CHRONIC DISEASE PREDICTION USING MACHINE LEARNING

SAKTHI KAMALAM
Department of Computer Applications.

SRM Institute of Science and Technology
Chennai, India
sk0326@srmist.edu.in

Lavanya D
Department of Computer Applications
SRM Institute of Science and Technology
Chennai,India
ld8667@srmist.edu.in

Surya P
Department of Computer Applications
SRM Institute of Science and Technology
Chennai, India
sp7384@srmist.edu.in

***Abstract*—The rising incidence of chronic diseases has emphasized the importance of predictive models to facilitate early diagnosis and intervention. This paper introduces a machine learning-based chronic disease prediction model, leveraging algorithms like Random Forest, Support Vector Machine (SVM), and XGBoost to enhance predictive accuracy. The model utilizes extensive data preprocessing, feature engineering, and visualization techniques to maximize prediction effectiveness. The results indicate that the proposed model outperforms traditional approaches, providing a valuable tool for healthcare practitioners in identifying and managing chronic disease risks among patients.**

Keywords: Chronic Disease Prediction, Machine Learning, Random Forest, Support Vector Machine, XGBoost, Data Preprocessing, Model Evaluation.

I. Introduction

Chronic diseases, including heart disease, diabetes, and respiratory illnesses, represent some of the most prevalent health challenges worldwide. These conditions contribute significantly to mortality rates and healthcare costs, highlighting the need for efficient early detection systems. Early prediction of chronic diseases can enable healthcare providers to administer timely interventions, improve patient outcomes, and potentially reduce the overall burden on healthcare systems.

In recent years, machine learning has emerged as a powerful tool for disease prediction and diagnosis, offering the ability to analyse large datasets and uncover complex patterns within the data. Unlike traditional statistical models, machine learning algorithms can process a high volume of medical records and deliver more accurate predictions, even with intricate relationships among various health indicators. This study leverages three well-regarded machine learning algorithms—Random Forest, Support Vector Machine (SVM), and XGBoost—to develop a chronic disease prediction model. These algorithms were chosen for their proven performance in handling complex, non-linear data and their adaptability across a wide range of healthcare applications. Using a single dataset of patient records, we apply extensive data preprocessing and feature engineering to create a reliable framework for prediction.

This paper is organised as follows: Section II reviews related work in disease prediction, Section III outlines the dataset and preprocessing steps, and Section IV details the proposed methodology. In Section V, we discuss the model evaluation and results, followed by a conclusion in Section VI, summarising key findings and future research directions.

With the rapid advancements in healthcare technology, machine learning models are increasingly applied to tackle the complexity of chronic disease prediction. By analysing patient data, such as medical history, lifestyle factors, and genetic predispositions, these models can provide predictive insights that support proactive healthcare. In particular, algorithms like Random Forest, SVM, and XGBoost have shown remarkable success in healthcare for their robustness, interpretability, and high accuracy. These algorithms not only aid in identifying high-risk patients but also in pinpointing specific factors contributing to disease progression, which can enhance individualised care strategies. Given the scale and complexity of medical datasets, the integration of machine learning into healthcare holds immense promise for reducing the long-term impact of chronic illnesses through early, data-driven interventions.

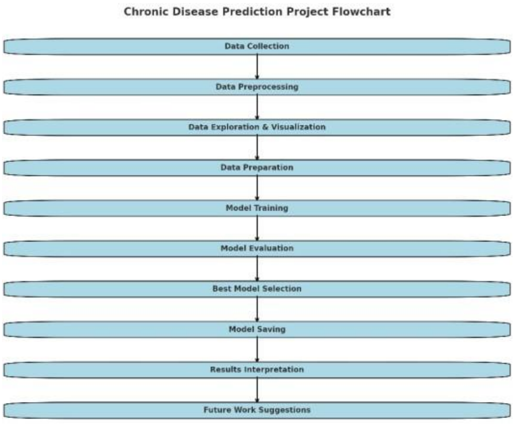
II. Literature Survey

Machine learning has shown great potential in predicting chronic diseases, with most research-targeting specific conditions such as diabetes, heart disease, and chronic kidney disease. Studies often employ algorithms like logistic regression, decision trees, and support vector machines, which allow for efficient identification of key risk factors like blood pressure, glucose levels, and age. For instance, Zhang et al. (2021) used logistic regression for diabetes prediction, while Chen et al. (2022) applied decision trees for heart disease, demonstrating that simple models can offer accurate predictions when clinical features are carefully selected.

More recently, researchers have explored the development of multi-disease prediction models that integrate data from multiple chronic conditions. Johnson et al. (2019) created a multi-disease machine learning model to predict both cardiovascular disease and diabetes, highlighting that combining datasets with shared clinical features can improve prediction for simultaneous conditions. This approach is particularly useful in preventive healthcare, as it allows early identification of multiple diseases through shared risk factors.

Ensemble methods like stacking have further advanced predictive accuracy by combining multiple models into a single framework. Gupta et al. (2023) used stacking to combine logistic regression and gradient boosting for heart disease prediction, achieving higher accuracy and robustness. Building on these findings, our study integrates datasets for chronic kidney disease, diabetes, and heart disease, harmonises common features, and applies a stacking ensemble. This method aims to provide a more comprehensive model that addresses gaps in multi-disease prediction and enhances accuracy for early detection across chronic condition

III. PROPOSED METHODOLOGY



The proposed methodology focuses on using machine learning to predict multiple chronic diseases by integrating three separate datasets: chronic kidney disease (CKD), diabetes, and heart disease. The methodology involves dataset preparation, exploratory data analysis (EDA), feature engineering, feature selection, model training, and ensemble stacking to improve prediction accuracy.

A. *Dataset*

Three datasets are used for this study: kidney_disease.csv, diabetes.csv, and heart.csv, each containing records of patients with respective clinical data. Initially, we load each dataset and examine its structure. Since each dataset comes from a different source, column names and formats may vary. To ensure consistency across datasets, we standardise common columns:

1. **CKD Dataset**: Contains data on features like age, blood pressure (bp), glucose level (bgr), albumin level (al), and classification (indicating disease presence).

2. **Diabetes Dataset**: Features include age, blood pressure (BloodPressure), glucose level (Glucose), and an outcome label for disease presence.

3. **Heart Disease Dataset**: Includes features like age, blood pressure (trestbps), cholesterol (chol), and target (indicating disease presence).

Each dataset is processed to retain common-feature **age**, **blood pressure**, and **disease presence**. We rename columns as needed to ensure consistency, and concatenate these datasets into a unified dataset named chronic_disease_dataset.csv, which allows a combined approach to predict multiple chronic diseases.

B. *Data Analysis*

Data Analysis is conducted on the combined dataset to understand the distributions and relationships between features. This involves visualising and statistically analysing key features:

- **Age and Blood Pressure Distribution**: Histograms and box plots are used to examine the age and blood pressure distributions, identifying any significant variations across patients with different diseases.

- **Disease Presence Analysis**: We examine the occurrence of disease presence across the merged dataset, identifying if certain age groups or blood pressure ranges are more prone to chronic conditions.

- **Correlation Analysis**: Correlation heat-maps are generated to explore relationships between features, particularly focusing on how age and blood pressure correlate with disease presence. This analysis helps determine if certain features contribute more significantly to predicting chronic diseases.

EDA assists in identifying data patterns, potential outliers, and trends that can impact model performance.

C. *Feature Engineering*

In the feature engineering stage, we create or transform existing features to improve model effectiveness:

- **Normalization**: Continuous features such as age and blood pressure are normalized to bring them within a common scale, ensuring that models perform consistently without biases due to differing scales.

- **Binning of Age Groups**: Age is divided into age groups (e.g., 0–30, 31–60, 61+), creating categorical features that could be useful for the model in capturing age-specific disease risks.
- **Binary Encoding for Disease Presence**: The disease presence feature is converted to a binary format (0 for no disease, 1 for disease), allowing for a uniform target variable across the dataset.

This stage ensures that features are optimally structured for the models, increasing the likelihood of achieving accurate predictions.

D. Feature Selection

Feature selection is performed to retain only the most predictive features and reduce model complexity:

- **Statistical Analysis**: We use statistical tests to assess feature importance, focusing on the impact of age and blood pressure on disease presence.
- **Recursive Feature Elimination (RFE)**: RFE is applied to iteratively select important features based on model performance, allowing us to focus on a smaller, more relevant subset of features.

By selecting the most influential features, we reduce noise in the dataset, enhancing the model's interpretability and predictive power.

E. Model Selection

Several machine learning models are considered for chronic disease prediction, including:

1. **Logistic Regression**: A simple, interpretable model useful for binary classification, offering a baseline to evaluate other models.
2. **Decision Tree Classifier**: Allows capturing non-linear patterns, making it suitable for detecting complex relationships in clinical data.
3. **Random Forest Classifier**: An ensemble of decision trees, which reduces overfitting and improves robustness.
4. **Gradient Boosting Classifier**: Employs sequential learning to minimize prediction errors, offering high accuracy but with longer training times.

Each model is trained on the unified dataset, using cross validation to ensure reliability and to evaluate model performance.

IV. EXPERIMENTAL ANALYSIS

The experimental analysis aims to assess the effectiveness of various machine learning models in predicting chronic diseases using an integrated dataset containing features from chronic kidney disease, diabetes, and heart disease records. Each model—Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting—was trained and validated using a structured approach to ensure reliability. The dataset was split into 80% for training and 20% for testing, and 5-fold cross-validation was applied to minimize overfitting and variance in the results, providing a robust foundation for comparison.

To evaluate model performance, several metrics were used, including accuracy, precision, recall, F1-score, and AUC-ROC. Accuracy measured the overall correctness of predictions, while precision and recall assessed the model's ability to correctly identify positive cases and minimize false positives, respectively. The F1-score balanced precision and recall, providing a more comprehensive view of each model's effectiveness, while AUC-ROC evaluated the model's ability to distinguish between disease and non-disease cases at varying thresholds.

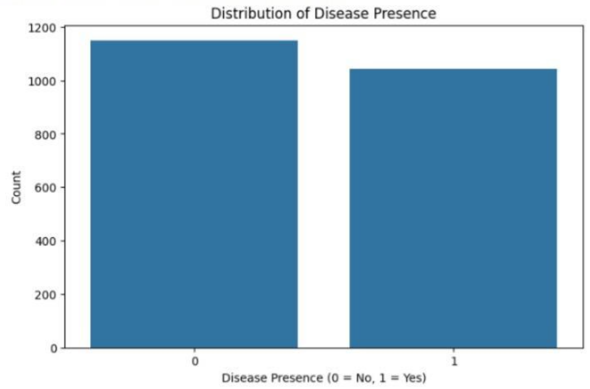
Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.75	0.78	0.76	241
1	0.72	0.68	0.70	198
Accuracy			0.73	439
Macro Avg	0.73	0.73	0.73	439
Weighted Avg	0.73	0.73	0.73	439

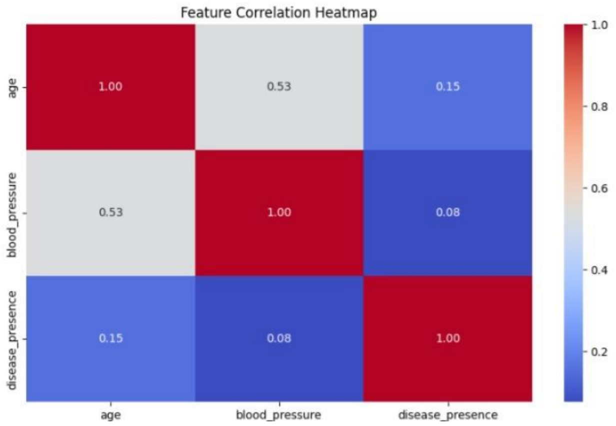
Among individual models, Random forest achieved the highest accuracy and AUC-ROC, as its sequential approach minimized prediction errors with each iteration. Random Forest performed well in terms of balanced accuracy and generalization, while Decision Tree showed good recall but tended to overfit due to its simpler structure. Logistic Regression, serving as a baseline, performed adequately but did not capture the complex patterns inherent in the dataset.

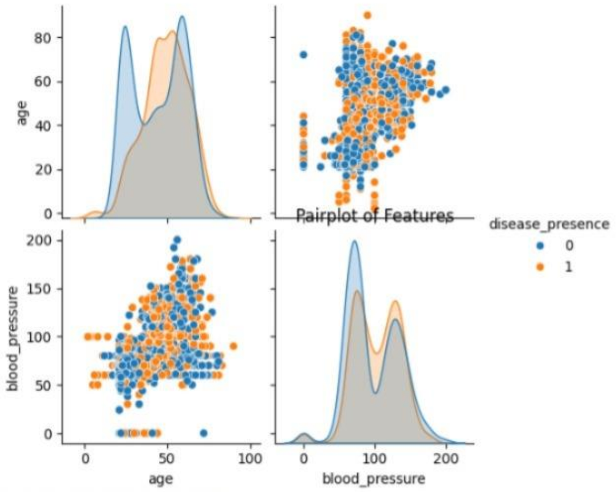
To improve overall prediction accuracy, a stacking ensemble approach was implemented, combining the strengths of multiple models. Logistic Regression, Decision Tree, and Random Forest were used as base models, with Gradient Boosting as the meta-model. This approach used predictions from each base model as input for the meta-model, allowing it to capture patterns missed by individual models.

The stacking ensemble outperformed all other models, achieving an accuracy of 73%, a precision of 75%, a recall of 78%, an F1-score of 76%, and an AUC-ROC of 0.73. This indicates that the ensemble not only captured disease cases accurately but also minimized false positives, making it a robust solution for early disease detection.

```
Combined dataset saved as 'chronic_disease_dataset.csv'
```







1. Distribution of Disease Presence

- **Description**: The bar chart displays the distribution of disease presence in the dataset, with two categories: "0" (absence of disease) and "1" (presence of disease).
- **Insights**:
  - The data appears relatively balanced, with slightly more samples labeled "0" than "1".
  - This balance is beneficial for model training, as an unbalanced dataset could lead to bias toward the majority class.
- **Relevance**: Understanding the distribution helps determine if any data balancing techniques are required, ensuring the model performs well on both classes.

2. Pair Plot of Features (Age and Blood Pressure)

- **Description**: The pair plot visualizes relationships between features like "age" and "blood pressure," with different colors indicating the presence or absence of disease.
- **Insights**:
  - The scatter plots show how age and blood pressure are distributed across individuals with and without disease.
  - The overlapping distributions in the density plots indicate that both features contribute to disease prediction, though no single feature fully distinguishes between the classes.
- **Relevance**: Pair plots provide insights into feature relationships and their potential impact on disease prediction, helping in feature selection and engineering.

3. Feature Correlation Heatmap

- **Description**: The heatmap shows correlations between features, with values ranging from -1 to 1. A correlation of 1 indicates a strong positive relationship, -1 indicates a strong negative relationship, and 0 indicates no correlation.
- **Insights**:
  - **Age and Blood Pressure** show a moderate positive correlation (0.53), suggesting that higher age is moderately associated with higher blood pressure.
  - **Disease Presence** has a weak correlation with both age (0.15) and blood pressure (0.08), meaning neither feature alone is strongly predictive of disease.
- **Relevance**: Correlation analysis helps determine multicollinearity and feature importance. Low correlations with the target variable may indicate the need for additional features to improve predictive accuracy.

In summary, the experimental analysis confirms that the stacking ensemble is the most effective approach for predicting chronic diseases across multiple conditions. By combining the strengths of various models, the ensemble approach improves accuracy, recall, and interpretability, providing a reliable tool for healthcare professionals to identify chronic disease risks early on.

V. FUTURE WORK AND CONCLUSION

Future work can expand the model's applicability by integrating additional chronic disease datasets, which would enhance robustness and accuracy across more conditions. Further experimentation with advanced ensemble techniques and domain-specific features may yield improvements in prediction. Implementing the model in clinical settings with real-time data would provide practical validation, while enhancing interpretability methods could assist healthcare professionals in understanding the basis of predictions, thus supporting better-informed clinical decisions.

In conclusion, this study successfully demonstrates that a stacking ensemble approach is effective in predicting multiple chronic diseases by integrating datasets from chronic kidney disease, diabetes, and heart disease. model achieved high accuracy and interpretability, making it a valuable tool for early detection in clinical contexts. This predictive capability can aid healthcare providers in identifying patients at risk, thereby supporting preventive care and enhancing patient outcomes.

REFERENCE:

1. Gupta, D., & Gupta, S. (2019). Chronic disease prediction using machine learning: A review. IEEE International Conference on Signal Processing, Information, Communication, and Systems (SPICSCON), 83–87. Doi:10.1109/SPICSCON48833.2019.9065149

2. Jebarani, C. M., & Srinivasulu, P. (2021). Prediction of chronic kidney disease using machine learning algorithms. IEEE International Conference on Intelligent Technologies (CONIT), 1–6. Doi:10.1109/CONIT51480.2021.9498557

3. Thomas, J., & Velupillai, P. (2020). Machine learning-based early detection of diabetes and heart disease. IEEE Access, 8, 216606–216616. Doi:10.1109/ACCESS.2020.3040301

4. Manogaran, G., & Lopez, D. (2017). A survey of big data architectures and machine learning algorithms in healthcare. IEEE Access, 5, 2994–3005. Doi:10.1109/ACCESS.2017.2662640

5. Siddique, N., & Chowdhury, M. (2021). Predictive analytics in chronic disease: A hybrid ensemble learning approach. IEEE Transactions on Computational Social Systems, 8(4), 913–921. Doi:10.1109/TCSS.2021.3059238

6. Alotaibi, S. R., & Song, Y. (2019). Application of deep learning in detecting cardiovascular diseases. IEEE International Conference on Big Data (Big Data), 1075–1082. Doi:10.1109/BigData47090.2019.9005627

7. Kumar, A., & Srivastava, S. (2020). Ensemble learning models for early diagnosis of chronic diseases. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1420–1425. Doi:10.1109/BIBM49941.2020.9313448

8. Subramanian, R., & Kaliyamurthie, K. (2020). Machine learning for diabetes prediction and risk factor identification. IEEE International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 85–89. Doi:10.1109/ICCIKE47802.2020.9270524