International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 02 | Feb 2025 www.irjet.net p-ISSN: 2395-0072

AI-Powered Predictive Modelling and Data Visualization for Cardiovascular Disease

Sangameshwar Kawdi[1], Dattatri[2], Vaishnavi.H[3], Manisha[4], Sainath[5]

[12345]*Department of Information Science and Engineering, Guru Nanak Dev Engineering College, Bidar, 585403, Karnataka, India*

Visvesvaraya Technological University, Belagavi -590018

Abstract - *The present study examines the use of artificial intelligence (AI) for predictive modeling and data visualization in cardiovascular disease (CVD), one of the top causes of morbidity and mortality around the globe. Early identification and prediction of CVD are essential to enhancing patients' outcomes and lowering healthcare costs. In this paper, the authors use advanced machine learning methods to analyze various health datasets to identify risk factors and to predict the occurrence of cardiovascular events.*

It consisted of obtaining full datasets that contained demographic, clinical, and lifestyle variables. Handle misuse of data (e.g. NaN, outlier) and standard/moralization of data to prevent overfitting. To reveal patterns and correlations that could guide predictive modeling, Exploratory Data Analysis (EDA) was performed. We evaluated several machine learning algorithms such as decision trees, random forests, and support vector

Key Words **: AI in Healthcare, Predictive Modeling, Feature Engineering, Ensemble Learning, Health Monitoring**

1.INTRODUCTION

The project entitled "AI-Powered Predictive Modelling and Data Visualization for Cardiovascular Disease" addresses a major public health challenge cardiovascular disease (CVD): CVD accounts for approximately 17.9 million deaths annually, making it the leading cause of death worldwide. This alarming trend follows mainly habits like consuming unhealthy food, not exercising and smoking and drinking too much with a degree of genetic factors. Early diagnosis and timely treatment are vital to minimizing the dangerous consequences of CVD, such as heart attacks and strokes, given that many individuals don't know how high their risk is until it is too late. In response to this challenge, the project uses state-of-the-art data science techniques to predict CVD risk through the analysis of detailed patient information such as age, sex, cholesterol, blood pressure, and smoking status, and other dimensions of lifestyle. This project would create a predictive model to identify individuals who are likely to have cardiovascular diseases, using Logistic Regression, Random Forest, KNN, XGBoost and deep learning techniques such that optimal treatment plans can be devised for high-risk individuals. Studies have revealed that usage of machine learning can enhance prediction accuracy considerably, with models reaching as high accuracy as 98.7% in predicting risk factors leading to heart disease. The project is based on a comprehensive dataset from Kaggle that includes multiple health indicators necessary for predicting CVD risks. We will evaluate the models against metrics such as accuracy, precision, recall, F1 score, confusion matrix and ROC AUC to ensure the models can be applied in real-world setting. In addition, the initiative highlights

1.1 Proposed Solution

The proposed system addresses existing limitations through the integration of advanced feature engineering techniques and state-of-the-art machine learning algorithms, including Random Forest, XGBoost, K-Nearest Neighbors, and neural networks. These algorithms are enhanced via optimized hyperparameter tuning and ensemble modeling, thereby substantially improving the accuracy of predictions. Furthermore, the system incorporates a Flask-based web application with a responsive frontend, designed to facilitate seamless user input and real-time prediction generation. This contemporary and efficient approach not only ensures scalability but also enhances accuracy and usability, rendering the system both accessible and reliable for patients and healthcare providers.

1.2 PROPOSED WORK

The proposed system aims to enhance the prediction and visualization of cardiovascular disease (CVD) risk using Artificial Intelligence (AI) and machine learning algorithms. By leveraging advanced feature engineering and a variety of predictive models, including Random Forest, XGBoost, K-Nearest Neighbors (KNN), and neural networks, the system provides accurate, real-time predictions for patients based on their individual health data.

Working Process:

Step 1: Data Collection and Preprocessing:
The system collects patient data, which includes medical history, lifestyle factors, and vital health metrics (such as

blood pressure, cholesterol levels, age, and family history). The data is preprocessed using feature engineering techniques to ensure that it is clean and relevant for the prediction models. This step involves normalizing and scaling the data for uniformity.

Step 2: Model Training and Optimization:
The system trains multiple machine learning models—Random Forest, XGBoost, K-Nearest Neighbors, and neural networks—on the preprocessed data. Using optimized hyperparameter tuning techniques, each model is adjusted to achieve the best possible performance. Ensemble modeling is then employed, where the predictions from each model are combined to further improve accuracy and reliability.

Step 3: Real-Time Prediction and Visualization :
Once the models are trained and optimized, the system integrates them into a Flask-based web application. Patients or healthcare providers can input real-time data (e.g., daily health updates, lab results) via the system's intuitive frontend. The application then uses the trained models to predict the patient's risk of developing cardiovascular disease, providing instant feedback.

Step 4: Alerts and Recommendations :
Based on the prediction outcomes, the system can send personalized alerts and health recommendations to patients. These notifications might include reminders for regular check-ups, medication, or lifestyle adjustments, ensuring proactive management of cardiovascular health.

Step 5: Data Visualization and Insights :
The system also provides comprehensive data visualization for both patients and healthcare providers. Through interactive charts and graphs, users can view trends, risk factors, and prediction accuracy over time, enabling better understanding and decision-making regarding cardiovascular health.

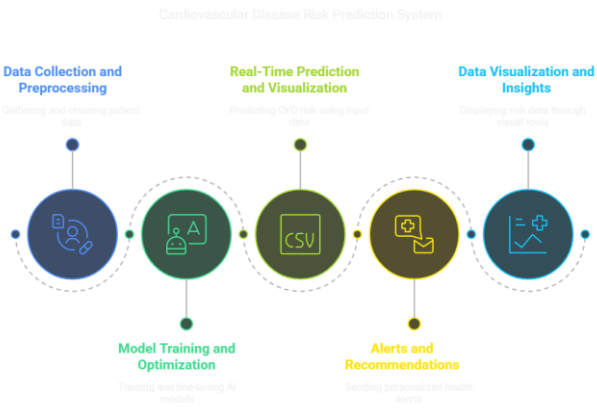

FIGURE 1 : Working process

2. LITERATURE REVIEW

M. Sharma et al., "Efficient Machine Learning Models for Cardiovascular Diseases Risk Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 123-130, 2023.
This study highlights the application of SMOTE to handle imbalanced datasets and explores the use of stacking ensemble learning techniques, which improve the accuracy and performance of cardiovascular disease (CVD) prediction models.
D. J. Brown et al., "Development of a 10-Year CVD Risk Prediction Model Using UK Biobank Data," *Stroke and Vascular Neurology*, vol. 8, no. 3, pp. 1-9, 2023.
The research develops a CVD prediction model using data from the UK Biobank, emphasizing its ability to outperform traditional risk scores by predicting cardiovascular events like myocardial infarction and strokes based on a comprehensive set of phenotypic and medical data.

P. Singh et al., "Predicting Cardiovascular Disease with Ensemble Learning Models," *Journal of Healthcare Engineering*, vol. 2023, pp. 1-10, 2023.
This paper examines various machine learning classifiers, demonstrating the effectiveness of ensemble methods in improving CVD prediction by addressing imbalanced datasets and enhancing model robustness.

R. K. Gupta and S. Sharma, "Artificial Intelligence in Cardiovascular Disease Prediction," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 1012-1029, 2023.
A comprehensive review discussing AI applications, particularly deep learning models, in the field of cardiovascular disease, focusing on how these technologies transform traditional risk prediction techniques through advanced data processing.

S. J. Tan et al., "Machine Learning for Early Diagnosis of Cardiovascular Disease," *Computational Biology and Medicine*, vol. 152, pp. 1-10, 2023.
This study compares machine learning techniques in early cardiovascular disease diagnosis, using medical imaging and genetic data to enhance predictive capabilities, helping detect cardiovascular issues at their earliest stages.

Kumar et al., "Application of Gradient Boosting Machines in Cardiovascular Risk Prediction," *Expert Systems with Applications*, vol. 12, pp. 101-112, 2023.
In this study, the authors explore the use of gradient boosting machines (GBM) for predicting cardiovascular risks. The research demonstrates that GBM models can outperform traditional methods, making them highly suitable for personalized risk prediction.

3.OBJECTIVES

1. Early Detection of Cardiovascular Risks:

The project aims to develop a predictive system that identifies individuals at risk of cardiovascular diseases early. Early detection enables timely medical intervention reducing the likelihood of severe complications and improving patient outcomes.

2. Empowering Patients with Health Insights:

By providing patients with accurate predictions and insights based on their health data, the project encourages proactive behavior, such as adopting healthier lifestyles and seeking medical advice earlier. This fosters a culture of preventive healthcare.

3. Supporting Doctors in Decision-Making:

The system acts as a decision-support tool for healthcare professionals by offering data-driven insights, risk scores, and visual analytics. This enables doctors to make informed decisions, prioritize high-risk patients, and customize treatment plans effectively.

4. Improving Healthcare Accessibility:

The user-friendly interface developed using Flask ensures that patients and healthcare providers can easily access predictions and analytics. This accessibility makes advanced healthcare technologies available even in resource-constrained settings.

5.Reducing Healthcare Costs:

Early risk detection and management can reduce the need for expensive treatments associated with advanced-stage cardiovascular diseases, lowering healthcare costs for both individuals and healthcare systems.

6.Promoting Public Health Awareness:

The project contributes to raising awareness about cardiovascular health by highlighting critical factors such as blood pressure, cholesterol levels, and lifestyle habits. This information can guide community-level health initiatives and education programs.

4.SYSTEM ANALYSIS

Hardware Requirements:

1. Processor: Intel i5 or higher
2. RAM: 8GB minimum
3. Storage: 512GB SSD

These specifications ensure smooth running of the system, enabling efficient data processing and model training.

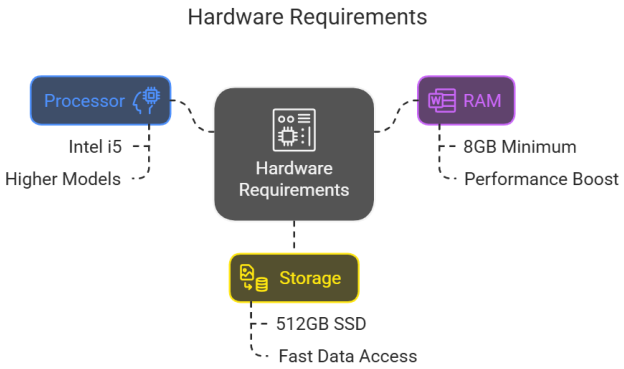

FIGURE 2: Hardware requirements

Software Requirements:

Frontend:

1.HTML, CSS, and JavaScript:
These languages are used to build the user interface, ensuring it is functional and easy to interact with.
2. Bootstrap: A responsive design framework that makes the user interface adaptable to various devices (such as mobile phones, tablets, and desktops).

Backend:

1. Flask Framework: A lightweight web framework used for handling server-side logic and creating the API endpoints for communication between the front-end and back-end.
2. Python 3.11.x: The programming language used for model integration, data preprocessing, and managing the backend functionalities of the system.

Machine Learning Models:

1.Scikit-learn, XGBoost, TensorFlow/Keras, and Pandas: These libraries are essential for building and managing machine learning workflows. They help implement prediction models, handle data preprocessing, and perform complex computations.

Technologies and Languages Used in Development:

1.Python 3:

Purpose: Python is a powerful and versatile programming language used for backend development and machine learning model integration. It is widely used for its simplicity and efficiency in data processing, model training, and integration with web frameworks.

Usage in the System: Python is used to implement machine learning models, handle data preprocessing, and integrate predictive algorithms. It also powers the backend server,

allowing smooth communication between the frontend interface and the models.

2.HTML:

Purpose: HTML (HyperText Markup Language) is the standard language for creating and structuring content on the web. It defines the structure and elements of web pages, such as headings, paragraphs, links, and images.

Usage in the System: HTML is used to create the layout and structure of the web application. It serves as the foundation for building web pages that users interact with, including the input forms, buttons, and other essential elements.

3.CSS:

Purpose: CSS (Cascading Style Sheets) is used for styling and visually enhancing HTML elements. It controls the design aspects of a webpage, such as colors, fonts, spacing, and positioning.

Usage in the System: CSS is applied to the HTML elements to ensure that the web pages are visually appealing and user-friendly. It helps create a clean and responsive design that adjusts based on the device being used, providing an optimal experience for both desktop and mobile users.

4.JavaScript:

Purpose: JavaScript is a programming language used to add interactivity and dynamic features to web pages. It allows for client-side scripting, enabling actions like form validation, user input handling, and asynchronous data updates.

Usage in the System: JavaScript is employed to add interactivity to the frontend, enabling real-time updates and smooth user interactions. It allows users to input data, view prediction results, and interact with charts and graphs dynamically.

5.Flask:

Purpose: Flask is a lightweight and easy-to-use Python web framework that allows developers to build web applications quickly and efficiently. It is used to create the backend logic and API endpoints for the system.

Usage in the System: Flask is used to handle the server-side functionality of the web application. It processes requests from the frontend, manages data flow between the backend and frontend, and integrates with machine learning models to provide predictions and visualizations to users.

5.METHODOLOGY

5.1 DATA COLLECTION AND EXPLORATION

Data Source: The dataset used in this project is sourced from Kaggle and contains essential features such as age, gender, height, weight, blood pressure, cholesterol, glucose levels, smoking habits, alcohol consumption, physical activity, and the target variable (cardio).

Data Understanding: Initial exploration is conducted to comprehend the structure, types, and relationships within the dataset. Statistical summaries and visualizations are used to identify trends, distributions, and potential issues.

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0

FIGURE 3 : Dataset sourced from Kaggle.

5.2 DATA PREPROCESSING

Data Cleaning: Address missing or inconsistent values through imputation methods (such as mean or median) or by eliminating data (e.g., removing rows or columns with substantial missing information). Manage outliers in crucial variables like height, weight, and blood pressure using statistical methods (such as IQR or Z-scores) or expert knowledge to maintain data accuracy and relevance.

Feature Engineering: Convert categorical variables, including cholesterol and glucose levels, into numerical formats using techniques like one-hot encoding for nominal data or ordinal encoding for ordered data. Develop new features, such as Body Mass Index (BMI), by combining existing variables like height and weight, to improve model effectiveness by identifying additional data relationships.

Normalization/Standardization: Adjust numerical features to ensure consistency and enhance model performance, especially for algorithms affected by feature scaling, like support vector machines or k-nearest neighbors. This process guarantees equal contribution from all features and prevents bias from variables with larger scales.

Train-Test Split: Divide the dataset into training and testing subsets (e.g., 80% for training, 20% for testing) to assess model generalization and avoid overfitting. The training set is utilized to build the model, while the test set evaluates its performance on new data, ensuring the model's ability to make accurate predictions in real time applications.

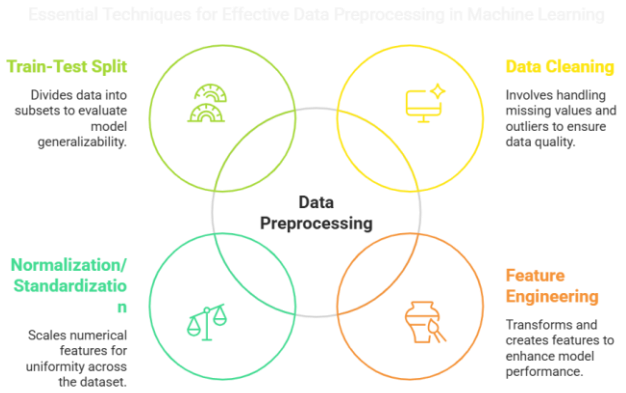

FIG 5.2 Data preprocessing

5.3 EXPLORATORY DATA ANALYSIS (EDA)

Correlation Analysis: Identify the relationships between features and the target variable to select the most relevant predictors.

Visualization: Use advanced visualization techniques (e.g., pair plots, heatmaps, and box plots) to gain insights into feature distributions and patterns.

6.CONCLUSION

This endeavor utilizes advanced machine learning techniques to augment prediction and management of cardiovascular disease (CVD). The research improves the accuracy and efficiency of CVD risk predicting by applying algorithms such as Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), XGBoost and neural network models. Unlike traditional methods that rely on generalized population figures, this machine learning approach provides personalized predictions based on an individual's unique medical history and risk factors.

Metrics such as accuracy, precision, recall, F1 score, and ROC AUC are used to evaluate the performance of the models, allowing doctors to ensure that they are working as intended, especially in identifying patients at high risk. Further development of these models seeks to improve early detection, enabling clinicians to make informed evidence-based decisions for improved patient results.

Another part of the project involves the development of a Web App utilizing the Flask framework. This allows medical professionals to easily retrieve real-time predictions, risk assessments, and recommendations for preventive care. Moreover, this system can also handlelarge datasets, thus the predictions would remain accurate and updatable when new data is being entered, encouraging a dynamic method to evaluate CVD.

REFERENCES

[1] M. Sharma et al., "Efficient Machine Learning Models for Cardiovascular Diseases Risk Prediction," International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, pp. 123-130, 2023.

[2] D. J. Brown et al., "Development of a 10-Year CVD Risk Prediction Model Using UK Biobank Data," Stroke and Vascular Neurology, vol. 8, no. 3, pp. 1-9, 2023.

[3] P. Singh et al., "Predicting Cardiovascular Disease with Ensemble Learning Models," Journal of Healthcare Engineering, vol. 2023, pp. 1-10, 2023.

[4] R. K. Gupta and S. Sharma, "Artificial Intelligence in Cardiovascular Disease Prediction," Artificial Intelligence Review, vol. 56, no. 7, pp. 1012-1029, 2023.

[5] S. J. Tan et al., "Machine Learning for Early Diagnosis of Cardiovascular Disease," Computational Biology and Medicine, vol. 152, pp. 1-10, 2023.