Effective Prediction of Chronic Diseases: An Interpretable Approach with Explainable Artificial Intelligence

Daniyal Asif [1] **Adithya Shetty** [1] **Anwar Shamim** [1]

Abstract

In the face of escalating global health challenges, early detection of chronic diseases have become paramount to enhance patient outcomes and mitigate long-term healthcare burdens. This project addresses the urgent need for early detection of chronic diseases by developing an interpretable machine learning (ML) framework for the prediction of heart disease (HD), chronic kidney disease (CKD), and liver disease (LD), using datasets from the Kaggle. By employing a combination of ensemble learning (EL) and deep learning (DL) techniques, alongside explainable artificial intelligence (XAI) methods such as local interpretable model agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP), our approach not only enhances predictive accuracy but also ensures model transparency. This enables medical professionals to understand and trust the predictive insights, facilitating timely and informed decision-making.

Github repo: https://github.com/markbc7/DataHealers

1. Introduction

Big data is reshaping the understanding of patient histories and care in healthcare. It relies on optimizing and using data collection tools correctly. This requires active involvement from patients and healthcare providers alike. Big data sources in healthcare range from electronic health records and medical imaging to genetic sequencing and wearable device data. It differs from traditional data by being more accessible, moving quickly across the health sector's digital landscape, and coming from a varied array of sources (Azmi et al., 2022; Kaur et al., 2018).

Chronic diseases such as HD, CKD, and LD are major global health issues. HD includes conditions like coronary artery disease and arrhythmias. CKD leads to a loss of kidney function, accumulating harmful wastes in the body. LD can be inherited or result from lifestyle choices, affecting digestion and waste removal. These diseases are significant, with HD causing 20.5 million deaths a year, LD 2 million, and CKD affecting 10% of the global population (Devarbhavi et al., 2023; Tsao et al., 2023; Kalantar-Zadeh et al., 2017).

ML is transforming disease prediction by analyzing large datasets efficiently. It helps manage data, offers fast processing, and enables early disease prediction. ML is vital for preventing hospital errors, improving health policy, and detecting diseases early to save lives (Siddique & Chow, 2022; Sendak et al., 2020). However, using ML in healthcare faces challenges. High dimensional, noisy data can complicate analysis. The black box nature of ML models also poses transparency issues, making healthcare providers and patients hesitant to trust them (Miotto et al., 2018; Rahmani et al., 2021).

This study introduces an interpretable machine learning approach using XAI. It focuses on diseases like HD, CKD, and LD. Our goal is to create an XAI model that not only predicts diseases accurately but also makes the decision-making process clear. XAI techniques like LIME and permutation importance help to make predictions by our model understandable. Such transparency is crucial for building trust between healthcare providers and patients.

2. Literature Review

Recently, there has been significant progress in the application of ML techniques within the healthcare sector, particularly in the areas of early diagnosis and preventive strategies. This section will present a comprehensive literature review of studies related to HD, CKD, and LD, focusing exclusively on the latest research. Our analysis will delve into the methodologies employed, the findings obtained, and the limitations identified in each study, providing a thorough discussion of their contributions to the field.

[1] Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Daniyal Asif <Daniyal.Asif@skoltech.ru>.

Final Projects of the Machine Learning 2020 Course, Skoltech, Moscow, Russian Federation, 2020.

2.1. Heart Disease

(Srinivasu et al., 2024) conducted a study using a Kaggle HD dataset which contain 5110 observations and 11 features. They used the Synthetic Minority Oversampling Technique (SMOTE) with Tomek links for balance the data. ANOVA was used for feature selection, and various ML models including Random Forest (RF) , XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) were employed. LIME and permutation importance were used for model interpretability. The study acknowledged limitations related to the dataset size and the omission of exploring other EL and DL techniques.

(Asif et al., 2023) amalgamated three datasets from Kaggle to create a comprehensive dataset. They utilized various ML models, including ETC, RF, XGBoost, and CatBoost. To optimize the model parameters, both GridSearch CV and RandomizedSearch CV methods were employed. Their findings indicated that the ETC, optimized with Randomized Search, yielded the best results. However, the study had some limitations, including the absence of XAI techniques for model interpretability, and the failure to explore other EL techniques such as stacking, boosting, and voting.

(Bhatt et al., 2023) take a dataset from Kaggle comprising 70,000 observations with 11 features. They initiated data cleaning by identifying and removing outliers, which reduced the dataset to 59,000 observations with 9 features. Subsequently, they converted all numeric values into categorical values and applied clustering. They evaluated various machine learning models, including RF, DTs, Multilayer Perceptron (MLP), and XGBoost, and tuned their HPO using GridSearch CV. The results indicated that the MLP outperformed the other models. Limitations of the study included the removal of 11,000 outliers without exploring alternative methods and XAI techniques for model interpretability.

(Ogundepo & Yahya, 2023) analyzed the Cleveland and Statlog HD datasets, employing various machine learning models and identifying the SVM model as the most effective for the task. The study encountered limitations due to the small sizes of the datasets and did not use XAI or EL techniques, suggesting potential areas for future research enhancement.

ML approaches have been extensively utilized in various studies for predicting HD. For example, (Bizimana & El-Latif, 2023) discovered that logistic regression (LR) yielded the best outcomes for heart disease prediction. (Almustafa, 2020) found that the k-nearest neighbor (KNN) algorithm provided the highest accuracy among the reviewed studies. (Shah et al., 2020) explored DTs, naïve Bayes, KNN, and RF models, concluding that KNN offered the most reliable prediction. (Shorewala, 2021) used the LASSO algorithm for feature selection, finding a dense neural network to lead in performance. Common challenges highlighted across these studies include the reliance on small datasets, which can limit the generalizability of the findings. Moreover, there was a noted lack of exploration into EL and DL techniques, which could potentially enhance model accuracy and robustness. Additionally, the absence of XAI techniques was consistently mentioned, indicating a gap in the interpretability and transparency of the models used for heart disease prediction.

2.2. Kidney Disease

(Arif et al., 2023) developed a robust machine learning model for the early detection of Chronic Kidney Disease (CKD). They addressed missing values through iterative imputation and scaled the data using a sequential data scaling approach. Feature selection was conducted using the Boruta algorithm, resulting in 23 features being chosen for model development. Grid Search CV was utilized for HPO. They achieved a 100% accuracy KNN model on the testing dataset and further validated their model through CV. Concerns were raised regarding the complexity of the model and the potential for overfitting, given the selection of 23 features. Additionally, the absence of XAI techniques was identified as a limitation of the study.

(Swain et al., 2023) aimed to develop a machine learning model for the early detection of CKD using the UCI dataset. They tackled missing data with imputation techniques and employed a sampling method to balance the dataset, followed by normalization to prepare the data for analysis. Feature selection was conducted using the chi-square test, which resulted in a set of nine features. Their findings indicated that the SVM model performed well in detecting early CKD. The study underscored the necessity for advanced imputation methods to minimize potential information loss due to the reduction of the feature set. It also highlighted the importance of incorporating XAI techniques to enhance model interpretability.

(Ullah & Jamjoom, 2023) focused on predicting CKD by employing a decision tree-based method for missing value imputation and a filter method for feature selection. They used various ML methods and found that KNN perform well. While the methodology provided valuable insights into CKD prediction, it did not explore data scaling, hyperparameter optimization (HPO), or XAI techniques.

(Farjana et al., 8–11 March, 2023) develop a ML model for CKD prediction using the UCI dataset. They used mean imputation for missing values and employing hold-out validation. Light gradient boosting emerged as the most effective algorithm. However, the study was limited by a lack of advanced imputation techniques, outlier handling, data scaling, feature selection, model HPO, and the integration of XAI techniques.

(Islam et al., 2023) approached CKD prediction using ML algorithms, applying mean and mode techniques for imputing missing data and using recursive feature elimination and principal component analysis for feature selection. The study identified the absence of scaling methods, HPO, and XAI techniques as notable limitations.

(Md et al., 2023) aimed to predict CKD through the analysis of patients' clinical records, employing predictive mean matching for imputation and K-means for data clustering. They used the XGBoost model and SHAP for feature selection but did not explore scaling methods, HPO, or XAI techniques.

(Kaur et al., 2023) utilized ML for CKD prediction. They used Little's MCAR test for missing values and ant colony optimization method for feature selection. They used various EL method and found that Bagging to be the most effective. However, this study highlighted the absence of scaling methods, cross-validation (CV), HPO, and XAI techniques.

2.3. Liver Disease

(Md et al., 2023) developed a ML model for predicting LD, using EL with advanced preprocessing steps. Their model included comprehensive preprocessing steps such as data balancing, feature scaling, multivariate imputation for missing values, and log1p transformation for skewed data. Important feature was selected by using univariate feature importance, and correlation analysis. They used various EL techniques including gradient boosting, XGBoost, Bagging, RF, extra tree classifier (ETC), and Stacking and found that the ETC and RF show notable performance. A key limitation identified was the absence of explainable AI techniques, which could enhance the transparency and interpretability of model.

(Dritsas & Trigka, 2023) proposed an innovative approach for LD prediction by applying the SMOTE with a 5-NN to address data imbalance. Feature importance was evaluated using pearson correlation, gain ratio, and RF methods. The research explored various ML models, highlighting a voting classifier that used the RF and AdaBoost as the most effective approach. The limitations were its singular focus on data scaling methods, lack of additional efforts to optimize the accuracy of model, and the absence of XAI techniques for improved insight into the decision-making process.

ML approaches have also been used in other studies to predict LD. For example (Ayeldeen et al., 2015) used DTs to classifying the degrees of liver fibrosis. (Hashem et al., 2017) used DTs, NB, SVM, ANN, and LR to predict LD. They found that DTs performed best among others. Other studies by (Ambesange et al., 2020), (Geetha & Arunachalam, 2021), (Gogi & Vijayalakshmi, 2018), (Ma et al., 2018), and (Durai et al., 2019) have used various models and feature selection techniques to predict the LD. These studies used the UCI LD dataset and aim to improve the accuracy of model. Common challenges across these studies include a need for more exploration of data scaling methods, balancing techniques, EL and DL approaches, and the integration of XAI techniques. Addressing these limitations could further enhance model performance and interpretability, providing valuable insights into LD diagnosis and contributing to the advancement of healthcare.

3. Problem Statement

Despite the widespread adoption of ML classification algorithms in research, their application often lacks comprehensive depth. Many studies predominantly utilize basic ML models, thereby neglecting the potential benefits offered by more sophisticated ensemble methods. Furthermore, critical data preprocessing techniques, especially advanced imputation strategies, are frequently overlooked. The absence of effective feature selection and transformation methods is also a notable gap. Moreover, the integration of XAI to elucidate model decisions is often disregarded, which compromises the interpretability and trustworthiness of these models.

This study aims to directly tackle these identified challenges, proposing innovative strategies to enhance the existing research framework. The primary innovations introduced in our work include:

1. We implement advanced data preprocessing methods, alongside sophisticated feature scaling and selection techniques, to significantly improve the model's performance.

2. Our approach includes the use of various ML and DL algorithms, with a particular focus on ensemble EL such as boosting, stacking, bagging, and voting classifier, to secure superior outcomes.

3. The effectiveness of our proposed model is thoroughly evaluated using a comprehensive suite of evaluation metrics, including accuracy, precision, recall, F1-score, and curve analysis, to ensure robust validation of performance.

4. We incorporate XAI techniques to enhance the transparency and interpretability of our models, aiming to foster increased trust and understanding in the model predictions.

4. Methodology

This study outlines a detailed process for the prediction of chronic diseases. To illustrate the different phases of the proposed methodology, Figure 1 offers a schematic diagram.

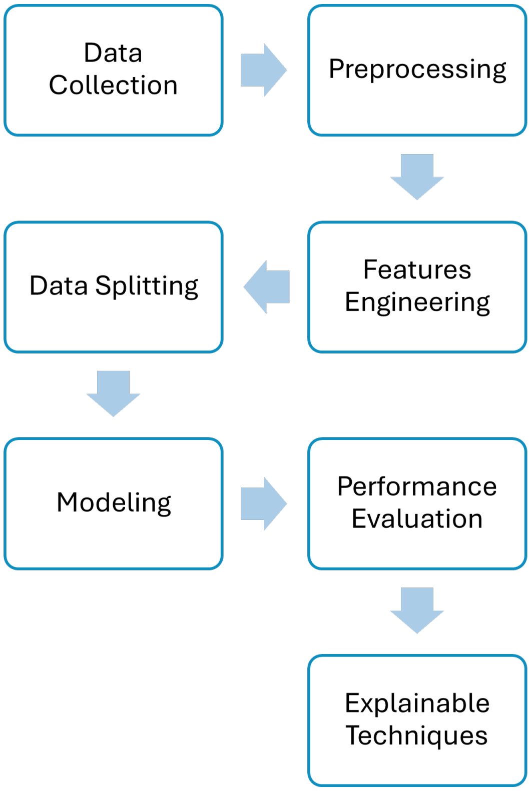


Figure 1. The proposed work flow.

4.1. Data Collection

To validate our proposed models, we obtained datasets from publicly available sources. For the HD model, we acquired data from Kaggle, consisting of 70,000 observations and 13 predictive features (HD). The target value is binary, indicating whether HD is present or the patient is normal. For the CKD model, we sourced data from the Kaggle, which includes 400 observations and 26 predictive features (CKD). The target value is also binary, denoting whether CKD is present or the patient is normal. Lastly, for the LD model, we obtained data from Kaggle, comprising 30,690 observations and 10 predictive features (LD). The target value is binary, signifying whether LD is present or the patient is normal.

4.2. Preprocessing

The medical dataset contains numerous issues, so using preprocessing enhances the quality of the data and improves the performance of the models. The preprocessing steps include cleaning the data, removing duplicates, encoding data, imputing missing values, and transforming data.

4.2.1. Data Cleaning

In the data cleaning step, we refine the data by eliminating incorrect values. We also check for duplicates and remove any that are found.

4.2.2. Data Encoding

To manage the combination of categorical and numerical values, we employed the label encoder module from the Scikit-learn library. This convert categorical features into numerical form, enabling effective integration and analysis within our models.

4.2.3. Data Imputation

To address the challenge of missing data, we employed iterative imputation. Iterative imputation is a statistical technique used to estimate missing values in a dataset by treating each feature with missing data as a target variable, while other features with complete data serve as predictors. The process begins by selecting one feature with missing values as the target and using the other complete features to predict and fill in the missing values. After imputing values for the initial target, the algorithm selects a new target feature, now incorporating the previously imputed values as part of the dataset, and repeats the prediction and imputation process. This cycle continues iteratively, refining the imputations with each round, until the algorithm converges or meets a predefined stopping criterion, ensuring a thorough and progressively refined estimation of missing values (Farhangfar et al., 2007).

4.2.4. Data Transformation

To address issues related to outliers and achieve data normalization, we employed a sequential approach as outlined in Algorithm 1. This process involves several steps: initially, robust scaling is applied, which mitigates the impact of extreme values by subtracting the median (Q_2) and dividing by the interquartile range ($Q_3 - Q_1$). Subsequently, z-score standardization is performed, resulting in a distribution that is standardized by subtracting the mean (μ) and dividing by the standard deviation (σ). Finally, min-max scaling is applied to ensure that the features fall within a specific range. This sequential process enhances robustness, standardizes the distribution, and ensures that the features are within a standardized range, thus improving model performance.

4.3. Feature engineering

Feature engineering is the process of leveraging domain knowledge to select, transform, and create the most pertinent features from raw data when constructing a predictive model using ML. This critical phase can markedly improve model performance by supplying data that is more informative and devoid of redundancy.

For the CKD dataset, which includes 26 features, we identified the most significant features. By utilizing mutual information, we selected the top 10 most important features. The details of the selected features and their descriptions are presented in Table 6.

For the HD dataset, we created some new important features. We used weight and height to calculate the body mass index (BMI) using the formula provided in Equation 1. Then, we utilized diastolic and systolic blood pressure readings to calculate the mean arterial pressure (MAP) using the formula given in Equation 2.

$$\text{BMI} = \frac{\text{weight (lb)}}{\text{height}^2(\text{in}^2)} \tag{1}$$

$$\text{MAP} = \frac{2 \times \text{Diastolic BP} + \text{Systolic BP}}{3}(\text{in}^2) \tag{2}$$

After generating new features, we applied binning to transform continuous variables like age into categorical ones for the HD dataset, enhancing our model performance and interpretability. This process facilitates a clearer understanding of the relationships between variables and outcomes. The definitive features and their descriptions are presented in Table 5. After converting numeric features into categorical variables, we applied clustering, a ML technique that groups instances based on similarity measures. We utilized k-modes clustering, which employs dissimilarity measures tailored for categorical data and replaces the means of the clusters with modes. This approach enables the algorithm to effectively handle categorical data.

For LD, given the dataset already contains only 10 features, we opted not to perform any additional feature engineering.

Algorithm 1 Sequential approach for data transformation.

```
Require: Dataset: $X$
Ensure: Transformed dataset: $X_{\text{transform}}$
  Procedure DataTransformation($X$)
  $X_{\text{robust}} \leftarrow \frac{X - Q_2}{Q_3 - Q_1}$
  $X_{\text{standard}} \leftarrow \frac{X_{\text{robust}} - \mu}{\sigma}$
  $X_{\text{transform}} \leftarrow \frac{X_{\text{standard}} - \min(X_{\text{standard}})}{\max(X_{\text{standard}}) - \min(X_{\text{standard}})}$
  return $X_{\text{transform}}$
  End Procedure
```

The features and their descriptions for the LD dataset are outlined in Table 4.

4.4. Data splitting

Data splitting is a crucial step in ML that significantly contributes to the reliable evaluation and generalization of models. This process involves dividing the dataset into training and testing subsets:

$$D = D_{\text{train}} + D_{\text{test}} \tag{3}$$

Here, D_{train} is used to train the ML model and D_{test} is employed to test the model. In this research, we allocated 70% of the data for training purposes and reserved 30% for testing.

4.5. Modeling

During the model training phase, we used various ML models and enhanced their performance using EL techniques. To optimize the hyperparameters, we employed randomized Search CV. This technique randomly selects combinations from a predefined hyperparameter space to evaluate model performance. It is efficient for large or varied hyperparameter sets as it avoids exhaustive searches.

4.5.1. Machine Learning Techniques

ML, a subset of AI, employs data-trained algorithms to create models that predict outcomes and classify information autonomously. In supervised ML, algorithms are trained with labeled data sets, providing a reference for data interpretation. Conversely, unsupervised ML involves training algorithms with unlabeled data, requiring them to identify patterns independently. Semi-supervised ML combines both approaches, starting with a small set of labeled data for initial guidance, followed by a larger volume of unlabeled data to refine the model. Reinforcement learning trains algorithms through trial and error within specific environments, using feedback to shape their development.

In our research, aimed at classifying diseases in patients, we utilize supervised ML techniques due to the need for precise and guided learning from labeled data. We apply various machine learning methods, including NB, LR, DTs, SVM, and KNN, to achieve this task.

4.5.2. Ensemble Learning Techniques

Ensemble methods are strategies designed to enhance the accuracy of predictive models by combining the predictions from multiple models rather than relying on a single one. These methods significantly boost the results by aggregating the predictions of various individual models trained on the same dataset, producing a consolidated final prediction. An illustration of this process, depicted in Figure 2, shows how the ensemble learning technique combines the outputs of individual classifiers to form a unified prediction. In our research, we employ several ensemble techniques, including RF, ETC, Bagging, Stacking, and Voting Classifier, as well as boosting algorithms like XGBoost, AdaBoost, and CatBoost.

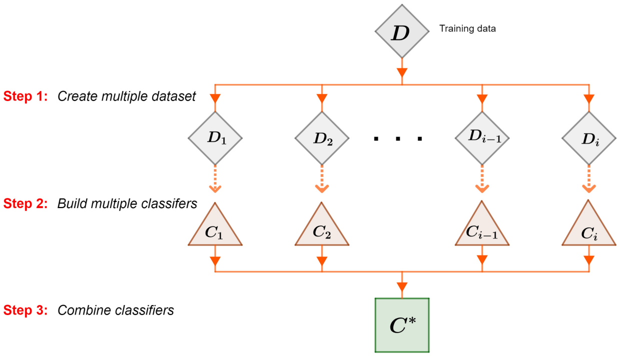

D
Training data
Step 1: Create multiple dataset
D_1
D_2
D_{i-1}
D_i
Step 2: Build multiple classifers
C_1
C_2
C_{i-1}
C_i
Step 3: Combine classifiers
C^*

Figure 2. The ensemble learning procedure.

4.6. Explainable AI Techniques

Explainable AI (XAI) refers to the methods and techniques used in artificial intelligence that provide insights into the functioning, decision-making processes, and outcomes of AI models, making them understandable to humans. This transparency is vital for assessing the model's accuracy, fairness, and potential biases, thereby fostering trust and confidence in AI systems, especially when deployed in critical decision-making scenarios. XAI aims to demystify the complex and often opaque nature of machine learning models, such as deep neural networks, enabling stakeholders to comprehend, trust, and effectively manage AI solutions, while promoting a responsible and ethical approach to AI development (Došilović et al., 2018) (Minh et al., 2022).

4.6.1. Local Interpretable Model-agnostic Explanations (LIME)

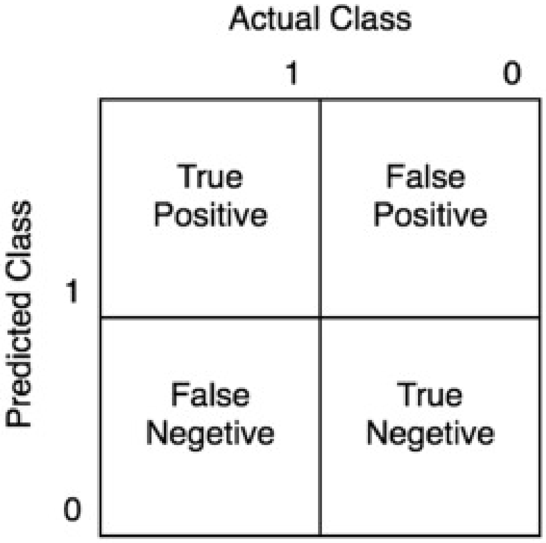
LIME is a technique designed to explain the predictions of any black-box machine learning model on a local scale, regardless of the model complexity or the type of problem it addresses (Ribeiro et al., 2016). LIME works by creating a new dataset with perturbed samples around a particular observation, using these samples to generate predictions from the original model, and then fitting a simpler, interpretable model (like linear regression) to these predictions, weighted by their proximity to the original observation. This process aims to provide insights into how the original model makes decisions at a local level, offering a transparent view of the specific features and their contributions that lead to a particular prediction.

4.6.2. Shapley Additive exPlanations (SHAP)

SHAP is a method derived from game theory (Shapley & Roth, 1988) that provides a consistent and objective way to quantify the impact of each feature on a machine learning model's prediction. By treating each feature as a "player" in a game where the prediction is the "payout," SHAP values distribute this payout among the features based on their contribution to the prediction. This is achieved by evaluating the model with all possible combinations (coalitions) of features with and without a particular feature, a process that is computationally intensive but can be approximated using techniques like Monte-Carlo sampling (E & I, 2014). SHAP values are model-agnostic, offering valuable insights into the importance and impact of features across various model types, enhancing the interpretability and transparency of machine learning predictions.

4.7. Performance Metrics

Various performance metrics, including accuracy, recall, precision, and F1-score, were utilized to evaluate the effectiveness of our model. The evaluation employed a confusion matrix, as depicted in Table 3, providing a detailed perspective on the classification results. True positives (TP) refer to cases where the disease was correctly identified in patients, while true negatives (TN) represent cases accurately classified as no disease. False negatives (FN) are instances wrongly predicted as no disease, indicating missed disease detection. Conversely, false positives (FP) are instances where the model incorrectly identified a disease presence.



		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negetive	True Negetive

Figure 3. The confusion matrix.

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+FP+TP+FN}} \times 100\% \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \tag{6}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{7}$$

5. Result and discussion

This research was conducted using Google Colab on a computer equipped with a Ryzen 7 4800H processor and 16 GB of RAM. We developed machine learning models for three diseases: CKD, HD, and LD. In this section, we present an analysis of our models' performance. Additionally, we delve into the interpretability of our models, employing LIME for local interpretability and SHAP for a comprehensive understanding of the feature influences across the models.

5.1. Performance Evaluation

5.1.1. Chronic Kidney Disease

For the CKD model, optimal hyperparameters obtained through randomized search CV with 5-fold CV and 50 splits are presented in Table 9. Table 1 showcases the performance metrics of all models. Among them, KNN yielded the best results with 99.16% accuracy, 100% precision, 98.68% recall, 98.68% F1 score, and 99.34% AUC-ROC score.

5.1.2. Heart Disease

For the HD model, optimal hyperparameters obtained through randomized search CV with 5-fold CV and 50 splits are presented in Table 8. Table 2 showcases the performance metrics of all models. Among them, soft voting classifier yielded the best results with 88.06% accuracy, 88.69% precision, 86.28% recall, 87.46% F1 score, and 88.03% AUC-ROC score.

5.1.3. Liver Diseases

For the LD model, optimal hyperparameters obtained through randomized search CV with 5-fold CV and 50 splits are presented in Table **??**. Table 3 showcases the performance metrics of all models. Among them, stacking model yielded the best results with 99.26% accuracy, 98.46% precision, 98.99% recall, 98.72% F1 score, and 99.20% AUC-ROC score.

5.2. Model Interpretability

Explainability of outcomes is paramount in interpreting ML models, a factor that gains heightened significance in the healthcare domain. Our study showcases the application of three distinct ML models, each optimized for a specific disease: the kNN for CKD, a Stacking classifier for LD, and a Soft Voting classifier for HD. The elucidation of these models predictive behavior is essential to foster trust and facilitate their integration into clinical decision-making processes.

To unravel the decision-making process at an individual prediction level, we harnessed the LIME tabular explainer for each model. Figures 4, 5 and 6 portray the LIME explanations, offering granular insights into the feature contributions for single predictions. This level of detail aids clinicians in understanding how each model weighs specific clinical features in its predictions for CKD, LD, and HD.

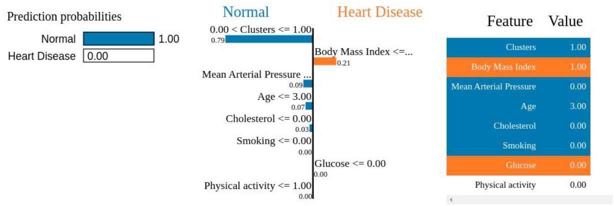


Figure 4. LIME explanations of patient having HD

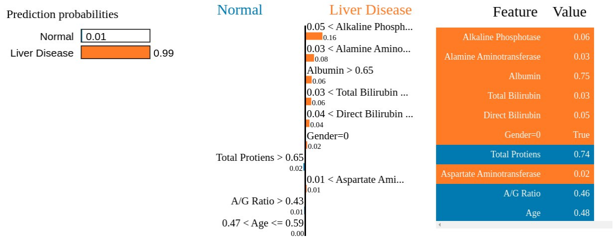


Figure 5. LIME explanations of patient having LD

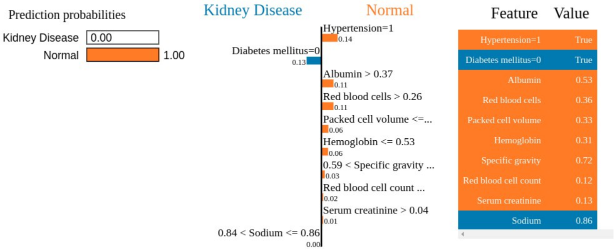


Figure 6. LIME explanations of patient having CKD

Table 1. The performance of the models for CKD

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Naïve Bayes	98.33	97.43	100	98.70	97.72
Decision Tree	99.10	98.70	100	99.34	98.86
Logistic Regression	99.10	100	98.60	99.30	99.30
K-Nearest Neighbors	99.16	100	98.68	99.33	99.34

Table 2. The performance of the models for HD

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Trees	88	88.6	86.27	87.46	88
CatBoost	87.87	87.85	86.91	87.38	87.84
Logistic Regression	81.35	77.91	85.67	81.61	81.49
Stacking	84.08	83.13	84.09	83.61	84.08
Soft Voting	88.06	88.69	86.28	87.46	88.03

Table 3. The performance of the models for LD

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	98.45	96.07	98.69	97.36	98.52
Bagging	98.84	97.47	98.57	98.02	98.76
K-Nearest Neighbors	90.74	80.43	89.97	84.93	90.51
XGBoost	97.65	93.93	98.27	96.05	97.84
Logistic Regression	62.53	42.84	87.18	57.45	69.82
Soft Voting	99.17	98.00	99.16	98.58	99.17
Stacking	99.26	98.46	98.99	98.72	99.20

Further, we employed SHAP (SHapley Additive exPlanations) to measure the global importance of features across the dataset for all three models. Figures 7, 8 and 9 showcase the SHAP values, providing a global perspective on which features exert the most significant influence on each model's output.

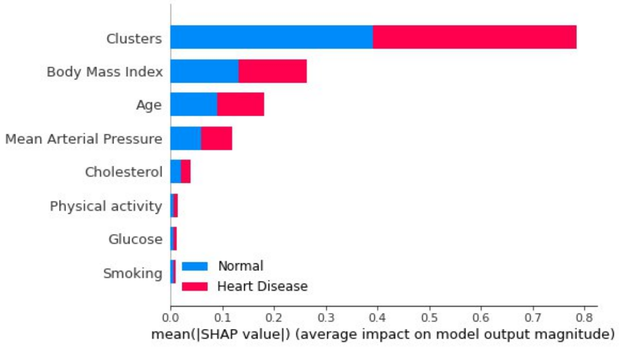


Figure 7. SHAP feature importance for HD

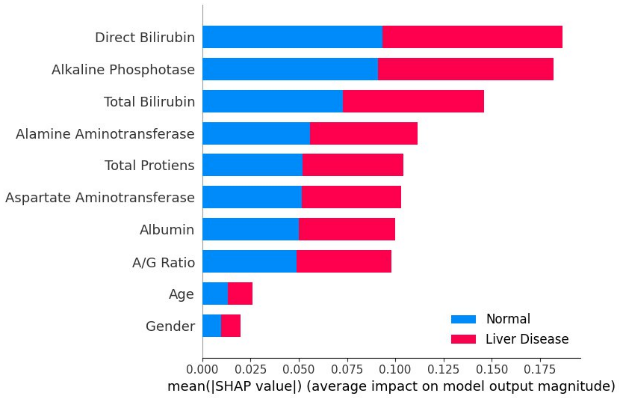


Figure 8. SHAP feature importance for LD

This multifaceted approach to model interpretability, employing both LIME and SHAP, provides a deep dive into the predictive dynamics of our models. It bridges the complex space between algorithmic predictions and human-centric clinical judgment. By doing so, it not only bolsters the healthcare professionals' trust in ML models but also paves the way for their practical adoption in medical diagnostics. The clarity and transparency achieved through these interpretability techniques are instrumental in validating the ML

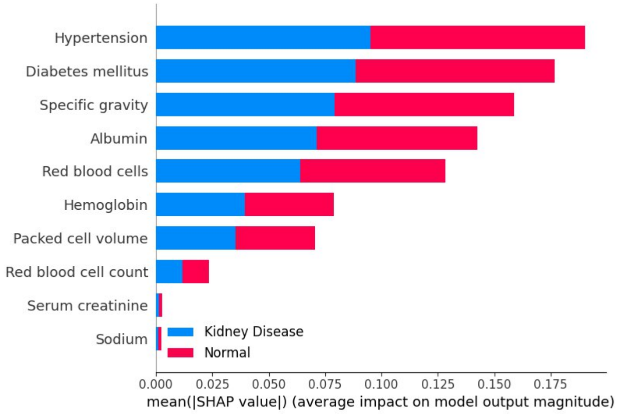


Figure 9. SHAP feature importance for CKD

models as reliable, understandable, and actionable tools in the prediction and management of CKD, LD, and HD.

6. Conclusion

In conclusion, our study advances disease prediction with ML, focusing on improving diagnostic accuracy and model interpretability. Through sophisticated preprocessing and an ensemble of classifiers, we demonstrate ML potential to enhance early detection and patient care. This approach promises to alleviate healthcare burdens and improve outcomes, marking a significant step forward in medical diagnostics.

The limitations of our study include a dataset with limited size and diversity, impacting model generalizability. Future research should expand datasets, incorporate clinical and genetic data, and conduct external validations in real-world settings to enhance accuracy and applicability. Additionally, exploring the model's adaptability to diverse data and its performance in prospective settings will be crucial. Addressing these issues will significantly improve the reliability and utility of predictive models for early liver disease detection in clinical practice.

References

Chronic kidney disease dataset. URL https://www.kaggle.com/datasets/mansoordaku/ckdisease.

Cardiovascular disease dataset. URL https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

Liver disease patient dataset. URL https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset.

Almustafa, K. M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC bioinformatics*, 21(1):1–18, 2020.

Ambesange, S., Vijayalaxmi, A., Uppin, R., Patil, S., and Patil, V. Optimizing liver disease prediction with random forest by various data balancing techniques.. In *IEEE international conference on cloud computing in emerging markets (CCEM)*, pp. 98–102. IEEE, 2020.

Arif, M. S., Mukheimer, A., and Asif, D. Enhancing the early detection of chronic kidney disease: a robust machine learning model. *Big Data and Cognitive Computing*, 7(3):144, 2023.

Asif, D., Bibi, M., Arif, M. S., and Mukheimer, A. Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms*, 16(6):308, 2023.

Ayeldeen, H., Shaker, O., Ayeldeen, G., and Anwar, K. Prediction of liver fibrosis stages by machine learning model: A decision tree approach. In *Proceedings of the 2015 Third World Conference on Complex Systems (WCCS) on Machine Learning (ICML 2000)*, Marrakech, Morocco, 2015.

Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., and Wang, G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering Physics*, 105: 103825, 2022.

Bhatt, C., Patel, P., Ghetia, T., and Mazzeo, P. Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2):88, 2023.

Bizimana, P.C.and Zhang, Z. A. M. and El-Latif, A.and Ahmed, A. An effective machine learning-based model for an early heart disease prediction. *BioMed Research International,*, 2023, 2023.

Devarbhavi, H., Asrani, S. K., Arab, J. P., Nartey, Y. A., P. E., and Kamath, P. S. Global burden of liver disease: 2023 update. *Journal of Hepatology*, 2023.

Došilović, F. K., Brčić, M., and Hlupić, N. Explainable artificial intelligence: A survey. *International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE.*, pp. 0210–0215, 2018.

Dritsas, E. and Trigka, M. Supervised machine learning models for liver disease risk prediction. *Computers*, 12: 19, 2023.

Durai, V., Ramesh, S., and Kalthireddy, D. Liver disease prediction using machine learning. *Int. J. Adv. Res. Ideas Innov. Technol*, 5(2):1584–1588, 2019.

E, S. and I, K. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*, 41 (3):647–65, 2014.

Farhangfar, A., Kurgan, L. A., and Pedrycz, W. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(5):692–709, 2007.

Farjana, A., Liza, F., Pandit, P., Das, M., Hasan, M., Tabassum, F., and Hossen, M. Predicting chronic kidney disease using machine learning algorithms. *In Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA*, pp. 1267–1271, 8–11 March, 2023.

Geetha, C. and Arunachalam, A. R. Evaluation based approaches for liver disease prediction using machine learning algorithms. In *International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4. IEEE, 2021.

Gogi, V. J. and Vijayalakshmi, M. N. Prognosis of liver disease: Using machine learning algorithms. In *International Conference on Recent Innovations in Electrical, Electronics Communication Engineering (ICRIEECE)*, pp. 1–4. IEEE, 2018.

Hashem, S., Esmat, G., Elakel, W., Habashy, S., Raouf, S., Elhefnawi, M., Eladawy, M., and ElHefnawi, M., . Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis c patients. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(3):861–868, 2017.

Islam, M. A., Majumder, M. Z. H., and Hussein, M. A. Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics*, 14:100189, 2023.

Kalantar-Zadeh, K., Jafar, T. H., Nitsch, D., Neuen, B. L., and Perkovic, V. Chronic kidney disease. *The lancet*, 398(10302):786–802, 2017.

Kaur, C., Kumar, M. S.and Anjum, A., Binda, M., Mallu, M. R., and Al Ansari, M. S. Chronic kidney disease prediction using machine learning. *Journal of Advances in Information Technology*, 14(2):384–391, 2023.

Kaur, P., Sharma, M., and Mittal, M. Big data and machine learning based secure healthcare framework. *Procedia computer science*, 132:1049–1059, 2018.

Ma, H., Xu, C. F., Shen, Z., Y., H., C., and Li, Y. M. Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in china. *BioMed research international*, 2018.

Md, A., Kulkarni, S., Joshua, C.J.; Vaichole, T., Mohan, S., and Iwendi, C. Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease. *Biomedicines*, 11:581, 2023.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pp. 1–66, 2022.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges.. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Ogundepo, E. and Yahya, W. Performance analysis of supervised classification models on heart disease prediction. *Innovations in Systems and Software Engineering*, 19(1): 129–144, 2023.

Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S.and Mehmood, Z., Haider, A., Hosseinzadeh, M., and Ali Naqvi, R. Machine learning (ml) in medicine: Review, applications, and challenges. *Mathematics*, 9(22): 2970, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. 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*, pp. 1135–1144, 2016.

Sendak, M., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K. andRatliff, W., and Balu, S. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, 10:19–00172, 2020.

Shah, D., Patel, S., and Bharti, S. K. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1:1–6, 2020.

Shapley, L. S. and Roth, A. E. The shapley value: essays in honor of lloyd s. shapley. *Cambridge University Press.*, 1988.

Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26:100655, 2021.

Siddique, S. and Chow, J. C. Machine learning in healthcare communication. *Encyclopedia*, 1(1):220–239, 2022.

Srinivasu, P. N., Sirisha, U., Sandeep, K., Praveen, S. P., Maguluri, L. P., and Bikku, T. An interpretable approach with explainable ai for heart stroke prediction. *Diagnostics*, 14(2):128, 2024.

Swain, D., Mehta, U., Bhatt, A.; Patel, H., Patel, K., Mehta, D., Acharya, B.; Gerogiannis, V., Kanavos, A., and Manika, S. A robust chronic kidney disease classifier using machine learning. *Electronics*, 12(1):212, 2023.

Tsao, C., Aday, A., Almarzooq, Z., Anderson, C., Arora, P., Avery, C., Baker-Smith, C., Beaton, A., Boehme, A., Buxton, A., and Commodore-Mensah, Y. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

Ullah, Z. and Jamjoom, M. Early detection and diagnosis of chronic kidney disease based on selected predominant features. *Journal of Healthcare Engineering*, pp. 3553216, 2023.

A. Additional Tables

Table 4. The feature description of the LD dataset.

Features	Description
Age	Age of patient in years
Gender	Gender of patient (male, female)
Total Bilirubin	Level of bilirubin in the patient's blood (mg/dl)
Direct Bilirubin	Level of bilirubin that has been conjugated in the liver (mg/dl)
Alkaline Phosphatase	Level of alkaline phosphatase in the patient's blood (IU/L)
Alamine Aminotransferase	Level of alamine aminotransferase in the patient's blood (U/L)
Aspartate Aminotransferase	Level of aspartate aminotransferase in the patient's blood (U/L)
Total Proteins	Level of proteins in the patient's blood (g/dL)
Albumin	Level of albumin in the patient's blood (g/dL)
A/G Ratio	Ratio of albumin to globulin in the patient's blood
Diagnosis	Presence of liver disease (yes, no)

Table 5. The feature description of the HD dataset.

Features	Description
Gender	Gender of patient (male, female)
Age	Age of patient in years (0, 1, 2, 3, 5, 6)
Body Mass Index	weight in kilograms divided by the square of height in meters (0, 1, 2, 3, 5)
Mean arterial pressure	average calculated blood pressure during a single cardiac cycle (0, 1, 2, 3, 5)
Cholesterol	level of Cholesterol in patient (1, 2, 3)
Glucose	level of glucose in patient (1, 2, 3)
Smoking	patient is a smoker (yes, no)
Alcohol intake	patient consumes alcohol (yes, no)
Physical activity	patient engages in physical activity (yes, no)
Diagnosis	Presence of cardiovascular disease (yes, no)

Table 6. The feature description of the CKD dataset.

Features	Description
Specific gravity	The ratio between urine density and water density
Albumin	Protein percentage in blood plasma (0, 1, 2, 3, 4, 5)
Red blood cells	Percentage of red blood cells in blood plasma (normal, abnormal)
Serum creatinine	Creatinine level in patient muscles
Sodium	Sodium mineral level in blood in mEq/L
Hemoglobin	Red protein responsible for oxygen transport in the blood in gms
Packed cell volume	Volume of blood cells in a blood sample
Red blood cell count	Count of red blood cells
Hypertension	Continuously high blood pressure condition (yes, no)
Diabetes mellitus	Impairment in insulin production or response (yes, no)

Table 7. The optimal hyperparameters for LD model.

Model	Optimal Hyperparameters
Decision Tree	min samples split: 5, min samples leaf: 1, max depth: 40, criterion: gini
Bagging	Base classifier: DTs
K-Nearest Neighbors	weights: distance, p: 1, n neighbors: 3, metric: minkowski, leaf size: 30, algorithm: auto
XGBoost	n estimators: 200, max depth: 5, learning rate: 0.05
Logistic Regression	solver: liblinear, penalty: l2, C: 1
Soft Voting	estimator: DTs, Bagging, XgBoost, KNN
Stacking	estimator: (DTs, Bagging, XgBoost, KNN), final estimator: LR

Table 8. The optimal hyperparameters for HD model.

Model	Optimal Hyperparameters
Decision Trees	criterion: entropy, min samples leaf: 2, min samples split: 5, max depth: 10
CatBoost	n estimators: 200, max depth: 5, learning rate: 0.05
Logistic Regression	solver: saga, penalty: l2, C: 0.001
Stacking	estimator: (LR, DTs, CatBoost), final estimator: DTs
Soft Voting	estimator: LR, DTs, CatBoost

Table 9. The optimal hyperparameters for CKD model.

Model	Optimal hyperparameters
Naïve Bayes	priors: None, var smoothing: 1×10^{-9}
Decision Tree	criterion: entropy, min samples leaf: 2, min samples split: 5, max depth: None
Logistic Regression	solver: liblinear, penalty: l1, C: 100
K-Nearest Neighbors	weights: uniform, p: 2, n neighbors: 5, metric: minkowski, leaf size: 30, algorithm: brute

B. Team member's contributions

Daniyal Asif

- Reviewing literature on LD and CKD
- Coding the algorithm for CKD and LD prediction
- Experimenting with model parameters on CKD and LD datasets
- Preparing this report
- Managing project administration

Adithya Shetty

- Reviewing literature on LD, CKD and HD
- Coding the algorithm for HD, CKD and LD prediction prediction
- Experimenting with model parameters on CKD, LD and HD datasets
- Preparing the GitHub Repository

Anwar Shamim

- Reviewing literature on HD
- Coding the algorithm for HD prediction
- Experimenting with model parameters on HD datasets
- Preparing the GitHub Repository

C. Reproducibility checklist

Answer the questions of following reproducibility checklist. If necessary, you may leave a comment.

1. A ready code was used in this project, e.g. for replication project the code from the corresponding paper was used.

   □ Yes.
   ☑ No.
   □ Not applicable.

   Students' comment: None

2. A clear description of the mathematical setting, algorithm, and/or model is included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

3. A link to a downloadable source code, with specification of all dependencies, including external libraries is included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

4. A complete description of the data collection process, including sample size, is included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

5. A link to a downloadable version of the dataset or simulation environment is included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

6. An explanation of any data that were excluded, description of any pre-processing step are included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

7. An explanation of how samples were allocated for training, validation and testing is included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

8. The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results are included in the report.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

9. The exact number of evaluation runs is included.

   ☑ Yes.
   □ No.
   □ Not applicable.

   Students' comment: None

10. A description of how experiments have been conducted is included.

    ☑ Yes.
    □ No.
    □ Not applicable.

    Students' comment: None

11. A clear definition of the specific measure or statistics used to report results is included in the report.

    ☑ Yes.
    □ No.
    □ Not applicable.

    Students' comment: None

12. Clearly defined error bars are included in the report.

    □ Yes.
    □ No.
    ☑ Not applicable.

    Students' comment: None

13. A description of the computing infrastructure used is included in the report.

    ☑ Yes.
    □ No.
    □ Not applicable.

    Students' comment: None