demonstrated superior performance in accurately classifying the data and capturing the overall patterns in the dataset.

Furthermore, an in-depth analysis of the distinct test set originating from Setif's Hospital, comprising instances of pneumococcal Meningitis and tuberculous Meningitis, revealed sig-

nificant performance metrics for the XGBoost model: (Accuracy: 0.7143, Precision: 1.0, Recall: 0.7143, F1-Score: 0.7857). These results underscore the efficacy of the XGBoost model in precisely categorizing cases within this subset, further affirming its robust performance across specific meningitis types.

Figure 3.3 shows ROC curves and AUC calculations for meningitis classes to compare the

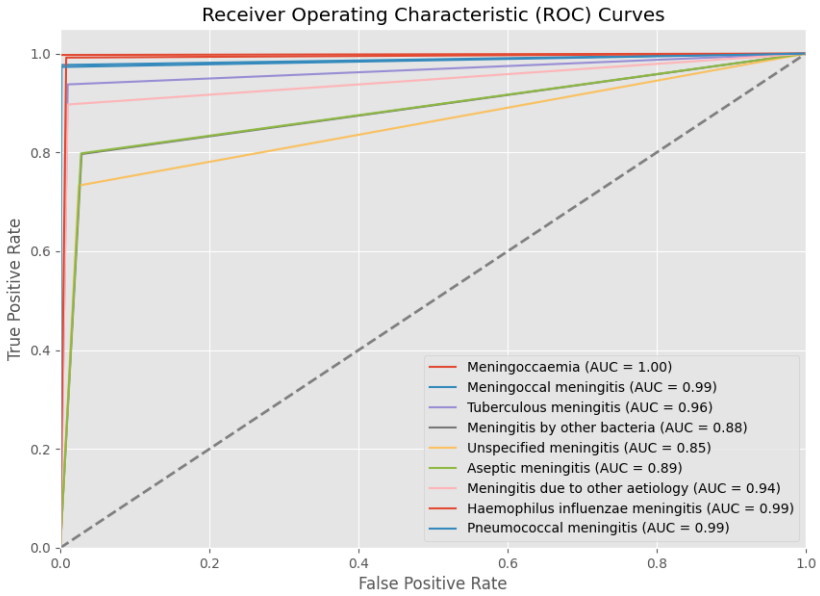


Figure 3.3: Area under the receiver operating characteristic curve (AUROC) of our multi-Class XGBoost model on validation data.

classification performance for each class. Therefore, we selected the XGBoost model as the optimal choice for our classification task due to its superior performance. To provide an in-depth understanding of the XGBoost model and the factors driving its excellent performance, we further delve into its interpretability in section 3.2.4. By utilizing Shapley values, we explore the contributions of individual features towards the model's predictions, unraveling the key insights and highlighting the factors that significantly influence the classification outcomes. This interpretability analysis enhances our understanding of the inner workings of the XGBoost model, shedding light on its decision-making process and reinforcing our confidence in selecting it as the optimal choice for our classification task.

3.3.2 XGBoost Global interpretability

In our study, we primarily focused on specific classes of Meningitis, namely meningococcaemia, meningococcal Meningitis, Tuberculous Meningitis, Aseptic Meningitis, Haemophilus influenzae Meningitis, and Pneumococcal Meningitis. To gain insights into the predictive performance of our XGBoost model, we generated a summary variable importance plot (Figure 3.4).

Table 3.4: Comparison of classifier performance metrics across various models.

Classifier	Metrics (Macro Avg)				
	Accuracy	Precision	Recall	F1-Score	AUROC One-vs-Rest
Decision Tree	0.725	0.826	0.725	0.7691	0.854
Random Forest	0.778	0.932	0.778	0.838	0.886
XGBoost	0.900	0.900	0.90	0.899	0.944

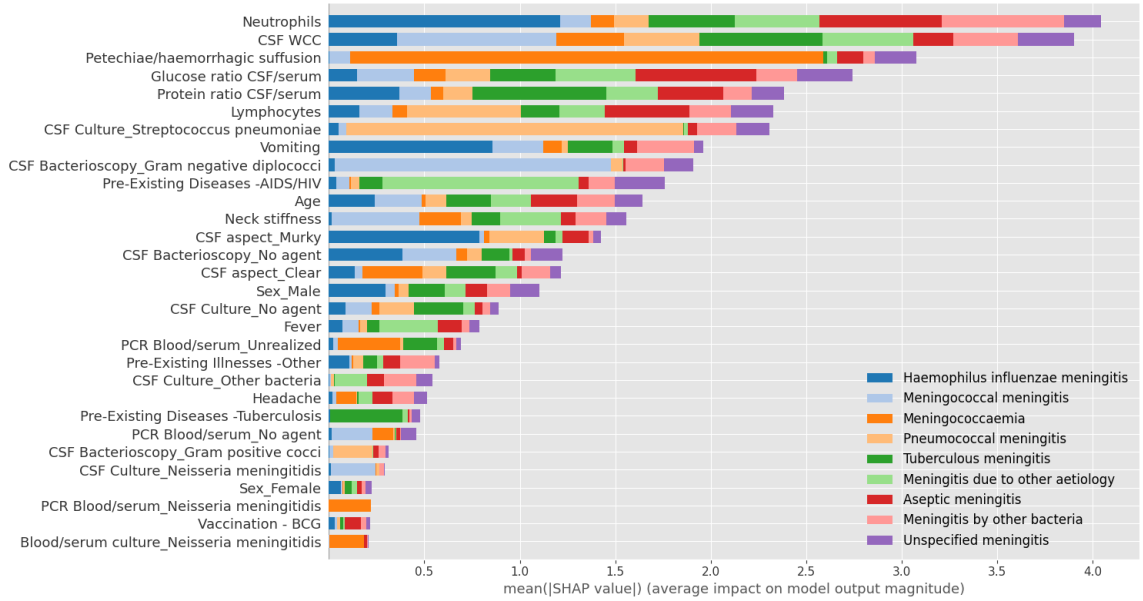


Figure 3.4: Variable importance plot for the XGBoost classifier. This figure shows the average contribution of each feature to the model's predictions, determined by the mean absolute SHAP value across all samples. Features are ranked according to the sum of SHAP value magnitudes across all samples.

Our analysis reveals that several features, including Neutrophils and Lymphocytes level, White Cell Count (WCC), Protein and Glucose ratio, the presence of Petechiae/Haemorrhagic suffusion, along with signs of vomiting and neck stiffness, as well as the identification of Gram-negative diplococci in cerebrospinal fluid (CSF) bacterioscopy and Streptococcus pneumoniae in CSF culture, are among the most influential factors affecting the model's predictions.

Furthermore, we observed that the level of Neutrophils hardly influences the classification of Haemophilus influenzae meningitis, Aseptic Meningitis, and Meningitis by other bacteria. WCC has the most significant influence on Meningococcal and Tuberculosis meningitis. We found that the presence of Petechiae/Haemorrhagic suffusion signs hardly influences the classification of Meningococcaemia. These findings align with the existing literature, which recognizes petechial purpuric exanthema as a classic sign of meningococcemia, present in approximately 40% to 80% of cases [172].

Glucose ratio substantially impacts Aseptic Meningitis, while Protein ratio has a greater influence on Tuberculosis meningitis. Lymphocytes have a more significant impact on Pneumococcal and Aseptic meningitis cases. The figure shows that Neisseria meningitidis and Gram-negative diplococcus in CSF are associated with invasive meningococcal disease [173] along with neck stiffness.

The Murky aspect of the CSF has the most impact on Haemophilus influenzae meningitis and Pneumococcal cases. Additionally, the presence of Streptococcus pneumoniae in CSF culture has the most impact on Pneumococcal meningitis cases. Furthermore, vomiting has a more pronounced effect on Haemophilus influenzae Meningitis than other meningitis types.

3.3.3 Features impact on the Meningitis diagnosis outcome

The Global Interpretability results (Section 3.3.2) demonstrate that our diagnosis outcomes typically align with expert knowledge. We provide additional diagrams constructed to depict the feature's importance and effect on each diagnosis outcome of Meningitis type.As shown in Figure 3.5, our findings emphasize the significance of petechiae/hemorrhagic suffusion as the most influential feature, increasing the likelihood of meningococcemia. This observation aligns with clinical reports, where this symptom is detected in approximately 50% to 60% of patients, strongly associating it with the disease. In cases of meningococcemia, confirmation of the diagnosis involves detecting the presence of Neisseria meningitidis in blood cultures [174]. The strong positive impact of finding this organism in blood cultures significantly increases the probability of meningococcemia for our model.

Moreover, our analysis reveals that clear cerebrospinal fluid appearance has a comparatively modest yet positive influence on the model's predictions. We also find that a low white cell count (WCC), low protein ratio, low neutrophil level, and elevated glucose ratio positively impact the diagnosis of meningococcemia. However, these results deviate from the typical diagnostic characteristics of the condition, characterized by an elevated WCC count in cerebrospinal fluid, increased protein levels, low glucose levels, and gram-negative diplococcus. Further research is essential to reconcile these disparities and gain a comprehensive understanding of the diagnostic indicators for meningococcemia.

The cerebrospinal fluid (CSF) analysis is instrumental in diagnosing meningococcal Meningitis, encompassing vital parameters like Gram stain, culture, glucose and protein levels, and cell count. Notably, CSF findings indicative of bacterial Meningitis frequently include low glucose and elevated protein levels. In some instances, Gram stains may reveal the presence of Gram-negative diplococci [175]. The presence of neutrophils in cerebrospinal fluid is a crucial indicator of a bacterial origin of the Meningitis.

Our analysis, represented in the beeswarm plot for Meningococcal Meningitis (Figure 3.6), highlights noteworthy patterns. Specifically, higher values of Gram-negative diplococci, signifying their presence, result in positive SHAP values. This implies that the absence of

this bacterial type in CSF bacterioscopy corresponds to a lower predicted Meningococcal meningitis class. Detecting these bacteria in CSF bacterioscopy is a robust indicator of a meningococcal meningitis diagnosis [175]. A similar trend is observed when Neisseria meningitidis is detected in blood culture. Furthermore, 'no agent' in the PCR test aligns with the absence of viral agents as the causative factor in meningococcal Meningitis. Additionally, our findings demonstrate that higher Glucose ratio values yield negative SHAP values, while lower values correspond to positive SHAP values. This pattern also extends to the age attribute. In contrast, the protein ratio and Neutrophil level exhibit the opposite effect.

Our findings are consistent with these clinical observations and diagnostic standards. Common symptoms of meningococcal Meningitis often involve a stiff neck, reduced cognitive function, and other signs of meningeal inflammation.

Protein ratio, white cell count, and Neutrophil levels significantly influence the diagnosis of tuberculous Meningitis. Elevated WCC, Neutrophils, and glucose levels negatively impact the prediction of this class. In contrast, high lymphocyte levels and protein ratios contribute positively to diagnosing tuberculous meningitis. Shap values also indicate that the presence of the BCG vaccine feature affects the model's predictions for tuberculosis meningitis. This vaccine consistently protects against the most severe forms of TB, including TB meningitis in children [176]. However, its effectiveness in preventing tuberculosis in adults is comparatively lower.

Figure 3.7 demonstrates that vomiting and neck stiffness signs negatively affect the diagnosis of tuberculous Meningitis. Tuberculous Meningitis is frequently observed in patients with tuberculosis and/or HIV/AIDS, either as a new occurrence of the disease or as a consequence of a prior tuberculosis infection. In cases of co-occurring tuberculosis and HIV, patients may currently have tuberculosis, indicating a co-infection, or have a history of tuberculosis. Although our dataset lacks specific information on the causative agent, these observations align with the well-established understanding that tuberculous Meningitis is primarily caused by Mycobacterium tuberculosis. It is essential to consider these factors when diagnosing patients with suspected tuberculous Meningitis.

Aseptic Meningitis is a condition characterized by negative bacterial cultures of CSF and can be caused by various aetiologies [177]. It is commonly, associated with viral Meningitis or prior antibiotic usage. Our analysis (Figure 3.8) has revealed factors that positively correlate with diagnosing aseptic Meningitis. These include elevated levels of glucose and lymphocytes in the CSF, along with negative results in culture and latex tests for bacterial agents. Conversely, a diagnosis of aseptic Meningitis has been associated with lower levels of neutrophils and reduced protein levels in the CSF. The presence of seizures and neck stiffness positively predicts this particular outcome.

The diagnosis of Haemophilus influenzae meningitis relies on a combination of clinical manifestations and specific diagnostic investigations, including laboratory testing and cere-

brospinal fluid analysis. In cases of Haemophilus influenzae meningitis, the cerebrospinal fluid often appears cloudy or turbid (murky) and contains an increased number of white blood cells, primarily neutrophils. Our Shapley value analysis (Figure 3.9) reinforces these findings, highlighting that high neutrophil values, relatively elevated white cell count (WCC), and a cloudy CSF appearance positively influence the prediction of the target class. Confirmation of the presence of Haemophilus influenzae type b (Hib), the most common causative organism, can be achieved through Gram staining and bacterial culture of the CSF. Haemophilus influenzae infections are most prevalent at the extremes of age, affecting infants, young children, and older adults. However, due to specific vaccines, the incidence of Haemophilus influenzae infections has significantly decreased in the general population. Nevertheless, it is still observed in patients aged 65 and older with underlying conditions. The dataset used for this analysis consists of medical records and clinical data from individuals aged 18 years and older. This age range was chosen to focus on the adult population and provide insights into cerebrospinal fluid profiles in cases of Haemophilus meningitis. Our analysis further confirms that the presence of Haemophilus influenzae, as indicated by culture results, is a positive factor in diagnosing Haemophilus meningitis.

Pneumococcal Meningitis is often marked by a cloudy appearance of the CSF, elevated white blood cell count with a predominance of neutrophils, a substantial reduction in glucose levels, occasionally reaching near zero, and high protein levels within the CSF. Moreover, CSF bacterioscopy and LATEX tests can confirm the presence of Streptococcus pneumoniae, the bacteria responsible for pneumococcal Meningitis. Our analysis (Figure 3.10) reaffirms the typical diagnostic indicators of pneumococcal Meningitis. This includes the positive influence of cloudy/Murky appearance of the CSF and decreased glucose levels on predicting the disease. Furthermore, the presence of Gram-positive cocci and Gram-positive diplococci in CSF bacterioscopy strongly contributes to diagnosing pneumococcal Meningitis. Consistent with these findings, identifying Streptococcus pneumoniae through culture and LATEX tests further supports the diagnosis of Pneumococcal Meningitis.

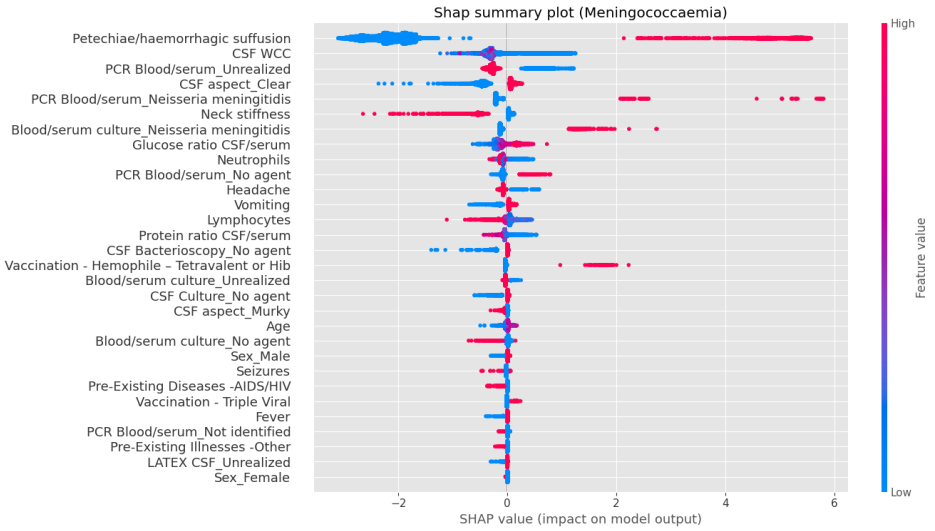


Figure 3.5: SHAP summary plot for Meningococcaemia Meningitis.

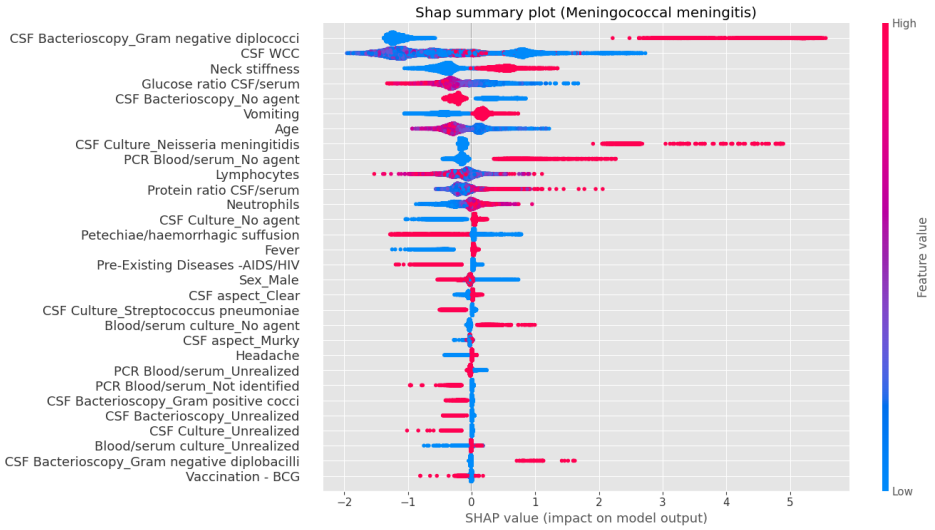


Figure 3.6: SHAP summary plot for Meningococcal Meningitis.

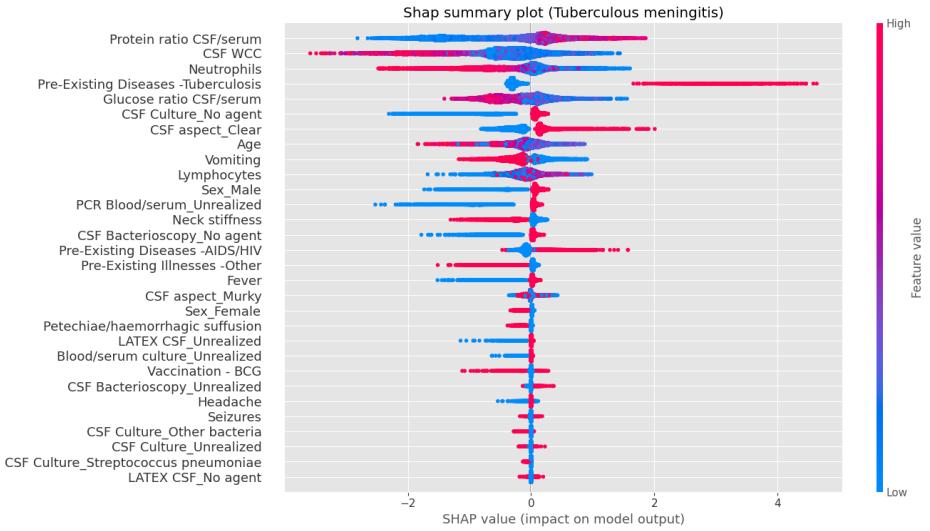


Figure 3.7: SHAP summary plot for Tuberculous Meningitis.

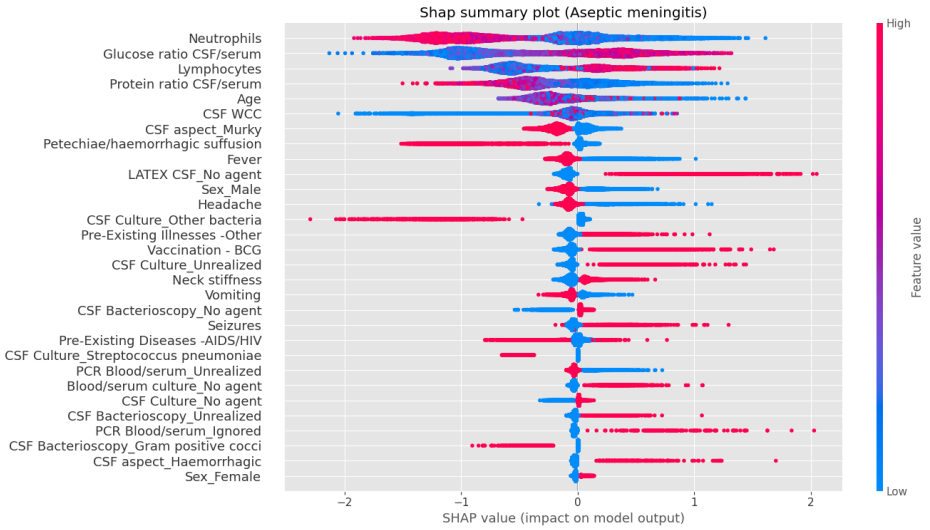


Figure 3.8: SHAP summary plot for Aseptic Meningitis.

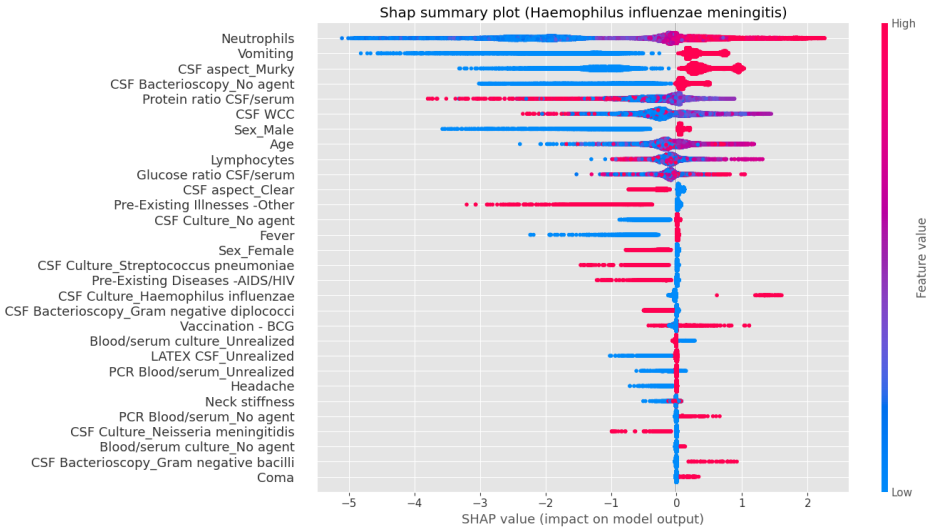


Figure 3.9: SHAP summary plot for Haemophilus influenzae Meningitis.

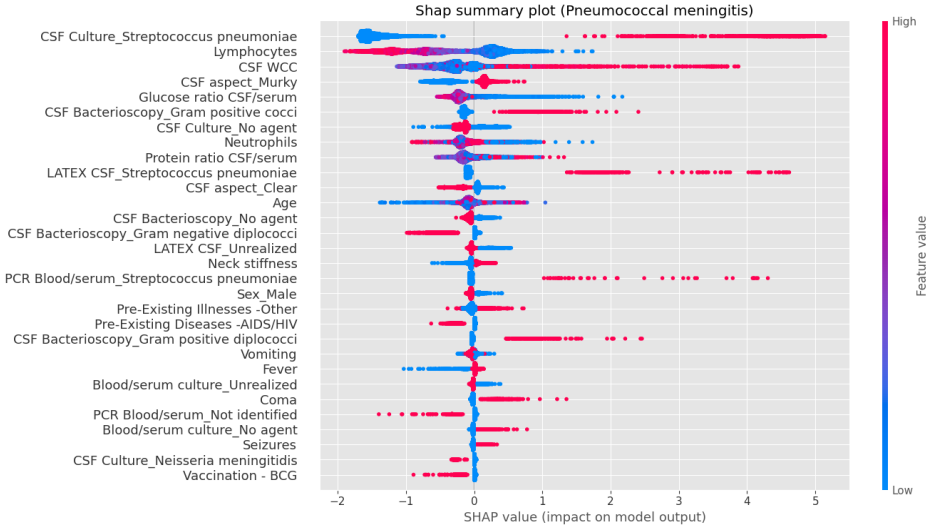


Figure 3.10: SHAP summary plot for Pneumococcal Meningitis.

3.3.4 Influence of Neutrophil and Lymphocyte Levels on Meningitis Predictions

Figure 3.11 illustrates the SHAP (SHapley Additive exPlanations) dependence plot, offering insights into the impact of Neutrophil and Lymphocyte levels on the model's predictions. It provides a clear visualization of how changes in these specific biomarkers impact the predictive outcome of the model. We focused our analysis on cases of meningococcal Meningitis, tuberculosis meningitis, aseptic Meningitis, and Haemophilus influenzae meningitis, as these categories displayed distinct interaction patterns with these features.

Analysis shows a common trend in the dependence plots, characterized by a monotonically increasing curve for meningococcal and Haemophilus influenzae Meningitis. This suggests

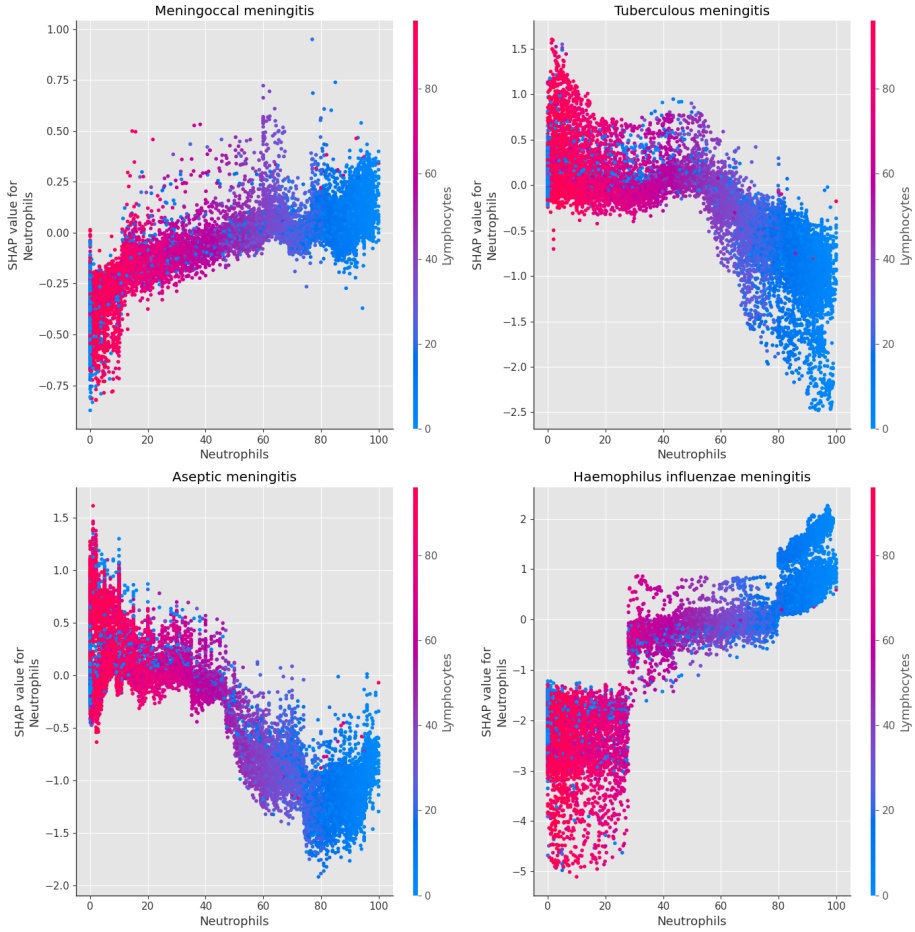


Figure 3.11: Dependence plot illustrating the impact of Neutrophil and Lymphocyte levels on predicting outcomes for meningococcal Meningitis, tuberculosis meningitis, aseptic Meningitis, and Haemophilus influenzae meningitis.

that changes in these features significantly affect the model's predictions, as evidenced by the positive Shapley values. We further identified specific ranges where variations in neutrophil and lymphocyte levels substantially impacted the predictions. This range typically falls within the ¿40% neutrophils and 40-60% lymphocytes range for meningococcal Meningitis. Within this range, these feature variations notably influence the model's predictions. Similarly, for Haemophilus influenzae meningitis, we observed a monotonic increasing pattern, particularly with a strong predominance of neutrophils. However, a significant threshold effect becomes evident when neutrophil levels reach approximately 80%, resulting in a noticeable and substantial increase in Shapley values. This highlights the impact of neutrophils on the model's predictions for both types of Meningitis. Conversely, Tuberculous and Aseptic meningitis exhibited a typical pattern characterized by a monotonically decreasing curve. Positive Shapley values in Tuberculous Meningitis were associated with neutrophil levels below 60% and high lymphocyte levels above 60%, signifying their influence on the model's predictions. In contrast, Shapley values consistently remained below 0 for Aseptic Meningitis, primarily for

neutrophil levels around 40% and lymphocyte levels below 50-60%. This suggests a predominance of lymphocytes influencing the model's predictions for Tuberculous and Aseptic meningitis.

3.4 Discussion

Our research focused on applying machine learning techniques to diagnose various types of acute Meningitis. Machine learning algorithms analyze large volumes of biomedical laboratory testing, clinical symptoms, or clinical records, enabling the identification of intricate disease patterns [178]. We have applied three tree-based decision models, and the XGBoost classifier outperformed the others, achieving 90% accuracy and a 94% AUROC. Through a comprehensive analysis using the SHapley Additive exPlanations technique, we identified key diagnostic indicators for different meningitis types, shedding light on the significance of specific clinical signs and laboratory biomarkers.

The results of the Global Interpretability (see Section 3.3.2) demonstrate that our diagnosis outcomes typically align with expert knowledge. The identified diagnostic indicators can aid healthcare professionals in making more precise and timely diagnoses and reinforce confidence and effectiveness in integrating AI-driven techniques into clinical practice. This can improve patient outcomes and reduce misdiagnoses, as early intervention for Meningitis diagnosis is crucial. Furthermore, we provide additional diagrams to depict the features' importance and effect on each diagnosis outcome of Meningitis type.

Our analysis identifies vital diagnostic indicators for various forms of Meningitis. Notably, petechiae/hemorrhagic suffusion is a significant predictor for meningococcemia (see Figure 3.5), while other factors exhibit distinctive patterns involving clear cerebrospinal fluid appearance and specific biomarkers. Also, the presence of Gram-negative diplococci and Neisseria meningitidis, as shown Figure 3.6, strongly supports the diagnosis of meningococcal meningitis [175], with specific biomarkers exhibiting contrasting effects. Tuberculous Meningitis is significantly influenced by attributes such as protein ratio, white cell count, Neutrophils, Lymphocytes, and the presence of the BCG vaccine. Aseptic meningitis diagnosis (see Figure 3.8) correlates with elevated glucose and lymphocyte levels, negative culture and latex tests, low neutrophil levels, and reduced protein levels, with seizures and neck stiffness contributing positively. Haemophilus influenzae meningitis diagnosis is associated with high neutrophil values, elevated white cell count, and a cloudy cerebrospinal fluid appearance, with the presence of Haemophilus influenzae in culture positively influencing the diagnosis (see Figure 3.9). Pneumococcal meningitis diagnosis is characterized by a cloudy cerebrospinal fluid appearance, low glucose levels, the presence of Gram-positive cocci in CSF bacterioscopy, and the identification of Streptococcus pneumoniae (Figure 3.10).

To gain deeper insights into the influence of specific biomarkers on the predictive out-

comes of our model, we focused on studying the impact of Neutrophil and Lymphocyte levels. We identified distinct ranges where variations in these biomarkers significantly affected the model's predictions. For meningococcal Meningitis, Neutrophil levels above 40% and Lymphocyte levels between 40% to 60% notably influenced the model's outcomes. Similarly, in the case of Haemophilus influenzae meningitis, Neutrophil predominance is observed with a significant threshold effect when it reaches 80%. This highlighted the critical role of Neutrophils in predicting both types of Meningitis. The case of Tuberculous and Aseptic meningitis showed a common decreasing pattern of Shapley values. For Tuberculous Meningitis, Neutrophil levels below 60% , and high Lymphocyte levels above 60% were associated with positive Shapley values. Conversely, Aseptic Meningitis consistently had Shapley values below 0, primarily when Neutrophil levels were around 40%, and Lymphocyte levels were below 50-60%.

While our study draws on a substantial dataset and has undergone rigorous validation, several limitations must be acknowledged. First, interpretability lacks standard evaluation metrics, making quantifying and comparing different models challenging.

3.5 Conclusion

One crucial aspect of clinical reasoning is differential diagnosis, where a list of potential problems causing a patient's signs and symptoms is developed. This process allows for a thorough investigation to rule out possibilities and confirm an accurate diagnosis. However, losing follow-up on investigations and patients leads to diagnostic delays or misdiagnoses. In this work, we developed an explainable AI automatic medical decision approach to highlight the importance of specific features in accurately diagnosing different types of Meningitis. The XGBoost model demonstrates a vital accuracy of 0.90 and AUROC of 0.944. We performed an evaluation test using collected data from Setif's hospital in Algeria to assess the model's efficiency in handling diverse and unseen real-world instances. The test set includes instances of pneumococcal Meningitis and tuberculous Meningitis and reveals notable performance metrics for the XGBoost model: (Accuracy: 0.7143, Precision: 1.0, Recall: 0.7143, F1-Score: 0.7857). To enhance our model's trustworthiness, we delve deeper into its workings using Shapley Additive Explanations. SHAP helps us break down the model's output by assessing the impact of each feature. This allows us to comprehend the significance of each feature, facilitating clear explanations to medical practitioners, aiding in their decision-making process, and ensuring consistent and reliable results. Our study identified critical biomarker ranges for meningitis diagnosis. For meningococcal Meningitis, Neutrophil levels $> 40\%$ and Lymphocyte levels $40-60\%$ were influential. Haemophilus influenzae meningitis was associated with Neutrophil predominance, specifically when reaching 80%, highlighting Neutrophils' significance. Positive Shapley values were associated with Neutrophil levels

approximately $< 60\%$ and Lymphocyte levels $\geq 60\%$ in Tuberculous Meningitis. Conversely, Aseptic Meningitis consistently had Shapley values below 0 when Neutrophil levels were around 40% and Lymphocyte levels were $< 50 - 60\%$.
Understanding the relative influence of these factors can help healthcare professionals improve diagnostic accuracy and optimize treatment strategies. Our model utilizes accurate and relevant attributes that closely resemble clinical standards. While most attributes align with these standards, we've also uncovered some disparities that highlight the need for further investigation to understand the diagnostic indicators of meningitis classes and reconcile the differences between our findings and the expected characteristics.

3.6 Future directions

Although our study benefits from a significant dataset and rigorous external validation, it must recognize several limitations. Notably, the absence of standardized evaluation metrics for interpretability poses challenges in quantifying and comparing interpretability across different models. Furthermore, our training set solely originates from a single country, focusing exclusively on the adult population. This limitation may restrict the generalizability of our models, as meningitis diagnosis and characteristics can vary significantly across diverse geographical regions and populations. To address this issue, we plan to expand the dataset collection globally, collaborate internationally for diverse data, and ensure adaptability to different healthcare systems and diagnostic practices. Furthermore, we aim to leverage external expertise through expert evaluation to validate the explanations provided by the model, ensuring alignment with domain-specific knowledge and expectations. Moreover, we plan to conduct validation studies across different populations to enhance model applicability on a broader spectrum of cases and diagnostic patterns and practices.

Chapter 4

Transparent AI Models for Meningococcal Meningitis Diagnosis: Evaluating Interpretability and Performance Metrics

*****

4.1 Introduction

Meningococcal meningitis is a severe bacterial infection that affects the meninges, caused by *Neisseria meningitidis* or *meningococcus*. *Neisseria meningitidis* accounts for approximately 11% of bacterial meningitis cases in individuals over 16 [179]. Criteria for suspicion of meningococcal meningitis include elevated serum leukocytes, relatively elevated granulocytes, and CSF leukocyte counts >800 cells/mm^3. CSF protein levels tend to be elevated, while glucose levels are reduced, and the presence of fever and meningeal signs such as sensitivity to light, nausea, vomiting, altered mental status, and a characteristic rash in some cases [180]. Although a positive result on a Gram stain for bacteria confirms the presence of bacterial meningitis, its sensitivity in accurately detecting the infection can range between 50% and 90% [179]. Without prompt antibiotic treatment, approximately 11% to 19% of cases result in severe complications and permanent sequelae, such as brain damage, amputation, hearing or vision loss, skin scarring, and neurological impairments [181]. Despite treatment, meningococcal infections have a high mortality rate (10%) and a strong epidemic potential [182].

Several studies have been proven effective in diagnosing different pathologies using ML and Deep Learning (DL) techniques. These techniques leverage advanced algorithms and com-

putational models to analyze complex datasets and extract meaningful patterns or features relevant to disease diagnosis. Nevertheless, ethical and legal considerations surround the use of AI in healthcare, particularly concerning issues related to the lack of transparency in decision-making processes and interpretability [183]. This challenges clinicians in understanding the process behind a diagnosis and trusting the results. The degree to which a physician can predict the outcome of a model or understand the reasons for its decisions is called interpretability. Local interpretability helps clinicians understand individual predictions, helps in final diagnosis decisions, and helps in initiating treatment. Global interpretability evaluates overall model behavior, enhancing trust and identifying biases. For this purpose, we assess ML models' local and global interpretability and explain how they perform the Meningococcal meningitis diagnosis.

The main contribution of this paper is as follows:

We propose an Explainable AI classification model for detecting Meningococcal meningitis among various types of meningitis. Our work assesses the model's reliability in distinguishing between cases. It calculates an importance value for each feature to reflect the extent to which the ML model relies on that particular feature using ELI5 [184] and LIME [29] techniques.

We conducted our experiments on the SINAN dataset for meningitis, and in consultation with infectious disease experts, we selected 34 features based on their informativeness and relevance to meningitis diagnosis. We compare the model transparency with clinicians' assessments, identifying key variables influencing the automatic diagnosis process. Our analysis reveals that explanations from ELI5 and LIME align closely with physicians' interpretations, particularly concerning bacterioscopy, culture results, and CSF biomarker levels in Meningococcal meningitis. Clinicians exhibit substantial trust and reliance on these AI-driven explanations, enhancing diagnostic precision and clinical decision-making. The rest of the chapter is organized as follows. Section 4.2 describes in detail the proposed method. Section 4.3 illustrates the experimental results. Section 4.4 summarizes the findings.

4.2 Methodology

Meningococcal meningitis is a severe bacterial infection predominantly caused by Neisseria meningitidis. This bacterium often resides harmlessly in the human nasopharynx but can, under certain conditions, invade the bloodstream and central nervous system, leading to either meningitis, septicemia, or both. Meningococcal meningitis can present as a fulminant condition where patients may rapidly deteriorate, displaying classic symptoms like fever, headache, neck stiffness, and altered mental status. However, these symptoms can vary significantly; for instance, only 30-40% of cases show all four classic symptoms, with additional symptoms such as nausea, vomiting, photophobia, and hypersensitivity to sound

commonly observed [96].

The illness can manifest differently across age groups, making diagnosis complex, particularly in the elderly where typical symptoms might be less pronounced [107]. In contrast, children and young adults often present more overt signs, including a distinctive petechial rash, which may indicate septicemia and severe systemic involvement, such as septic shock and multiorgan failure—a characteristic more common in meningococcal disease than other forms of bacterial meningitis [96]. Meningococcal meningitis exhibits different epidemiological patterns globally. While serogroups B and C are prevalent in developed regions like Europe and North America, serogroup A has historically dominated Africa's "Meningitis Belt," spanning from Nigeria to Somalia. Outbreaks have been documented periodically, often reaching epidemic levels, underscoring the disease's public health significance. Various meningococcal strains have exhibited resilience to antibiotic treatment over decades, with only minimal resistance, likely due to the bacteria's adaptability in gene exchange with other nasopharyngeal microbes [185].

Our AI approach aims to improve the diagnosis and management of meningococcal meningitis through an explainable AI framework designed to enhance transparency and support clinical decision-making. We conducted our experiments using the SINAN database, obtained from the Brazilian Government's Health Information System on Notifiable Diseases, which comprises data from 15,275 patients diagnosed with meningitis cases. Specifically, we focused on cases notified with Meningococcal meningitis (N = 460) to study how our model can predict the disease compared to other types of meningitis. To ensure unbiased learning, we balanced the dataset by randomly sampling an equal number of rows as 'Meningococcal meningitis' for the 'Other' category, including various meningitis types. This step is vital as imbalanced data can lead to difficulty modeling minority classes and subsequently lower model precision [186]. We incorporated 34 features, including demographic details, medical history, pre-existing conditions, and clinical signs, selected in consultation with infectious disease experts for their relevance to meningitis diagnosis. Table 4.1 provides a detailed description of the diagnostic variables used in the study, along with their respective values.

Table 4.1: Description of predictive indicators for the diagnosis

Diagnostic examination	**Features and values**
Demographic Informations	Age and Sex: Numerical, Male or Female: True or False
Medical History	AIDS/HIV, Tuberculosis, ARI, Trauma, Hospital Infection, Other Diseases/Illnesses: True or False
Vaccination History	Meningococcal C Conjugate, BCG, Triple Viral, Hemophilus Influenzae Type B (Hib), Pneumococcus: True or False
Symptoms and Signs	Headache, Fever, Vomiting, Seizures, Neck stiffness, Petechiae/haemorrhagic suffusion, Kernig/Brudzinski, Coma: True or False
CSF Chemocytological examination	Lymphocytes, CSF white blood cell (WBC), Neutrophils, Protein ratio CSF/serum, Glucose ratio CSF/serum: Numerical
CSF Aspects	CSF aspect: Clear, Purulent, Haemorrhagic, Cloudy, Xanthochromic, Other, Ignored
CSF Microbiological Examination	CSF Bacterioscopy: Gram negative diplococci, Gram positive diplococci, Gram positive bacilli, Gram negative bacilli, Gram positive cocci, Gram negative cocci, Gram negative diplobacilli, Cocobacilli, Other Bacteria, No agent, Unrealized, Ignored
Microbiological culture CSF and Blood	CSF culture and Blood/serum culture: Neisseria meningitidis, Haemophilus influenzae, Streptococcus pneumoniae, Other bacteria, No agent, Unrealized, Ignored
Latex agglutination tests	LATEX CSF and LATEX Blood/serum: Neisseria meningitidis, Haemophilus influenzae, Streptococcus pneumoniae, Streptococci(sp, pigeons, alpha, hemolytic, faecalis, agalactia), Cryptococci, Other Bacteria, No agent, Unrealized, Ignored
Polymerase Chain Reaction test	PCR Blood/serum: Neisseria meningitidis, Haemophilus influenzae, Streptococcus pneumoniae, Other viruses, No agent, Unrealized, Ignored, Not identified

We conducted data cleaning and typo elimination to ensure data consistency and accuracy, which could otherwise affect the analysis or interpretation. Collaboration with a medical team facilitated feature selection, ensuring the chosen features were relevant and meaningful for

the medical diagnosis task, specifically infectious diseases. One-hot encoding was then applied to categorical variables to prepare them for further analysis. This encoding allowed us to assess the impact of each category within categorical variables on the diagnostic outcome. In the current work, we compared eight machine learning models, including logistic regression, K-nearest neighbors, Support Vector Machine (SVM), decision tree, gradient boosting, AdaBoost, random forest, and LightGBM classifier. Boosting ensemble techniques, such as Gradient Boosting and AdaBoost, were particularly effective because they combined multiple weak learners into a strong predictive model. In these techniques, each subsequent classifier prioritizes instances previously misclassified by earlier ones, thereby enhancing classification accuracy [169]. Sensitivity (recall) metrics were prioritized to ensure the model's effectiveness in identifying true positive cases and minimizing the risk of undetected infections. False negatives, where the model incorrectly predicts a negative result for a positive case, can have severe consequences in medical diagnosis, potentially leading to delayed or missed treatments for patients with meningitis. AUROC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to correctly distinguish between patients with Meningococcal meningitis (positive cases) and those with other types of meningitis (negative cases). A higher AUROC value, closer to 1, indicates a better-performing model with higher sensitivity and specificity. The Receiver Operating Characteristic (ROC) curve is used to evaluate the performance of our classifiers. It offers a graphical representation of the trade-off between sensitivity (true positive rate (TPR)) represented on the y-axis and 1-specificity (false positive rate (FPR)) depicted on the x-axis across various threshold values for class assignment. Sensitivity measures the proportion of true positive predictions correctly identified by the model, while specificity measures the proportion of true negatives correctly identified [187].

Explainability aims to bridge the gap between the complex inner workings of machine learning models and human understanding. Gradient Boosting models are typically black-box models as they involve complex feature interactions. Various explainability techniques are employed to assess the transparency of our model. These techniques help clinicians understand not only what decision the model made but also why it made that decision. Global explainability methods offer insights into the overall behavior of the model across the dataset. Techniques such as feature importance scores or ELI5 (Explain Like I'm 5) provide a comprehensive overview of how various features contribute to the model's predictions on average. The ELI5 aids healthcare professionals in identifying the most influential factors guiding the model's diagnostic decisions. However, ELI5 is particularly effective for linear or tree-based models, such as Gradient Boosting, which made it valuable for our analysis [188]. The permutation importance technique is also applied to gain insights into the model's behavior across the dataset. It involves shuffling the values of individual features in the unseen dataset and observing the impact on the model's performance. It evaluates feature

importance based on how much the model's performance deteriorates when the feature values are randomly permuted [21]. Local explainability methods, on the other hand, provide insights into individual predictions. For instance, techniques like LIME (Local Interpretable Model-agnostic Explanations) generate locally faithful explanations by approximating the decision boundary around a specific instance. This model-agnostic approach is advantageous as it allows healthcare professionals to gain insights into the underlying reasons behind one particular diagnostic prediction for a given medical case, irrespective of the model's complexity or architecture [21].

4.3 Experiment and Results

This work was conducted on the Kaggle platform using Python's scikit-learn[1] library for machine learning model development and optimization. The model interpretation was performed using ELI5[2] and LIME[3] libraries to analyze feature importances and individual prediction explanations.

Exploratory data analysis serves as a preliminary step in understanding the characteristics of a dataset, revealing key features and potential relationships for further study. Figure 4.1 shows a balanced class distribution, ensuring fair predictions in all classes and enhancing the reliability of the model.

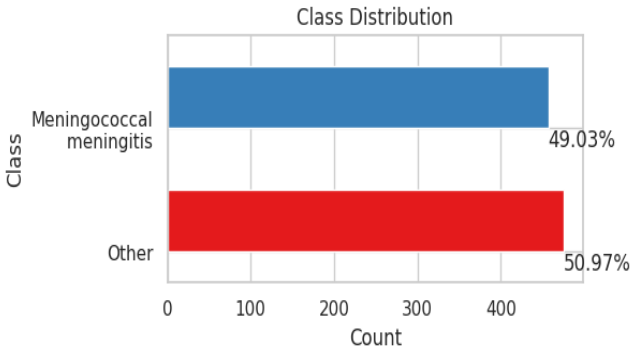


Figure 4.1: Bar plot displaying class distribution between Meningococcal meningitis and other types.

Figure 4.2 illustrates the relationships between numerical features extracted from cerebrospinal fluid samples in cases of Meningococcal meningitis and other types of meningitis. Kernel density estimates (KDE) are plotted in the lower triangle and marginal plots for numerical characteristics, including age, neutrophils, glucose ratio CSF / serum, CSF WBC, protein ratio CSF / serum and lymphocytes. These plots display the probability density of each numerical feature with Meningoccal meningitis and other types of meningitis, re-

[1]Scikit-learn library: https://scikit-learn.org/stable/
[2]ELI5 library: https://eli5.readthedocs.io/en/latest/overview.html
[3]LIME library: https://lime-ml.readthedocs.io/en/latest/

vealing insights into their distribution and spread. Peaks in density indicate regions with higher data concentration, while multiple peaks suggest the presence of distinct subpopulations. This observation is particularly evident in the ”other” category, which encompasses various meningitis types and immune responses. Pearson correlation coefficients are also displayed to indicate the strength of the relationships between these features. Higher neutrophil percentages align with lower lymphocyte percentages in meningococcal meningitis cases, suggesting an inverse relationship. The strong negative linear correlation of ($r = -0.68$) supports this observation. This helps understand the underlying data distribution and identify potential patterns or differences between Meningococcal meningitis cases and other types of meningitis.

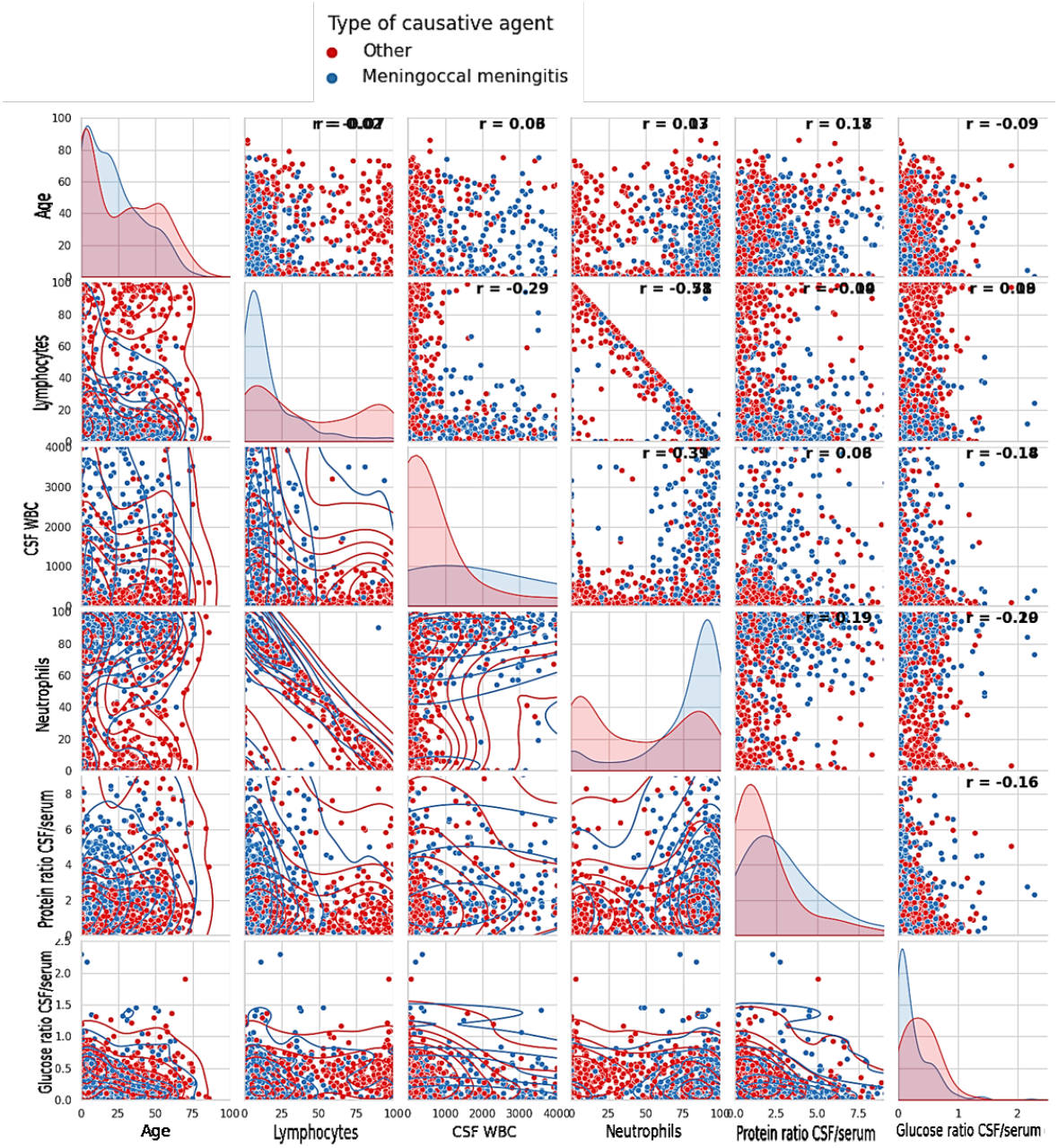


Figure 4.2: Pair plot showing numerical features from cerebrospinal fluid samples in Meningoccal meningitis vs. other types. Includes KDE plots and Pearson correlations, offering insights into data distribution and patterns.

This study assesses machine learning models' reliability, accuracy, and effectiveness in detecting meningococcal meningitis. Eight different machine learning models were trained on a subset of data comprising 70% of the dataset and tested using the remaining 30%. GridSearch cross-validation is utilized for hyperparameter tuning to identify the optimal model configuration. This technique involves applying cross-validation within the grid search process, utilizing 6 folds to assess the model's performance across various data subsets. Table 4.2 summarizes the performance metrics and hyperparameters of different classifiers used in this study. It provides information on hyperparameters, best parameters,

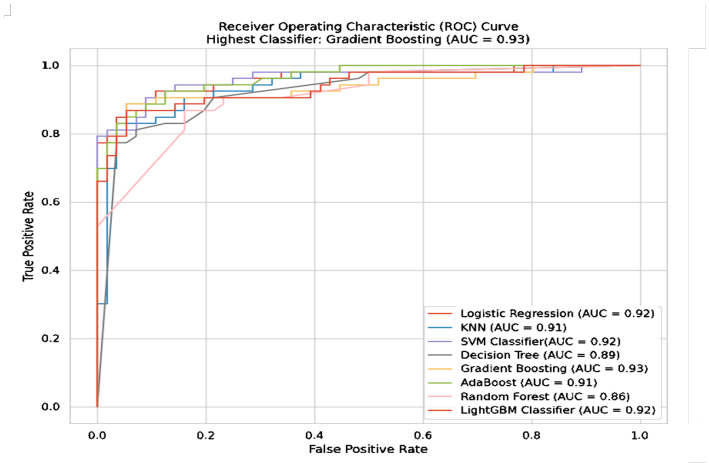


Figure 4.3: Receiver operating characteristic curves of different classifiers

and various evaluation metrics, including training and validation accuracy, precision, recall, AUROC, and F1 Score. The gradient Boosting model demonstrated superior performance in accurately identifying meningococcal meningitis cases compared to other classifiers (Recall = 0.826, AUROC = 0.930). Figure 4.3 displays Receiver Operating Characteristic curves for different classifiers used in this study, along with their corresponding Area Under the Curve (AUC) values to evaluate the performance of the classifiers. It plots the TPR against the FPR at various threshold settings.

Table 4.2: Performance metrics and hyperparameters for different classifiers

Classifier	**Hyperparameters**	**Best Parameters**	**Training Accuracy**	**Training Precision**	**Training Recall**	**Training AUROC**	**Training F1 Score**	**Validation Accuracy**	**Validation Precision**	**Validation Recall**	**Validation AUROC**	**Validation F1 Score**
Logistic Regression	C: [0.001, 0.01, 0.1, 1, 10, 100] penalty: ['l1', 'l2', 'elasticnet']	C: 1 penalty: 'l2'	0.904	0.936	0.863	0.968	0.898	**0.872**	**0.918**	**0.812**	**0.924**	**0.862**
KNN	n_neighbors: [5, 10, 15, 20] weights: ['uniform', 'distance'] p: [1, 2]	n_neighbors: 20 p: 1 weights: 'distance'	1.000	1.000	1.000	1.000	1.000	0.823	0.828	0.804	0.907	0.816
SVC	C: [0.001, 0.01, 0.1, 1, 10] kernel: ['linear', 'rbf'] gamma: ['scale', 'auto']	C: 0.1 gamma: 'auto' kernel: 'rbf'	0.912	0.965	0.851	0.970	0.904	0.872	0.932	0.797	0.925	0.859
Decision Tree	criterion: ['gini', 'entropy'] max depth: [None, 4, 5, 10, 15] min samples split: [2, 5, 10, 15] min samples leaf: [1, 2, 4]	criterion: 'gini' max depth: 10 min samples leaf: 2 min samples split: 15	0.930	0.957	0.898	0.982	0.926	0.826	0.845	0.790	0.888	0.816
Gradient Boosting	number of estimators: [50, 100, 120] learning rate: [0.01, 0.1, 0.2] max depth: [3, 4, 5] subsample: [0.8, 0.9, 1.0]	learning rate: 0.1 max depth: 4 number of estimators: 120 subsample: 0.8	0.996	0.996	0.996	0.996	0.996	**0.879**	**0.919**	**0.826**	**0.930**	**0.870**
AdaBoost	number of estimators: [50, 100] learning rate: [0.01, 0.1, 0.2, 0.5, 1] base_estimator max depth: [1, 2, 3, 4]	base_estimator max depth: 3 learning rate: 0.01 number of estimators: 80	0.904	0.939	0.860	0.976	0.898	0.840	0.855	0.812	0.911	0.833
Random Forest	number of estimators: [50, 80, 100] max depth: [None, 5] min samples split: [5, 10] min samples leaf: [2, 4] bootstrap: [True, False] max features: ['sqrt', 'log2'] min impurity decrease: [0.0, 0.1, 0.2]	bootstrap: True max depth: 3 max features: 'sqrt' min impurity decrease: 0.1 min samples leaf: 2 min samples split : 5 number of estimators: 120	0.788	0.798	0.761	0.882	0.779	0.791	0.816	0.739	0.855	0.776
LightGBM	C: [0.001, 0.01, 0.1, 1, 10] kernel: ['linear', 'rbf'] gamma: ['scale', 'auto']	learning rate: 0.1 max depth: 10 number of estimators: 100 subsample: 0.8	1.00	1.00	1.00	1.00	1.00	0.865	0.903	0.812	0.919	0.855

To enhance the understanding of the Gradient Boosting model's output, explainability techniques such as Eli5 and LIME are employed. These techniques provide insights into the

model's predictions, facilitating model debugging, validation, and assessment of trustworthiness. Figure 4.4 showcases permutation importance analysis using ELI5, explaining the gradient boosting model's output globally. It presents feature weights, highlighting the most influential features with the highest weights. These values indicate each feature's average impact on predictions, with higher weights signifying greater importance as changes in these features lead to larger decreases in model performance. Identifying *Neisseria meningitidis* and gram-negative diplococci through culture and smear examination of CSF has the most significant impact on the model's ability to predict Meningococcal meningitis, alongside the detection of *Neisseria meningitidis* bacteria in the latex agglutination test of the CSF. Conversely, detecting other bacteria like Streptococcus pneumoniae or Haemophilus influenzae in CSF cultures is relevant for distinguishing Meningococcal meningitis from other types of meningitis. White cell counts in CSF and age also provide valuable insights to our model. These findings align with real-world observations, as *Neisseria meningitidis* typically presents as gram-negative diplococci under microscopy [189].

Weight	Feature
0.0921 ± 0.0153	CSF Culture-Neisseria meningitidis
0.0857 ± 0.0136	CSF Bacterioscopy-Gram negative diplococci
0.0314 ± 0.0354	CSF WBC
0.0250 ± 0.0235	LATEX CSF-Neisseria meningitidis
0.0157 ± 0.0215	Age
0.0136 ± 0.0083	CSF Culture-Streptococcus pneumoniae
0.0114 ± 0.0029	PCR Blood/serum-No agent
0.0100 ± 0.0029	LATEX CSF-Streptococcus pneumoniae
0.0093 ± 0.0073	CSF Culture-Haemophilus influenzae
0.0071 ± 0.0120	Lymphocytes
0.0064 ± 0.0053	Culture Blood/serum-No agent
0.0057 ± 0.0035	CSF Culture-Unrealized
0.0043 ± 0.0029	CSF Bacterioscopy-Gram negative cocci
0.0043 ± 0.0029	CSF aspect-Ignored
0.0043 ± 0.0029	CSF Culture-Other bacteria
0.0043 ± 0.0029	CSF aspect-Clear
0.0043 ± 0.0114	CSF Bacterioscopy-No agent
0.0029 ± 0.0029	Pre-Existing Diseases -AIDS/HIV-Ignored
0.0021 ± 0.0035	Headache
0.0021 ± 0.0035	Fever
0.0021 ± 0.0035	LATEX CSF-Unrealized
0.0014 ± 0.0035	Petechiae/haemorrhagic suffusion
0.0014 ± 0.0073	LATEX CSF-Haemophilus influenzae
0.0007 ± 0.0029	CSF aspect-Haemorrhagic
0.0007 ± 0.0029	CSF Bacterioscopy-Gram positive cocci
	... 75 more ...

Figure 4.4: Permutation Importance Analysis: Importance of Features in Gradient Boosting Model Predictions using ELI5

Further analysis evaluates the role of non-laboratory features in diagnosing meningococcal meningitis, aiming to determine their diagnostic reliability. Figure 4.5 illustrates the results of permutation importance analysis using the ELI5 library, excluding microbiological bacterioscopy examinations, CSF and blood culture, latex agglutination and PCR tests. The CSF Chemocytological examination tests are identified as paramount features in detecting meningococcal meningitis, along with the potential impact of vaccination status and certain symptoms like neck stiffness, petechiae/hemorrhagic suffusion, vomiting, and Kernig/Brudzinski signs. Conversely, features related to CSF aspect (Xanthochromic, Purulent) and pre-existing illnesses (AIDS/HIV, Trauma, Tuberculosis) exhibit lower importance scores but remain informative for detecting the meningococcal meningitis.

While global explanation offers insights into the overall model behavior based on feature

Weight	Feature
0.0607 ± 0.0186	CSF WBC
0.0200 ± 0.0167	Glucose ratio CSF/serum
0.0171 ± 0.0165	Neutrophils
0.0164 ± 0.0132	Protein ratio CSF/serum
0.0100 ± 0.0095	Lymphocytes
0.0057 ± 0.0132	Neck stiffness
0.0057 ± 0.0097	Petechiae/haemorrhagic suffusion
0.0057 ± 0.0116	Vaccination -BCG
0.0050 ± 0.0073	Vaccination -Pneumococcus
0.0036 ± 0.0090	CSF aspect-Xanthochromic
0.0029 ± 0.0053	Vomiting
0.0029 ± 0.0029	Kernig/Brudzinski
0.0029 ± 0.0029	Pre-Existing Diseases -AIDS/HIV
0.0021 ± 0.0035	Vaccination -Hemophile –Tetravalent or Hib
0.0007 ± 0.0029	CSF aspect-Purulent
0.0007 ± 0.0029	Pre-Existing Illnesses -Trauma
0 ± 0.0000	Pre-Existing Illnesses -ARI
0 ± 0.0000	Pre-Existing Illnesses -Hospital Infection
0.0000 ± 0.0045	Pre-Existing Diseases -Tuberculosis
0 ± 0.0000	CSF aspect-Other
0 ± 0.0000	Pre-Existing Illnesses -Other
0 ± 0.0000	Vaccination -Meningococcal C Conjugate
0 ± 0.0000	Vaccination -Triple Viral
0 ± 0.0000	Headache
0 ± 0.0000	Seizures
0 ± 0.0000	Coma
0 ± 0.0000	CSF aspect-Clear
0 ± 0.0000	CSF aspect-Haemorrhagic
0 ± 0.0000	CSF aspect-Cloudy
-0.0014 ± 0.0057	Fever
-0.0029 ± 0.0177	Age

Figure 4.5: Permutation Importance Analysis: Permutation Importance Analysis Results Excluding Non-Laboratory Tests using ELI5

importance, local explanation in the other hand, delves into individual predictions, showcasing the specific contributions of features for two instances—one categorized as Meningococcal meningitis (Fig. 4.6b) and the other as 'Other' type (Fig. 4.6a). The model predicted the specific instance as Meningococcal meningitis with a likelihood of 0.986. Important features contributing to this prediction include the presence of *Neisseria meningitidis* and gram-negative diplococci in the CSF culture and smear examination. Additionally, the absence of other bacteria, such as Streptococcus pneumoniae or Haemophilus influenzae, along with low lymphocyte percentage (13%) and Glucose ratio in CSF (0.010) and high levels of Neutrophils (71%) and WBC (10000 cells/mm^3) significantly influenced the model's decision, indicating typical characteristics of Meningococcal meningitis.

y=**Other** (probability **0.974**, score **-3.633**) top features

Contribution[?]	Feature	Value
+0.483	PCR Blood/serum-No agent	0.000
+0.464	Lymphocytes	90.000
+0.452	Protein ratio CSF/serum	0.700
+0.393	CSF WBC	3.000
+0.334	CSF Bacterioscopy-Gram negative diplobacilli	0.000
+0.276	Neutrophils	0.000
+0.272	Culture Blood/serum-No agent	0.000
+0.176	PCR Blood/serum-Neisseria meningitidis	0.000
+0.170	Petechiae/haemorrhagic suffusion	1.000
+0.154	CSF Culture-Ignored	0.000
+0.126	CSF Bacterioscopy-Gram negative diplococci	0.000
+0.105	CSF Bacterioscopy-Gram positive bacilli	0.000
+0.070	LATEX Blood/serum-No agent	0.000
+0.068	CSF Culture-Neisseria meningitidis	0.000
+0.067	CSF aspect-Cloudy	0.000
+0.062	LATEX CSF-Neisseria meningitidis	0.000
+0.051	LATEX CSF-Unrealized	1.000
	... 18 more positive ...	
	... 24 more negative ...	
-0.053	Glucose ratio CSF/serum	0.600
-0.073	Age	12.000
-0.124	PCR Blood/serum-Unrealized	0.000

(a) ELI5 explanation for an instance classified as 'Other' type

y=**Meningococcal meningitis** (probability **0.986**, score **4.277**) top features

Contribution[?]	Feature	Value
+0.610	CSF Bacterioscopy-Gram negative diplococci	1.000
+0.503	Lymphocytes	13.000
+0.391	PCR Blood/serum-Haemophilus influenzae	0.000
+0.347	CSF Bacterioscopy-No agent	0.000
+0.343	Glucose ratio CSF/serum	0.010
+0.321	CSF Culture-Neisseria meningitidis	1.000
+0.303	CSF Culture-Haemophilus influenzae	0.000
+0.248	CSF WBC	10000.000
+0.197	CSF Culture-Streptococcus pneumoniae	0.000
+0.171	Age	60.000
+0.162	Neutrophils	71.000
+0.120	LATEX CSF-No agent	0.000
+0.115	Petechiae/haemorrhagic suffusion-Ignored	0.000
+0.096	Sex_Female	1.000
+0.093	LATEX CSF-Haemophilus influenzae	0.000
+0.082	LATEX CSF-Unrealized	0.000
+0.081	LATEX CSF-Ignored	0.000
+0.074	CSF aspect-Cloudy	1.000
	... 31 more positive ...	
	... 14 more negative ...	
-0.094	Protein ratio CSF/serum	0.200
-0.128	CSF Bacterioscopy-Gram negative cocci	0.000

(b) ELI5 explanation for an instance classified as Meningococcal meningitis

Figure 4.6: Comparison of ELI5 explanations for two instances

Similarly, the likelihood of predicting the instance as having another type of meningitis is 0.974. The top features include PCR Blood/serum indicating identification of an infectious agent (negative value), along with a high level of Lymphocytes (90%), high CSF Protein ratio (0.7 g/L, which is above the normal range [0.15 g/l - 0.4 g/l] [190], low CSF WBC

(3 cells/mm^3), low neutrophils (0%), absence of *Neisseria meningitidis* in the PCR test, and the presence of Petechiae/hemorrhagic suffusions strongly support the classification as another type of meningitis. The presence of petechiae and hemorrhagic suffusions, along with the absence of *Neisseria meningitidis* in the PCR test, could indicate other conditions or infections that share similar clinical presentations with meningococcal meningitis. Therefore, the model's prediction may reflect the presence of these symptoms alongside other clinical or laboratory findings that support a diagnosis of non-meningococcal meningitis.

We extended our analysis by employing LIME explanations to interpret individual instances. Figure 4.7 and 4.8 display the individual feature contributions for the same instances as in the ELI5 comparison.

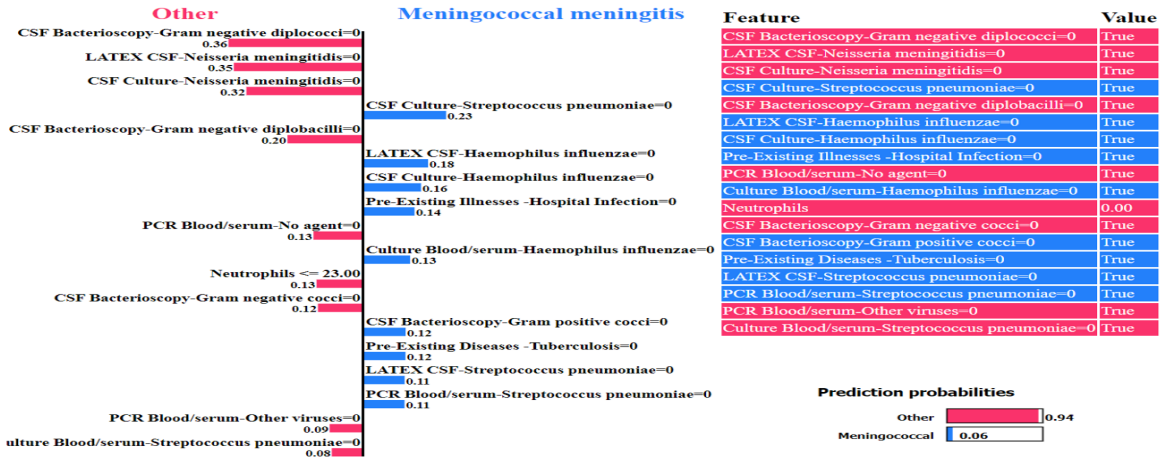


Figure 4.7: LIME explanation for an instance classified as 'Other' type

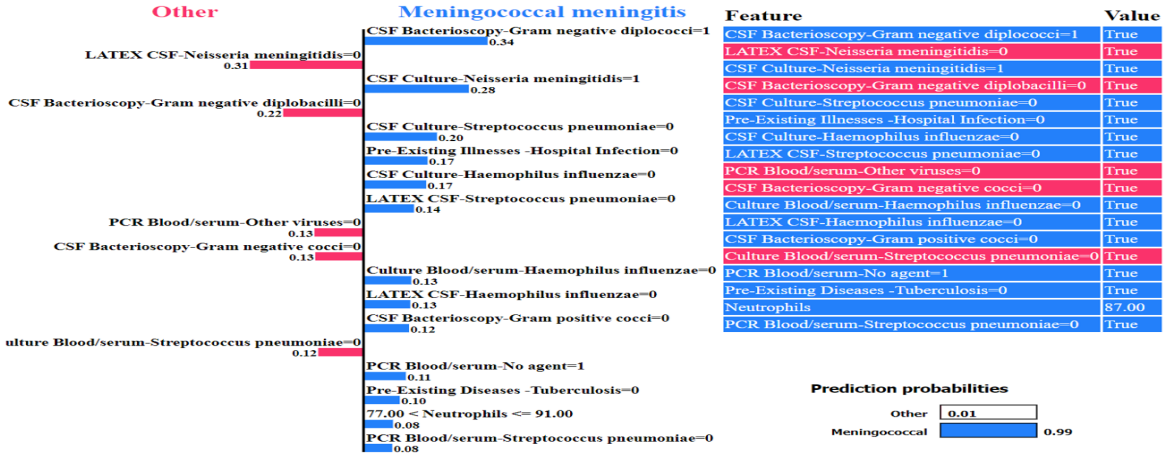


Figure 4.8: LIME explanation for an instance classified as Meningococcal meningitis

The model predicted the instance as diagnosed with Meningococcal meningitis (Fig. 4.8) with a high confidence probability of 0.99. Similar to ELI5 explanations, the detection of *Neisseria meningitidis* and gram-negative diplococci (value 1 indicates a positive test

result) through culture and smear examination of CSF, and the absence of other bacterial agents such as Streptococcus pneumoniae or Haemophilus influenzae in the CSF culture and LATEX agglutination test, along with a high percentage of Neutrophils (77 < Neutrophils ≤ 91), are strongly associated with Meningococcal Meningitis causes and have a significant positive impact. The model predicted the instance as being diagnosed with another type of meningitis with a probability of 0.94 (Figure 4.7). The top feature contributing the most to the discrimination of meningococcal meningitis is the absence of the causative pathogen of this disease, which includes negative tests for the presence of *Neisseria meningitidis* and gram-negative diplococci. Additionally, a PCR Blood/serum indicating the identification of an infectious agent and low Neutrophil levels (<= 23%) suggest that a different pathogen may cause the infection.

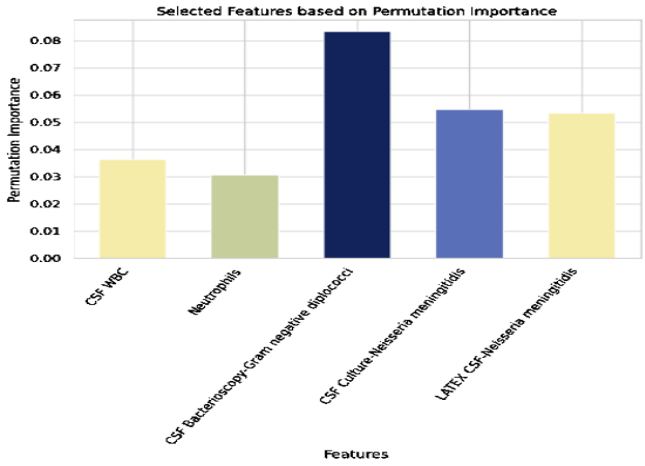


Figure 4.9: Permutation importance plot showing key predictors of Meningococcal Meningitis.

The permutation importance figure 4.9 reveals key features contributing significantly to the model's predictive performance. CSF WBC, neutrophil levels, detection of *Neisseria meningitidis* by CSF culture and latex agglutination test, and gram-negative diplococci detected by CSF bacterioscopy stand out as the most influential features. Among these, the detection of gram-negative diplococci appears to have the most significant impact on the model's predictions. This observation is consistent with findings from both ELI5 and LIME explanations. Understanding the prominent role of these features enhances our comprehension of the model's behavior and decision-making process.

4.4 Discussion and conclusion

The primary objective of this study is to develop a reliable AI model capable of accurately identifying cases of Meningococcal meningitis among various types of meningitis. Our results indicate that the Gradient Boosting model exhibits promising potential in distinguishing Meningococcal cases, achieving notable performance metrics: accuracy (0.88), precision

(0.92), recall (0.83), AUROC (0.93), and f1-score (0.87). The model's success in identifying Meningococcal meningitis is attributed to several key factors. The presence of the causative pathogen, *Neisseria meningitidis*, identified through CSF culture and latex agglutination, along with the detection of gram-negative diplococci in smear examination of CSF, emerged as significant contributors to the model's predictive capability. These pathogens are known to be responsible for Meningococcal meningitis, thereby validating the model's focus on this specific etiology [189]. Additionally, white cell counts in CSF and patient age further influence the model's performance. However, the reliance on invasive laboratory tests is evident in diagnosing meningococcal meningitis. By excluding these tests and focusing on clinical indicators, the significance of features such as CSF Chemocytological examination tests, particularly WBC count, vaccination status, and specific symptoms like neck stiffness and petechiae, are effective in identifying cases of meningococcal meningitis within the model. Local explanation highlights a high percentage of neutrophils in CSF as a characteristic feature of Meningococcal meningitis. Moreover, LIME analysis revealed that low lymphocyte percentages and elevated white blood cell counts are associated with accurate predictions of Meningococcal meningitis, providing further insight into the disease's diagnostic biomarkers. The consistency between our findings and prior research underscores the importance of the Neutrophil-to-Lymphocyte Ratio as a predictor in distinguishing between viral and bacterial meningitis within machine learning algorithms [46]. Our study targets explicitly Meningococcal meningitis, providing a focused analysis compared to broader studies encompassing various meningitis etiologies. This underscores the need for trustworthy AI models in clinical decision-making. As AI's role in healthcare grows, ensuring model reliability, transparency, and interpretability becomes increasingly crucial. Further research could explore additional biomarkers to enhance the model's accuracy and robustness.

Chapter 5

Does AI model resonate like a medical expert?: A novel concept-based model explanations for Meningitis diagnosis

*****

5.1 Introduction

Pneumococcal meningitis, majorly caused by *Streptococcus pneumoniae*, is a leading cause of acute bacterial meningitis in adults, responsible for 75–80% of all cases globally [191,192]. With a high mortality rate of up to 15% overall, and 33% among critically ill patients [193, 194], the case fatality rate varies by region, ranging from 17–30% globally [195], 20–37% in high-income countries, and up to 51% in low-income countries [196]. Additionally, 38% of survivors experience neurological aftereffects, such as mental retardation, seizures, hearing loss, and cerebral palsy in children, while older individuals may develop hydrocephalus, cerebellar dysfunction, and paresis. These neurological complications are primarily driven by excessive inflammation, and a CSF glucose level below 25 mg/dL is linked to a worse prognosis in pneumococcal meningitis [68].

Despite medical advancements, diagnosing pneumococcal meningitis remains challenging due to its symptom overlap with other types of meningitis. This study uses deep learning models for diagnosing pneumococcal meningitis, integrating biological tests, clinical symptoms, and patient medical history through the Testing Concept Activation Vectors (TCAV) method. Deep learning models have shown promise in disease diagnosis by processing large datasets with high accuracy. However, they often operate as ”black boxes,” making it difficult to understand the factors influencing their predictions [39, 197]. This lack of interpretability raises concerns in clinical settings, where transparency is crucial for gaining healthcare pro-

fessionals' trust [198]. To address this, we propose Bio-TCAV (Biological Testing Concept Activation Vectors) , a concept-based model explanation tailored for infectious disease diagnosis. Bio-TCAV interprets deep learning predictions by linking them to clinically relevant concepts, enhancing diagnostic transparency. Our contribution is summarized as follows:

- Unlike many explanation tools (e.g., SHAP, LIME) based on approximations, it may not fully capture the model's decision-making process. We collaborate with domain experts to introduce an explanation method that evaluates whether the model's explanations align with clinical knowledge. The proposed methodology defines medical concept vectors based on the clinician's knowledge to enhance clinical interpretability, as medical practitioners often need feature-level explanations to trust and act on model outputs.

- We defined concept vectors by analyzing the core indicators for the diagnostic assessment such as demographics, medical history, vaccination records, clinical symptoms, CSF chemo-cytological analysis (analyzed in Section 2.5), CSF appearance, CSF bacterioscopy, cultures, latex agglutination, and PCR testing are considered. This approach allows the concept to become more nuanced, reflecting real-world meningitis presentations.

- We introduced Bio-TCAV, which interprets the model's decision-making process by evaluating the influence of high-level medical concepts, such as biomarkers, clinical symptoms, and patient history, on the model's predictions. The method involves extracting activations from the deep learning model and training concept classifiers for each medical concept. By comparing the activation of these concepts with random baseline concepts, Bio-TCAV generates concept importance scores. Welch's t-test is then applied to confirm the statistical significance of these differences, ensuring that the identified concepts are critical in distinguishing between pneumococcal meningitis and control cases.

- Our experiments, which leverage both the SINAN database and real-world data from Setif's hospital in Algeria, show the model's diagnostic performance while providing infectious disease experts with transparent insights into the factors driving the diagnosis of pneumococcal meningitis.

5.2 The proposed methodology

While highly effective, deep neural networks are often criticized for their complexity and lack of interpretability. To address this, our proposed methodology for diagnosing pneumococcal meningitis cases is outlined in Figure 5.1.

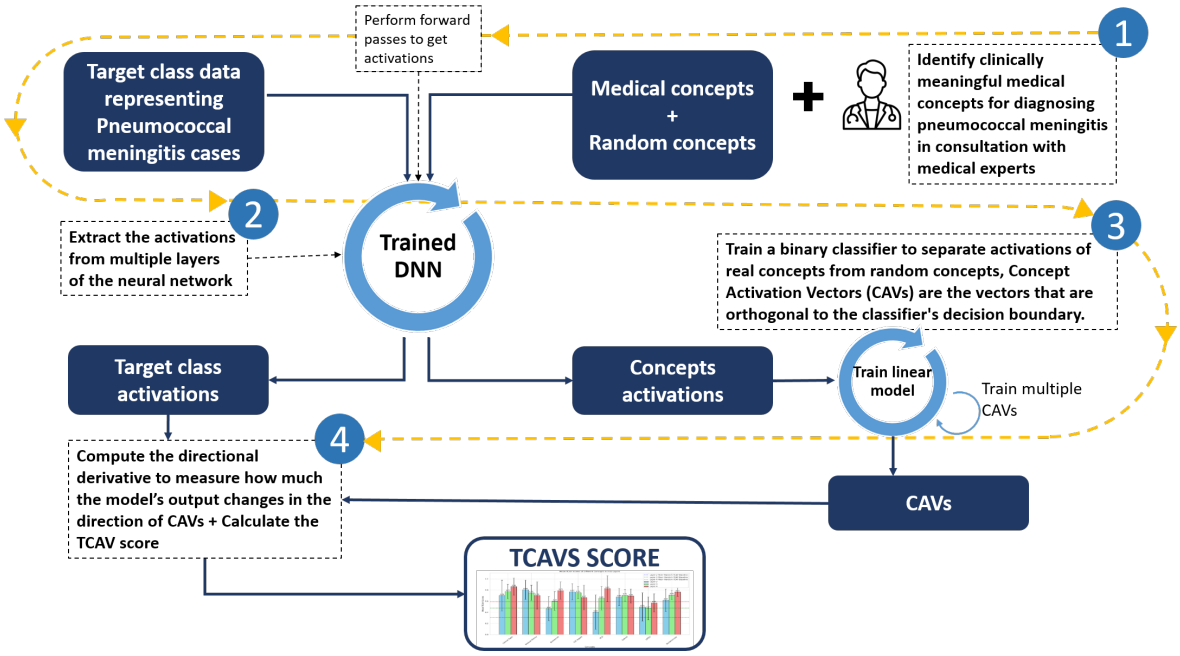


Figure 5.1: Flowchart showcasing the methodology for pneumococcal meningitis diagnosis, employing Bio-TCAV to elucidate model understanding.

5.2.1 Domain knowledge-based feature selection

Feature selection for this study was conducted in collaboration with infectious disease experts at Setif Hospital, Algeria. Leveraging their clinical expertise and practical experience in meningitis diagnosis, we identified a targeted set of variables with high diagnostic relevance for distinguishing pneumococcal meningitis from other types. Our previous study validated the selected features through clinical practice and statistical evaluation (Section 2.5). Table 5.1 summarizes the selected features of the dataset.

5.2.2 Features engineering and concept definition

In the following phase, we defined clinically meaningful medical concepts based on the diagnostic approaches applied by infectious disease practitioners. In collaboration with infectious disease practitioners, we defined clinically meaningful medical concepts based on diagnostic criteria and pathways used to differentiate meningitis types. A prior study analyzed cerebrospinal fluid biomarkers, including glucose and protein levels, white blood cell count, and neutrophil/lymphocyte percentages (Section 2.5), revealing distinct biological patterns associated with underlying pathology and patient age. Discussions with medical experts contextualized these findings within diagnostic workflows, highlighting their practical relevance. Building on this, we integrated additional diagnostic inputs, such as patient history, clinical signs, and pathogen detection methods such as Gram staining, culture, and PCR, to refine the concepts for enhanced specificity in diagnosing pneumococcal meningitis and

Table 5.1: Predictive indicators for diagnostic assessment

Examination Type	**Features**
Demographics	Age, Sex
Medical History	Presence of conditions like AIDS/HIV, Tuberculosis, Acute Respiratory Infection (ARI), Trauma, Hospital Infections, and Other Illnesses
Vaccination Records	Indicate if the the patient has received specific vaccinations for Meningococcal C Conjugate, BCG, Triple Viral, Hemophile Tetravalent or Hib, Pneumococcus
Clinical Symptoms	Informs if this clinical manifestation has occurred: Headache, Fever, Vomiting, Seizures, Neck Stiffness, Petechiae/Hemorrhagic Suffusion, Kernig/Brudzinski Signs, Coma
CSF Chemocytological Analysis	Includes CSF White Blood Cells (WBC), Neutrophils and Lymphocytes levels, Glucose and Protein ratio
CSF Appearance	Appearance of the CSF obtained through the puncture: Clear, Purulent, Haemorrhagic, Cloudy, Xanthochromic, Other, and Ignored
CSF Bacterioscopy	Etiological agents identified in the examination: Gram-negative bacilli, Gram-positive bacilli, Coccobacilli, Gram-negative cocci, Gram-positive cocci, Gram-negative diplobacilli, Gram-negative diplococci, Gram-positive diplococci, Other Bacteria, including No Agent, Unrealized, or Ignored
Cultures (CSF & Blood)	The etiological agent identified in the examination: Neisseria meningitidis, Haemophilus influenzae, Streptococcus pneumoniae, and Other Bacteria (categorized as Agent Present), No Agent, Unrealized, or Ignored
Latex Agglutination	LATEX Tests for CSF and Blood/Serum: Detecting Neisseria meningitidis, Haemophilus influenzae, Streptococcus pneumoniae, various Streptococci species, Cryptococci, Other bacteria, No Agent, Unrealized, or Ignored
PCR Testing	PCR Analysis of Blood/Serum: Mumps, Measles, Herpes Simplex, Chickenpox/Herpes Zoster, Rubella, Influenza, Dengue, Adenovirus, Echovirus, Coxsackie Virus, Other Enteroviruses, West Nile Virus, Other Arboviruses, Other viruses, Neisseria meningitidis, H. influenzae, Streptococcus pneumoniae, Other bacteria, Cryptococcus/Torula, Other fungi, Toxoplasma, Other Parasites, with possible outcomes as No Agent, Unrealized, Ignored, or Not Identified

distinguishing it from other etiologies. Table 5.2 summarizes the conditions defining each concept in the dataset.

Table 5.2: Defined medical concepts related to Pneumococcal meningitis condition

Defined Concept	**Medical condition**
Biomarkers Concept	Neutrophils > 50% AND Lymphocytes < 50% AND CSF WBC > 500 cells/mm^3 AND Glucose ratio < 0.4 AND (1 < Protein ratio < 4)
Clinical Signs Concept	Seizures OR Coma reported
Medical History Concept	Prior Trauma-Related Health Issues OR Pneumococcus vaccine status = not received
CSF Aspect Concept	CSF aspect = Purulent OR Cloudy
PCR Concept	Positive PCR for Streptococcus pneumoniae in Blood/Serum
Culture Concept	Positive Culture for Streptococcus pneumoniae in Blood/Serum OR CSF
LATEX Concept	LATEX Test Positive for Streptococcus pneumoniae in Blood/Serum OR CSF
Bacterioscopy Concept	CSF Bacterioscopy Results Indicate Gram Positive Cocci

5.2.3 Model implementation

We developed a deep neural network for binary classification to diagnose pneumococcal meningitis. The model architecture consists of a feedforward structure with four hidden layers containing 512, 256, 128, and 64 densely connected neurons. To introduce non-linearity and mitigate the risk of overfitting, Leaky Rectified Linear Unit (Leaky ReLU) activation functions were employed, which allow a small gradient in the negative section instead of being completely zero [199]. Batch normalization was applied to enhance generalization and stabilize the learning process.

The output layer consists of a single neuron with a sigmoid activation function, yielding a probability score between 0 and 1 to indicate the likelihood of the input belonging to the pneumococcal meningitis class. The model leverages the interplay of its hidden layers to perform complex, non-linear transformations on high-dimensional data. However, the model leverages the sequential processing of its hidden layers, combining linear operations, weights, biases, and non-linear activations to transform high-dimensional input data into a feature space suitable for binary classification. These transformations capture complex patterns and interactions within the data that traditional methods often miss. Yet, this complexity results in a 'black-box' model where the relationship between inputs and predictions remains opaque.

This lack of transparency challenges clinical applicability, as medical professionals require clear, feature-level explanations to trust and act on predictions [200]. Understanding which features and how contribute to the classification outcome can help bridge the gap between computational efficacy and clinical applicability, ensuring that decisions made with the assistance of the model align with medical reasoning and standards. To address interpretability challenges in deep learning, we propose the Bio-TCAV explanation model tailored for pneumococcal meningitis diagnosis.

5.2.4 Bio-TCAV explanation approach for diagnosis

To improve the interpretability of our deep neural network (DNN) model, we implemented the Bio-TCAV method, which incorporates domain-specific knowledge to align model behavior with clinically relevant concepts defined in Table 5.2. This approach allows us to link model predictions with medically meaningful features, enhancing transparency and understanding of the decision-making process. The detailed steps of this method are described in Algorithm 1.

5.2.4.1 Activation extraction

For each predefined concept, we computed activations from various layers of the DNN. This process involved forwarding input samples associated with each concept through the network and capturing the output activations from neurons in the specified layers. These activations

represent the model's learned encoding of the input data, reflecting how features associated with each concept are processed and represented at different hierarchical levels of the network. By examining these activations, we gain insight into how the DNN processes and identifies concept-specific patterns.

5.2.4.2 Concept classifier training

To differentiate activations associated with relevant medical concepts from unrelated or spurious patterns, we trained logistic regression models as concept classifiers. These classifiers were trained using two sets of activations:

- **Positive Set:** Activations from input samples corresponding to a predefined medical concept (Table 5.2.).
- **Negative Set:** Activations from input samples associated with random concepts designed to be unrelated to pneumococcal meningitis.

The random concepts represent noise or non-specific patterns and serve as a baseline for comparison. The logistic regression model learns a hyperplane in the activation space that separates the positive and negative sets. The normal vector to this hyperplane is defined as the Concept Activation Vector (CAV), which points toward the direction of activations corresponding to the predefined medical concept.

5.2.4.3 Concept Activation Vectors (CAVs)

CAVs quantitatively capture the influence of medical concepts on the model's decision-making process by identifying the direction in which concept-specific activations deviate from random noise. By analyzing these CAVs, we assess the sensitivity of the model to each concept, enabling an understanding of the network's internal mechanisms in terms of medically significant features.

5.2.4.4 Reliability and statistical significance

To ensure robust and reliable results, we performed 39 independent training runs for each CAV. This iterative approach minimizes variability and provides averaged results that are less sensitive to random fluctuations [198]. To validate the significance of the resulting TCAV scores, we applied Welch's t-test, a statistical method suited for comparing groups with unequal variances and sample sizes [132]. A p-value threshold of < 0.01 was used to ensure statistical significance and minimize the likelihood of Type I errors [31]. This rigorous statistical testing confirmed that the observed effects were unlikely to result from random chance, thereby strengthening the reliability of our findings.

Algorithm 1: Bio-TCAV pseudo algorithm

Input: D: Full dataset for all classes
$C = \{C_1, C_2, \ldots, C_k\}$: Set of medical/biological concepts, where $C_i \subseteq D$
$R = \{R_1, R_2, \ldots, R_n\}$: Set of random concepts, where $R_k \subseteq D$
M: Pre-trained AI model
D_t: Dataset for the studied class t
Output: TCAV Scores for each concept C_i
Step 1: Extract Layer Activations for the Studied Class and Concepts
foreach *layer* $L_j \in M$ **do**
  $A_{t,j} \leftarrow \texttt{Activations}(D_t, L_j)$ // Extract activations for the studied class dataset D_t from L_j;
  foreach *concept* $C_i \in C$ **do**
    $A_{C_i,j} \leftarrow \texttt{Activations}(C_i, L_j)$ // Extract activations for subset C_i from L_j;
  end
  foreach *random concept* $R_k \in R$ **do**
    $A_{R_k,j} \leftarrow \texttt{Activations}(R_k, L_j)$ // Extract activations for subset R_k from L_j;
  end
end
Step 2: Train Concept Activation Vectors (CAVs)
foreach *layer* $L_j \in M$ **do**
  foreach *concept* $C_i \in C$ *and* **do**
    // Perform N runs of logistic regression to train CAVs:
    for $n = 1$ *to* N **do**
      $A_{R_k,j} \leftarrow$ Select random activations from pre-extracted random activations list from layerL_j;
      $CAV^{(n)}_{C_i,j} \leftarrow \texttt{BinaryClassifier}_n(A_{C_i,j}, A_{R_k,j})$;
      // $A_{R_k,j}$ is from pre-extracted random activations. All trained CAVs are stored: $\{CAV^{(1)}_{C_i,j}, CAV^{(2)}_{C_i,j}, \ldots, CAV^{(N)}_{C_i,j}\}$;
    end
  end
end

Algorithm 1: Bio-TCAV Pseudo algorithm (Continued)

Step 3: Compute TCAV Scores

foreach *layer* $L_j \in M$ **do**

Compute the gradient of the model's output logit f_t for the target class t of each layer L_j:

$$\nabla_{A_{t,j}} f_t = \frac{\partial f_t}{\partial A_{t,j}}$$

foreach *concept* $C_i \in C$ **do**

for $n = 1$ *to* N **do**

Compute the directional derivative along each $CAV^{(n)}_{C_i,j}$

$$S^{(n)}_{C_i,t,j} = \nabla A_{t,j} f_t \cdot CAV^{(n)}_{C_i,j}$$

Calculate the TCAV score for each run:

$$\text{TCAV}^{(n)}_{C_i,t,j} = \frac{\left|\{x \in D_t : S^{(n)}_{C_i,t,j}(x) > 0\}\right|}{|D_t|}$$

end

end

end

Step 4: Statistical significance testing

foreach *layer* $L_j \in M$ **do**

foreach *concept* $C_i \in C$ **do**

// Perform Welch's t-test to compare the means:

$$H_0 : \mu_{C_i} = \mu_{R_k}, \qquad H_1 : \mu_{C_i} > \mu_{R_k}$$

// where μ_{C_i} and μ_{R_k} are the mean TCAV scores for C_i and R_k, respectively;

if *p-value* < 0.01 **then**

Reject H_0; // Concept C_i is considered meaningful;

else

Fail to reject H_0; // No significant difference observed, concept C_i is not considered meaningful;

end

end

end

5.3 Experimental setting and results

This study was conducted on the Kaggle platform. We used PyTorch[1] for building and training neural network model, Scikit-learn[2] for implementing metrics and scoring tasks, and

[1]PyTorch library: https://pytorch.org

[2]Scikit-learn library: https://scikit-learn.org/stable/

the ttest_ind[3] function from scipy.stats for conducting statistical analysis. This study evaluates the reliability, accuracy, and effectiveness of DNN models in detecting Pneumococcal meningitis. The data is divided into three distinct sets: a training set (70%), a validation set (15%), and a test set (15%), which are used for training, tuning, and evaluating the model, respectively.

5.3.1 Data analysis exploration

We conducted our experiments using the SINAN database from the Brazilian Government's Health Information System on Notifiable Diseases. Our analysis focused on 340 cases of Pneumococcal meningitis to evaluate the model's predictive performance for this specific disease. To ensure a fair comparison, we also included 344 cases of various other types of meningitis. To avoid bias, we balanced the dataset by randomly sampling an equal number of rows from the 'Other' category to match the number of Pneumococcal meningitis cases. Additionally, we collected a set of data containing 17 cases from Setif Hospital in Algeria, consisting of 10 cases of pneumococcal meningitis and 7 cases of other meningitis conditions. This addition aims to diversify our samples and provide a broader perspective on the performance and generalizability of our model, particularly when evaluating its effectiveness on unseen data. We performed a series of data preparation steps on the datasets, including removing duplicates, addressing missing values, handling outliers, and encoding categorical features to enhance representation and facilitate further interpretability.

Figure 5.2 presents a pair plot illustrating the relationships between several clinical features: Age, Neutrophils, Glucose ratio CSF/serum, CSF WBC, Protein ratio CSF/serum, and Lymphocytes. The diagonal plots show the distribution of each feature. Notably, Neutrophils and Lymphocytes exhibit differences between the two groups, with higher concentrations observed in certain pneumococcal meningitis cases. Low Glucose levels are particularly pronounced in pneumococcal cases. Bimodality is observed in the age distribution, with two peaks indicating notably higher rates of pneumococcal meningitis in very young children and elderly individuals (>50 years).

The off-diagonal scatter plots display pairwise feature relationships, with Pearson correlation coefficients (r) indicated at the top of each plot. A strong negative correlation (r = -0.75) is observed between Neutrophils and Lymphocytes, suggesting that as neutrophil levels rise, lymphocyte levels decrease, especially in pneumococcal meningitis. Additionally, the Glucose ratio CSF/serum tends to be lower in pneumococcal meningitis, and it shows a moderate negative correlation (r = -0.33) with the Protein ratio CSF/serum, indicating an inverse relationship between these features.

[3]T-test (Welch's t-test): https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

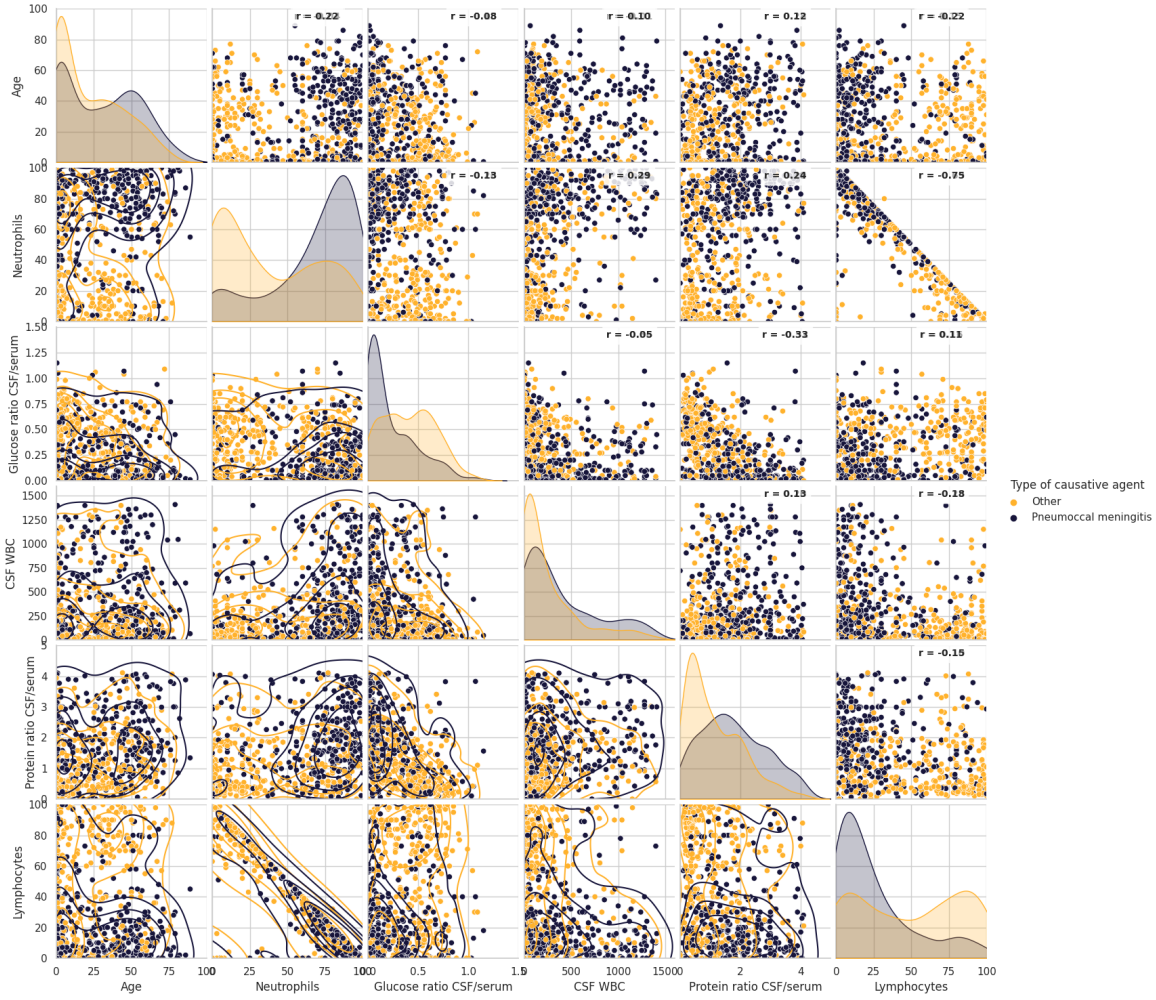


Figure 5.2: Bivariate plot of clinical CSF parameters distinguishing pneumococcal meningitis (blue) from other causes (orange), showing feature distributions in the diagonal and Pearson correlations for pairwise relationships (r).

5.3.2 Results and discussion

We employed multiple performance metrics to evaluate the model's effectiveness in the binary classification task. Accuracy measures the overall percentage of correctly classified instances. Precision measures the proportion of correctly identified pneumococcal meningitis cases out of all predicted positive cases, reflecting the model's reliability in confirming true positives. Recall or sensitivity determines how well the model detected actual pneumococcal meningitis cases, highlighting its ability to identify true positives among all existing cases. The F1 score provides a balanced view of precision and recall, offering an aggregate measure that considers false positives and false negatives. We also used the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) to evaluate the model's ability to distinguish pneumococcal meningitis from other conditions, with higher values indicating better discriminatory power. Table 5.3 summarizes the model's performance metrics for

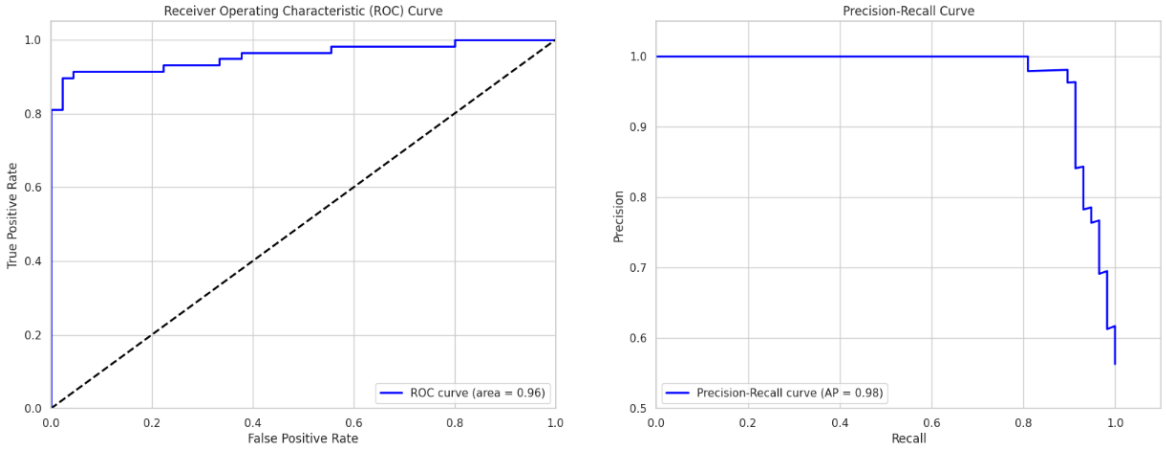


Figure 5.3: (a) ROC curve illustrating the model's true positive rate versus false positive rate. (b) The precision-recall curve showing precision versus recall.

diagnosing pneumococcal meningitis. The model demonstrates high accuracy (0.9223), with precision (0.9464) and recall (0.9138), f1-score (0.9298), and AUROC (0.9236). Figure 5.3 presents the ROC curves, AUC calculations for the pneumococcal meningitis class, and precision-recall curves, demonstrating the model's discriminative power and reliability.

Table 5.3: Model's performance metrics

Metrics	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUROC**
Other	-	0.89	0.93	0.91	-
Pneumococcal meningitis	-	0.95	0.91	0.93	-
Overall performance	0.9223	0.9464	0.9138	0.9298	0.9236

Additional unseen data collected from Setif Hospital in Algeria was used to ensure the model's reliability, robustness, and applicability to real-world tasks. The model was tested on data from 10 patients diagnosed with pneumococcal meningitis and 7 patients with other types of meningitis, correctly identifying all 10 cases of pneumococcal meningitis and 5 cases out of 7 of other conditions. The evaluation metrics reflect an accuracy (0.88), precision (0.83), indicates a low rate of false positives, ensuring high reliability in predicting pneumococcal meningitis cases. The recall of (1) highlights the model's effectiveness in identifying true positive cases, minimizing the likelihood of missing actual pneumococcal meningitis patients, and F1 score (0.87), indicating high performance in predicting Pneumococcal meningitis on this dataset. Cohen's Kappa score was calculated to provide a more nuanced understanding of the agreement between the model's predictions and the true diagnoses. A score of 0.75, which indicates substantial agreement, reinforces the model's reliability in clinical settings where diagnostic accuracy is critical. To enhance the understanding of the inner workings of the DNN model, we employed Bio-TCAV, a biologically informed extension of the TCAV technique, to interpret and explain deep neural network models. Bio-TCAV reveals the

influence of specific medical concepts, such as clinical signs, biomarkers, and patient history, on the model's predictions. This interpretability transforms the deep neural network from a black box into a more transparent tool for medical professionals. By analyzing the activations corresponding to defined medical concepts and randomly generated concepts, we captured the internal representations of the input data as they propagated through the model's four hidden layers. These activations provided valuable insights into how the model associates medical concepts with its predictions, highlighting the relevance and impact of clinically significant features in diagnosing pneumococcal meningitis. The TCAV score indicates the strength of each concept's influence on the model's predictions for pneumococcal meningitis, offering a clear measure of concept significance. Figure 5.4 shows the variation in concept importance across different layers. We exclude the final layer of the DNN from TCAV analysis due to its lack of non-linearities. In this layer, the gradient of the logit is determined solely by the model weights, independent of the activations. As a result, the directional derivatives for all inputs become identical, producing TCAV scores that are always either 0 or 1 for individual CAVs. Since TCAV relies on analyzing the distribution of these directional derivatives across different data points, this uniformity eliminates meaningful variation. Therefore, TCAV is only performed on non-linear layers, as these ensure a diverse and informative range of directional derivatives [31]. Therefore, TCAV is only applied to layers with non-linear transformations. The plot highlights differences in TCAV scores between layers, providing insights into how concepts influence the model's behavior at different depths. PCR, CSF, blood culture tests, and LATEX and bacterioscopy results emerged as the most influential concepts with a score of 1. This aligns with real-world medical practices, where these tests are critical for confirming the presence of *Streptococcus pneumoniae*, the causative agent of pneumococcal meningitis, and establishing a definitive diagnosis. Unsurprisingly, these concepts yield the highest scores, representing conclusive evidence in the medical field and the model's decision-making process. Clinical signs also achieved a high TCAV score of 0.92 in layers 2 and 3, increasing from 0.72 in layer 1. This increase in TCAV scores over the layers indicates that the model progressively relies on these concepts at deeper layers. They become more relevant as the model refines its decision-making process. The CSF aspect demonstrated a notable influence, attaining a TCAV score of 0.88. Similarly, Medical History showed significant contributions, with scores of 0.79 in layer 3, 0.54 in layer 2, and 0.74 in layer 1. This fluctuation in TCAV scores across layers suggests its influence grows as the model processes deeper layers. The biomarkers showed the lowest TCAV score of 0.56 in layer 3, a decrease from 0.79 in layer 1. This suggests that biomarkers contribute less to the model's decision-making process in diagnosing pneumococcal meningitis as the layers deepen. This lower score likely reflects that these biomarkers are common across various types of meningitis, reducing their specificity for pneumococcal meningitis and making them less reliable for differentiation. However, this concept may still be meaningful, as it passes the

statistical significance test with a p-value (< 0.01), indicating that it is meaningfully different from random concepts, regardless of the magnitude of the TCAV score.

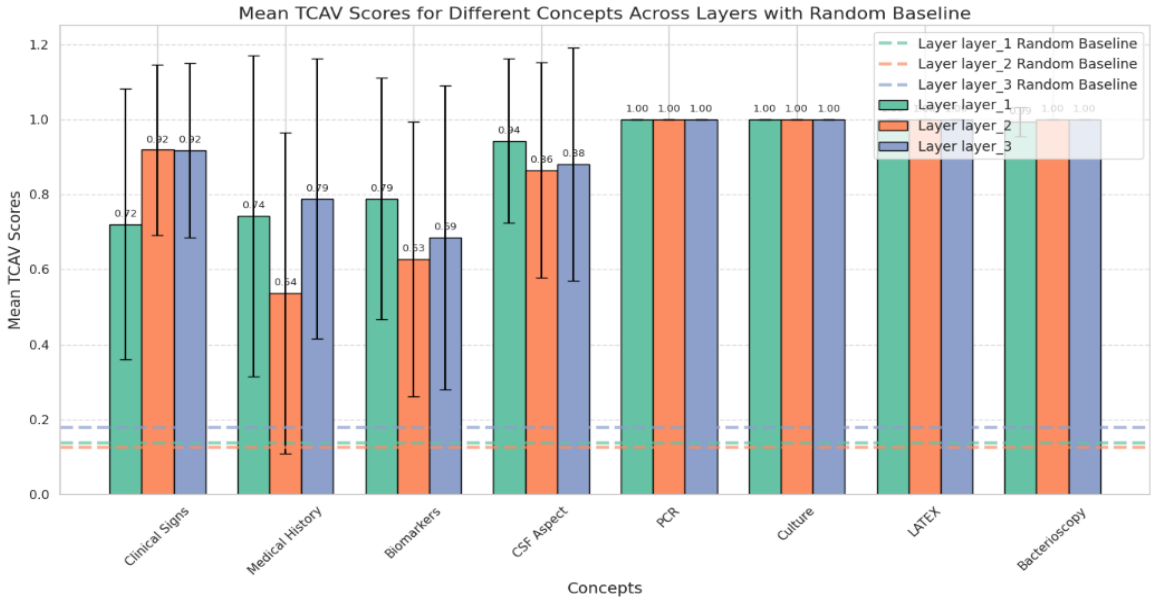


Figure 5.4: Comparison of mean TCAV scores across three neural network layers (Layer 1, Layer 2, and Layer 3) for medical concepts. Bars represent the mean TCAV score for each concept, with error bars showing the standard deviation. Dashed lines indicate the mean random TCAV baselines for each layer.

We performed a Welch's t-test to evaluate the statistical significance of the differences between the TCAV scores for concept CAVs and random CAVs. This test helps determine whether the observed differences in TCAV scores are statistically significant. Table 5.4 presents the results of Welch's t-test for various medical concepts across different model layers. All medical concepts show statistically significant differences from random concepts, implying their relevance in the model's decision-making.

Table 5.4: Welch's t-test p-values for different medical concepts across model layers. Lower p-values (<0.01) indicate a statistically significant difference from random concepts.

Medical Concepts	**Welch's t-test p-value**		
	Layer 1	**Layer 2**	**Layer 3**
Clinical Signs Concept	4.76e-11	1.14e-20	8.53e-18
Medical History Concept	4.46e-10	7.64e-06	9.16e-11
Biomarkers Concept	7.57e-14	8.81e-09	7.56e-08
CSF Aspect Concept	1.03e-21	4.46e-17	1.38e-14
PCR Concept	9.84e-20	1.48e-19	1.74e-17
Culture Concept	9.84e-20	1.48e-19	1.74e-17
LATEX Concept	9.84e-20	1.48e-19	1.74e-17
Bacterioscopy Concept	7.49e-20	1.48e-19	1.74e-17

5.4 Conclusion

This study demonstrates the effectiveness of deep learning models, with the model showing strong performance across multiple evaluation metrics, achieving high accuracy (92.23%), precision (94.64%), recall (91.38%), F1-score (92.98%), and AUROC (92.36%). Testing the model on real-world clinical data from Algeria confirmed its robustness and reliability. The model was tested on data from 10 patients diagnosed with pneumococcal meningitis and 7 patients with other types of meningitis, successfully identifying all 10 cases of pneumococcal meningitis and 5 cases of other types of meningitis. The model achieved an accuracy of 0.88, precision of 0.83, recall of 1, and an F1 score of 0.87, demonstrating strong performance in predicting pneumococcal meningitis. Additionally, a Cohen's Kappa score of 0.75 indicates substantial agreement between the model's predictions and the actual clinical diagnoses, confirming the robustness of the model. To improve the model's interpretability, we employed Bio-TCAV, which offered valuable insights into the influence of specific medical concepts on the model's predictions. Key concepts such as clinical signs, medical history, and CSF aspects significantly impacted the model's decisions, while biomarkers played a more limited role. The model finds it challenging to rely heavily on biomarkers for differentiation, as their presence may not be as specific or distinctive in pneumococcal meningitis cases compared to other types of meningitis. Statistical tests further validated the significance of these concepts in the model's decision-making process. Our findings suggest that integrating deep learning models with interpretability techniques like TCAV holds great potential for improving diagnostic accuracy in clinical settings. This approach enhances diagnostic models' performance and gives clinicians meaningful insights into the underlying factors influencing predictions, enabling more informed clinical decisions. By highlighting how relevant medical concepts affect model outcomes, Bio-TCAV can foster trust among medical practitioners and stakeholders by demonstrating that the model considers clinically meaningful factors. However, validating these findings on larger and more diverse datasets is important to ensure generalizability and confirm its clinical applicability across various patient populations and conditions.

Conclusion

*****

The findings of this thesis offer significant insights into the immunological and biochemical profiles associated with different types of meningitis, analyzed through both traditional laboratory parameters and advanced machine learning techniques. By examining neutrophil and lymphocyte counts, CSF white cell counts, glucose and protein ratios, and stratifying these data across age groups, this study highlights critical diagnostic differences among bacterial, viral, and specific meningitis pathogens. Key observations, such as the marked increase in neutrophil levels in bacterial meningitis cases and elevated lymphocyte levels in viral infections, underscore the distinct immune responses each infection type elicits. These patterns support and refine existing diagnostic benchmarks, especially within bacterial etiologies, aligning with diagnostic criteria established in previous research.

The analysis of variance of cerebrospinal fluid biomarkers revealed significant differences in CSF WCC (cells/mm^3), Neutrophils (%), Lymphocytes (%), CSF/blood Glucose and Protein ratios in children and adults ($p < 0.05$). However, differences were observed solely in neutrophil levels among elderly subjects. In bacterial meningitis, neutrophil levels increased across all ages, with adults showing higher levels. For meningococcal meningitis, adults had a median neutrophil level of 78.5% [60%, 90%] compared to 57% [24.5%, 85.25%] in children. The CSF white cell count (WCC) median was also higher in adults (339 cells/mm^3) compared to children (224 cells/mm^3). Viral meningitis cases exhibited higher lymphocyte levels across all age groups, with medians of 69% [34%, 87%] in children, 82% [61%, 93%] in adults, and 77% [55%, 90%] in the elderly. The glucose ratio was lower in bacterial meningitis (<0.3) and higher in viral cases (>0.5). Protein ratios were elevated in bacterial meningitis, indicating increased blood-brain barrier permeability. These results demonstrate a distinct biological profile for different causative agents, modulated by patient age.

Further, this thesis bridges clinical understanding with innovative AI-driven diagnostic tools by developing and applying an explainable AI model with notable accuracy and predictive metrics. The XGBoost and Gradient Boosting models, alongside interpretability tools like SHAP, ELI5 and LIME, reveal the underlying mechanisms by which clinical features influence diagnostic outcomes. The AI model's performance demonstrates potential in diagnosing different meningitis types and meningococcal meningitis. It shows adaptability to diverse, real-world cases, as evidenced by model testing on data from Setif's hospital. Identi-

fying essential biomarkers and the role of CSF findings like neutrophil levels and white blood cell counts enhance the robustness of the model. At the same time, the local explanations provide transparency, which is critical for clinical decision-making.

In the second part of the study, the Extreme Gradient Boosting model was chosen for its superior performance metrics (Accuracy: 0.90, AUROC: 0.94, and F1-score: 0.98). Setif's hospital data revealed notable performance metrics (Accuracy: 0.7143, F1-Score: 0.7857). This study's findings showcase each feature's contribution to the model's predictions and diagnosis. It also reveals critical biomarker ranges associated with distinct types of Meningitis. Significant diagnostic effect was found for Meningococcal Meningitis with elevated neutrophil levels (40%) and balanced lymphocyte levels (40%–60%). Tuberculous Meningitis demonstrated low neutrophil levels (60%) and elevated lymphocyte levels (60%). H. influenzae meningitis exhibited a predominance of neutrophils (80%), while Aseptic meningitis showed lower neutrophil levels (40%) and lymphocyte levels within the range of 50%–60%.

In the third contribution, the gradient-boosting model exhibits superior performance metrics, including accuracy (0.88), precision (0.92), recall (0.83), AUROC (0.93), and f1-score (0.87). Using ELI5 and LIME, we elucidated the importance of key features in the model's predictions. Our results highlight key factors contributing to model success, such as Neisseria meningitidis identified through cerebrospinal fluid (CSF) culture and latex agglutination, gram-negative diplococci in CSF smear examination, white cell counts in CSF, and patient age in the identification of meningococcal meningitis. Local explanations reveal the presence of neutrophils in CSF as a characteristic feature of Meningococcal meningitis. LIME analysis indicates the significance of low lymphocyte percentages and elevated white blood cell counts in predicting this condition.

The final contribution focused on developing a novel explainable approach for the black box model for the infectious disease diagnosis. our deep learning model achieved strong predictive performance in identifying Pneumococcal meningitis cases, with high accuracy (92.23%), precision (94.64%), recall (91.38%), F1-score (92.98%), and AUROC (92.36%). Its robustness was confirmed through validation on real-world clinical data from Algeria, correctly identifying all 10 pneumococcal meningitis cases and 5 out of 7 other meningitis cases, resulting in an accuracy of 0.88, precision of 0.83, recall of 1, F1-score of 0.87, and a Cohen's Kappa of 0.75—indicating substantial agreement with clinical diagnoses. To enhance interpretability, we applied Bio-TCAV, which highlighted the importance of clinical signs (TCAV score: 0.92), medical history (0.79), and CSF aspect (0.88) in driving model predictions. Biomarkers had a comparatively limited influence (0.56). Diagnostic tests such as PCR, CSF and blood cultures, LATEX, and bacterioscopy showed maximal influence (TCAV score = 1), consistent with their established clinical relevance. Statistical validation using Welch's t-test confirmed these findings were significant. This biologically

informed interpretability approach aligns model behavior with medical reasoning, enhancing transparency and trust in AI-assisted diagnostic tools.

Overall, this thesis illustrates the promising role of AI in refining meningitis diagnostics, complementing traditional lab-based methods with improved precision and explainability. It underscores the importance of integrating machine learning in clinical settings to support timely, accurate diagnoses, especially in resource-limited regions. Future research could explore additional biomarkers, such as CSF lactate, enhancing the model's accuracy and generalizability across various demographic and clinical environments, ultimately advancing the management of meningitis on a broader scale.

Bibliography

*****

[1] I. Fong, Ed., *Challenges in Infectious Diseases*, 1st ed., ser. Emerging Infectious Diseases of the 21st Century. New York, NY: Springer Science+Business Media, 2013, vol. 1. [Online]. Available: https://doi.org/10.1007/978-1-4614-4496-1

[2] N. K. Tran, S. Albahra, L. May, S. Waldman, S. Crabtree, S. Bainbridge, and H. Rashidi, "Evolving applications of artificial intelligence and machine learning in infectious diseases testing," *Clinical chemistry*, vol. 68, no. 1, pp. 125–133, 2022.

[3] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, "Health intelligence: how artificial intelligence transforms population and personalized health," *NPJ digital medicine*, vol. 1, no. 1, p. 53, 2018.

[4] E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, 1st ed. USA: Basic Books, Inc., 2019.

[5] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, Eds., *Explainable AI in Healthcare and Medicine*, 1st ed., ser. Studies in Computational Intelligence. Cham: Springer, 2021, vol. 1. [Online]. Available: https://doi.org/10.1007/978-3-030-53352-6

[6] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[7] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[8] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 248–266.

[9] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[10] L. Gianfagna and A. Di Cecco, *Explainable AI with Python*, 1st ed. Cham: Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-68640-6

[11] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, 2022.

[12] "What is Explainable AI (XAI)? — IBM — ibm.com," https://www.ibm.com/se-en/watson/explainable-ai, [Accessed 09-11-2024].

[13] G. Rubeis, *Ethics of Medical AI*, 1st ed., ser. The International Library of Ethics, Law and Technology. Cham: Springer, 2024, eBook Packages: Religion and Philosophy, Philosophy and Religion (R0). Hardcover published: 21 March 2024. eBook published: 20 March 2024. Softcover published: 07 April 2025. Includes 1 b/w illustration. [Online]. Available: https://doi.org/10.1007/978-3-031-55744-6

[14] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[15] D. Killock, "Ai outperforms radiologists in mammographic screening," *Nature Reviews Clinical Oncology*, vol. 17, no. 3, pp. 134–134, 2020.

[16] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable ai for healthcare 5.0: opportunities and challenges," *IEEE Access*, vol. 10, pp. 84 486–84 517, 2022.

[17] T. Alsuliman, D. Humaidan, and L. Sliman, "Machine learning and artificial intelligence in the service of medicine: Necessity or potentiality?" *Current research in translational medicine*, vol. 68, no. 4, pp. 245–251, 2020.

[18] S. Haddadin and D. Knobbe, "Robotics and artificial intelligence—the present and future visions," *Algorithms and Law", Cambridge University Press, Cambridge, UK*, pp. 20–23, 2020.

[19] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[20] U. Kamath and J. Liu, *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*, 1st ed. Cham: Springer, 2021, eBook Packages: Computer Science (R0). Hardcover published: 16 December 2021. eBook published: 15 December 2021. Softcover published: 17 December 2022. Includes 33 b/w and 161 color illustrations. [Online]. Available: https://doi.org/10.1007/978-3-030-83356-5

[21] C. Molnar, *Interpretable Machine Learning*, 3rd ed., 2025, available online at https://christophm.github.io/interpretable-ml-book. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[22] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[23] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[24] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154 096–154 113, 2019.

[25] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., ser. Chapman and Hall/CRC Monographs on Statistics and Applied Probability. New York: Routledge, 2019, originally published in 1989. eBook published January 22, 2019.

[26] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[27] S. Sarkar, T. Weyde, A. d. Garcez, G. G. Slabaugh, S. Dragicevic, and C. Percy, "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in *CEUR Workshop Proceedings*, vol. 1773. CEUR Workshop Proceedings, 2016.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[30] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186. [Online]. Available: https://doi.org/10.1145/3375627.3375830

[31] A. Nicolson, L. Schut, J. A. Noble, and Y. Gal, "Explaining explainability: Understanding concept activation vectors," *arXiv preprint arXiv:2404.03713*, 2024.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

[33] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.

[35] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of biomedical informatics*, vol. 113, p. 103655, 2021.

[36] E. Kazim and A. Koshiyama, "Explaining decisions made with ai: a review of the co-badged guidance by the ico and the turing institute," *Available at SSRN 3656269*, 2020.

[37] M. E. Webb, A. Fluck, J. Magenheim, J. Malyn-Smith, J. Waters, M. Deschênes, and J. Zagami, "Machine learning for human learners: opportunities, issues, tensions and threats," *Educational Technology Research and Development*, vol. 69, no. 4, pp. 2109–2130, 2021.

[38] U. Kose, O. Deperlioglu, J. Alzubi, and B. Patrut, *Deep Learning for Medical Decision Support Systems*, 1st ed., ser. Studies in Computational Intelligence. Singapore: Springer Singapore, 2020, originally published as hardcover on June 18, 2020, and as eBook on June 17, 2020. Softcover edition released on June 19, 2021.

[39] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[40] G. D'Angelo, R. Pilla, C. Tascini, and S. Rampone, "A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees," *Soft Computing*, vol. 23, no. 22, pp. 11 775–11 791, 2019.

[41] K. Zaccari and E. C. Marujo, "Machine learning for aiding meningitis diagnosis in pediatric patients," *International Journal of Medical and Health Sciences*, vol. 13, no. 9, pp. 411–419, 2019.

[42] L. Šeho, H. Šutković, V. Tabak, S. Tahirović, A. Smajović, E. Bečić, A. Deumić, L. S. Bećirović, L. G. Pokvić, and A. Badnjević, "Using artificial intelligence in diagnostics of meningitis," *IFAC-PapersOnLine*, vol. 55, no. 4, pp. 56–61, 2022.

[43] V.-M. Lélis, E. Guzmán, and M.-V. Belmonte, "A statistical classifier to support diagnose meningitis in less developed areas of brazil," *Journal of medical systems*, vol. 41, pp. 1–10, 2017.

[44] V. M. Lelis, E. Guzman, and M.-V. Belmonte, "Non-invasive meningitis diagnosis using decision trees," *IEEE Access*, vol. 8, pp. 18 394–18 407, 2020.

[45] E. Guzman, M.-V. Belmonte, and V. M. Lelis, "Ensemble methods for meningitis aetiology diagnosis," *Expert Systems*, vol. 39, no. 8, p. e12996, 2022.

[46] A.-F. A. Mentis, I. Garcia, J. Jiménez, M. Paparoupa, A. Xirogianni, A. Papandreou, and G. Tzanakaki, "Artificial intelligence in differential diagnostics of meningitis: a nationwide study," *Diagnostics*, vol. 11, no. 4, p. 602, 2021.

[47] P. G. C. D. Pinheiro, L. I. C. Pinheiro, R. Holanda Filho, M. L. D. Pereira, P. R. Pinheiro, P. J. L. Santiago, and R. Comin-Nunes, "An application of machine learning in the early diagnosis of meningitis," in *The International Research & Innovation Forum*. Springer, 2022, pp. 97–106.

[48] S. O. Alile and M. E. Bello, "A machine learning approach for diagnosing meningococcal meningitis," *Int. J. Sci. Res. in Computer Science and Engineering Vol*, vol. 8, no. 3, 2020.

[49] J. Amann, E. Vayena, K. E. Ormond, D. Frey, V. I. Madai, and A. Blasimme, "Expectations and attitudes towards medical artificial intelligence: A qualitative study in the field of stroke," *Plos one*, vol. 18, no. 1, p. e0279088, 2023.

[50] D. Burema, N. Debowski-Weimann, A. von Janowski, J. Grabowski, M. Maftei, M. Jacobs, P. Van Der Smagt, and D. Benbouzid, "A sector-based approach to ai ethics: Understanding ethical issues of ai-related incidents within their sectoral context," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 705–714.

[51] L. Lucaj, P. van der Smagt, and D. Benbouzid, "Ai regulation is (not) all you need," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1267–1279.

[52] B. K. Choi, Y. J. Choi, M. Sung, W. Ha, M. K. Chu, W.-J. Kim, K. Heo, K. M. Kim, and Y. R. Park, "Development and validation of an artificial intelligence model for

the early classification of the aetiology of meningitis and encephalitis: a retrospective observational study," *eClinicalMedicine*, vol. 61, 2023.

[53] Y. Yang, Y.-M. Wang, C.-H. R. Lin, C.-Y. Cheng, C.-M. Tsai, Y.-H. Huang, T.-Y. Chen, and I.-M. Chiu, "Explainable deep learning model to predict invasive bacterial infection in febrile young infants: A retrospective study," *International Journal of Medical Informatics*, vol. 172, p. 105007, 2023.

[54] H. Sial, F. Carandell, S. Ajanovic, J. Jimenez, R. Quesada, F. Santos, W. C. Buck, M. Sidat, U. S. Consortium, Q. Bassat *et al.*, "Novel ai-driven infant meningitis screening from high resolution ultrasound imaging," *medRxiv*, pp. 2024–08, 2024.

[55] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, and R. Chadaga, "A machine learning and explainable artificial intelligence triage-prediction system for covid-19," *Decision Analytics Journal*, p. 100246, 2023.

[56] M. Laatifi, S. Douzi, H. Ezzine, C. E. Asry, A. Naya, A. Bouklouze, Y. Zaid, and M. Naciri, "Explanatory predictive model for covid-19 severity risk employing machine learning, shapley addition, and lime," *Scientific Reports*, vol. 13, no. 1, p. 5481, 2023.

[57] F. Mercaldo, M. P. Belfiore, A. Reginelli, L. Brunese, and A. Santone, "Coronavirus covid-19 detection by means of explainable deep learning," *Scientific Reports*, vol. 13, no. 1, p. 462, 2023.

[58] Y. Shi, C. Zhang, S. Pan, Y. Chen, X. Miao, G. He, Y. Wu, H. Ye, C. Weng, H. Zhang *et al.*, "The diagnosis of tuberculous meningitis: advancements in new technologies and machine learning algorithms," *Frontiers in Microbiology*, vol. 14, p. 1290746, 2023.

[59] K. A. Thakoor, S. C. Koorathota, D. C. Hood, and P. Sajda, "Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2456–2466, 2020.

[60] D. Mincu, E. Loreaux, S. Hou, S. Baur, I. Protsyuk, M. Seneviratne, A. Mottram, N. Tomasev, A. Karthikesalingam, and J. Schrouff, "Concept-based model explanations for electronic health records," in *Proceedings of the Conference on Health, Inference, and Learning*, 2021, pp. 36–46.

[61] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, and K. Curran, "Interpretability of a deep learning model in the application of cardiac mri segmentation with an acdc challenge

dataset," in *Medical imaging 2021: image processing*, vol. 11596. SPIE, 2021, pp. 861–872.

[62] K. E. Nelson and C. M. Williams, *Infectious disease epidemiology: theory and practice*. Jones & Bartlett Publishers, 2014.

[63] M. M. Wagner, L. S. Gresham, and V. Dato, "Case detection, outbreak detection, and outbreak characterization," *Handbook of Biosurveillance*, p. 27, 2007.

[64] G. A. Soper, "The curious career of typhoid mary," *Bulletin of the New York Academy of Medicine*, vol. 15, no. 10, p. 698, 1939.

[65] R. Riffenburgh, *Statistics in Medicine*, 3rd ed. San Diego, California, USA: Academic Press, 2012, copyright © 2012 Elsevier Inc. All rights reserved.

[66] M. Martcheva and O. Prosper, "Unstable dynamics of vector-borne diseases: Modeling through delay-differential equations," *Dynamic Models of Infectious Diseases: Volume 1: Vector-Borne Diseases*, pp. 43–75, 2013.

[67] T. Z. Nigussie, T. T. Zewotir, and E. K. Muluneh, "Seasonal and spatial variations of malaria transmissions in northwest ethiopia: Evaluating climate and environmental effects using generalized additive model," *Heliyon*, vol. 9, no. 4, 2023.

[68] F. S. Southwick, Ed., *Infectious Diseases: A Clinical Short Course*, 4th ed. New York: McGraw-Hill Education, 2020, accessed May 23, 2025. [Online]. Available: https://accessmedicine.mhmedical.com/content.aspx?bookid=2816&sectionid=240346380

[69] R. Hasbun, "The acute aseptic meningitis syndrome," *Current infectious disease reports*, vol. 2, no. 4, pp. 345–351, 2000.

[70] B. Shukla, E. A. Aguilera, L. Salazar, S. H. Wootton, Q. Kaewpoowat, and R. Hasbun, "Aseptic meningitis in adults and children: Diagnostic and management challenges," *Journal of clinical virology*, vol. 94, pp. 110–114, 2017.

[71] H. Rudolph, H. Schroten, and T. Tenenbaum, "Enterovirus infections of the central nervous system in children: an update," *The Pediatric Infectious Disease Journal*, vol. 35, no. 5, pp. 567–569, 2016.

[72] B. R. Lee, A. Sasidharan, C. J. Harrison, and R. Selvarangan, "Positive impact of routine testing for enterovirus and parechovirus on length of hospitalization and antimicrobial use among inpatients ≤ 6 months of age," *Journal of clinical microbiology*, vol. 59, no. 1, pp. 10–1128, 2020.

[73] R. Izumita, K. Deuchi, Y. Aizawa, R. Habuka, K. Watanabe, T. Otsuka, and A. Saitoh, "Intrafamilial transmission of parechovirus a and enteroviruses in neonates and young infants," *Journal of the Pediatric Infectious Diseases Society*, vol. 8, no. 6, pp. 501–506, 2019.

[74] S. Owatanapanich, R. Wutthanarungsan, W. Jaksupa, and U. Thisyakorn, "Risk factors for severe enteroviral infections in children," *J Med Assoc Thai*, vol. 99, no. 3, pp. 322–30, 2016.

[75] K. Benschop, J. Schinkel, R. Minnaar, D. Pajkrt, L. Spanjerberg, H. Kraakman, B. Berkhout, H. Zaaijer, M. Beld, and K. Wolthers, "Human parechovirus infections in dutch children and the association between serotype and disease severity," *Clinical infectious diseases*, vol. 42, no. 2, pp. 204–210, 2006.

[76] L. Olijve, L. Jennings, and T. Walls, "Human parechovirus: an increasingly recognized cause of sepsis-like illness in young infants," *Clinical microbiology reviews*, vol. 31, no. 1, pp. 10–1128, 2018.

[77] P. G. E. Kennedy and A. Chaudhuri, "Herpes simplex encephalitis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. 3, pp. 237–238, 2002.

[78] K. L. Tyler, "Acute viral encephalitis," *New England Journal of Medicine*, vol. 379, no. 6, pp. 557–566, 2018.

[79] K. L. Tyler *et al.*, "Herpes simplex virus infections of the central nervous system: encephalitis and meningitis, including mollaret's," *HERPES-CAMBRIDGE-*, vol. 11, pp. 57A–64A, 2004.

[80] Q. Kaewpoowat, L. Salazar, E. Aguilera, S. H. Wootton, and R. Hasbun, "Herpes simplex and varicella zoster cns infections: clinical presentations, treatments and outcomes," *Infection*, vol. 44, pp. 337–345, 2016.

[81] J. D. Beckham and K. L. Tyler, "Arbovirus infections," *CONTINUUM: Lifelong Learning in Neurology*, vol. 21, no. 6, pp. 1599–1611, 2015.

[82] L. D. Kramer, L. M. Styer, and G. D. Ebel, "A global perspective on the epidemiology of west nile virus," *Annu. Rev. Entomol.*, vol. 53, no. 1, pp. 61–81, 2008.

[83] I. Lafri, C. M. Prat, I. Bitam, P. Gravier, M. Besbaci, F. Zeroual, M. H. Ben-Mahdi, B. Davoust, and I. Leparc-Goffart, "Seroprevalence of west nile virus antibodies in equids in the north-east of algeria and detection of virus circulation in 2014," *Comparative immunology, microbiology and infectious diseases*, vol. 50, pp. 8–12, 2017.

[84] I. Lafri, A. Hachid, and I. Bitam, "West nile virus in algeria: a comprehensive overview. new microbes new infect. 2018; 27: 9–13."

[85] N. P. Lindsey, J. E. Staples, J. A. Lehman, and M. Fischer, "Medical risk factors for severe west nile virus disease, united states, 2008–2010," *The American journal of tropical medicine and hygiene*, vol. 87, no. 1, p. 179, 2012.

[86] S. Jaijakul, C. A. Arias, M. Hossain, R. C. Arduino, S. H. Wootton, and R. Hasbun, "Toscana meningoencephalitis: a comparison to other viral central nervous system infections," *Journal of clinical virology*, vol. 55, no. 3, pp. 204–208, 2012.

[87] T. Lenhard, D. Ott, N. J. Jakob, M. Pham, P. Bäumer, F. Martinez-Torres, and U. Meyding-Lamadé, "Predictors, neuroimaging characteristics and long-term outcome of severe european tick-borne encephalitis: a prospective cohort study," *PloS one*, vol. 11, no. 4, p. e0154143, 2016.

[88] C. Fevola, S. Kuivanen, T. Smura, A. Vaheri, H. Kallio-Kokko, H. C. Hauffe, O. Vapalahti, and A. J. Jääskeläinen, "Seroprevalence of lymphocytic choriomeningitis virus and ljungan virus in finnish patients with suspected neurological infections," *Journal of Medical Virology*, vol. 90, no. 3, pp. 429–435, 2018.

[89] N. G. Martin, M. A. Iro, M. Sadarangani, R. Goldacre, A. J. Pollard, and M. J. Goldacre, "Hospital admissions for viral meningitis in children in england over five decades: a population-based observational study," *The Lancet Infectious Diseases*, vol. 16, no. 11, pp. 1279–1287, 2016.

[90] R. Hasbun, N. Rosenthal, J. Balada-Llasat, J. Chung, S. Duff, S. Bozzette, L. Zimmer, and C. C. Ginocchio, "Epidemiology of meningitis and encephalitis in the united states, 2011–2014," *Clinical infectious diseases*, vol. 65, no. 3, pp. 359–363, 2017.

[91] A. Dubot-Pérès, O. Sengvilaipaseuth, A. Chanthongthip, P. Newton, and X. de Lamballerie, "How many patients with anti-jev igm in cerebrospinal fluid really have japanese encephalitis?" *Lancet Infectious Diseases*, vol. 15, no. 12, 2015.

[92] M. Puccioni-Sohler, N. Roveroni, C. Rosadas, F. Ferry, J. M. Peralta, and A. Tanuri, "Dengue infection in the nervous system: lessons learned for zika and chikungunya," *Arquivos de Neuro-psiquiatria*, vol. 75, no. 2, pp. 123–126, 2017.

[93] S. H. Waterman, H. S. Margolis, and J. J. Sejvar, "Surveillance for dengue and dengue-associated neurologic syndromes in the united states," *The American journal of tropical medicine and hygiene*, vol. 92, no. 5, p. 996, 2015.

[94] M. Jacobs, A. Rodger, D. J. Bell, S. Bhagani, I. Cropley, A. Filipe, R. J. Gifford, S. Hopkins, J. Hughes, F. Jabeen *et al.*, "Late ebola virus relapse causing meningoencephalitis: a case report," *The Lancet*, vol. 388, no. 10043, pp. 498–503, 2016.

[95] R. Hasbun, M. Bijlsma, M. C. Brouwer, N. Khoury, C. M. Hadi, A. van der Ende, S. H. Wootton, L. Salazar, M. M. Hossain, M. Beilke *et al.*, "Risk score for identifying adults with csf pleocytosis and negative csf gram stain at low risk for an urgent treatable cause," *Journal of infection*, vol. 67, no. 2, pp. 102–110, 2013.

[96] R. Hasbun, Ed., *Meningitis and Encephalitis*, 1st ed., ser. Medicine (R0). Cham: Springer Cham, 2018, published: 19 July 2018, Copyright Springer International Publishing AG 2018.

[97] S. Jaijakul, L. Salazar, S. H. Wootton, E. Aguilera, and R. Hasbun, "The clinical significance of neutrophilic pleocytosis in cerebrospinal fluid in patients with viral central nervous system infections," *International Journal of Infectious Diseases*, vol. 59, pp. 77–81, 2017.

[98] V. Shrikanth, L. Salazar, N. Khoury, S. Wootton, and R. Hasbun, "Hypoglycorrhachia in adults with community-acquired meningitis: etiologies and prognostic significance," *International Journal of Infectious Diseases*, vol. 39, pp. 39–43, 2015.

[99] A. Spanos, F. E. Harrell, and D. T. Durack, "Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations," *Jama*, vol. 262, no. 19, pp. 2700–2707, 1989.

[100] L. E. Nigrovic, R. Malley, and N. Kuppermann, "Meta-analysis of bacterial meningitis score validation studies," *Archives of disease in childhood*, vol. 97, no. 9, pp. 799–805, 2012.

[101] N. T. Huy, N. T. Thao, D. T. Diep, M. Kikuchi, J. Zamora, and K. Hirayama, "Cerebrospinal fluid lactate concentration to distinguish bacterial from aseptic meningitis: a systemic review and meta-analysis," *Critical care*, vol. 14, pp. 1–15, 2010.

[102] K. Sakushima, Y. Hayashino, T. Kawaguchi, J. L. Jackson, and S. Fukuhara, "Diagnostic accuracy of cerebrospinal fluid lactate for differentiating bacterial meningitis from aseptic meningitis: a meta-analysis," *Journal of Infection*, vol. 62, no. 4, pp. 255–262, 2011.

[103] P. Sormunen, M. J. Kallio, T. Kilpi, and H. Peltola, "C-reactive protein is useful in distinguishing gram stain–negative bacterial meningitis from viral meningitis in children," *The Journal of pediatrics*, vol. 134, no. 6, pp. 725–729, 1999.

[104] A. Viallon, F. Zeni, C. Lambert, B. Pozzetto, B. Tardy, C. Venet, and J.-C. Bertrand, "High sensitivity and specificity of serum procalcitonin levels in adults with bacterial meningitis," *Clinical infectious diseases*, vol. 28, no. 6, pp. 1310–1316, 1999.

[105] S. Schwarz, M. Bertram, S. Schwab, K. Andrassy, and W. Hacke, "Serum procalcitonin levels in bacterial and abacterial meningitis," *Critical care medicine*, vol. 28, no. 6, pp. 1828–1832, 2000.

[106] D. Van de Beek, J. De Gans, L. Spanjaard, M. Weisfelt, J. B. Reitsma, and M. Vermeulen, "Clinical features and prognostic factors in adults with bacterial meningitis," *New England Journal of Medicine*, vol. 351, no. 18, pp. 1849–1859, 2004.

[107] M. Weisfelt, D. Van De Beek, L. Spanjaard, J. B. Reitsma, and J. De Gans, "Community-acquired bacterial meningitis in older people," *Journal of the American Geriatrics Society*, vol. 54, no. 10, pp. 1500–1507, 2006.

[108] J. Best and S. Hughes, "Evidence behind the who guidelines: hospital care for children—what are the useful clinical features of bacterial meningitis found in infants and children?" *Journal of tropical pediatrics*, vol. 54, no. 2, pp. 83–86, 2008.

[109] F. McGill, R. Heyderman, B. Michael, S. Defres, N. Beeching, R. Borrow, L. Glennie, O. Gaillemin, D. Wyncoll, E. Kaczmarski *et al.*, "The uk joint specialist societies guideline on the diagnosis and management of acute meningitis and meningococcal sepsis in immunocompetent adults," *Journal of Infection*, vol. 72, no. 4, pp. 405–438, 2016.

[110] D. van de Beek, C. Cabellos, O. Dzupova, S. Esposito, M. Klein, A. Kloek, S. Leib, B. Mourvillier, C. Ostergaard, P. Pagliano *et al.*, "Escmid guideline: diagnosis and treatment of acute bacterial meningitis," *Clinical microbiology and infection*, vol. 22, pp. S37–S62, 2016.

[111] J. Khumalo, M. Nicol, D. Hardie, R. Muloiwa, P. Mteshana, and C. Bamford, "Diagnostic accuracy of two multiplex real-time polymerase chain reaction assays for the diagnosis of meningitis in children in a resource-limited setting," *PLoS ONE*, vol. 12, no. 3, p. e0173948, 2017.

[112] A. L. Leber, K. Everhart, J.-M. Balada-Llasat, J. Cullison, J. Daly, S. Holt, P. Lephart, H. Salimnia, P. C. Schreckenberger, S. DesJarlais, S. L. Reed, K. C. Chapin, L. LeBlanc, J. K. Johnson, N. L. Soliven, K. C. Carroll, J.-A. Miller, J. D. Bard, J. Mestas, M. Bankowski, T. Enomoto, A. C. Hemmert, and K. M. Bourzac, "Multicenter evaluation of biofire filmarray meningitis/encephalitis panel for detection of bacteria,

viruses, and yeast in cerebrospinal fluid specimens," *Journal of Clinical Microbiology*, vol. 54, no. 9, pp. 2251–2261, 2016.

[113] M. C. Brouwer, G. E. Thwaites, A. R. Tunkel, and D. van de Beek, "Dilemmas in the diagnosis of acute community-acquired bacterial meningitis," *The Lancet*, vol. 380, no. 9854, pp. 1684–1692, 2012.

[114] M. Radetsky, "Fulminant bacterial meningitis," *The Pediatric Infectious Disease Journal*, vol. 33, no. 2, pp. 204–207, 2014.

[115] S. Kastenbauer, F. Winkler, and H.-W. Pfister, "Cranial ct before lumbar puncture in suspected meningitis," *New England Journal of Medicine*, vol. 346, no. 16, pp. 1248–1251, 2002, author reply -51.

[116] L. Salazar and R. Hasbun, "Cranial imaging before lumbar puncture in adults with community-acquired meningitis: clinical utility and adherence to the infectious diseases society of america guidelines," *Clinical Infectious Diseases*, vol. 64, no. 12, pp. 1657–1662, 2017.

[117] M. Glimaker, B. Johansson, O. Grindborg, M. Bottai, L. Lindquist, and J. Sjolin, "Adult bacterial meningitis: earlier treatment and improved outcome following guideline revision promoting prompt lumbar puncture," *Clinical Infectious Diseases*, vol. 60, no. 8, pp. 1162–1169, 2015.

[118] N. Proulx, D. Fréchette, B. Toye, J. Chan, and S. Kravcik, "Delays in the administration of antibiotics are associated with mortality from adult acute bacterial meningitis," *QJM: An International Journal of Medicine*, vol. 98, no. 4, pp. 291–298, 2005.

[119] R. Hasbun, J. Abrahams, J. Jekel, and V. J. Quagliarello, "Computed tomography of the head before lumbar puncture in adults with suspected meningitis," *New England Journal of Medicine*, vol. 345, no. 24, pp. 1727–1733, 2001.

[120] M. Glimaker, B. Johansson, H. Halldorsdottir *et al.*, "Neuro-intensive treatment targeting intracranial hypertension improves outcome in severe bacterial meningitis: an intervention-control study," *PLoS ONE*, vol. 9, no. 3, p. e91976, 2014.

[121] J. T. Kanegaye, P. Soliemanzadeh, and J. S. Bradley, "Lumbar puncture in pediatric bacterial meningitis: defining the time interval for recovery of cerebrospinal fluid pathogens after parenteral antibiotic pretreatment," *Pediatrics*, vol. 108, no. 5, pp. 1169–1174, 2001.

[122] B. Michael, B. F. Menezes, J. Cunniffe *et al.*, "Effect of delayed lumbar punctures on the diagnosis of acute bacterial meningitis in adults," *Emergency Medicine Journal*, vol. 27, no. 6, pp. 433–438, 2010.

[123] B. Shmaefsky and H. Babcock, *Meningitis*. Infobase Publishing, 2010.

[124] T. Akaishi, K. Tarasawa, K. Fushimi, N. Yaegashi, M. Aoki, and K. Fujimori, "Demographic profiles and risk factors for mortality in acute meningitis: A nationwide population-based observational study," *Acute Medicine & Surgery*, vol. 11, no. 1, p. e920, 2024.

[125] "Meningitis — who.int," https://www.who.int/news-room/fact-sheets/detail/meningitis, [Accessed 15-09-2024].

[126] T. He, S. Kaplan, M. Kamboj, and Y.-W. Tang, "Laboratory diagnosis of central nervous system infection," *Current infectious disease reports*, vol. 18, pp. 1–12, 2016.

[127] S. M. de Almeida, S. M. P. Furlan, A. M. M. Cretella, B. Lapinski, K. Nogueira, L. L. Cogo, L. R. R. Vidal, and M. B. Nogueira, "Comparison of cerebrospinal fluid biomarkers for differential diagnosis of acute bacterial and viral meningitis with atypical cerebrospinal fluid characteristics," *Medical Principles and Practice*, vol. 29, no. 3, pp. 244–254, 2020.

[128] "Meningitis — who.int," https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON439, [Accessed 15-09-2024].

[129] D. d. I. d. S. D. Ministério da Saúde, Secretaria de Vigilância em Saúde, "SINAN (Sistema de Informação de Agravos de Notificação)," https://datasus.saude.gov.br/transferencia-de-arquivos/#, 2022, brazil's database for notifiable diseases, gathers data from epidemiological reports at health facilities across Brazil.

[130] R. Herold, H. Schroten, and C. Schwerk, "Virulence factors of meningitis-causing bacteria: enabling brain entry across the blood–brain barrier," *International journal of molecular sciences*, vol. 20, no. 21, p. 5393, 2019.

[131] N. M. Razali, Y. B. Wah *et al.*, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[132] Y. Zhou, Y. Zhu, and W. K. Wong, "Statistical tests for homogeneity of variance for clinical trials and recommendations," *Contemporary Clinical Trials Communications*, vol. 33, p. 101119, 2023.

[133] J. Valladares-Neto, "Effect size: A statistical basis for clinical practice," *Revista Odonto Ciência*, vol. 33, no. 1, pp. 84–90, 2018.

[134] S. A. Rusticus and C. Y. Lovato, "Impact of sample size and variability on the power and type i error rates of equivalence tests: A simulation study," *Practical Assessment, Research, and Evaluation*, vol. 19, no. 1, p. 11, 2019.

[135] P. Chavanet, C. Schaller, C. Levy, J. Flores-Cordero, M. Arens, L. Piroth, E. Bingen, and H. Portier, "Performance of a predictive rule to distinguish bacterial and viral meningitis," *Journal of Infection*, vol. 54, no. 4, pp. 328–336, 2007.

[136] M. Coeckelbergh, *AI Ethics*, 1st ed. The MIT Press, 2020, paperback edition published on April 7, 2020; eBook edition published on March 13, 2020. [Online]. Available: https://mitpress.mit.edu/9780262538190/ai-ethics/

[137] P. Humphreys, "The philosophical novelty of computer simulation methods," *Synthese*, vol. 169, pp. 615–626, 2009.

[138] J. M. Durán and N. Formanek, "Grounds for trust: Essential epistemic opacity and computational reliabilism," *Minds and Machines*, vol. 28, pp. 645–666, 2018.

[139] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.

[140] C. Beisbart and T. Räz, "Philosophy of science at sea: Clarifying the interpretability of machine learning," *Philosophy Compass*, vol. 17, no. 6, p. e12830, 2022.

[141] S. Kundu, "Ai in medicine must be explainable," *Nature medicine*, vol. 27, no. 8, pp. 1328–1328, 2021.

[142] J. M. Durán and K. R. Jongsma, "Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai," *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329–335, 2021.

[143] A. J. London, "Artificial intelligence and black-box medical decisions: accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.

[144] R. L. Pierce, W. Van Biesen, D. Van Cauwenberge, J. Decruyenaere, and S. Sterckx, "Explainability in medicine in an era of ai-based clinical decision support systems," *Frontiers in genetics*, vol. 13, p. 903600, 2022.

[145] J. Browning and M. Theunissen, "Putting explainable ai in context: institutional explanations for medical ai," *Ethics and Information Technology*, vol. 24, no. 2, 2022.

[146] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi, "Clinical applications of machine learning algorithms: beyond the black box," *Bmj*, vol. 364, 2019.

[147] C. Rudin and J. Radin, "Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition," *Harvard Data Science Review*, vol. 1, no. 2, pp. 1–9, 2019.

[148] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, 2016.

[149] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[150] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[151] J. Rueda, J. D. Rodríguez, I. P. Jounou, J. Hortal-Carmona, T. Ausín, and D. Rodríguez-Arias, ""just" accuracy? procedural fairness demands explainability in ai-based medical resource allocations," *AI & society*, vol. 39, no. 3, pp. 1411–1422, 2024.

[152] A. Ferrario, M. Loi, and E. Viganò, "In ai we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions," *Philosophy & Technology*, vol. 33, no. 3, pp. 523–539, 2020.

[153] W. Samek, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[154] A. Holzinger, M. Plass, K. Holzinger, G. C. Crisan, C.-M. Pintea, and V. Palade, "A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop," *arXiv preprint arXiv:1708.01104*, 2017.

[155] M. J. Lucas, M. C. Brouwer, and D. van de Beek, "Neurological sequelae of bacterial meningitis," *Journal of Infection*, vol. 73, no. 1, pp. 18–27, 2016.

[156] "Meningitis," https://www.who.int/news-room/fact-sheets/detail/meningitis, 2023, accessed: May 6, 2023.

[157] N. C. Bahr, G. Meintjes, and D. R. Boulware, "Inadequate diagnostics: the case to move beyond the bacilli for detection of meningitis due to mycobacterium tuberculosis," *Journal of medical microbiology*, vol. 68, no. 5, p. 755, 2019.

[158] S. Ashique, N. Mishra, S. Mohanto, A. Garg, F. Taghizadeh-Hesary, B. J. Gowda, and D. K. Chellappan, "Application of artificial intelligence (ai) to control covid-19 pandemic: Current status and future prospects," *Heliyon*, 2024.

[159] Z. Yang, P. Bogdan, and S. Nazarian, "An in silico deep learning approach to multi-epitope vaccine design: a sars-cov-2 case study," *Scientific reports*, vol. 11, no. 1, p. 3238, 2021.

[160] M. Ghaderzadeh, A. Hosseini, F. Asadi, H. Abolghasemi, D. Bashash, and A. Roshanpoor, "Automated detection model in classification of b-lymphoblast cells from normal b-lymphoid precursors in blood smear microscopic images based on the majority voting technique," *Scientific Programming*, vol. 2022, pp. 1–8, 2022.

[161] Z. Fasihfar, H. Rokhsati, H. Sadeghsalehi, M. Ghaderzadeh, and M. Gheisari, "Ai-driven malaria diagnosis: developing a robust model for accurate detection and classification of malaria parasites," *Iranian Journal of Blood and Cancer*, vol. 15, no. 3, pp. 112–124, 2023.

[162] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi, and A. H. Gandomi, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522002505

[163] M. Arquam, A. Singh, and H. Cherifi, "Impact of seasonal conditions on vector-borne epidemiological dynamics," *IEEE Access*, vol. 8, pp. 94 510–94 525, 2020.

[164] K. A. Qureshi, R. A. S. Malick, M. Sabih, and H. Cherifi, "Complex network and source inspired covid-19 fake news classification on twitter," *IEEE Access*, vol. 9, pp. 139 636–139 656, 2021.

[165] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[166] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *ACM SIGKDD Explorations Newsletter*, vol. 3, no. 1, pp. 27–32, 2001.

[167] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should always check for them)," *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, p. 6, 2004.

[168] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, p. 54, 2023.

[169] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning: Methods and Applications*, 1st ed. Springer New York, 2012. [Online]. Available: https://doi.org/10.1007/978-1-4419-9326-7

[170] Z. Sun, Y. Tao, S. Li, K. K. Ferguson, J. D. Meeker, S. K. Park, S. A. Batterman, and B. Mukherjee, "Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons," *Environmental Health*, vol. 12, no. 1, pp. 1–19, 2013.

[171] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[172] R. S. Batista, A. P. Gomes, J. L. D. Gazineo, P. S. B. Miguel, L. A. Santana, L. Oliveira, and M. Geller, "Meningococcal disease, a clinical and epidemiological review," *Asian Pacific Journal of Tropical Medicine*, vol. 10, no. 11, pp. 1019–1029, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1995764517308714

[173] N. G. Rouphael and D. S. Stephens, *Neisseria meningitidis: Biology, Microbiology, and Epidemiology*. Totowa, NJ: Humana Press, 2012, pp. 1–20. [Online]. Available: https://doi.org/10.1007/978-1-61779-346-2_1

[174] A. S. Juwairriyyah, M. A. Ameer, and P. G. Gulick, "Meningococcemia - statpearls - ncbi bookshelf," Aug 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK534849/

[175] N. Nguyen and D. Ashong. (2022) Neisseria meningitidis. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK549849/

[176] K. Du Preez, J. Seddon, H. Schaaf, A. Hesseling, J. Starke, M. Osman, C. Lombard, and R. Solomons, "Global shortages of bcg vaccine and tuberculous meningitis in children," *The Lancet Global Health*, vol. 7, no. 1, pp. e28–e29, 2019.

[177] H. Kaur, E. M. Betances, and T. B. Perera. (2024, Feb.) Aseptic meningitis. PMID: 32491344. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32491344/

[178] E. Capobianco and M. Dominietto, "From medical imaging to radiomics: role of data science for advancing precision health," *Journal of personalized medicine*, vol. 10, no. 1, p. 15, 2020.

[179] R. Hasbun, "Progress and challenges in bacterial meningitis: a review," *Jama*, vol. 328, no. 21, pp. 2147–2154, 2022.

[180] T. Flægstad, "Meningococcal meningitis," in *Disorders of Consciousness - A Review of Important Issues*. IntechOpen, 2020, [Internet]. [Online]. Available: http://dx.doi.org/10.5772/intechopen.90687

[181] R. Yogev and T. Tan, "Meningococcal disease: the advances and challenges of meningococcal disease prevention," *Human vaccines*, vol. 7, no. 8, pp. 828–837, 2011.

[182] "Meningococcal meningitis - pasteur.fr," https://www.pasteur.fr/en/medical-center/disease-sheets/meningococcal-meningitis, [Accessed 03-04-2024].

[183] B. Y. Kasula, "Ethical implications and future prospects of artificial intelligence in healthcare: A research synthesis," *International Meridian Journal*, vol. 6, no. 6, pp. 1–7, 2024.

[184] "Welcome to ELI5's documentation!; ELI5 0.11.0 documentation — eli5.readthedocs.io," https://eli5.readthedocs.io/en/latest/, [Accessed 02-01-2025].

[185] M. Frosch and M. C. J. Maiden, Eds., *Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management*. Hoboken, NJ: Wiley-Blackwell, November 2008. [Online]. Available: https://www.wiley.com/en-us/Handbook+of+Meningococcal+Disease%3A+Infection+Biology%2C+Vaccination%2C+Clinical+Management-p-9783527614455

[186] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced big data classification: outcomes and challenges," *Complex & Intelligent Systems*, vol. 3, pp. 105–120, 2017.

[187] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, Inc., Oct. 2022. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

[188] V. Vishwarupe, P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, "Explainable ai and interpretable machine learning: A case study in perspective," *Procedia Computer Science*, vol. 204, pp. 869–876, 2022.

[189] M. Pizza and R. Rappuoli, "Neisseria meningitidis: pathogenesis and immunity," *Current opinion in microbiology*, vol. 23, pp. 68–72, 2015.

[190] A. P. Hrishi and M. Sethuraman, "Cerebrospinal fluid (csf) analysis and interpretation in neurocritical care for acute neurological conditions," *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, vol. 23, no. Suppl 2, p. S115, 2019.

[191] E. M. Mascini and R. J. Willems, "Chapter 166 - streptococci, enterococci and other catalase-negative cocci," in *Infectious Diseases (Third Edition)*, 3rd ed., J. Cohen, S. M. Opal, and W. G. Powderly, Eds. London: Mosby, 2010, pp. 1645–1659. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323045797001660

[192] M. W. Bijlsma, M. C. Brouwer, E. S. Kasanmoentalib, A. T. Kloek, M. J. Lucas, M. W. Tanck, A. van der Ende, and D. van de Beek, "Community-acquired bacterial meningitis in adults in the netherlands, 2006–14: a prospective cohort study," *The Lancet infectious diseases*, vol. 16, no. 3, pp. 339–347, 2016.

[193] M. Auburtin, M. Wolff, J. Charpentier, E. Varon, Y. Le Tulzo, C. Girault, I. Mohammedi, B. Renard, B. Mourvillier, F. Bruneel *et al.*, "Detrimental role of delayed antibiotic administration and penicillin-nonsusceptible strains in adult intensive care unit patients with pneumococcal meningitis: the pneumorea prospective multicenter study," *Critical care medicine*, vol. 34, no. 11, pp. 2758–2765, 2006.

[194] D. van de Beek, M. Brouwer, R. Hasbun, U. Koedel, C. G. Whitney, and E. Wijdicks, "Community-acquired bacterial meningitis," *Nature reviews Disease primers*, vol. 2, no. 1, pp. 1–20, 2016.

[195] "Meningitis - who.int," https://www.who.int/news-room/fact-sheets/detail/meningitis, [Accessed 03-04-2024].

[196] M. C. Brouwer, A. R. Tunkel, and D. van de Beek, "Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis," *Clinical microbiology reviews*, vol. 23, no. 3, pp. 467–492, 2010.

[197] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[198] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[199] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, 2022.

[200] A. A. Biswas, "A comprehensive review of explainable ai for disease diagnosis," *Array*, vol. 22, p. 100345, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005624000110

Bibliography

*****

Self References

[1] A. Messai, A. Drif, A. Ouyahia, M. Guechi, M. Rais, L. Kaderali, and H. Cherifi, "Towards XAI agnostic explainability to assess differential diagnosis for meningitis diseases," *Machine Learning: Science and Technology*, vol. 5, no. 2, pp. 025052, 2024.

[2] A. Messai, A. Drif, A. Ouyahia, M. Guechi, M. Rais, L. Kaderali, and H. Cherifi, "Transparent AI Models for Meningococcal Meningitis Diagnosis: Evaluating Interpretability and Performance Metrics," in *Proc. 2024 IEEE 12th Int. Conf. on Intelligent Systems (IS)*, pp. 1–8, 2024.

[3] A. Messai, A. Drif, A. Ouyahia, M. Rais, M. Guechi, L. Kaderali, and H. Cherifi, "A comprehensive investigation into the ranges of laboratory tests present in cerebrospinal fluid across various types of meningitis within different age categories," *BATNA Journal of Medical Sciences*, vol. 11, no. 3, pp. 347–356, 2024.

[4] A. Messai, A. Drif, A. Ouyahia, M. Guechi, M. Rais, L. Kaderali, and H. Cherifi, "ARTIFICIAL INTELLIGENCE SYSTEM FOR INFECTIOUS DISEASE DIAGNOSIS," Poster presented at the 7th Study Day on Medical Physics, 2024. [Available upon request].

[5] A. Messai, A. Drif, A. Ouyahia, M. Guechi, M. Rais, L. Kaderali, and H. Cherifi, "An Interpretable Artificial Intelligence System for Infectious Disease Diagnosis," Poster presented at the 1st Study Day on Developing Scientific Research Skills Using Artificial Intelligence Techniques, 2024. [Available upon request].