Machine Learning Approach for Predicting Heart and Diabetes Diseases

Dipen Khatri and Shreela Sapkota

***Abstract*— Heart disease and diabetes are among the leading causes of morbidity and mortality worldwide, posing significant challenges to global healthcare systems. The increasing availability of electronic medical data has paved the way for leveraging machine learning techniques to enhance disease prediction and diagnosis. In this study, we propose a comprehensive framework for predicting diabetes and heart disease using advanced machine learning algorithms. By employing a structured data science workflow, we perform extensive exploratory data analysis, data preprocessing, and feature engineering on publicly available datasets. Our approach incorporates multiple classification models, including K-Nearest Neighbors, Logistic Regression, Random Forest, and Neural Networks, alongside ensemble methods such as hard and soft voting to optimize prediction accuracy. Rigorous hyperparameter tuning and cross-validation ensure robust model performance. We evaluate the models using metrics such as ROC-AUC scores, confusion matrices, and cross-validated classification reports. Additionally, feature importance analysis reveals key contributors to disease prediction, offering valuable insights for healthcare professionals. This research demonstrates the potential of machine learning to predict diabetes and heart disease with high accuracy, aiding early diagnosis and informed medical decision-making.**

***Index Terms*— Heart disease, diabetes, machine learning, classification mdoels, feature engineering, prediction, healthcare**

I. Introduction

CHRONIC illnesses, including heart disease and diabetes, remain among the leading causes of global mortality and morbidity, with millions of individuals affected every year. These conditions not only impose significant economic burdens but also require timely diagnosis and intervention to mitigate severe complications. Traditional diagnostic methods, although effective, often rely heavily on manual processes, subjective interpretations, and limited data, leading to variability in outcomes. With the growing availability of large-scale electronic medical datasets, there is an opportunity to improve disease prediction and diagnostic accuracy through machine learning (ML) techniques.

ML has revolutionized healthcare by providing advanced tools to analyze large, complex datasets, uncover hidden patterns, and improve clinical decision-making. Key metrics such as blood pressure, cholesterol, glucose, and body mass index (BMI) play a crucial role in predicting heart disease and diabetes, necessitating the use of robust computational methods for accurate, early diagnosis and effective intervention.

This study aims to predict the likelihood of heart disease and diabetes using various ML techniques. We used publicly available datasets, specifically the UCI Cleveland Heart Disease dataset and the PIMA Indian Diabetes dataset, to develop and evaluate predictive models. The selected data sets, heart disease and diabetes data sets, offer a diverse range of characteristics, including clinical parameters and demographic information. Our work encompasses all stages of the data science process, including data pre-processing, exploratory data analysis (EDA), feature selection, Modeling, Hyperparameter Tuning, and performance evaluation, ensuring that the data sets are optimized for predictive modeling.

We explore multiple ML algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Decision Trees, among others, to identify the most effective models for these prediction tasks. In addition, we employ ensemble methods, such as voting classifiers, to combine the strengths of individual models and enhance predictive performance. Hyperparameter tuning is performed using techniques like GridSearchCV and RandomizedSearchCV to further optimize the model parameters. Models are rigorously evaluated using metrics such as ROC-AUC scores, confusion matrices, and classification reports to ensure reliability and robustness. Feature importance analysis is also performed to identify the most critical variables that influence predictions, providing information that could aid healthcare professionals in clinical decision-making.

The major contribution of the paper are as follows:

- Comparing different ML algorithms and showing how tuning their parameters can enhance performance. and one more highest accuracy of prediction.
- Creating a simple and effective framework to predict heart disease and diabetes using advanced ML techniques.
- Provide a new method to uncover hidden patterns in medical data.

This paper highlights the effectiveness of ML in predicting diabetes and heart disease and provide a scalable, cost-effective solution for healthcare systems worldwide. By com-

This paper was submitted on 7th January 2025. This work received no specific grant from any funding agency in the public, commercial, or not-for- profit sectors.

Dipen Khatri is currently a student in the Department of Computer Engineering at Kathmandu University, Dhulikhel, Nepal (e-mail: dk23021320@ku.edu.np).

Shreela Sapkota is currently a student in the Department of Computer Engineering at Kathmandu University, Dhulikhel, Nepal (e-mail: ss44021320@gmail.com).

The Associate Editor for this paper is Assistant Professor Dr. Rajani Chulyadyo, Department of Computer Science and Engineering, School of Engineering, Kathmandu University, Nepal (e-mail: rajani.chulyadyo@ku.edu.np).

bining rigorous data science techniques with cutting-edge ML models, this work aims to contribute to the growing field of AI-powered healthcare analytics and offer a practical tool for early detection and risk assessment of chronic diseases.

II. Literature Review

The prediction of heart and diabetic diseases is a vital area of research due to their increasing prevalence and significant impact on global mortality rates. ML techniques have gained substantial traction in this domain, leveraging large datasets to identify patterns and improve prediction accuracy. This survey examines various studies, methodologies, and tools applied to address these challenges.

A foundational study [1] established the interrelation between diabetes, hypertension, and cardiovascular diseases, noting that diabetes significantly increases the risk of hypertension, a major contributor to heart disease. The study also highlighted the role of ML models in predicting these comorbidities by identifying key risk factors like BMI, blood sugar levels, and cholesterol. The findings underscored the importance of integrating multiple parameters into predictive models for enhanced accuracy. Building on this, research [2] focused on the Cleveland heart disease dataset to explore clustering techniques. K-means clustering was applied to segment data based on attributes such as chest pain type(cp), resting blood pressure (trestbps), and cholesterol levels. The study also utilized visualization tools like Tableau to interpret clusters, providing a user-friendly way to analyze risk patterns. This approach laid the groundwork for integrating unsupervised learning methods in disease prediction.In another significant contribution, researchers [3] combined ensemble techniques like AdaBoost, Sigmoid SVC, and Decision Trees to develop a Voting Classifier. This model achieved accuracy rates of 88.57% for heart disease prediction and 80.95% for diabetes prediction. The study demonstrated the strength of ensemble models in handling complex datasets and highlighted their potential for real-world applications.Algorithm benchmarking was the focus of another study [4], which evaluated methods like KNN, Naïve Bayes, and Decision Trees using WEKA. The study concluded that Decision Trees achieved the highest accuracy for heart disease prediction. The use of WEKA provided a standardized platform for testing algorithms, enabling researchers to make meaningful comparisons.The study in [5] proposes an ensemble approach using various classification algorithms, including KNN, Adaptive Boosting (AB), Gradient Boosting (XB), Decision Tree, and Random Forest. After analyzing the performance of different algorithms, the research concluded that the proposed system demonstrated promising results, particularly in terms of Area Under the Curve (AUC). The best combination for prediction was found to be an ensemble of AB and XB classifiers. Feature selection and extraction techniques were central to a study [6] that proposed a predictive framework combining Chi-square tests for feature selection and Principal Component Analysis (PCA) for dimensionality reduction. The model was tested using XGBoost, achieving high accuracy rates. This research demonstrated the importance of preprocessing steps in improving model performance, especially for datasets with high-dimensional features.The role of diabetes type in predicting COVID-19 outcomes was examined in a study [7] that explored the relationship between diabetes and severe COVID-19 complications. The findings revealed that type 1 and type 2 diabetes influenced outcomes differently, emphasizing the need for personalized models that account for such variations. This study also illustrated how pandemic contexts can provide new insights into chronic disease management.Logistic regression, a fundamental ML technique, was analyzed in a study [8] that tested its performance on multiple datasets. The model achieved a precision of 96%, emphasizing its utility in binary classification tasks like disease diagnosis. However, when compared to boosting techniques, logistic regression was outperformed, with Adaptive Boosting achieving a precision of 98.8%. These findings highlighted the limitations of traditional methods in the face of advanced ensemble techniques. The study in [9] presents an approach for predicting heart disease using various algorithms, including Artificial Neural Networks (ANN), Random Forest, and Support Vector Machine (SVM). With the application of 3-fold cross-validation, the SVM algorithm achieved a maximum accuracy of 83.17%. The Decision Tree algorithm, when applied with 37 splits and 6 leaf nodes, resulted in an accuracy of 79.12%, which increased slightly to 79.54% with 5-fold cross-validation. The Random Forest algorithm achieved the highest accuracy of 85.81%, outperforming all other algorithms.The importance of cross-validation techniques was emphasized in a study [10] that utilized ROC curves to assess model performance. Random Forest, SVM, and Decision Trees were tested, with Random Forest achieving the highest accuracy of 85.81%. The study highlighted the need for rigorous evaluation metrics to ensure the reliability of predictive models. A study [11] applied KNN, Neural Networks, and SVM to an Algerian dataset, with Neural Networks achieving an accuracy of 93%. These findings highlight the potential of deep learning methods for medical predictive analytics, particularly when coupled with comprehensive datasets.

The limitations of the existing system are as follows: Variability in datasets, such as differences in size, features, and demographics, limits the generalizability of models, as many studies use public datasets that may not represent diverse populations. Overfitting is a common issue, especially in advanced models, where high accuracy on training data does not always translate to good performance on new data. Computational complexity is another challenge, as models like Neural Networks require significant resources for training and tuning. Additionally, the lack of interpretability in many ML models makes it difficult for healthcare professionals to fully trust and use these predictions. Finally, hybrid models, though promising, require substantial expertise and resources for development and deployment, making them less accessible for widespread use. The absence of standardized evaluation metrics further complicates the comparison and validation of different models. Moreover, data privacy concerns often restrict access to real-world datasets, hindering model training and testing in practical scenarios.

III. Proposed Methodology

This research presents a comprehensive ML-based approach to predict heart disease and diabetes using advanced classification algorithms and ensemble modeling techniques. The methodology is designed to address the challenges posed by raw datasets, such as missing values, feature redundancies, and imbalanced data distributions. By employing data preprocessing, feature engineering, hyperparameter tuning, and evaluation strategies, this approach ensures that the prediction models achieve high accuracy and reliability. The methodology is broadly divided into system architecture, dataset handling, and the integration of multiple ML models, which are further elaborated below.

A. System Architecture

The system architecture is designed to efficiently transform raw data into actionable insights. It begins with preprocessing steps to address missing values and prepare the data for analysis. EDA is performed to uncover patterns, and correlations, which informs the subsequent feature selection process. Feature selection is carried out using correlation analysis and domain knowledge to isolate the most impactful predictors. Once features are finalized, the data is divided into training and testing sets to evaluate the models' generalization capability effectively.

After feature selection, the data is split into training and testing subsets to ensure that the models are evaluated on unseen data. Multiple ML classifiers are trained, and the final system enables users to input their symptoms or medical details, which are processed through the trained models to generate predictions for heart disease or diabetes with high precision and accuracy.

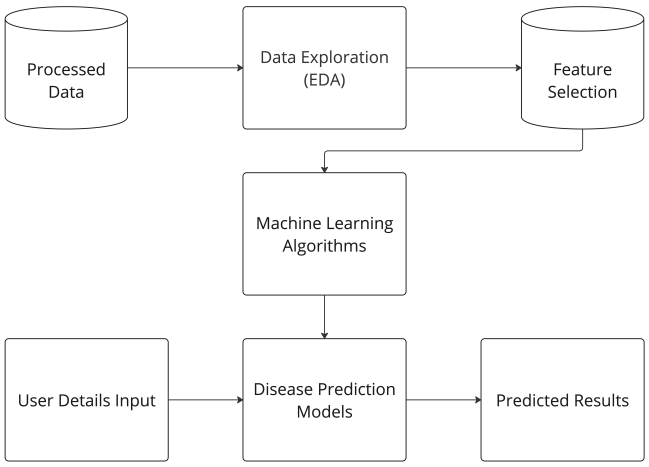

Fig. 1: System Architecture

B. Dataset handling

The datasets utilized in this research are pivotal to the success of the prediction models. We rely on publicly available datasets from repositories such as UCL and Kaggle, specifically the Pima Indian Diabetes Dataset and the UCI Heart Disease Dataset. These datasets were chosen due to their detailed and relevant medical attributes, such as glucose levels, cholesterol levels (chol), blood pressure, BMI, age, and medical history. The process of handling these datasets is depicted in Fig. 2.

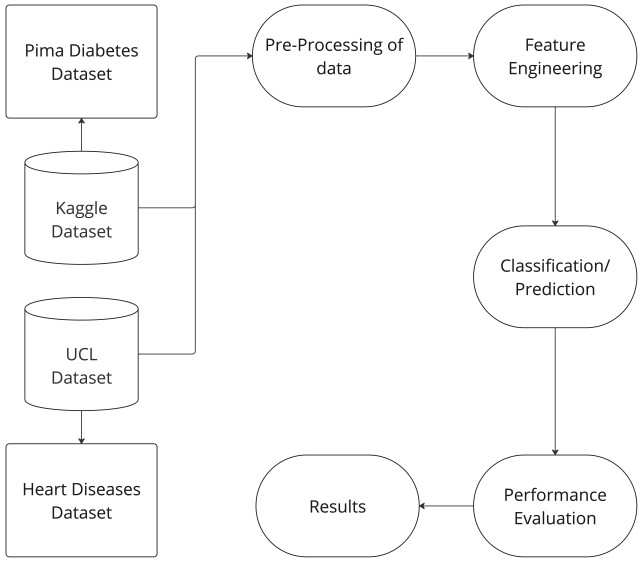

Fig. 2: Dataset Handling

C. ML Models and Ensemble Techniques

This research implements an array of ML models, both individually and in ensembles, to predict heart disease and diabetes. The base models include traditional classifiers such as KNN, Logistic Regression, Decision Trees, Random Forest, Naive Bayes (Gaussian and Bernoulli), and ANN. SVC is used with multiple kernels (linear, radial basis function, polynomial, and sigmoid) to evaluate its effectiveness across different hyperparameter configurations.

Hyperparameter optimization plays a critical role in improving the performance of these models. Techniques such as GridSearchCV and RandomizedSearchCV are applied to systematically explore the hyperparameter space of each algorithm. For instance, parameters like the number of neighbors in KNN, regularization strength in Logistic Regression, tree depth in Decision Trees, and kernel parameters in SVC are tuned to achieve the best possible accuracy and precision.

To further enhance predictive capabilities, ensemble techniques are implemented, combining the strengths of individual models. Hard Voting Classifiers determine predictions based on majority voting, while Soft Voting Classifiers weigh the probabilities generated by base models to produce the final output. Ensemble models such as Random Forest, Logistic Regression, KNN, are utilized to capture complex patterns within the data. These ensemble approaches are particularly effective in improving the robustness and accuracy of predictions by mitigating the weaknesses of individual classifiers.

The methodology integrates these components into a seamless workflow. Preprocessed data is used to train and test ML models, with hyperparameter tuning conducted to refine

their configurations. The best-performing models are selected for deployment, where users can input their medical details or symptoms. The system processes this information through the trained models to generate real-time predictions for heart disease or diabetes risk, providing accurate and actionable insights. This comprehensive approach ensures scalability, precision, and reliability, making it a valuable tool for early disease detection in healthcare.

IV. Implementation Details & Results

The methodology was implemented in Jupyter Notebook using Python libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, following a proper data science folder structure. The analysis focused on heart disease and diabetes datasets, applying EDA and advanced ML algorithms. Techniques such as correlation matrices, feature importance analysis, data visualization, and model evaluation (before and after hyperparameter tuning) were used.

In the heart disease dataset, 6 records with missing values in ca and thal were dropped, reducing the total data points to 297. For classification purposes, the original labels representing varying levels of heart disease were consolidated into binary classes: 0 for no disease and 1 for the presence of disease. Similarly, for the diabetes dataset, missing values were handled, and the features were scaled for optimal model performance.

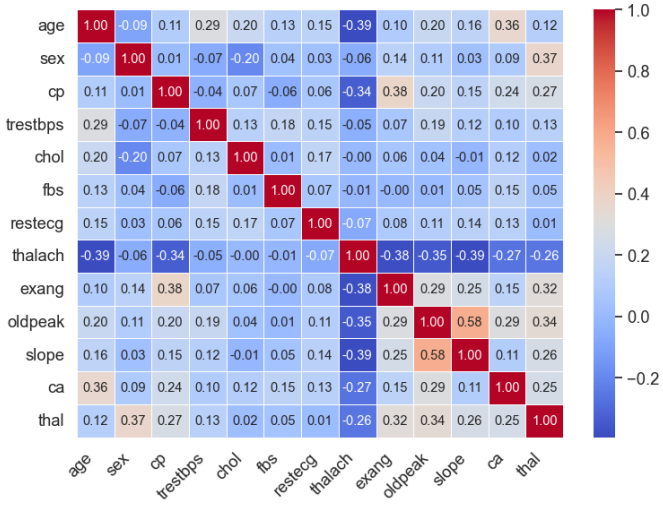

Fig. 3: Heart Correlation Matrix

The heart correlation matrix is a heatmap that displays the Pearson correlation coefficients between various features in the heart disease dataset. The diagonal elements (correlation of a feature with itself) show perfect correlation (value = 1), indicated by the darkest red. Features with positive correlations are shaded in varying intensities of red, while negative correlations are in blue. For instance, "slope" and "oldpeak" show a moderately strong negative correlation (-0.58), suggesting an inverse relationship. Similarly, "thalach" (maximum heart rate) negatively correlates with "age" (-0.39), implying that younger individuals tend to have higher heart rates. Such insights help in identifying potential relationships and redundancies among features, enabling effective feature selection and dimensionality reduction in the ML pipeline.

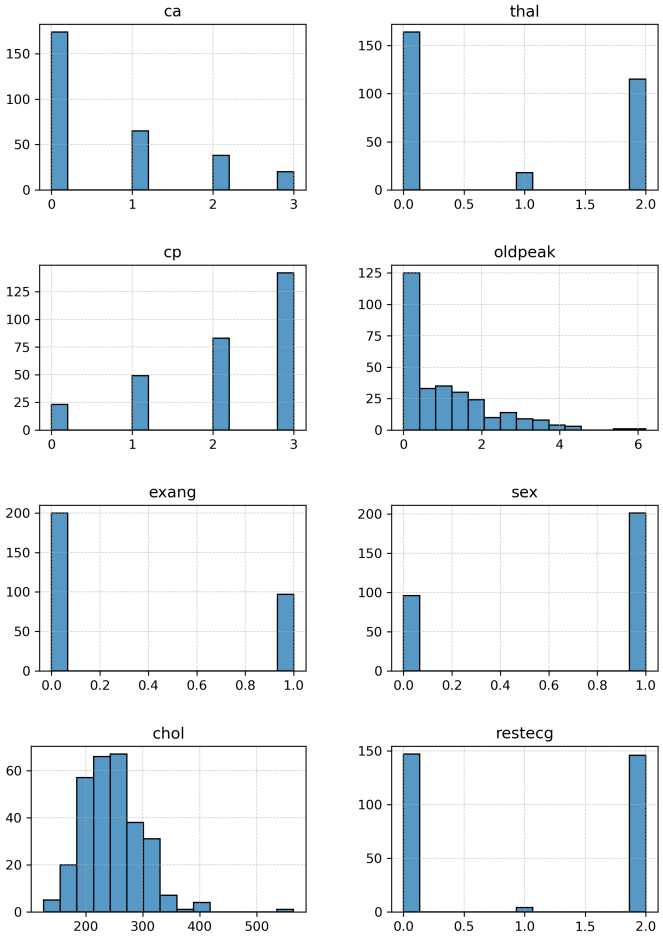

Fig. 4: Distribution of Variables in the Heart Disease Dataset

The distribution of variables in the heart disease dataset, as shown in Fig. 4, provides valuable insights into the characteristics of each feature. The histograms highlight the frequency of values for key variables. For instance, the "ca" feature (number of major vessels colored by fluoroscopy) shows a highly imbalanced distribution, with most values concentrated at 0 and fewer observations for 1, 2, and 3. Similarly, "thal" (thalassemia) has the majority of its values at 0 and 2, with sparse occurrences of 1. The "cp" feature includes four categories, with type 3 being the most frequent, indicating an imbalance in the categorical data. The "oldpeak" variable (ST depression induced by exercise) exhibits a right-skewed distribution, with most values concentrated near 0, suggesting potential outliers at higher values. The continuous feature "chol" displays a slightly skewed normal distribution, whereas binary variables like "sex" (male/female) and "exang" (exercise-induced angina) are heavily imbalanced, with one category dominating the dataset. The "restecg" variable (resting electrocardiographic results) also reflects an imbalanced distribution with distinct but unevenly distributed categories. These observations highlight the need for careful preprocessing steps, such as addressing imbalances, scaling continuous features, and encoding categorical variables, to ensure the

dataset is well-prepared for ML models.

TABLE I: Pre-Tuning Metrics for Heart Disease Dataset

Algorithm	Accuracy	Precision	Recall
KNN	58.33%	58.44%	58.33%
Logistic Regression	91.67%	92.79%	91.67%
Random Forest	88.33%	89.32%	88.33%
Decision Tree	78.33%	78.91%	78.33%
SVC	55.00%	54.48%	55.00%
ANN	73.33%	76.76%	73.33%
Naive Bayes (Gaussian)	86.67%	89.33%	86.67%
Naive Bayes (Bernoulli)	90.00%	90.14%	90.00%
VotingClassifier (Hard)	83.34%	85.67%	83.34%
VotingClassifier (Soft)	86.67%	89.33%	86.67%

TABLE II: Hyperparameter Tuning: RandomizedSearchCV for Heart Disease

Algorithm	Accuracy	Precision	Recall
KNN	66.67%	67.99%	66.67%
Logistic Regression	91.67%	92.79%	91.67%
Random Forest	88.33%	89.32%	83.33%
Decision Tree	78.33%	78.91%	78.33%
SVC (Linear)	86.67%	89.33%	86.67%
SVC (RBF)	88.33%	89.32%	88.33%
SVC (Poly)	78.33%	78.48%	78.33%
SVC (Sigmoid)	56.67%	56.33%	56.67%
ANN	73.33%	76.76%	73.33%
Naive Bayes (Gaussian)	86.67%	89.33%	86.67%
Naive Bayes (Bernoulli)	90.00%	90.14%	90.00%
VotingClassifier	86.67%	89.33%	86.67%

The experimental evaluation of various ML algorithms for heart disease prediction revealed compelling insights across different optimization approaches. The initial pre-tuning assessment (Table I) demonstrated that Logistic Regression achieved superior performance with 91.67% accuracy, 92.79% precision, and 91.67% recall, significantly outperforming other baseline models. Subsequent hyperparameter optimization through RandomizedSearchCV (Table II) yielded notable improvements, particularly for the KNN algorithm, which saw an increase from 58.33% to 66.67% in accuracy. The Support Vector Classification (SVC) algorithm, when optimized with different kernels, showed varying degrees of improvement, with the RBF kernel achieving 88.33% accuracy. The most comprehensive optimization using GridSearchCV (Table III) further refined these results, with SVC (Linear) reaching 90.00% accuracy and 91.58% precision, while the Artificial Neural Network (ANN) demonstrated improved performance at 81.67% accuracy compared to its initial 73.33%. Notably, Logistic Regression maintained its superior performance

TABLE III: Hyperparameter Tuning: GridSearchCV for Heart Disease

Algorithm	Accuracy	Precision	Sensitivity
KNN	63.33%	63.23%	63.33%
Logistic Regression	91.67%	92.79%	91.67%
Random Forest	88.33%	89.32%	83.33%
Decision Tree	78.33%	78.91%	78.33%
SVC (Linear)	90.00%	91.58%	90.00%
SVC (RBF)	88.33%	90.43%	88.33%
SVC (Poly)	73.33%	73.52%	73.33%
SVC (Sigmoid)	53.33%	28.44%	53.33%
ANN	81.67%	82.38%	81.67%
Naive Bayes (Gaussian)	86.67%	89.33%	86.67%
Naive Bayes (Bernoulli)	90.00%	90.14%	90.00%
VotingClassifier	86.67%	89.33%	86.67%

consistently across all three scenarios, suggesting its robust generalization capabilities for heart disease prediction. The Naive Bayes (Bernoulli) classifier also demonstrated remarkable stability, maintaining 90.00% accuracy throughout the optimization process. These findings indicate that while sophisticated algorithms and extensive hyperparameter tuning can enhance model performance, traditional algorithms like Logistic Regression and Naive Bayes offer a compelling balance of accuracy, precision, and computational efficiency for heart disease prediction.

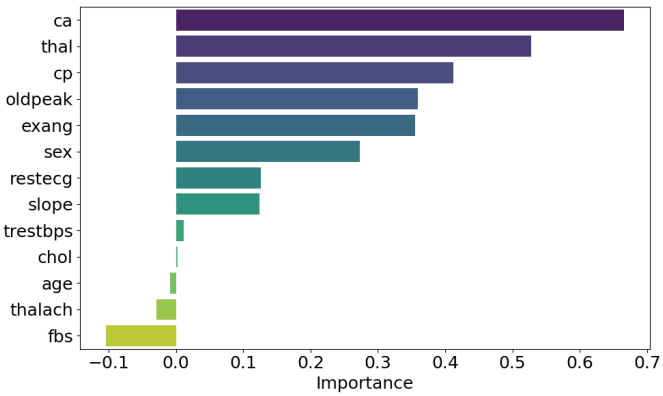

Fig. 5: Feature Importance of Heart Disease Dataset using Logistic Regression

Fig. 5, showing the logistic regression analysis, revealed significant variations in feature importance for heart disease prediction. The coronary artery calcium score (ca) emerged as the most influential predictor with an importance value of 0.65, followed by thal at 0.55 and cp at approximately 0.4. Intermediate importance was observed in features such as oldpeak and exang, both showing values around 0.35. The model attributed moderate importance to sex (0.25), while resting ECG results (restecg) and slope of peak exercise ST segment demonstrated lower importance values of approximately 0.15.

Notably, conventional risk factors such as trestbps, chol, and age exhibited minimal importance in the model's predictive capability. Interestingly, thalach and fasting blood sugar (fbs) showed slight negative importance values, suggesting potential inverse relationships with the prediction outcome. This hierarchical distribution of feature importance underscores that anatomical and diagnostic markers carry greater predictive weight compared to traditional health metrics in this particular logistic regression model. These results suggest that reliance on anatomical indicators may enhance the accuracy of heart disease prediction models. Furthermore, the findings challenge the conventional focus on traditional risk factors, calling for a reevaluation of their role in predictive modeling for heart disease.

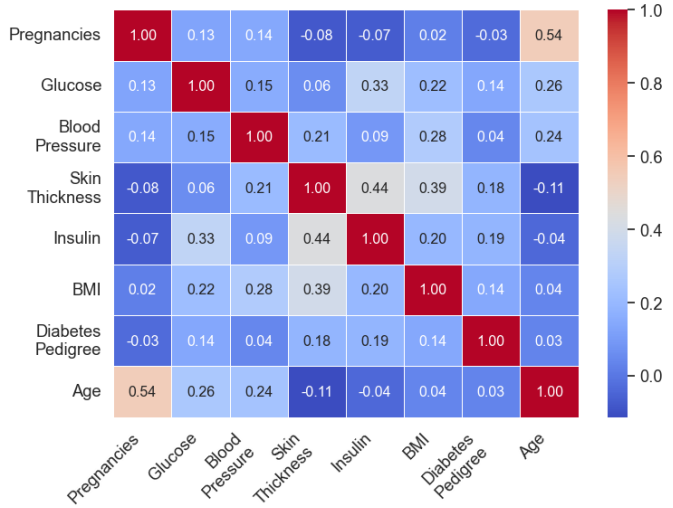

Fig. 6: Diabetes Correlation matrix

The correlation matrix for the diabetes dataset, presented in Fig. 6, provides critical insights into the relationships between the features under consideration. It highlights several noteworthy interdependencies that can inform feature selection and the design of predictive models. For example, "Age" exhibits a moderate positive correlation with "Pregnancies" (0.54), which is biologically plausible as older individuals in the dataset are likely to have had more pregnancies. "Glucose" shows a moderate correlation with "Insulin" (0.33), indicating that individuals with higher glucose levels may also have elevated insulin levels, a potential marker for insulin resistance or diabetes progression. Similarly, "BMI" correlates positively with both "Glucose" (0.22) and "Skin Thickness" (0.39), suggesting that individuals with higher body mass indices tend to have elevated glucose levels and thicker skinfold measurements, factors associated with obesity and diabetes. The generally weak correlations between most features, such as "Blood Pressure" and "Diabetes Pedigree Function," indicate minimal linear relationships, which suggests that nonlinear relationships or interactions may play a significant role in diabetes prediction. These findings suggest the potential need for advanced analytical methods to capture underlying complexities. This analysis underscores the importance of careful consideration of feature contributions during model development, as certain features may provide complementary information rather than direct correlations, improving overall predictive performance.

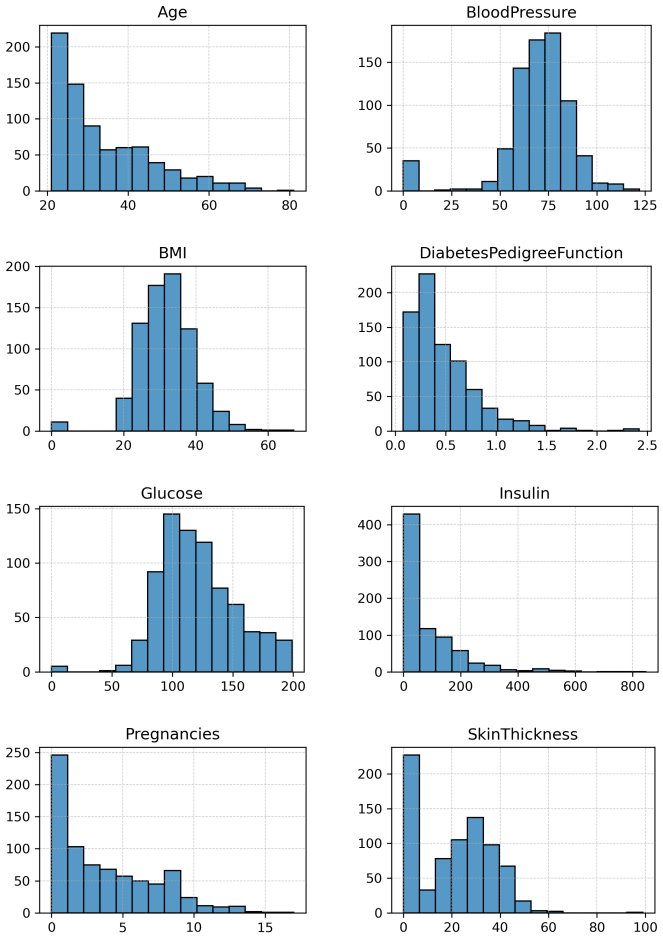

Fig. 7: Distribution of Variables in the Diabetes Disease Dataset

The distribution plots of the variables in the diabetes dataset, as illustrated in Fig. 7, reveal important patterns and characteristics essential for understanding the data. The "Age" variable is right-skewed, with the majority of the individuals falling between 20 and 50 years of age, while older age groups are less frequent, reflecting the dataset's demographic composition. The "Pregnancies" variable also shows a right-skewed distribution, with most values clustered between 0 and 5, and fewer occurrences of higher pregnancy counts, which may influence its significance in model development. Similarly, "Diabetes Pedigree Function," a measure of hereditary diabetes risk, is heavily right-skewed, with a majority of values below 1.0, indicating a low genetic predisposition for most individuals.

The continuous variables "BMI" and "Glucose" are more centrally distributed, though "BMI" demonstrates a near-normal distribution centered around 30, reflecting typical values associated with overweight or obesity. In contrast, "Insulin" and "Skin Thickness" exhibit highly skewed distributions with the presence of potential outliers, suggesting variability in these clinical measures across individuals. The

"Blood Pressure" variable has a unimodal distribution centered around 70-80 mmHg, but with a slight skewness toward lower values, which may require scaling during preprocessing. These observations indicate the need for preprocessing techniques such as handling skewness, scaling features, and addressing outliers to ensure balanced input data for ML models. Such data preparation steps will be critical to improving model performance and interpretability in predicting diabetes outcomes.

TABLE IV: Model Accuracies for Diabetes (Before Parameter Tuning)

Algorithm	**Accuracy**	**Precision**	**Recall**
KNN	66.88%	66.22%	66.88%
Logistic Regression	71.43%	70.65%	71.43%
Random Forest	77.27%	76.76%	77.27%
Decision Tree	73.38%	72.50%	73.38%
SVC	72.08%	71.06%	72.08%
ANN	67.53%	68.44%	67.53%
Naive Bayes (Gaussian)	70.78%	71.79%	70.78%
Naive Bayes (Bernoulli)	64.94%	60.37%	64.94%
VotingClassifier (Hard)	73.38%	72.56%	73.38%
VotingClassifier (Soft)	73.38%	72.56%	73.38%

TABLE V: Hyperparameter Tuning: RandomizedSearchCV for Diabetes

Algorithm	**Accuracy**	**Precision**	**Recall**
KNN	70.78%	69.53%	70.78%
Logistic Regression	71.43%	70.65%	71.43%
Random Forest	77.27%	76.76%	77.27%
Decision Tree	73.38%	72.50%	73.38%
SVC (Linear)	70.78%	69.89%	70.78%
SVC (RBF)	73.38%	72.53%	73.38%
SVC (Poly)	72.08%	71.14%	72.08%
SVC (Sigmoid)	64.94%	42.17%	64.94%
ANN	68.83%	67.12%	68.83%
Naive Bayes (Gaussian)	70.78%	71.79%	70.78%
Naive Bayes (Bernoulli)	64.94%	60.37%	64.94%
VotingClassifier	74.03%	73.49%	74.03%

The baseline performance evaluation of various ML algorithms for diabetes prediction, prior to parameter tuning, revealed interesting patterns as shown in Table IV. Random Forest demonstrated superior performance even without parameter optimization, achieving the highest accuracy of 77.27%, precision of 76.76%, and recall of 77.27%. Decision Tree and both variants of VotingClassifier (Hard and Soft) showed promising results with identical metrics of 73.38% accuracy. Logistic Regression and SVC displayed moderate performance levels with accuracies of 71.43% and 72.08% respectively. The Naive Bayes implementations showed varying results, with the Gaussian variant (70.78% accuracy) outperforming the Bernoulli variant (64.94% accuracy). The KNN algorithm and ANN showed relatively lower performance, with accuracies of 66.88% and 67.53% respectively.

The implementation of RandomizedSearchCV (Table V) for diabetes prediction maintained similar performance patterns across different algorithms. Random Forest retained its position as the top performer with unchanged metrics (77.27% accuracy, 76.76% precision, and 77.27% recall), suggesting its robust initial parameterization. Decision Tree and VotingClassifier (both Hard and Soft variants) continued to show consistent performance with 73.38% accuracy. SVC demonstrated moderate performance with 72.08% accuracy. The simpler algorithms like KNN and Naive Bayes (Bernoulli) showed relatively lower performance, with accuracies of 66.88% and 64.94% respectively. Artificial Neural Network (ANN) achieved modest results with 67.53% accuracy.

TABLE VI: Hyperparameter Tuning: GridSearchCV for Diabetes

Algorithm	**Accuracy**	**Precision**	**Sensitivity**
KNN	70.13%	68.89%	70.13%
Logistic Regression	71.43%	70.65%	71.43%
Random Forest	77.27%	76.76%	77.27%
Decision Tree	73.38%	72.50%	73.38%
SVC (Linear)	71.43%	70.65%	71.43%
SVC (RBF)	73.38%	72.50%	73.38%
SVC (Poly)	74.68%	73.97%	74.68%
SVC (Sigmoid)	64.94%	42.17%	64.94%
ANN	70.78%	69.90%	70.78%
Naive Bayes (Gaussian)	70.78%	71.79%	70.78%
Naive Bayes (Bernoulli)	64.94%	60.37%	64.94%
VotingClassifier	74.68%	74.10%	74.68%

The GridSearchCV approach yielded notably different results for some algorithms. SVC with Polynomial kernel and VotingClassifier emerged as the leading models, both achieving 74.68% accuracy. This represented a significant improvement for the polynomial SVC compared to its RandomizedSearchCV performance. Linear SVC maintained consistent performance with 71.43% accuracy. The Decision Tree, Naive Bayes models maintained similar performance levels across both tuning methods, with the Gaussian variant performing better than the Bernoulli variant.

The SVM with Polynomial kernel demonstrated noteworthy performance improvements through hyperparameter tuning. In the GridSearchCV implementation, it achieved an accuracy of 74.68% with a precision of 73.97%, marking a significant improvement from its initial and RandomizedSearchCV performance. This enhancement suggests that the grid search approach was more effective in optimizing the polynomial kernel's hyperparameters. The model's consistent performance

across different metrics (accuracy, precision, and sensitivity) indicates its robust predictive capabilities for diabetes classification.

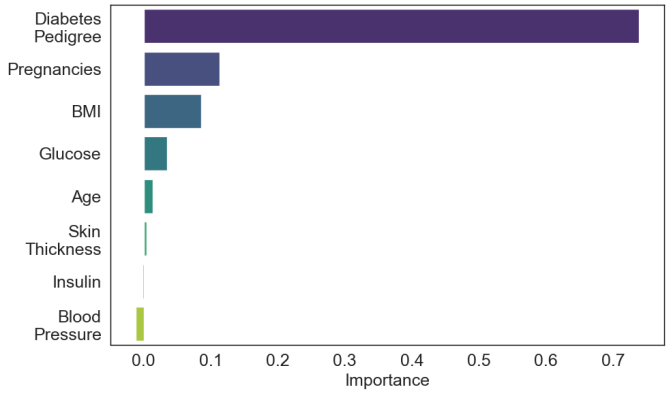

Fig. 8: Feature Importance of Diabetes Dataset using SVM with Polynomial Kernel

The feature importance analysis for diabetes prediction shown in Fig. 8 revealed a clear hierarchy of influential factors. Diabetes Pedigree emerged as the most significant predictor with an importance value of approximately 0.7, substantially higher than all other features. Pregnancies and BMI showed moderate importance with values around 0.1, while Glucose levels demonstrated lower but notable importance. Age and Skin Thickness exhibited minimal predictive value. Interestingly, despite their clinical significance, Insulin and Blood Pressure displayed the lowest importance values in the model's predictions.

V. Evaluation of Model Performance

The performance of the proposed ML models was rigorously evaluated using a combination of techniques to ensure robustness and generalizability. For heart disease prediction, the Randomized Search Logistic Regression model demonstrated the highest accuracy, while for diabetes prediction, the Grid Search SVC with a polynomial kernel emerged as the top-performing model. The evaluation focused on ROC-AUC analysis, confusion matrix visualization, classification reports, and cross-validated metrics for these models.

A. ROC Curve and AUC Scores

The evaluation of model performance prominently featured the Receiver Operating Characteristic (ROC) curve analysis, which is a graphical representation of a model's ability to distinguish between classes. The AUC score, derived from the ROC curve, quantifies the overall performance, where a higher AUC indicates better discriminative capability.

Fig. 9 shows the ROC curve for heart disease prediction, where the top-performing model, the Randomized Search Logistic Regression, achieved an impressive AUC score of 0.93. This high AUC score reflects the model's strong ability to correctly classify cases of heart disease and non-heart disease. The steep initial rise in the ROC curve demonstrates a high true positive rate with minimal false positives. The curve's proximity to the top-left corner underscores the model's high sensitivity and specificity, making it a reliable choice for heart disease prediction.

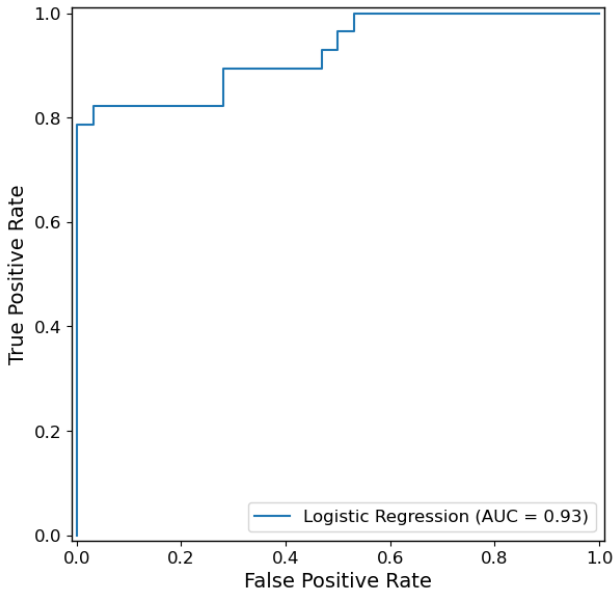

Fig. 9: ROC Curve with AUC Score for Heart Disease

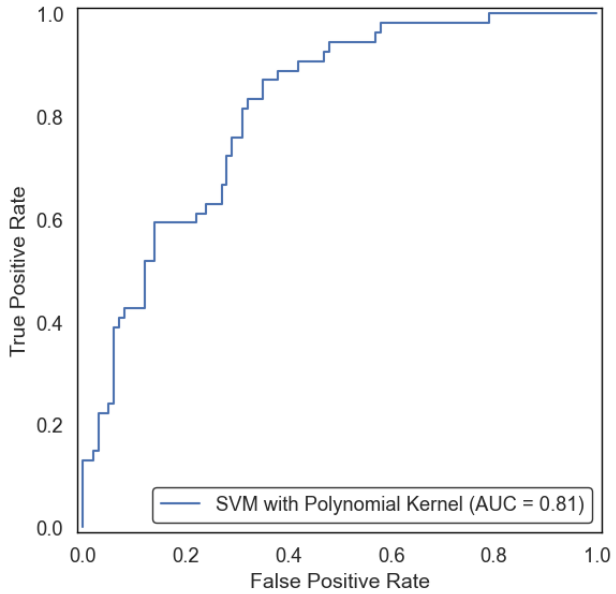

Fig. 10: ROC Curve with AUC Score for Diabetes

Fig. 10 illustrates the ROC curve for diabetes prediction, where the Grid Search SVC with a polynomial kernel emerged as the best model, achieving an AUC score of 0.81. Although slightly lower than the heart disease model, this score highlights the model's ability to effectively distinguish between diabetic and non-diabetic cases. The ROC curve shows a moderate rise, reflecting a balance between true positive and false positive rates. The curve's position, slightly further from the top-left corner compared to the heart disease model, indicates relatively lower sensitivity and specificity. However, it still

demonstrates the model's overall precision and reliability in predicting diabetes.

B. Confusion Matrix

Confusion matrices provide a detailed visual summary of model performance, showing the number of true positives, true negatives, false positives, and false negatives for each dataset. They are essential for evaluating how well models classify each class and highlight the extent of misclassifications. By analyzing confusion matrices, one can assess the strengths and weaknesses of the models, gaining valuable insights into their accuracy and reliability in differentiating between positive and negative classes.

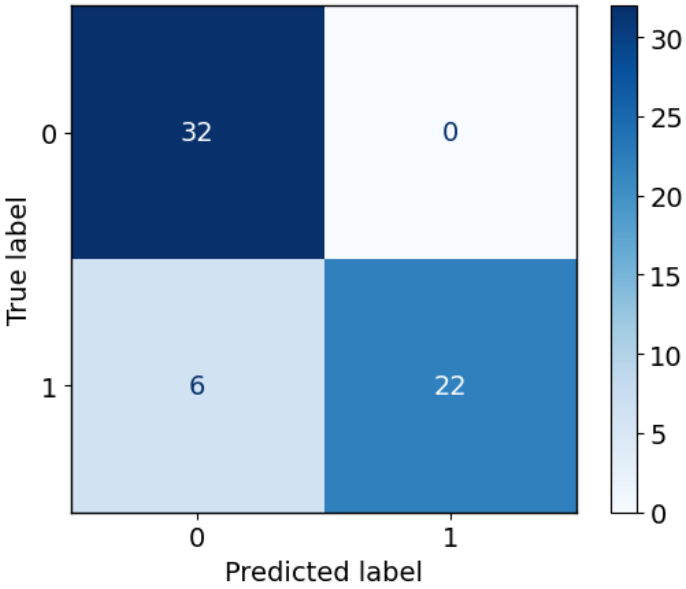

Fig. 11: Confusion Matrix of Heart Disease Dataset

In Fig. 11, the confusion matrix for heart disease prediction demonstrates the performance of the Randomized Search Logistic Regression model. The model accurately classified 32 cases of Class 0 (No Heart Disease) as true negatives and 22 cases of Class 1 (Heart Disease) as true positives. Misclassifications were minimal, with only 6 false negatives (heart disease cases misclassified as non-heart disease) and no false positives. This highlights the model's strong ability to predict both positive and negative cases accurately, resulting in high sensitivity and specificity.

In Fig. 12, the confusion matrix for diabetes prediction reflects the performance of the Grid Search SVC with a polynomial kernel. The model correctly identified 88 cases of Class 0 (No Diabetes) as true negatives and 27 cases of Class 1 (Diabetes) as true positives. However, 12 false positives (non-diabetic cases misclassified as diabetic) and 27 false negatives (diabetic cases classified as non-diabetic) were observed. While the model performs effectively in identifying non-diabetic cases, the higher number of misclassifications suggests there is room for improvement, particularly in detecting diabetic cases.

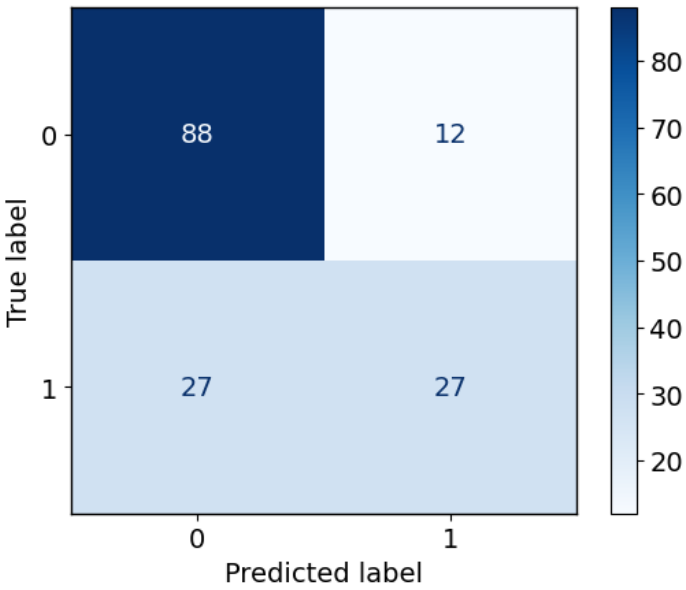

Fig. 12: Confusion Matrix of Diabetes Dataset

C. Classification Report

The classification reports provide a detailed summary of performance metrics, including precision, recall, and F1-score, for evaluating the models' accuracy and reliability. These metrics highlight the strengths and weaknesses of the models in minimizing false positives and false negatives, offering valuable insights into their effectiveness for heart disease and diabetes prediction tasks.

TABLE VII: Classification Report for Heart Disease

	precision	**recall**	**f1-score**	**support**
0	0.842105	1.000000	0.914286	32.000000
1	1.000000	0.785714	0.880000	28.000000
accuracy	0.900000	0.900000	0.900000	0.900000
macro avg	0.921053	0.892857	0.897143	60.000000
weighted avg	0.915789	0.900000	0.898286	60.000000

For heart disease prediction, the Randomized Search Logistic Regression model demonstrated strong and balanced performance, as shown in Table VII. For Class 0 (No Heart Disease), the model achieved a precision of 0.84, a perfect recall of 1.00, and an F1-score of 0.91, indicating its ability to accurately identify non-heart disease cases. For Class 1 (Heart Disease), the model achieved a precision of 1.00, a recall of 0.79, and an F1-score of 0.88, reflecting its effectiveness in detecting heart disease cases with minimal false positives. The model's overall accuracy of 90%, along with high macro and weighted averages, underscores its reliability and robustness.

For diabetes prediction, the Grid Search SVC with a polynomial kernel performed well, as outlined in Table VIII, particularly in identifying non-diabetic cases (Class 0). The model achieved a precision of 0.76, a recall of 0.88, and an F1-score of 0.82 for Class 0. However, its performance for Class

TABLE VIII: Classification Report for Diabetes Disease

	precision	recall	f1-score	support
0	0.765217	0.880000	0.818605	100.000000
1	0.692308	0.500000	0.580645	54.000000
accuracy	0.746753	0.746753	0.746753	0.746753
macro avg	0.728763	0.690000	0.699625	154.000000
weighted avg	0.739652	0.746753	0.735164	154.000000

1 (Diabetes) was slightly weaker, with a precision of 0.69, a recall of 0.50, and an F1-score of 0.58, indicating challenges in detecting diabetic cases accurately. The overall accuracy of 74.67% and moderate weighted averages suggest room for improvement, particularly in addressing class imbalance.

VI. Conclusion

The study effectively demonstrates the application of ML algorithms for predicting heart disease and diabetes. For heart disease prediction, Logistic Regression achieved the highest accuracy of 91.67% while for diabetes, the Random Forest reached 77.27%. Feature importance analysis highlighted key predictors like "ca" for heart disease and "Diabetes Pedigree Function" for diabetes. The system successfully correlates both diseases, enabling faster predictions using ensemble techniques. The analysis also identified key features contributing to disease prediction, offering valuable insights for healthcare decision-making. Overall, these results underscore the effectiveness of ML in enhancing early diagnosis and risk assessment in healthcare.

Acknowledgment

We express our sincere gratitude to Assistant Professor Dr. Rajani Chulyadyo for assigning this project and providing valuable guidance throughout the process. Her insights, encouragement, and support have been instrumental in the successful completion of this work. We truly appreciate your effort to improve our learning experience and promote our academic growth.

References

[1] De Almeida-Pititto, B., Dualib, P. M., Zajdenverg, L., Dantas, J. R., De Souza, F. D., Rodacki, M., & Bertoluci, M. C. (2020). Severity and mortality of COVID 19 in patients with diabetes, hypertension and cardiovascular disease: a meta-analysis. Diabetology & Metabolic Syndrome, 12(1). https://doi.org/10.1186/s13098-020-00586-4

[2] Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart Disease Prediction using Exploratory Data Analysis. Procedia Computer Science, 173, 130–139. https://doi.org/10.1016/j.procs.2020.06.017

[3] Dhande, B., Bamble, K., Chavan, S., & Maktum, T. (2022b). Diabetes & Heart Disease Prediction Using Machine Learning. ITM Web of Conferences, 44, 03057. https://doi.org/10.1051/itmconf/20224403057

[4] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An Overview of heart disease Prediction. International Journal of Computer Applications, 17(8), 43–48. https://doi.org/10.5120/2237-2860

[5] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516–76531. https://doi.org/10.1109/access.2020.2989857

[6] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00515

[7] Barron, E., Bakhai, C., Kar, P., Weaver, A., Bradley, D., Ismail, H., Knighton, P., Holman, N., Khunti, K., Sattar, N., Wareham, N. J., Young, B., & Valabhji, J. (2020). Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. The Lancet Diabetes & Endocrinology, 8(10), 813–822. https://doi.org/10.1016/s2213-8587(20)30272-2

[8] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science, 165, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

[9] Rindhe, Baban.U. et al. (2021) 'Heart disease prediction using machine learning', International Journal of Advanced Research in Science, Communication and Technology, pp. 267–276. doi:10.48175/ijarsct-1131.

[10] Sharma, H., & Rizvi, M. A. (2017b). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99–104. https://doi.org/10.17762/ijritcc.v5i8.1175

[11] Salhi, D. E., Tari, A., & Kechadi, M. (2021). Using machine learning for heart disease prediction. In Lecture notes in networks and systems (pp. 70–81). https://doi.org/10.1007/978-3-030-69418-0_7