Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

Priyanshu Maniyar	*221AI023*
Rohil Sharma	*221AI033*
Sachin Choudhary	*221A034*
Tushar Kanda	*221AI042*

Introduction & Motivation :

- Heart disease causes **millions of deaths** globally each year.

- **Early detection** is crucial but challenging due to healthcare limitations.

- Automated tools using machine learning (ML) can assist in predicting heart disease risks.

- **ML models analyze patient data** to improve diagnosis and offer real-time predictions.

- Challenges include selecting key features, handling unbalanced data, and making results easy to understand.

- The project aims to **create a cost-effective solution that improves heart disease prediction.**

- It uses **clustering** to group **patients by risk factors for personalized care.**

- Interactive visualizations help users see how factors like cholesterol and blood pressure impact their risk.

- **The goal is to provide a user-friendly platform with actionable insights to promote better health decisions.**

Literature Review:

The application of machine learning in heart disease prediction has garnered significant attention in recent years due to its potential to provide more accurate and timely diagnosis. Several studies have explored various ML techniques to predict heart disease risk, with many focusing on feature selection and model optimization. For instance, Rajkumar and Reena (2010) demonstrated that decision trees and Naive Bayes classifiers could be used effectively for heart disease prediction, achieving substantial accuracy when trained on medical datasets. However, these early approaches often lacked the complexity needed to address imbalanced datasets and explainability in predictions, which are crucial for practical healthcare applications.

More recent work has emphasized the importance of feature selection in improving prediction performance. Studies such as that by Ghosh and Ghosh (2021) have employed advanced feature selection methods, including chi-square and mutual information, to refine the input variables used by machine learning models. This has led to improved accuracy in classifiers like Random Forest and SVM. Moreover, the integration of techniques like the Synthetic Minority Oversampling Technique (SMOTE) to balance data has been shown to significantly enhance prediction outcomes, particularly in datasets with skewed class distributions, as highlighted by Kaur et al. (2019). Despite these advancements, there remains a gap in the literature regarding personalized patient profiling and the use of clustering techniques to create more tailored prediction models. This project aims to fill that gap by integrating clustering methods like K-means and DBSCAN to develop patient subgroups and interactive visualizations, providing both predictive accuracy and personalized healthcare insights.

Problem Statement:

The basic idea is to develop a machine learning-based heart disease prediction system that provides accurate, personalized risk assessments and interactive insights for both healthcare providers and patients.

Methodology

1. **Data Collection and Preprocessing:**

   - Gather and clean datasets from various sources, ensuring quality and completeness.

   - Handle missing values, normalize data, and encode categorical features.

2. **Feature Selection:**

   - Apply feature selection methods such as chi-square, ANOVA, and mutual information to identify the most relevant features for heart disease prediction.

   - Create feature subsets (e.g., SF-1, SF-2, SF-3) based on the results.

3. **Model Refinement** :

   - Fine-tune existing machine learning models, including Naive Bayes, SVM, XGBoost, and others, to optimize their performance on the dataset.

   - Evaluate model performance using metrics such as accuracy, sensitivity, specificity, precision, F1 score, and AUC.

4. **Deep Learning Exploration:**

   - Implement and test deep learning techniques, such as neural networks, on the dataset to explore potential improvements in prediction accuracy.

5. **Handling Imbalanced Data:**

   - Apply the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalances and improve model performance.

6. Explainable AI Integration:

- Apply explainable AI (XAI) methods, such as SHAP (SHapley Additive exPlanations), to all models to interpret and visualize the factors influencing predictions.

7**. Model Comparison and Selection:**

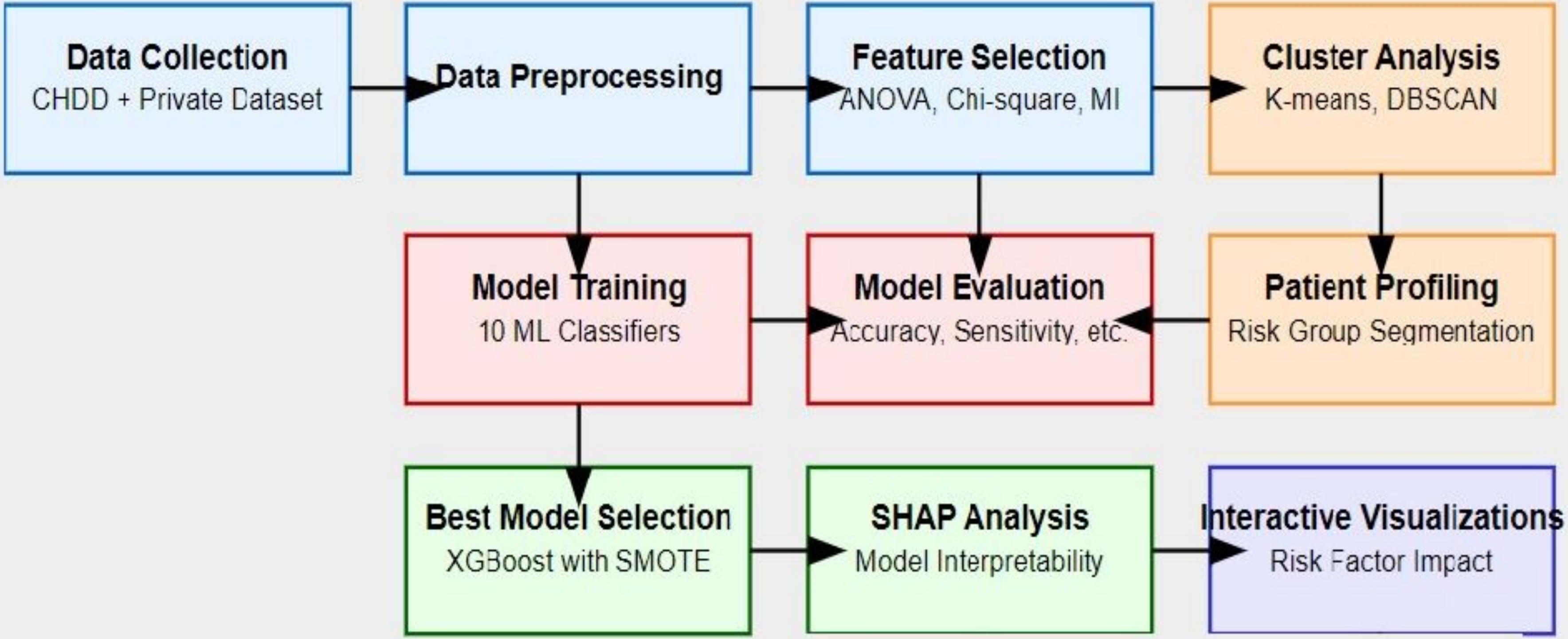
- Compare the performance of refined ML models, deep learning models, and explainable AI insights to determine the most effective approach for heart disease prediction.

8. **Interactive Visualization Development:**

- Create interactive visualizations to allow users to explore how different health factors impact their risk of heart disease.

9. **Validation and Testing:**

- Validate the final models using private and public datasets, and conduct cross-validation to ensure robustness and generalizability.


Data Collection
CHDD + Private Dataset
Data Preprocessing
Feature Selection
ANOVA, Chi-square, MI
Cluster Analysis
K-means, DBSCAN
Model Training
10 ML Classifiers
Model Evaluation
Accuracy, Sensitivity, etc.
Patient Profiling
Risk Group Segmentation
Best Model Selection
XGBoost with SMOTE
SHAP Analysis
Model Interpretability
Interactive Visualizations
Risk Factor Impact

Conclusion

Enhanced Prediction Accuracy: The project successfully refined existing machine learning models and explored deep learning methods, resulting in improved accuracy and performance for heart disease prediction.

Increased Interpretability: The integration of explainable AI (XAI) techniques provided clear insights into the reasoning behind predictions, enhancing the transparency and trustworthiness of the model outputs.

Personalized Risk Assessment: By employing feature selection and clustering techniques, the project developed a system that offers personalized risk assessments tailored to different patient profiles.