Heart Attack Analysis Using Ensemble Machine Learning

Dr. Raviprakash ML
Professor
Dept of CSE, KIT Tiptur

Inchara H P
incharahp3008@gmail.com
Dept of CSE, KIT Tiptur

S Pallavi
pallavihaps123@gmail.com
Dept of CSE, KIT Tiptur

Sheetal K M
ssheetalkm@gmail.com
Dept of CSE, KIT Tiptur

Nitish N Kotumuchagi
nitishnkotumuchagi@gmail.com
Dept of CSE, KIT Tiptur

***Abstract*—Heart disease remains one of the leading causes of mortality worldwide, emphasizing the urgent need for accurate and timely diagnosis. While traditional diagnostic methods have proven effective, advancements in machine learning (ML) offer promising avenues for enhanced detection and prevention strategies. This paper presents a comprehensive review of existing ML-based approaches for heart disease detection, ranging from classical statistical methods to cutting-edge deep learning techniques. We begin by outlining the various risk factors associated with heart disease, including hypertension, cholesterol levels, and lifestyle choices. Subsequently, we delve into the evolution of ML in healthcare, highlighting its transformative impact on diagnostic accuracy and patient care. Leveraging large-scale datasets and feature engineering, traditional ML algorithms such as Support Vector Machines (SVM) and Random Forests have demonstrated notable success in identifying cardiac abnormalities. However, recent breakthroughs in deep learning, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have revolutionized heart disease detection by extracting intricate patterns and temporal dependencies from raw data sources.**

I. Introduction

The human body. It's imperative for individuals to prioritize heart care, given its significance. Various diseases are interconnected with heart health, underscoring the necessity for predicting heart attacks. Currently, many patients succumb to heart attacks, often identified only in the advanced stages. This unfortunate trend stems from insufficient instrumentation to accurately predict heart attacks using efficient algorithms. Healthcare industries face challenges in accurately predicting heart attacks at an early stage. The complexity of health history statistics and the inconsistency of real-world physical data make this task difficult.[1]

Researchers are exerting considerable effort to develop a prototype capable of accurately predicting heart attacks in the early stages, yet they encounter challenges in creating a suitable model. Each proposed structure has its own set of strengths and weaknesses. Machine learning systems have been trained to comprehend and utilize data effectively, marking the intersection of both machine intelligence and technology. Machine learning, as explained, learns from regular patterns in data.

Different researchers have focused on reducing cardiovascular features and extracting nonlinear features using discriminant analysis. Fisher's method was employed in the experiment to address overfitting issues and enhance training speed.

This study aims to optimize machine learning models for predicting heart disease while addressing the challenge of overfitting, particularly within Logistic Regression. By drawing random samples from the complete dataset, overfitting issues can be mitigated effectively. Additionally, the model training is conducted on data samples sourced from the UCI Machine Learning repository. Consequently, the primary objective of this research is to enhance the accuracy of heart disease prediction.[2]

Over recent decades, research endeavors have underscored the critical role of data mining in augmenting clinical diagnosis, particularly in forecasting ailments such as heart disease. Factors ranging from prevalent conditions like diabetes, hypertension, elevated cholesterol levels, to subtler indicators like abnormal heart rates, collectively contribute to the complex mosaic of cardiovascular health. However, amidst this abundance of data, challenges persist. Incomplete or fragmented medical records often obscure the path to accurate diagnosis, impeding the efficacy of traditional diagnostic approaches.[3]

Additionally, crucial diagnostic tools like electrocardiograms and CT scans, vital for identifying coronary heart disease, are frequently inaccessible due to their high cost, particularly in low- and middle-income countries. Thus, early detection of heart disease is imperative to alleviate both its physical and financial toll on individuals and healthcare systems. A World Health Organization (WHO) report forecasts that by 2030, the total number of CVD-related deaths will surge to 23.6 million, primarily attributable to heart disease and stroke. Hence, leveraging data mining and machine learning techniques to predict the likelihood of developing heart disease becomes pivotal for saving lives and mitigating the societal economic burden.[4]

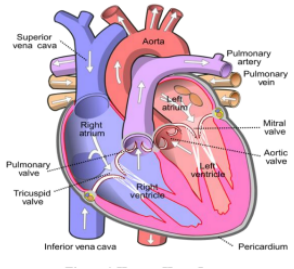

Fig. 1. Heart

age	49	64	43	69
cp	1	3	2	0
trestbps	120	150	172	135
chol	239	219	283	233
fbs	0	1	0	1
thalach	178	163	174	114
exang	0	1	0	1
old peak	1.4	0.6	1.8	0.8
thal	1	2	1	2
target	0	1	0	1

TABLE I
DATASET

In the medical realm, an extensive volume of data is generated daily through data mining techniques, revealing concealed patterns applicable to clinical diagnosis. Consequently, data mining assumes a critical role in the medical domain, as evidenced by research conducted over recent decades. Numerous factors, including diabetes, hypertension, elevated cholesterol levels, and abnormal heart rate, must be factored in when forecasting heart disease. Frequently, medical data may be incomplete, impacting the accuracy of heart disease prediction.[5]

II. Literature Survey

In recent years, the healthcare sector has experienced notable progress in leveraging data mining and machine learning techniques. These methodologies have been widely embraced and have showcased effectiveness across various healthcare domains, particularly in medical cardiology. The substantial accumulation of medical data has provided researchers with an unprecedented opportunity to innovate and validate new algorithms in this sphere. Heart disease persists as a primary cause of mortality in developing nations, and there's a growing emphasis on identifying risk factors and early indicators of the disease. The integration of data mining and machine learning techniques holds promise for enhancing early detection and preventive measures against heart disease.[6]

The study conducted by Narain et al. (2016) aims to develop an innovative machine-learning-based system for predicting cardiovascular disease (CVD) with higher precision compared to the widely utilized Framingham risk score (FRS). By leveraging data from 689 individuals displaying CVD symptoms and validating it with the Framingham research dataset, the proposed system utilizes a quantum neural network to learn and identify patterns associated with CVD. The system was experimentally validated and compared with the FRS. Results indicated that the proposed system achieved an impressive accuracy rate of 98.57% in forecasting CVD risk, significantly surpassing the FRS's accuracy of 19.22% and other existing techniques. The study suggests that this approach could serve as a valuable tool for healthcare professionals in predicting CVD risk, aiding in the formulation of more effective treatment plans, and facilitating early diagnosis.[7]

In a study by Drod et al. (2022), the aim was to utilize machine learning (ML) techniques to pinpoint the most significant risk factors for cardiovascular disease (CVD) among patients with metabolic-associated fatty liver disease (MAFLD). Blood biochemical analysis and subclinical atherosclerosis assessment were conducted on 191 MAFLD patients. ML approaches such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA) were employed to construct a model identifying those at highest risk of CVD. The study identified hypercholesterolemia, plaque scores, and duration of diabetes as the most crucial clinical characteristics. The ML technique performed well, accurately identifying 85.11% of high-risk patients and 79.17% of low-risk patients, with an AUC of 0.87. This study suggests that ML methods can effectively detect MAFLD patients with widespread CVD based on straightforward patient criteria.[8]

In another study by Alotalibi (2019), the objective was to explore the utility of ML techniques for predicting heart failure disease. The study utilized a dataset from the Cleveland Clinic Foundation and implemented various ML algorithms including decision trees, logistic regression, random forest, naive Bayes, and support vector machine (SVM) to develop prediction models. A 10-fold cross-validation approach was utilized during model development. Results indicated that the decision tree algorithm achieved the highest accuracy in predicting heart disease at 93.19%, followed closely by the SVM algorithm at 92.30%. This research underscores the potential of ML techniques as effective tools for predicting heart failure disease, with the decision tree algorithm showing promise for future research endeavors.[9]

In one study, researchers designed an electronic health record (EHR) model using sequential modeling with the utilization of a neural network. The EHR was employed for conducting experiments and predicting heart disease. The researchers utilized word vectors and hot encryption for modeling diagnostic scenarios and predicting cardiac failure. Additionally, an extended memory model based on the

network was applied. The study emphasized the importance of considering the sequential nature of healthcare through results analysis. This sequential aspect includes tracking a person's behavior such as health-related activities, changes in healthcare providers during sickness, exercise routines, diet routines, and more.

In another study, it was suggested that Principal Component Analysis (PCA) can serve as an effective dimensionality reduction technique for handling data with high dimensions and variance. PCA enables the preservation of more information through the creation of new components. When dealing with high-dimensional data, many researchers opt to employ PCA. In a separate study, five unsupervised dimensionality reduction techniques, both linear and nonlinear, were utilized alongside neural networks as a classifier to classify cardiac arrhythmia. Remarkably, with a minimum of 10 components, a remarkable F1 score of 99.83% was achieved using fast independent component analysis (FastICA), which was employed for Independent Component Analysis (ICA) in the context of breast cancer diagnosis.

In an effort to enhance performance, another researcher employed PCA approaches to time-frequency representations, aiming to reduce heart sounds. Additionally, a scale-invariant feature, Principle Component Analysis-K-Nearest Neighbor (PCA-KNN), was implemented in medical imaging for scaling purposes. This novel approach for diverse medical images achieved an accuracy of 83.6%, utilizing 200 images for training the machine. Moreover, a gray-level threshold of 150 was utilized as a result of PCA and Return on Investment (ROI), effectively reducing X-ray picture characteristics.

This project aims to predict whether a person is at risk of a heart attack and provide recommendations based on the prediction. The Random Forest algorithm has shown promising accuracy rates in achieving this goal. Below is a sample of the dataset used.

III. Methodology

In this study, a comparison of different machine learning techniques is conducted to predict the ten-year risk of coronary heart disease based on patients' medical data. The proposed methodology is illustrated in the following flowchart.

The input for the study is a dataset related to cardiac disease, which undergoes preprocessing by replacing missing values with the respective column means. Four distinct methods are employed in this research, as depicted in Figure 3. The study evaluates the accuracy metrics of these machine learning models, with the aim of using the best-performing model for prediction purposes.

The study is structured as follows:

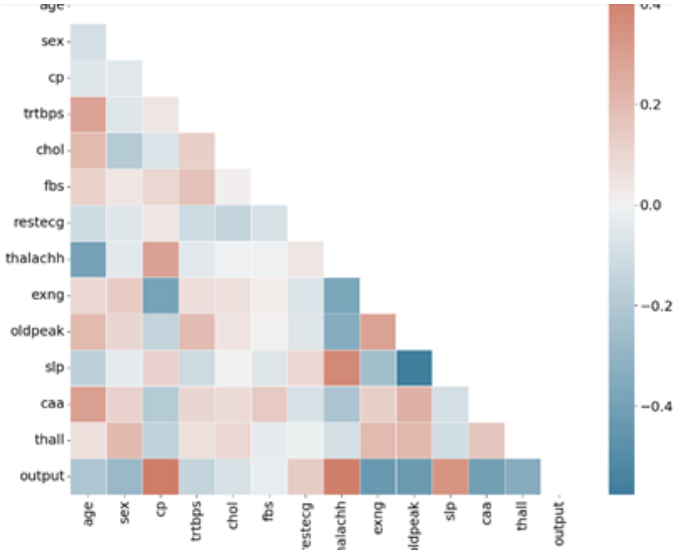

Fig. 2. Diagonal correlation matrix graph

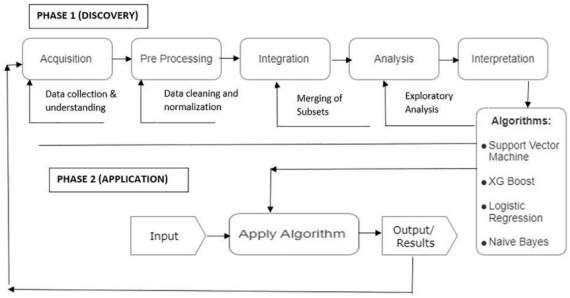

Fig. 3. table

Exploratory data analysis (EDA) is employed to detect errors, identify relevant data, and assess relationships between variables. This study focuses on risk factors related to heart disease and aims to predict heart attacks.

Machine learning classifiers such as logistic regression, support vector machines (SVM), naïve Bayes, and XGBoost are utilized. The selection of these classifiers is based on a comprehensive literature review of previous experiments in heart disease prediction, considering their performance attributes.

The experiment is conducted using the Cleveland dataset, comprising 294 tuples with 14 attributes. Figure 3 illustrates the process flow.

Data acquisition is the initial step, involving the evaluation of physical conditions and conversion of numeric data into computer-usable samples. The data is sourced from the UCI ML repository, providing multiple attributes to analyze heart disease risk factors.

1. **Pre-processing**: - Addressing data issues like missing values, outliers, and redundancy to clean the dataset. a. Cleaning the collected data through techniques such as missing values replacement, outlier detection, and duplicate removal. b. Replacing missing values with mean values. c. Identifying outliers using Boxplots and understanding data ranges. d. Removing duplicates using dictionary functions.

2. **Integration**: - Combining libraries and subsets by importing modules in Python and merging them for experiments. a. Using preprocessed data for experiments. b. Integrating cleaned data to apply machine learning (ML) algorithms.

3. **Analysis**: - Conducting exploratory data analysis (EDA) to understand data attribute relationships. a. EDA focuses on learning from data, pattern identification, and decision-making with minimal human intervention. b. Analyzing data attribute relationships, comparing variables, and using visualizations like boxplots and heatmaps.

4. **Intervention**: - Developing decision-making strategies through literature review and understanding of previous experimental studies. a. Conducting a detailed literature survey to identify promising ML models for optimizing results in heart disease prediction. b. Selecting promising models based on their performance in similar domains.

5. **Application of ML algorithms**: - Utilizing machine learning models for predictions, including SVM, Naïve Bayes, Logistic Regression, and XGBoost. a. Applying SVM using scikit-learn's svm extension in Python. b. Using Naïve Bayes classifier from scikit-learn's neighbors library. c. Employing Logistic Regression with sklearn's linear model class in Python. d. Leveraging XGBoost, a boosting algorithm, for optimized results using weak classifications.

The dataset used in this study, as outlined in [23], consists of 70,000 patient records with 12 unique features, detailed in Table 2. These features encompass age, gender, systolic blood pressure, and diastolic blood pressure. The target class, labeled "cardio," signifies whether a patient has cardiovascular disease (denoted as 1) or is considered healthy (denoted as 0).

We propose utilizing binning as a technique to transform continuous inputs, like age, into categorical inputs. This approach aims to enhance the performance and interpretability of classification algorithms. By grouping continuous inputs into distinct bins or categories, the algorithm can differentiate between various data classes based on specific input variable values. For example, if we categorize age into "Young," "Middle-aged," and "Elderly," a classification algorithm can effectively classify data based on age groups.

```
#read csv file
df=pd.read_csv('/content/drive/MyDrive/heart /heart.csv')
df
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Fig. 4. Datasets for input

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Fig. 5. DataFrame

IV. Conclusion

the utilization of the Random Forest algorithm for heart disease detection marks a pivotal milestone in predictive healthcare analytics. By amalgamating insights from multiple decision trees, this approach not only boosts early detection accuracy but also fosters proactive interventions, potentially saving lives. The envisioned enhancements, such as family notification systems and seamless integration with healthcare facilities, underscore the potential for technology to revolutionize patient care. Moreover, the versatility of machine learning algorithms highlights their broader applicability across various medical domains, promising a future where predictive analytics serve as indispensable tools

in personalized medicine and proactive health management.

the integration of machine learning algorithms like Random Forest into healthcare systems signifies a paradigm shift towards data-driven and personalized medicine. By harnessing vast amounts of patient data, these algorithms can uncover intricate patterns and correlations that might elude human perception, thereby augmenting diagnostic capabilities and treatment strategies. Furthermore, as these algorithms continue to evolve and incorporate new functionalities, such as real-time data analysis and online consultations, they hold the potential to bridge gaps in healthcare accessibility and affordability, especially in underserved communities. Ultimately, the convergence of cutting-edge technology and medical expertise heralds a future where predictive analytics not only aid in disease detection but also empower individuals to take proactive steps towards optimizing their health and well-being.

.

REFERENCES

[1] Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904.

[2] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.

[3] Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

[4] L, G., & Raviprakash M L. (2023). Machine Learning Defence Mechanism for Securing the Cloud Environment. International Journal of Advanced Scientific Innovation, 5(1). https://doi.org/10.5281/zenodo.7712783

[5] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.

[6] G. L, R. M L, G. H B, K. Mathada and M. B, "Intelligent Resume Scrutiny Using Named Entity Recognition with BERT," 2023 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2023, pp. 01-08, doi: 10.1109/ICDSNS58469.2023.10245304.

[7] L. Girish, M. L. Raviprakash, D. K. Thara, T. S. Prathibha and T. V. Rashmi, "Stock Market Time Series Forecasting using Long Short-Term Memory," 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2023, pp. 1-6, doi: 10.1109/ICIICS59993.2023.10421259.

[8] Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. J. Occup. Health 2016, 58, 216–219. [Google Scholar] [CrossRef] Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. Procedia Comput. Sci. 2016, 85, 962–969.

[9] Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Int. J. Comput. Appl. 2011, 17, 43–48.

[10] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554.

[11] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.

[12] Girish L, & Raviprakash M L. (2022). Data analytics in SDN and NFV: Techniques and Challenges. International Journal of Advanced Scientific Innovation, 4(8). https://doi.org/10.5281/zenodo.7657569

[13] Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. Comput. Methods Programs Biomed. 2013, 111, 62–71.

[14] K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. Int. J. Comput. Appl. 2011, 19, 6–12.

[15] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability.