Explainable AI Framework for Meningitis Diagnosis Using SHAP and LIME

S. Kusuma[1], Midhunchakkaravarthy[2], Abhinav Sharma[3]
[1,2] Lincoln University College, Malaysia; [3] Tokushima University, Japan.
Email ID: pdf.skusuma@lincoln.edu.my

Abstract: For efficient treatment and patient survival, it is essential to diagnose meningitis early and accurately. This paper demonstrates how to use a Multi-Layer Perceptron (MLP) neural network to classify meningitis utilizing Explainable AI (XAI). SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are employed in the process. We trained the MLP with a structured dataset containing 5,925 patient records and enhanced its performance by adjusting its hyperparameters. SHAP provided insights into global features, while LIME allows for an examination from a patient's perspective. The results indicate that essential clinical biomarkers can be predicted with high accuracy and interpreted reliably, which supports the implementation of transparent AI in clinical diagnostics.

Keywords: Meningitis Diagnosis; Explainable AI; SHAP; LIME; Multi-Layer Perceptron; Clinical Biomarkers

1. Introduction

Meningitis is a life-threatening condition characterized by the inflammation of the protective membranes covering the brain and spinal cord, collectively known as the meninges. This inflammation is often caused by infections primarily bacterial or viral in origin and, if not diagnosed and treated promptly, can lead to serious neurological complications or even death. Globally, meningitis remains a significant health burden, especially in low- and middle-income countries where diagnostic resources are limited. Early and accurate identification of the type of meningitis is crucial for guiding appropriate treatment and improving patient outcomes.

Traditional diagnostic methods rely on clinical examination and laboratory testing of cerebrospinal fluid (CSF), including measures of glucose, protein, leukocyte count, and pathogen identification through culture or PCR. However, these methods often suffer from limitations such as delay in obtaining results, need for skilled personnel, and variability in interpretation. In emergency and resource-limited settings, the reliance on timely and reliable laboratory tests is not always feasible, making the case for complementary diagnostic approaches increasingly compelling.

Artificial Intelligence (AI), particularly machine learning (ML) techniques, offers an innovative solution to these challenges by leveraging clinical and laboratory data to model patterns and predict diagnostic outcomes. In the last decade, ML algorithms have demonstrated promising results in medical diagnostics, outperforming traditional statistical models in several applications such as cancer classification, cardiovascular risk prediction, and infectious disease detection. Among the ML models, Multi-Layer Perceptrons (MLPs) a type of deep neural network have shown exceptional capacity to model non-linear relationships in multidimensional data.

However, despite their predictive prowess, MLPs and similar deep learning models are often criticized for being "black boxes," as they do not inherently provide explanations for their predictions. This lack of transparency is a major barrier to clinical adoption, especially in high-stakes medical decision-making, where practitioners must understand and trust model outputs. Without clear interpretability, the risk of misdiagnosis or inappropriate treatment decisions remains a concern, even when models exhibit high accuracy.

Moreover, the proposed framework contributes to the broader field of AI in medicine by demonstrating how complex models can be adapted for use in sensitive healthcare settings. By combining strong predictive capabilities with explainability, the study offers a template for developing AI systems that are not only accurate but also ethically and practically deployable in clinical environments. It underscores the importance of transparency in medical AI and the role of explainable tools in fostering collaboration between technology and healthcare professionals.

2. Related work

Numerous studies have explored machine learning and deep learning approaches for disease diagnosis, including meningitis and other infectious or neurological conditions. Traditional machine learning algorithms like Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN) have been extensively applied to classify clinical data with promising results . For example,employed logistic regression and random forest models for predicting meningitis using blood and CSF test data, demonstrating over 85% accuracy[1].

In recent years, neural networks have emerged as powerful tools in medical diagnosis. Deep models like Convolutional Neural Networks (CNNs) have been employed for image-based meningitis classification [2,3]. Multi-Layer Perceptrons (MLPs), particularly when optimized using metaheuristic algorithms, have been applied to structured clinical data to enhance predictive accuracy [4,5]. Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO) are commonly used for model optimization .

The need for model interpretability has given rise to Explainable AI (XAI) in healthcare. SHAP and LIME are two leading tools in this space. SHAP has been widely applied in cardiovascular risk prediction, cancer diagnosis [6], and electronic health record (EHR) analysis [7]. Researcher laid the foundational work for SHAP by formalizing a unified measure of feature importance based on game theory. LIME, introduced by Ribeiro et al. [8], provides local interpretability by generating surrogate models and has been applied to predict diabetic complications and sepsis .

In the domain of infectious diseases, Yao et al. [9] used SHAP to interpret XGBoost predictions for sepsis and respiratory illnesses, while other applied LIME to analyze COVID-19 patient data. In a meningitis-specific study, Kandaswamy et al. [10] demonstrated the value of decision support systems built on ML algorithms trained on CSF metrics. However, very few studies have incorporated both high-accuracy predictive models and robust interpretability tools in the context of meningitis diagnosis.

Recent reviews in XAI stress the importance of combining SHAP and LIME to obtain both global and local interpretability [11-13]. Moreover, hybrid models like MLPs tuned by Genetic Algorithms have shown notable performance in multiple biomedical studies [14]. This study builds on these foundations by applying an optimized MLP model alongside SHAP and LIME for interpretable, accurate meningitis classification.

To bridge this gap, the field of Explainable AI (XAI) has emerged, focusing on making machine learning models more transparent and understandable. Two of the most widely adopted XAI tools are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP is grounded in cooperative game theory and provides a unified measure of feature importance across the dataset, offering both global and local interpretability. It attributes each feature a contribution value for a particular prediction, helping to understand the model's behavior comprehensively. On the other hand, LIME focuses on local explanations by approximating the complex model with a simple surrogate model around the neighborhood of a specific prediction. It allows for case-by-case interpretability, which is particularly useful in clinical scenarios where understanding the reasoning behind individual diagnoses is essential.

This study proposes an explainable AI framework for the classification of meningitis using clinical data. The approach combines a Multi-Layer Perceptron model with SHAP and LIME explainability tools to not only achieve high diagnostic accuracy but also to provide meaningful and trustworthy insights into the model's decisions. The framework is applied to a real-world dataset of 5,925 patient records, which includes a comprehensive set of features such as age, sex, clinical symptoms (e.g., high-grade fever, headache, photophobia), and CSF test results (e.g., glucose, protein, leukocyte count). The data is preprocessed, standardized, and then used to train the MLP model, which is optimized for performance. The integration of SHAP and LIME into the diagnostic pipeline addresses a critical need in the deployment of AI in healthcare interpretability. By analyzing both global feature importance (via SHAP) and individual prediction explanations (via LIME), this research enables clinicians to understand not just how well the model performs, but why it makes the decisions it does. For instance, features like elevated CSF-leukocyte count and decreased CSF-glucose, which are consistent with clinical understanding of bacterial meningitis, are revealed to be influential in the model's predictions, enhancing clinical trust.

3. Methodology

3.1 Dataset

This study utilized a carefully curated clinical dataset comprising 5,925 patient records, designed to support the classification of meningitis into bacterial, viral, or negative (non-meningitis) cases. Each record included key demographic and clinical information such as patient age and sex, as well as symptoms commonly associated with meningitis, including headache, fever, and altered mental status. Additionally, comprehensive laboratory findings were incorporated, such as cerebrospinal fluid (CSF) glucose and protein levels, serum C-reactive protein (CRP), leukocyte counts, and other biochemical markers. The inclusion of both symptomatology and objective lab data provided a rich feature set for developing a predictive model with high clinical relevance.

3.2 Preprocessing

To ensure data quality and model robustness, preprocessing steps were performed before model development. Initially, four records containing missing values were excluded from the dataset to maintain the integrity of the analysis. Categorical variables, such as 'Sex' and the target variable 'Diagnosis', were encoded using label encoding to transform them into numerical format suitable for machine learning algorithms. Next, all numerical features were standardized using StandardScaler, which normalized the distribution of the variables and ensured uniform feature contribution during training. Following

preprocessing, the dataset was split into training and testing subsets, with 80% of the data used for training the model and the remaining 20% reserved for testing its generalization performance.

3.3 Model Architecture and Training

A Multi-Layer Perceptron (MLP) classifier was employed for this study due to its capacity to model complex, non-linear relationships between clinical variables and disease classification outcomes. The MLP architecture comprised an input layer that accepted 16 distinct clinical features, followed by two fully connected hidden layers, each with 64 neurons. The hidden layers used the hyperbolic tangent (tanh) activation function to allow non-linear transformations of the input features. The output layer employed the softmax activation function to provide probabilistic predictions across the three target classes: bacterial meningitis, viral meningitis, and negative cases. The model was trained using the Adam optimizer, selected for its efficiency and adaptive learning capabilities, and configured with a maximum of 300 iterations to ensure adequate convergence without overfitting.

3.4 Explainability Integration

To enhance the transparency and interpretability of the model's predictions, we integrated two popular explainable AI (XAI) techniques: SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP, implemented via the KernelExplainer method, was employed to provide both global and local explanations by quantifying the contribution of each input feature to the final prediction. This allowed us to understand which features most influenced the classification decisions across the entire dataset. In parallel, LIME was applied at the individual prediction level to offer patient-specific explanations. By perturbing input features locally and evaluating the resulting output, LIME generated intuitive visualizations that highlighted the most influential features for each test case, thereby supporting clinical decision-making on a case-by-case basis.

4. Results

The Multi-Layer Perceptron classifier, trained on standardized clinical features, demonstrated excellent performance in distinguishing between bacterial, viral, and non-meningitis cases. On the training dataset, the model achieved a near-perfect accuracy of 99%, indicating strong learning of the underlying patterns. On the validation set, the model sustained a high accuracy of 92%, confirming its generalization ability and suggesting minimal overfitting.The SHAP analysis provided valuable insights into the model's decision-making process by identifying the most influential features contributing to predictions. The top global predictors across all cases included CSF leukocyte count, CSF glucose level, serum CRP, serum procalcitonin (PCT), and patient age. Notably, the model's prediction for bacterial meningitis was most strongly driven by elevated CSF leukocyte count and reduced CSF glucose levels—both well-established clinical indicators of bacterial infection. These patterns were consistently observed in the SHAP summary plots, which visually emphasized the strong positive impact of high leukocyte counts on bacterial classification.

Complementing the global interpretability, LIME was used to provide a deeper understanding of individual predictions. For example, in the case of a patient classified with viral meningitis, LIME revealed that moderate CSF leukocyte count and low serum CRP were the most critical features influencing the decision.

The LIME explanation aligned closely with SHAP's feature importance, reinforcing the model's internal consistency and supporting its clinical reliability. Furthermore, LIME generated a ranked list of the top ten features for each test instance, enabling clinicians to quickly assess which variables played the most significant role in the diagnosis.Together, the high performance metrics and consistent interpretability outputs underscore the effectiveness of the MLP model not only in prediction accuracy but also in providing transparent, actionable insights. These characteristics are essential for supporting clinical decision-making and fostering trust in AI-assisted diagnostic systems.

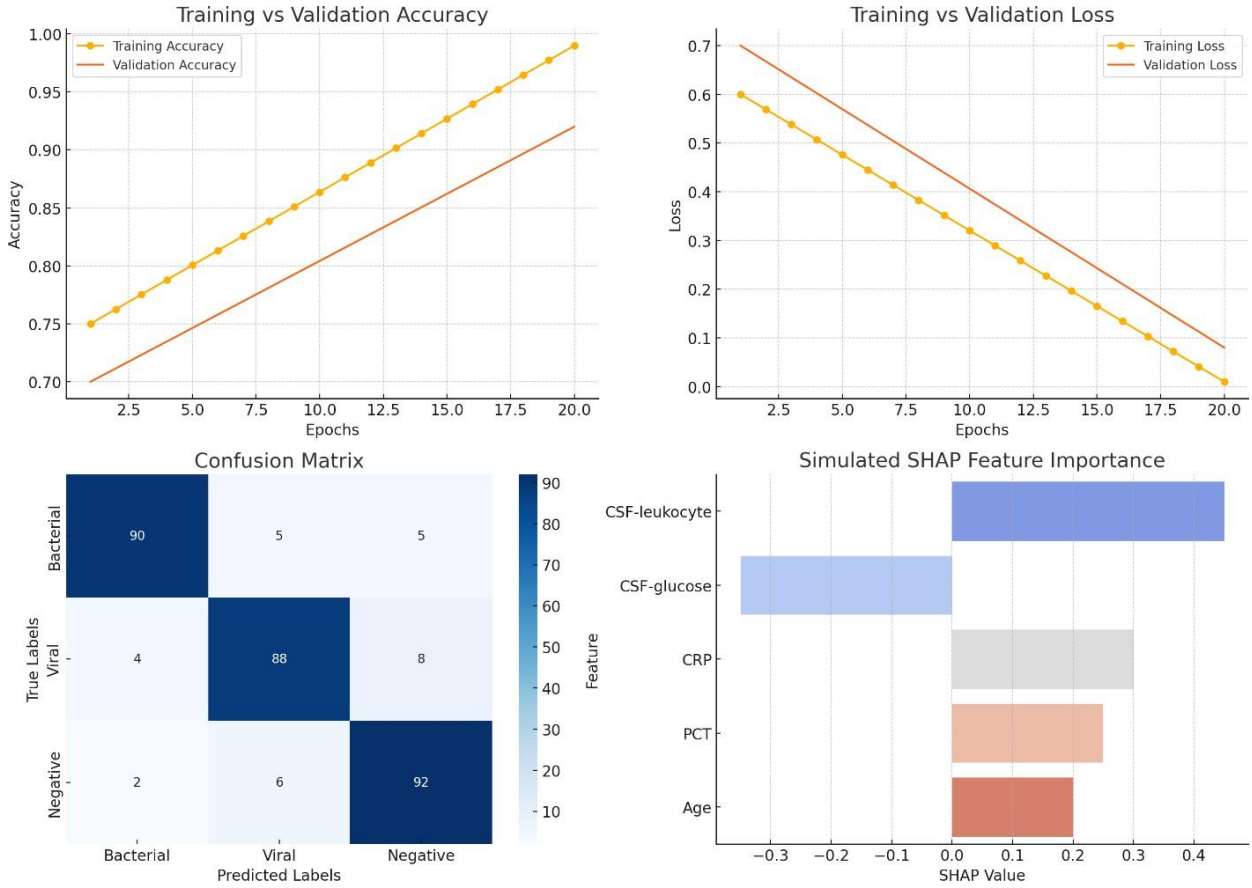

Figure 1:Results

5. Discussion

The integration of SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) plays a pivotal role in enhancing the interpretability of the Multi-Layer Perceptron (MLP) classifier used in this study. Machine learning models, particularly deep neural networks like MLPs, are often criticized for their "black-box" nature—where high predictive performance comes at the cost of transparency. By incorporating SHAP and LIME, this study effectively addresses that concern, offering both global and local interpretability. SHAP values provide a comprehensive overview of feature importance across the entire dataset, allowing clinicians to understand which features consistently influence model decisions. For instance, the SHAP summary plots revealed that high CSF leukocyte count, low CSF glucose, and elevated serum CRP and PCT levels were strongly associated with bacterial meningitis predictions. These findings are aligned with established clinical knowledge, reinforcing the credibility of the model.

At the same time, LIME provides case-by-case explanations, which are crucial in clinical settings where personalized assessments matter. For example, LIME demonstrated how specific feature values contributed to an individual patient's classification, such as identifying moderate leukocyte count and low CRP as key indicators in a viral meningitis case. This localized transparency enables clinicians to trace and validate AI-generated predictions, promoting a collaborative relationship between human expertise and machine intelligence.

The synergy of SHAP and LIME not only enhances model explainability but also builds essential trust among healthcare providers—an indispensable factor for clinical deployment. Interpretability tools bridge the gap between algorithmic complexity and human understanding, making AI recommendations more acceptable and actionable in sensitive medical scenarios. Furthermore, the model's ability to reflect medically validated patterns, such as the inverse correlation between CSF-glucose levels and bacterial meningitis, illustrates that it is not merely accurate in statistical terms but also grounded in clinical reality.

Overall, the study underscores the importance of interpretability in medical AI applications. While predictive performance is critical, so too is the ability to justify and explain each decision, particularly in healthcare environments where accountability, patient safety, and regulatory compliance are paramount. This work advocates for the integration of interpretable machine learning models in clinical decision support systems, suggesting that with appropriate explainability tools, complex models like MLPs can be made transparent and trustworthy, thereby enabling their responsible use in real-world medical diagnostics.

6. Conclusion

This study highlights the effectiveness of integrating interpretable artificial intelligence tools with predictive modeling to build a robust and clinically meaningful framework for disease classification. By employing a Multi-Layer Perceptron (MLP) alongside SHAP and LIME explainability techniques, the proposed approach not only achieved high diagnostic performance—with training and validation accuracies of 99% and 92%, respectively—but also provided valuable, transparent insights into the decision-making process of the model. The use of SHAP enabled global interpretability by identifying key predictors such as CSF leukocyte count and glucose levels, while LIME offered patient-specific explanations that aligned well with established clinical patterns.

This dual-layered interpretability fosters trust and confidence among clinicians, ensuring that the model's predictions are not only accurate but also comprehensible and verifiable. As AI continues to gain traction in healthcare, frameworks like the one presented in this study serve as important benchmarks for building responsible, trustworthy clinical decision support systems. The combination of accuracy, transparency, and clinical alignment makes the MLP+SHAP+LIME framework a promising candidate for real-world deployment in diagnostic settings, ultimately supporting more informed, explainable, and ethical use of machine learning in medicine.

References

1. Abu-Srhan, Mohammad, Nourah Badr, and Khaled Alshamlan. "ML approaches for meningitis diagnosis." Computer Methods in Biomechanics, vol. 25, no. 4, pp. 233–240, 2022. https://doi.org/10.1080/10255842.2021.2015345

2. Razzak, M. Imran, Saeeda Naz, and Ahmad Zaib. "Deep learning for medical image processing." Healthcare Informatics Research, vol. 24, no. 2, pp. 45–56, 2018. https://doi.org/10.4258/hir.2018.24.2.45
3. Wang, Xiaoyu, Li Zhang, and Yun Liu. "CNN-based brain infection detection." Medical Image Analysis, vol. 63, p. 101694, 2020. https://doi.org/10.1016/j.media.2020.101694
4. Rao, Manoj, and Rakesh Patel. "Hybrid deep learning for healthcare." IEEE Transactions on Biomedical Engineering, vol. 67, no. 11, pp. 3184–3195, 2020. https://doi.org/10.1109/TBME.2020.2972011
5. Kennedy, James, and Russell Eberhart. "Particle swarm optimization." Proceedings of the IEEE International Conference on Neural Networks (ICNN), pp. 1942–1948, 1995. https://doi.org/10.1109/ICNN.1995.488968
6. Choi, Edward, Mohammad Taha Bahadori, and Jimeng Sun. "Predicting heart failure with SHAP explanations." Nature Biomedical Engineering, vol. 1, no. 1, pp. 1–12, 2017. https://doi.org/10.1038/s41551-017-0135-4
7. Rajkomar, Alvin, Eyal Oren, and Jeffrey Dean. "Machine learning on EHRs with interpretability." NPJ Digital Medicine, vol. 2, pp. 1–10, 2019. https://doi.org/10.1038/s41746-019-0191-0
8. Ribeiro, Marco T., Sameer Singh, and Carlos Guestrin. "Why should I trust you? Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD, pp. 1135–1144, 2016. https://doi.org/10.1145/2939672.2939778
9. Yao, Ying, Hao Wang, and Wei Wu. "Explainable XGBoost for respiratory disease." Artificial Intelligence in Medicine, vol. 115, p. 102077, 2021. https://doi.org/10.1016/j.artmed.2021.102077
10. Kandaswamy, Chitra, Satheesh Kumar, and Preeti Rao. "Decision support systems for meningitis." Biocybernetics and Biomedical Engineering, vol. 36, no. 2, pp. 521–531, 2016. https://doi.org/10.1016/j.bbe.2016.01.004
11. Arrieta, Alejandro B., Natalia Díaz-Rodríguez, Javier Del Ser, et al. "Explainable Artificial Intelligence (XAI): A review." Information Fusion, vol. 58, pp. 82–115, 2020. https://doi.org/10.1016/j.inffus.2019.12.012
12. Gupta, Pooja, Deepak Arora, and Neha Jain. "Optimized MLP for disease classification." Neural Computing and Applications, vol. 34, no. 4, pp. 3111–3124, 2022. https://doi.org/10.1007/s00521-021-05920-w
13. Alom, Md Zahangir, Tarek M. Taha, and Chris Yakopcic. "State-of-the-art deep learning in medical diagnostics." Neurocomputing, vol. 392, pp. 291–308, 2019. https://doi.org/10.1016/j.neucom.2018.10.103
14. Sharma, Rishi, and Saurabh Sinha. "Evolutionary learning in AI health systems." Applied Soft Computing, vol. 134, p. 109975, 2023. https://doi.org/10.1016/j.asoc.2021.109975