Gallstone Disease Prediction with Explainable Artificial Intelligence

Augustus Clark Raphael P. Rodriguez
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
aprodriguez7@up.edu.ph

James Angelo R. Dela Cruz
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
jrdelacruz@up.edu.ph

Harry William R. Acosta II
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
hracosta@up.edu.ph

Jasper Anthony G. Perillo
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
jgperillo@up.edu.ph

Ma. Sheila A. Magboo
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
ORCID: 0000-0002-6221-7892

Vincent Peter C. Magboo
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
ORCID: 0000-0001-8301-9775

***Abstract*—Gallstone Disease is a concerningly common gastrointestinal disease worldwide that can lead to further complications if left untreated. Hence, early and accurate diagnosis is essential in improving patient quality of life. In recent times, development has progressed in Gallstone diagnosis, allowing for relatively accurate diagnoses; however, high costs and inaccuracies in specific demographics continue to impede its reliability. This study proposes the use of a data-driven machine learning risk prediction model leveraging a dataset comprising a broad range of clinical and demographic variables. The study examined seven machine learning models: Random Forest, Support Vector Machines, Gradient Boosting, Logistic Regression, AdaBoost, Decision Trees, and Bagging Classifier. The results showed that the Gradient Boost model exhibited the best performance among the models evaluated. The study incorporates Explainable Artificial Intelligence, augmented upon the best-performing model, to enhance transparency and improve understanding and trust across a spectrum of model complexities. Similarly, the study validated the effectiveness of integrating machine learning models augmented with explainers as an automated and reliable method for handling Gallstone classification tasks and risk assessment, as well as for integrating such tools into the clinical decision-making process.**

***Index Terms*—Gallstone Disease, Explainable AI, Machine Learning, Prediction, Clinical Decision Support**

I. Introduction

Gallstone disease (GSD) is a common gastrointestinal disease worldwide. It is the term used for the development of calculi in the gallbladder and other parts of the biliary tract. The stones frequently result in cholecystitis, pancreatitis, and biliary tract obstruction. Recent epidemiological studies found gallstones in about 6.1% of the global population, showing greater susceptibility among South Americans, females, and older people compared to Asian people [1].Multiple known risk factors contribute to the pathogenesis of GSD. They include lifestyle, sex, age, metabolic syndrome, and obesity. A study on Chinese adults has shown that metabolically healthy and unhealthy obesity were at the highest risk. [2]. A study in Taiwan also confirmed that central obesity, measured by the waist-height ratio, predicts GSD, especially among females [3].

There is an advancement in predictive modeling aimed at enhancing the early detection of GSD through non-invasive techniques. A recent study has developed a machine-learning (ML) model that utilizes bioimpedance and laboratory examination, achieving an accuracy of approximately 85.42% in gallstone prediction. The primary predictive factors were vitamin D, C-reactive protein, total body water, and lean body mass [4]. Despite these advancements, population-specific models are still necessary to enhance early detection and intervention programs. This study aims to develop and validate a risk prediction model for gallstone disease based on a moderately sized dataset with a broad range of clinical and demographic variables. To enable proactive GSD management and alleviate the health burden of the disease, this study tries to identify the primary predictors and build an accurate model.

II. Relevant Literature

One research conducted discovered that by utilizing ridge regression-based machine learning models, the presence of hepatocellular carcinoma in individuals with chronic viral hepatitis indicates the potential use of such models in early intervention strategies [5]. Researchers have found that machine learning (ML) algorithms based on hemoglobin level parameters and specific gravity can predict chronic kidney disease (CKD) with a detection rate exceeding 97%, which is clinically significant [6]. A study demonstrated that ML plays a pivotal role in cervical cancer prediction by using imbalanced data and various sampling methods [7]. Designing

models to handle unbalanced data effectively results in a higher detection rate of high-risk individuals. These findings demonstrate the effectiveness and applicability of ML methods in various medical disorders, highlighting their role in the development of modern healthcare.

Several studies have utilized machine learning to enhance gallstone disease prediction by incorporating clinical and laboratory data. A model has been developed based on bioimpedance and metabolic markers from 319 subjects, identifying vitamin D deficiency, elevated C-reactive protein (CRP), total body water, and lean mass as critical predictors. Their gradient boosting classifier achieved an accuracy of 85.42%, comparable to traditional ultrasonography, but with the added advantages of being non-invasive and cost-effective. Additional findings suggest a connection between gallstone formation and systemic inflammation, obesity-related factors, and alterations in body composition, including bone mass and the total body fat ratio. These approaches highlight the value of integrating physiological, biochemical, and demographic features to improve the early detection of gallstones. The study's results demonstrate that incorporating comprehensive datasets into machine learning models for gallstone prediction is crucial. These integrative methodologies are valuable in improving diagnostic capabilities.

Integrating Explainable Artificial Intelligence (XAI) within healthcare systems meets the need for transparency and confidence in AI-driven decisions. Traditional AI models often act like a "black box" and tend to be uninterpretable, making it difficult for clinicians to determine the cause of some predictions. Healthcare professionals can validate and trust AI predictions using XAI techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These XAI techniques provide insights into the decision-making processes of the model [8]. This transparency is significant, as it helps identify potential biases and ensures that AI recommendations are aligned with ethical standards and medical expertise, thereby facilitating greater adoption in clinical settings.

In a systematic review, a study noted the use of XAI in neurodegenerative disease prediction, highlighting the application of SHAP to identify the most significant cognitive and imaging features related to Alzheimer's disease, which enables early and precise diagnosis [9]. Another study also used SHAP and LIME on breast cancer prediction models. They demonstrated how explainers facilitated increased transparency by identifying the most informative biomarkers that resulted in a patient's diagnosis [10]. These studies suggest the potential applicability of XAI in enhancing clinical trust and decision-making precision in machine learning models used in the healthcare industry. Although the use of XAI in gallstone disease research is in its infancy, it has potential. Few experiments have been conducted specifically using SHAP and LIME to predict and classify gallstone disease, despite the increasing number of studies utilizing machine learning for medical diagnosis. Researchers have developed machine learning models to predict the development of gallstones using various clinical and demographic variables. Using XAI on these models can render them more interpretable, allowing clinicians to determine which factors are most significant in predicting gallstone risk. More individualized patient care and targeted prevention interventions may result from this information. The application of XAI in gallstone research can lead to more transparent and efficient diagnostic tools as the field of research develops. To help clinicians understand what features (e.g., demographics, laboratory work, symptoms) contribute the most to predictions, the current study uses SHAP and LIME to explain model predictions in gallstone classification. This study fills a previously underrepresented field within medical research by maximizing the transparency of machine learning models and their applications in diagnosis.

With their strong classification performance, machine learning algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, and Logistic Regression have become widely used throughout the healthcare industry. For example, a study that utilized explainable machine learning-based prediction demonstrated the accuracy of such models in predicting diabetic nephropathy [11]. Another study demonstrated how ensemble techniques, such as Gradient Boosting Decision Trees, can effectively handle tabular medical data with high accuracy [12]. Such studies affirm the use of these algorithms in gallstone disease prediction, where imbalanced datasets and numerous patient factors are typical liabilities. Through the elimination of redundant input data, feature selection techniques, therefore, play a significant role in making models more accurate. Among the most common methods used are correlation analysis and ANOVA F-value, which have been effective in improving diagnostic models in different medical studies. Through a focus on the most significant clinical factors, a study involving feature selection for heart disease demonstrated that utilizing these methods in predicting heart disease yielded more interpretable and practical models [13]. Through the improvement of model transparency in medical use, XAI techniques such as SHAP and LIME are becoming increasingly vital. However, a review on XAI using LIME and SHAP noted their significance in improving clinician trust and comprehension in complex models [14]. On the other hand, another study on XAI successfully utilized both tools in diabetes prediction to identify key features that impact model decisions [15]. Combining SHAP and LIME realizes the need for interpretable outcomes in gallstone disease, making predictions accurate and comprehensible to medical professionals.

III. Materials and Methods

The study follows a structured pipeline composed of a defined set of stages outlined in Figure 1.The researchers downloaded the dataset from the archive provided by the UC Irvine Machine Learning Repository. Exploratory Data Analysis was employed to understand the dataset and prepare it for further analysis and evaluation. Categorical features were encoded using one-hot encoding, while numerical features, such as bioimpedance and laboratory measurements, were standardized to ensure uniform scales during training.

Pre-processing steps included verifying missing values and segregating features.

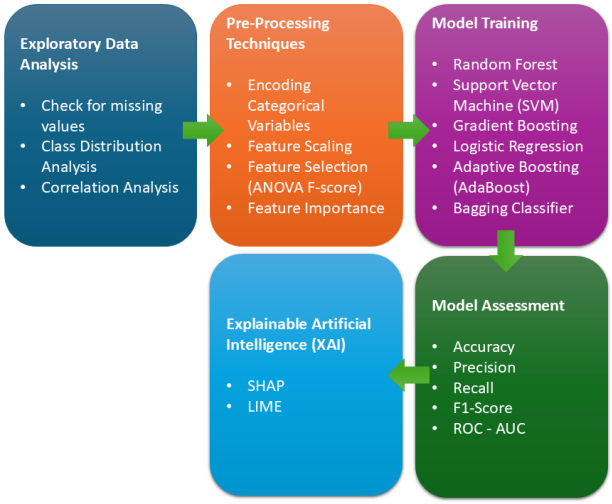

Fig. 1. Machine Learning Pipeline for Gallstone Disease Prediction Study

A. Dataset Specification

This study sourced the dataset from a publicly available repository containing clinical and laboratory data for the classification of gallstone disease. It consists of 319 patient records with no missing values. Each instance comprises multiple real-valued features derived from clinical examinations and laboratory tests, with the target variable labeled "Gallstone Status," indicating whether gallstones are present or absent. Initial exploratory data analysis confirmed the dataset's balanced class distribution.

B. Pre-processing Steps

Before modeling, the dataset underwent a series of preprocessing steps. First, the researchers performed data quality checks to ensure the absence of missing values and duplicate records. Numeric features were normalized using standard scaling to transform the values to have zero mean and unit variance, improving model convergence. Since the dataset was balanced, the researchers deemed oversampling techniques, such as SMOTE, unnecessary and omitted them. Exploratory data analysis identified outliers via boxplots, but the researchers retained them in the dataset due to their clinical relevance, preserving the full spectrum of patient variability.

C. Feature Selection and Class Imbalance Handling

The researchers perform feature selection using the ANOVA F-value method, implemented via `SelectKBest`, to reduce dimensionality and improve model performance. This technique evaluates the variance of each feature between the two classes and selects the subset with the highest discriminative power. The number of selected features was optimized through cross-validation experiments to strike a balance between performance and model interpretability.

D. Model Configuration and Machine Learning Models

The study employed a 70-30 stratified split to divide the dataset into training and testing sets, ensuring a balanced representation of the target classes. Numerical features were standardized using standard scaling. The researchers standardized numerical features using standard scaling and performed feature selection using the ANOVA F-value method with the `SelectKBest`. Applying a threshold at the 75th percentile to retain the most relevant features. The researchers evaluated seven machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Logistic Regression (LR), AdaBoost (AB), Decision Tree (DT), and Bagging Classifier (BC). Models were trained using 5-fold stratified cross-validation, a process in which the training dataset is divided into five equally sized subsets, maintaining the same class distribution in each fold. It effectively reduces the risk of overfitting and offers a more reliable estimate of how the model will perform on unseen data. Performance was assessed through multiple metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic (AUC), providing a comprehensive understanding of classifier effectiveness.

E. Explainable AI

To facilitate interpretability and clinical trust, the study leveraged both SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to provide explanations for model predictions. SHAP computes the contribution of each feature to individual predictions, enabling both local and global interpretability. In contrast, LIME provides intuitive, sample-specific explanations by approximating the model locally with interpretable surrogates. This combined approach provides transparent insights into the decision-making process of machine learning models, enabling clinicians to understand how specific clinical and laboratory parameters affect gallstone classification outcomes. The integration of SHAP and LIME enhances model transparency and aids in bridging the gap between complex predictive algorithms and practical medical decision-making.

Results and Discussion

As mentioned previously, the study evaluated seven models—Gradient Boosting, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, Bagging Classifier, and AdaBoost—for their performance in predicting Gallstone disease. The study assessed model performance using accuracy, precision, recall, F1 score, and AUC. Table I presents the performance metrics for the evaluated models, with GB demonstrating the highest recall performance (0.75), effectively identifying the highest proportion of positive cases, a practical ability in the context of the study. LR follows closely with the highest accuracy (0.77) and AUC (0.86). RF and SVM yielded moderate performance, with RF demonstrating a good balance across its metrics. At the same time, SVM suffers from low recall (0.53) and high precision (0.94),

indicating that the model favors fewer false positives (incorrectly diagnosing a patient with GSD) while risking a high occurrence of false negatives. The next pair of models, DT and BC, also demonstrated moderate performance with consistent but lower discriminative power. Lastly, AdaBoost presented the weakest performance with no discernible compromises. Overall, the models demonstrated moderate to good predictive performance, consistent with the study's target. Models that prioritized recall, accompanied by balanced metrics, were the preferred choice for the prediction task and were thus prioritized.

ML Model	Accuracy	Precision	Recall	F1 Score	AUC
RF	0.75	0.79	0.69	0.73	0.79
SVM	0.75	0.94	0.53	0.68	0.82
GB	0.72	0.71	0.75	0.73	0.81
LR	0.77	0.81	0.69	0.75	0.86
AB	0.69	0.70	0.66	0.68	0.79
DT	0.73	0.74	0.72	0.73	0.73
BC	0.73	0.76	0.69	0.72	0.75

TABLE I
PERFORMANCE METRICS FOR THE GALLSTONE DISEASE PREDICTION

The model with the best relevant performance, GB, alongside the strongest simple baseline model, DT, was chosen to be augmented with a SHAP explainer. The GB model applied LIME to explain a positive and a negative sample. Having the best complex and simple baseline models augmented with SHAP allows the study to cover the spectrum of the models' complexity and interpretability, yielding more profound insight into model behavior and explanations. As seen in Figure 2, the two most prominent features were Lean Mass (LM) (%) and Extracellular Fluid/Total Body Water (ECF/TBW), which followed closely. Figure 3 reinforces these findings by presenting the same hierarchy of feature importance, additionally providing insight into how the value of each feature contributes to the prediction. Notable observations indicate that lower feature values for both positively impact the outcome, effectively stating an increased risk of GSD at low values. Conversely, higher feature values negatively impact the outcome, decreasing patient risk.

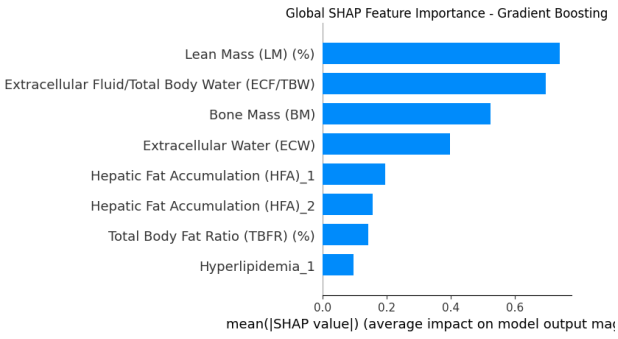

Fig. 2. SHAP Global Feature Importance Bar Plot for GB

Figure 4 illustrates the explained feature importance for the DT model, highlighting a similar yet slightly different feature importance hierarchy, with LM still dominating in importance,

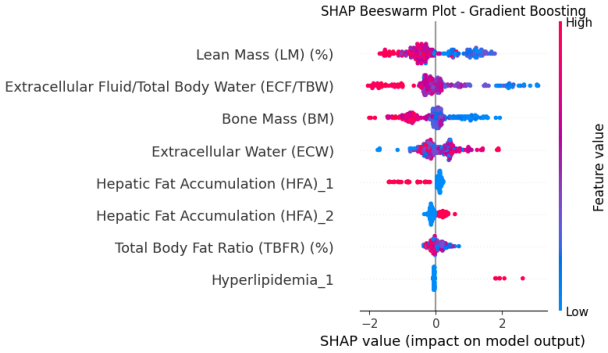

Fig. 3. Beeswarm Plot for GB

followed by Bone Mass (BM) and ECF/TBW, which have close, if not equal, importance values. Figure 5 demonstrates a similar relationship between feature value and impact on the SHAP outcome. Both models have consistently shown that low LM feature values, above all, increase GSD risk predictions, with BM and ECF/TBW holding similar importance in the DT model. Figure 5 shows a more concentrated SHAP value distribution in comparison to the relatively higher spread seen in Figure 3, highlighting the differences in each model's complexity and decision boundaries. The findings reinforce the importance of augmenting both models with SHAP, showcasing complementary perspectives that balance interpretability and predictive nuance, further enhancing understanding and trust in the model predictions.

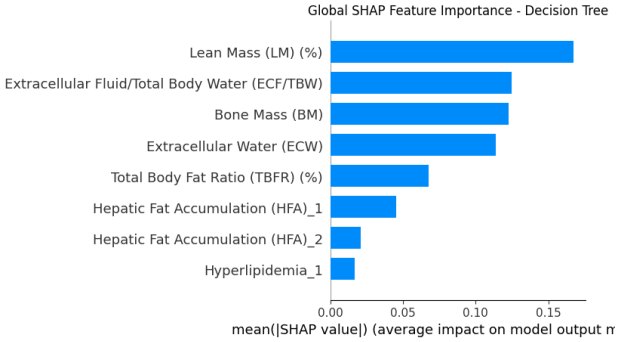

Fig. 4. SHAP Global Feature Importance Bar Plot for DT

Figure 6 shows a LIME instance on a positive sample observation from the GB test set. The model predicted the chosen instance as positive for GSD with a 73% probability, highlighting that the reduced LM and BM values contributed to its positive classification. The sample also reinforces the finding that the higher ECF/TBW feature value contributed to risk mitigation.

On the other hand, Figure 7 shows a LIME instance on a negative sample observation from the GB test set. The model predicted the chosen instance as negative for GSD with a high confidence of 94%. Showcasing a consistent relationship between the impact of the feature values on prediction and that of the global observation. For the sample in particular, the

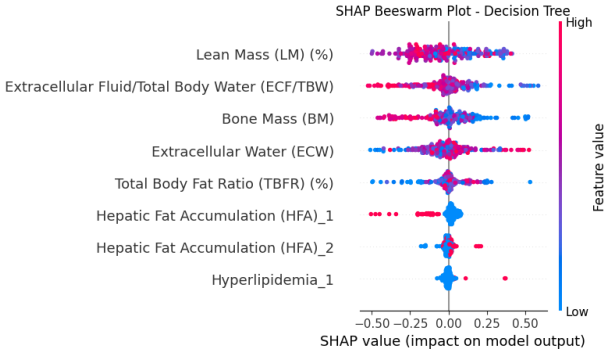

Fig. 5. Beeswarm Plot for DT

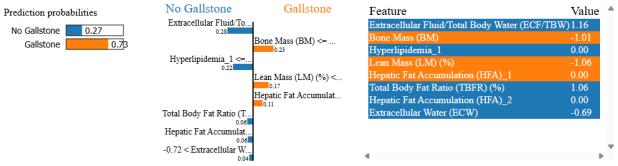

Fig. 6. LIME Output with GB on a positive sample

elevated levels of ECF/TBW and Extracellular Water (ECW) contributed to the negative classification. However, this sample had lower values for the LM and BM features, which had a positive impact on the risk prediction, yet the protective factors outweighed this influence.

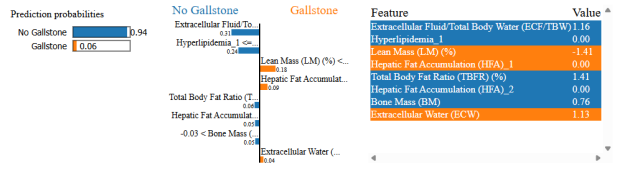

Fig. 7. Lime Output with GB on a negative sample

The interpretability analyses, using SHAP and LIME on the chosen GB and DT models, showed recurring feature importance patterns, with LM, ECF/TBW, and BM playing key roles in GSD prediction. Notably, the projected risk of GSD increased with the lower values of these features, reinforcing clinical relevance. The complementary explainability insight provided a more profound understanding of model behavior and increased trust in the predictions by elucidating the impact of features at both global and individual sample levels. These findings support the use of the GB model augmented with explainability models as an effective and highly interpretable approach for GSD risk prediction.

Conclusions and Recommendations

This study evaluated the predictive accuracy of a selection of machine learning algorithms for GSD using various physiological and clinical characteristics. Specifically, the study aimed to address the scarcity of research on interpretable gallstone prediction models and enhance transparency in model decision-making by incorporating explainable AI techniques, such as SHAP and LIME. All assessed models, with scores ranging from 70% to 80% across all recorded criteria, performed moderately well. The Gradient Boosting model yielded the best predictive performance in the context of the study's goal, presenting the highest recall at 75% whilst balancing the other metrics. SHAP and LIME augmentations provide transparent insights into feature contributions, reinforcing clinical interpretability and further supporting this. These findings are highly indicative of the effectiveness of data-driven tools in facilitating early diagnosis and personalized risk assessment, thereby making strides toward improved patient management.

Future studies should consider expanding the dataset size and diversity through relevant means, such as collecting data from multiple centers, to improve model generalizability and robustness further. Researchers may employ different feature engineering approaches by incorporating advanced data, such as imaging data, genetic markers, or longitudinal clinical measurements, which can reveal deeper patterns. Additionally, various pre-processing measures should be taken into consideration to improve model performance in conjunction with the diverse dataset. In terms of modeling, future studies should consider a different pool of models accompanied by additional ensemble and stacking techniques. Lastly, researchers may consider further external validation of the models through independent assessments or prospective trials aimed at testing their clinical utility and decision-making impact.

References

[1] X. Wang et al., "Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis," Clinical Gastroenterology and Hepatology, vol. 22, no. 8, Feb. 2024, doi: https://doi.org/10.1016/j.cgh.2024.01.051.

[2] J. Zhang et al., "Association between metabolically healthy overweight/obesity and gallstones in Chinese adults," Association between metabolically healthy overweight/obesity and gallstones in Chinese adults, vol. 20, no. 1, Mar. 2023, doi: https://doi.org/10.1186/s12986-023-00741-4.

[3] GlobalRPH, "Obesity Classification And The Risk Of Gallstones," GlobalRPH, Nov. 14, 2023. https://globalrph.com/2023/11/obesity-classification-risk-of-gallstones

[4] İrfan Esen, H. Arslan, Selin Aktürk Esen, Mervenur Gülşen, Nimet Kültekin, and Oğuzhan Özdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," Medicine, vol. 103, no. 8, pp. e37258–e37258, Feb. 2024, doi: https://doi.org/10.1097/md.0000000000037258.

[5] G. L.-H. Wong et al., "Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis," JHEP Reports, vol. 4, no. 3, p. 100441, Mar. 2022, doi: https://doi.org/10.1016/j.jhepr.2022.100441.

[6] Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," Journal of Pathology Informatics, p. 100189, Jan. 2023, doi: https://doi.org/10.1016/j.jpi.2023.100189.

[7] Mădălina Maria Muraru, Zsuzsa Simó, and László Barna Iantovics, "Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods," Applied Sciences, vol. 14, no. 22, pp. 10085–10085, Nov. 2024, doi: https://doi.org/10.3390/app142210085.

[8] Razan Alkhanbouli, Hour, F. Alhosani, and M. Can, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," BMC Medical Informatics and Decision Making, vol. 25, no. 1, Mar. 2025, doi: https://doi.org/10.1186/s12911-025-02944-6.

[9] V. Vimbi, N. Shaffi, and M. Mahmud, “Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer’s disease detection,” Brain informatics, vol. 11, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s40708-024-00222-1.

[10] S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre and M. Narvekar, ”A Study of LIME and SHAP Model Explainers for Autonomous Disease Predictions,” 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/IBSSC56953.2022.10037324.

[11] J.-M. Yin, Y. Li, J.-T. Xue, G.-W. Zong, Z.-Z. Fang, and L. Zou, “Explainable machine learning-based prediction model for diabetic nephropathy,” arXiv.org, 2023. https://arxiv.org/abs/2309.16730

[12] Y. A. Yarkın and A. Kalayci, “Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data,” arXiv.org, 2024. https://arxiv.org/abs/2410.03705

[13] B. Ahmad, J. Chen, and H. Chen, “Feature selection strategies for optimized heart disease diagnosis using ML and DL models,” arXiv.org, 2025. https://arxiv.org/abs/2503.16577

[14] A. S. Shaikh, R. M. Samant, K. S. Patil, N. R. Patil, and A. R. Mirkale, “Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain,” International Journal of Computer Applications, vol. 185, no. 45, pp. 18–23, 2023, Accessed: May 30, 2025. [Online]. Available: https://www.ijcaonline.org/archives/volume185/number45/32992-2023923263

[15] M. Panda and M. S. Ranjan, “Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP,” arXiv.org, 2023. https://arxiv.org/abs/2311.05665