e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 9, September 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

6381 907 438 6381 907 438 ijmrset@gmail.com www.ijmrset.com

© 2024 IJMRSET | Volume 7, Issue 9, September 2024| DOI: 10.15680/IJMRSET.2024.0709020

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Models for Prediction of Cardiovascular Disease by Health and Demographic Data

Amrutha D, Shifa A M, Vachana A C, Prof. Bharathi Ramesh

PG Student, Department of M.C.A, Surana College (Autonomous), Kengeri, Bangalore, India

Assistant Professor, Department of M.C.A, Surana College (Autonomous), Kengeri, Bangalore, India

ABSTRACT: In the modern world, cardiovascular diseases are the primary cause of death. Consequently, there is a growing need for early and accurate CVD prediction in order to enable prompt interventions with successful treatment.The research looks at the creation and assessment of several machine learning models intended to forecast probability depending on health and demographic variables such as age, gender, cholesterol, and kinds of chest pain. Convolutional neural networks, logisticc regression, random forest classifiers, decision tree classifiers, and linear regression are some of the machine learning models developed for this purpose. We evaluated both models' performances using mean squared error and accuracy.In fact, the CNN's accuracy exceeded that of all other models, indicating that it has great potential as a CVD prediction tool. Additionally evaluated, the Random Forest model showed a high capacity to strike a balance between interpretability and accuracy. This emphasizes how important machine learning research is to improving CVD prediction and provides insightful information for the creation of new models in the future.

KEYWORDS: Cardiovascular disease, Machine learning, CNN, Predictive modeling, Accuracy, Mean squared error

I. INTRODUCTION

The broader field of artificial intelligence gives rise to machine learning, while predictive analytics is undergoing a revolution as algorithms learn from massive datasets to produce precise predictions and identify patterns that might not be seen with other conventional statistical techniques[1].With machine learning, it has been possible to write algorithms that can process and analyze much larger volumes of highly variable data, such as medical record history, genetic information, lifestyle factors, and biomarkers [2]. Conventional models rely on a small set of input variables and predetermined equations. These algorithms will become more predictive as exposure rises, making it possible to identify intricate linkages and subtle patterns among CVD risk factors [3]. This is especially helpful for predicting CVD because complex interplay between genetic predisposition, environmental factors, and lifestyle choices drive underlying illnesses[4].Decision Trees, Random Forests, Logistic Regressions, and CNN are ML models that integrate a wide range of variables with subtle relationship discrimination to provide high prediction accuracy[5]. Better capacity makes it possible to identify individuals who are at risk early on, allowing medical professionals to provide targeted and individualized interventions. In order to maximize the effectiveness of the use of health resources and patient outcomes, it therefore improves risk assessments, leading to more individualized approaches to disease management[6].A broad spectrum of disorders affecting the heart and blood arteries are included in the category of cardiovascular diseases. Heart failure, stroke, and coronary artery disease are a few of these. When taken as a whole, they rank among the world's major causes of death, accounting for about 17.9 million deaths annually[7].[8] The World Health Organization claims that this is an impressive statistic that illustrates the widespread impact of CVD on global health and the significant strain it places on healthcare systems. Heart disease is fairly common. Multiple factors contribute to the pathophysiology, so it's important to avoid, identify, and treat the diseases early on.Several purely statistical models have been used to evaluate traditional risk for CVD. The most well-known of these is the Framingham Risk Score, which calculates probability based on a very small number of risk factors, including age, blood pressure, and cholesterol levels[9].Despite the fact that these models have proved extremely helpful and instructive, they typically lack the resolution necessary to account for the intricate interplay between genetic,

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

environmental, and lifestyle factors that contribute to the development of CVD[10]. This constraint has been highlighted by the requirement for sophisticated predictive instruments capable of comprehending the intricacy associated with cardiovascular risk [11].As such, ML technology is changing how CVD prediction is approached. One may argue that machine learning is a branch of artificial intelligence concerned with creating algorithms that are able to learn from data and forecast future events[12].

ML algorithms, as opposed to conventional statistical techniques, are able to process massive amounts of data, identify intricate patterns, and produce predictions that are more accurate. This capacity is especially crucial for cardiovascular disease (CVD), since the illness frequently results from the intricate interaction of multiple factors, including as comorbidities, genetic predisposition, and lifestyle decisions, all of which have a major impact on disease risk[13].While machine learning in the various forms of CVD predictions is one area where this holds great promise for the advancement of healthcare, large volumes of medical history, demographic data, lifestyle choices, and biomarkers are analyzed by ML algorithms for far more personalized risks than can be achieved with current typical methodologies [14].Improved prognoses from medical professionals enable considerably earlier identification of persons at risk and focused therapies to stop cardiovascular problems from getting worse [15].Because machine learning can provide personalized therapy and preventive methods based on the unique profile of each patient, it can greatly aid in the prediction of cardiovascular disease (CVD). By focusing therapies on people who are most likely to benefit, personalized medicine aims to improve both individual outcomes and the effective use of healthcare resources[16].[17]While this offers great promise for machine learning in the prediction of CVD, a number of obstacles still need to be overcome, including the creation of interpretable models that clinicians can trust in, the integration of models into current processes in healthcare, and the availability of high-quality and diverse datasets. These obstacles will surely be resolved in order to fully utilize ML capabilities for the advancement of cardiovascular disease research and treatment[18].Evaluating how well different machine learning models predict cardiovascular illness will add to the corpus of knowledge already available in this field[19].This study examines the most effective methods for CVD prediction by evaluating several models, including CNNs, Decision Trees, Random Forests, Logistic Regressions, and Linear Regressions. This also serves to highlight the possibility that applying machine learning could result in early detection and better results for the many individuals afflicted by cardiovascular illnesses [20].

II. LITERATURE REVIEW

In order to achieve the best CVD prediction outcome, Khalil 2024 proposed a hybrid classification model that included and contrasted multiple ML techniques. Its effectiveness was demonstrated in enhancing heart health predictions using suitable classifiers. This work has shown the benefits of combining several techniques to obtain better prediction outcomes[14].

The RF model outperformed all other models in Suleiman et al. (2023)'s investigation for CVD predictions, obtaining an accuracy of 90\%, an F1-score of 1.00, and an AUC of 1.00. Once again, RF is the best performer, thus it will be interesting to see how these models can be improved and validated for patient populations[15].Five models were examined using machine learning (ML) to assess the risks of CVD in a dataset comprising 1,189 patients in one study. The authors point out that future research will probably continue this trend because machine learning (ML) can handle larger datasets. Deep learning methods are also being explored as a way to improve the prediction model's accuracy[16].

For instance, Bagadi et al.'s study looked at how elastic feature selection affected the model's ability to predict CVD. Their comprehensive examination of ensemble and deep learning approaches revealed noticeably better prediction accuracy than that of conventional methods. On the other hand, despite obstacles like poor data quality and algorithmic transparency, interdisciplinary cooperation and upcoming technology developments will lead to better CVD management[17].In general[18], Pe et al. presented a CVD prediction application in 2021 employing 14 parameters, although the industry typically uses 10. This work demonstrated the impact of additional characteristics, such as diabetes, on the prediction of heart disease and highlighted the potential for increased diagnostic accuracy through the integration of more complex parameters into more sophisticated algorithms.In order to predict CVD, Arunachalam et al. 2020 investigated a variety of ML classifiers, including Multi-Layer Perceptrons, RF, SVM, Gradient Boosting,

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Extra Trees, and Logistic Regression. He stated that the accuracy of MLP and SVM utilizing fourteen attributes was 91.7\%. The SVM and MLP perform the best out of all the approaches in this publication, however further research on ensemble methods and parameter analysis is encouraged[19].

Using 4,238 records from the Framingham Health Study, Chauhan demonstrated the results of five ML classifiers: RF, KNN, SVM, Decision Trees, and Logistic Regression. The most accurate method was found to be logistic regression, which has an accuracy range of 0.850-0.900. This shows how predictive modeling may be used to diagnose conditions like hypertension a little earlier[20].In 2020, a study[21] analysis was carried out for the prediction of CVD utilizing a variety of ML techniques, including support vector machines, KNNs, decision trees, and ANNs. The Kaggle dataset was used to make the prediction. The greatest accuracy achieved with ANN, based on TensorFlow Keras, is 85.24\%. This strategy has outperformed previous strategies. According to the study, having a suitable prediction system is essential, and using efficient deep learning algorithms greatly enhances the process of predicting heart disease.In 2024, Bhavsar and Patel examined a few machine learning algorithms for the prediction of CVD, with a focus on data processing and feature selection. Additional future options that have been discussed include applying deep learning to larger datasets in order to build better methodologies, which will lead to more accurate results and more therapeutic applications (Explainable AI, for example) [22].

In 2022, Reddy conducted research on a variety of machine learning algorithms that are utilized in CVD diagnosis. The work focused on accuracy, sensitivity, and specificity and used a variety of methodologies to achieve these goals. Numerous models and their diverse features, ranging from blood pressure to family history, were explored. The review highlights how crucial it is for healthcare systems to integrate data in order to enhance their predictive performance[23].

Ullah et al. (2024) presented a scalable machine learning system with 100\% accuracy for the early identification of CVD using Random Forest and Extra Tree classifiers. The study employed optimized features, including FCBF, MrMR, and Relief, the computation of which was integrated. The FCBF's effective performance in big datasets, together with the potential for future breakthroughs in early CVD detection [24].According to this theory, machine learning classifiers were used in the Pal et al. 2022 study to predict CVD. Since K-Nearest Neighbors was the classification technique used, KNN was the most accurate algorithm, followed by Random Forest and LDA. Research has shown that the addition of new features improves model precision, albeit at the expense of computing time. After that, it suggested more study on cutting-edge techniques and worked with hospitals to create models that reflected those findings [25]. By highlighting various model performances, Kaushik et al. conducted a review of a few machine learning techniques in the context of CVD predictions.

This highlighted even more how crucial accurate disease prediction is, and how using these ML techniques to clinical practice can greatly enhance the accuracy of diagnosis results. In 2023, a study used ML classifiers to predict cardiovascular illness. The outcomes showed that very high accuracy was achieved by models like Random Forest. This study came to the conclusion that in order to further increase the prediction accuracy for clinical decision assistance, sophisticated methodologies must be used in conjunction with ideal feature selection. A review research [28] examined a number of ML techniques that were applied to the IEEE and Kaggle datasets in order to predict CVD. Random Forest outperformed the others with a maximum accuracy of 94\%, demonstrating the efficacy of ML. However, adding additional clinical data will always be necessary in order to enhance patient outcomes.

The use of machine learning models—more especially, the K-NN and multi-layer perceptron (MLP) algorithms—for the early detection of cardiovascular disease (CVD) is investigated in this work. The study uses a dataset that was preprocessed to remove outliers and null values from the University of California, Irvine (UCI) repository. Training accounted for 80\% and testing for 20\% of the 303 data samples. The MLP model outperforms the K-NN model with an accuracy of 82.47\% and an area under the curve (AUC) of 86.41\%. The work highlights the potential of the MLP model for accurate automatic identification of CVD, suggesting that other diseases could benefit from the early diagnostic strategy[12].[9]The study looks into how machine learning (ML) techniques can be used to forecast cardiovascular illnesses (CVDs), which are the main cause of death worldwide Support Vector Machines (SVM), Random Forests, K-Nearest Neighbors (KNN), and Naive Bayes are some of these methods. It is encouraging that ML

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

models have shown an accuracy range of 80\% to 99\%. This article also addresses the significance of feature selection and data preparation in enhancing prediction accuracy. All things considered, ML algorithms are hailed as practical tools for improving the diagnosis and management of CVD.[13]The project investigates the use of machine learning (ML) approaches for cardiovascular disease (CVD) prediction in order to improve early detection and diagnosis. The Catboost model is a shining illustration of the potential of machine learning models; it achieved great accuracy (90.94\%) and precision (92.3\% F1-score) in detecting early-stage heart disease. The study shows that machine learning (ML) can reduce the need for extensive clinical testing and save healthcare costs by using large health data sets for more efficient sickness management and prediction.

Table.1 Key findings of the research paper

Reference	Study Focus	Models Compared	Key Findings and Future Work
Khalil, Md. (2024)[20]	Hybrid classification for CVD prediction	Decision Tree, Random Forest, Logistic Regression, SVM, KNN	Highlights model effectiveness in CVD prediction. Select appropriate classifiers.
Suleiman,Aminu, Luka, Stephen,Ibrahim,Muhammad (2023) [21]	CVD prediction	Random Forest, others	RF outperformed other models. Validate and optimize across populations.
Islam,Taminul, Vuyia, Adifa, Hasan, Mahadi, Rana, Md. Masum (2023) [22]	CVD risk analysis Decision	Tree, Random Forest, Logistic Regression, SVM, KNN	ML handles large datasets. Use larger datasets, deep learning.
Bagadi, Kalapraveen and others[23]	CVD risk prediction	Ensemble Methods, Deep Learning	Improvement over conventional methods. Address data quality, transparency.
Arunachalam,Siddhika (2020) [25]	CVD forecasting	SVM, MLP, Logistic Regression, Decision Trees, Random Forest	SVM, MLP achieved superior performance. Improve models with ensemble methods.
Chauhan, Yash (2020)[26]	CVD prediction	Logistic Regression, KNN, SVM, Decision Trees, Random Forest	LR had the highest accuracy. Early diagnosis of hypertension.
Bhavsar, Maitri and Patel, Manish (2024) [28]	Early CVD detection	Random Forest, SVM, KNN, Naive Bayes, Gradient Boosting	Effective feature selection needed. Use Explainable AI, larger datasets.
Reddy, Anuradha (2022)[29]	CVD diagnosis	Various ML models	Highlights importance of integrating data. Identify effective
Ullah, Tahseen and others (2024)[30]	Early CVD detection	Extra Tree, Random Forest	Optimal feature selection. Enhance early diagnosis
Pal, Madhumitha and others (2022) [31]	Early CVD detection	MLP, K-NN	MLP outperformed K-NN. Recommends MLP for automated detection
Kaushik, Rohit and others (2023)[32]	CHD prediction in India	Random Forest	Supports clinical decisionmaking. Focus on managing high-dimensional data
M.Ramu and others (2023)[33]	CVD risk prediction	Seven classifiers (including Random Forest)	RF achieved highest accuracy. Enhance accuracy, integrate clinical info.

© 2024 IJMRSET | Volume 7, Issue 9, September 2024| DOI: 10.15680/IJMRSET.2024.0709020

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. OBJECTIVES

The objectives of the research study will be to:
• CVD with CNN, Decision Tree, Random Forest, and Regression models.
• Determine the risk factors for CVD and use machine learning to detect and treat the disease early.

IV.METHODOLOGY

The prediction of cardiovascular disease is done using four different algorithms, and the outcomes are also contrasted. The architecture diagram for cardiovascular disease prediction is displayed in Fig. 1
• Decision Tree Classifier
• Random Forest Classifier
• Logistic Regression
• Linear Regression
• Convolutions Neural Network

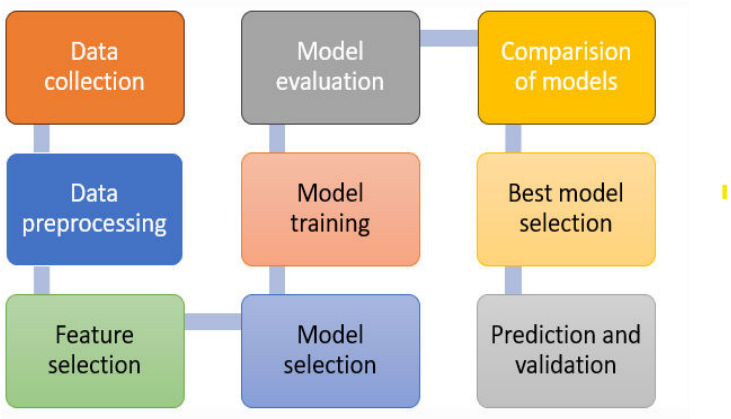

Fig 1.Methodology of proposed research

V. DATA DESCRIPTION

There are an aggregate of 912 observations in this data set, which is obtained from Kaggle, hence health and demographic data on the following 16 variables: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, Smoke Status, Weight, Diet, Medical History, Family History, and ST Slope. Hence, this target variable will be HeartDisease, explaining whether the patient does or does not have heart disease.

Table.2 Categorical variable encoding

Features	Description	Encoding Type
Age	The age of the patient in years	Numerical
Sex	Gender of the patient	0=Female, 1=Male
ChestPainType	Type of chest pain experienced	Categorical: Typical angina, Atypical angina, Non-anginal pain, Asymptomatic
RestingBP	Resting blood pressure in mm Hg	Numerical
Cholesterol	Serum cholesterol level in mg/dL	Numerical
FastingBS	Fasting blood sugar level	1 = 120 mg/dL, 0 = 120 mg/dL

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

RestingEOG	Results of the resting electrocardiogram	Categorical: Normal, STT wave abnormality, Left ventricular hypertrophy
MaxHR	Maximum heart rate achieved during exercise	Numerical
ExerciseAngina	Exercise-induced angina	1=Yes, 0=No
Oldpeak	ST depression induced by exercise relative to rest	Numerical
Smoke_Status	Smoking status of the patient	0=Non-smoker, 1=Smoker
Weight	Weight of the person in kilograms	Numerical
Diet	Dietary habits of the patient	0=Unhealthy, 1=Healthy
Medical_History	Presence of medical history relevant to CVD	0=No, 1=Yes
Family_History	Family history of CVD	0=No, 1=Yes
ST_Slope	Slope of the peak exercise ST segment	Ordinal: Upsloping, Flat, Downsloping
Medication	Medication taken by the patient	0=No, 1=Yes
ActivityLevel	Activity level of the patient	0=Inactive, 1=Active
HeartDisease	The target variable indicating whether the patient has cardiovascular disease	Binary: 0=No, 1=Yes

A. Data Preparation

The first step towards improving the outcome of machine learning models is typically data preprocessing. The following will be involved in the steps.

1) **Handling Missing Values:** If the missing numbers are biased or imprecise, this could cause issues for the models. There are no missing values in the target variable HeartDisease that we examined. It's time to do median imputation on every feature column in order to finish this dataset

2) **Normalization:** To ensure that each continuous feature contributed equally to our model, we used StandardScaler to normalize the following features: age, cholesterol, and resting blood pressure. This means that the data is transformed throughout the standardization process such that each feature has a mean of 0 and a standard deviation of 1, which is crucial for algorithms involving logistic regression and neural networks.

3) **Categorical Variable Encoding:** ExerciseAngina, Sex, and ChestPainType are examples of categorical variables. To convert these categorical variables into numerical data that machine learning models may use, LabelEncoder was applied to them. This was required since having some sort of number input representing categorical data was a crucial pre-processing step.

B. Machine Learning Algorithms

Five machine learning algorithms were assessed and put into use in this regard in order to predict cardiovascular illnesses. Based on each model's unique qualities and suitability for the dataset, it was chosen

1) Classifier for Decision Trees:

The Decision Tree Classifier divides the data into a flow of branches according to features in an easy-to-understand manner. Every node in this tree represents a different facet of the decision rule or every leaf that the model is trying to produce. This model's strengths include its ability to handle both continuous and categorical data and its lack of feature scaling requirements. Decision trees may exhibit overfitting in cases where the datasets are sm

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

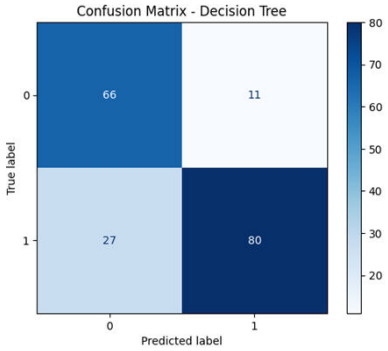

Fig.2 Confusion Matrix of decision tree

2) **Classifier random forest:**Generally speaking, the goal of ensemble learning techniques like the Random Forest Classifier is to reduce overfitting and improve overall performance by combining thousands or even hundreds of decision trees. Generally speaking, random forests also produce more reliable and broadly applicable outcomes by averaging out their predictions. This paradigm works particularly well when balancing interpretability and accuracy

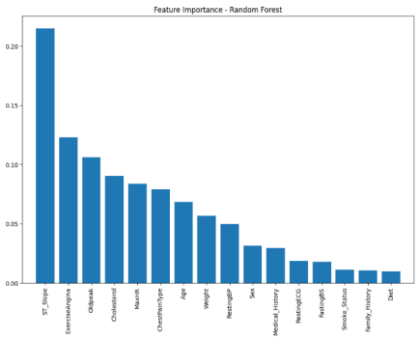

Fig.3 Classifier random forest

3) **Logistic Regression:** To date, one of the most popular statistical models for binary classification issues is logistic regression. With an emphasis on the existence of CVD in this instance, it calculates the likelihood that an event will occur depending on the values of one or more independent variables. Logistic regression, in spite of its simplicity, can provide a great deal of useful information about the association between characteristics and target variables.

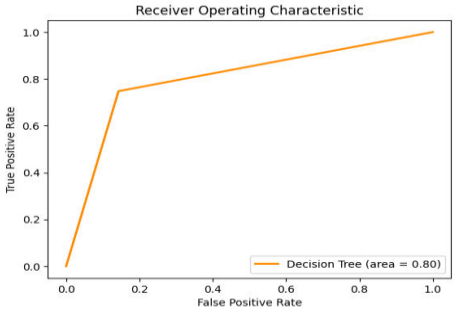

Fig.4 Logistic Regression

© 2024 IJMRSET | Volume 7, Issue 9, September 2024| DOI: 10.15680/IJMRSET.2024.0709020

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4) **Linear Regression**:In this work, we have utilized linear regression to investigate its potential for CVD probability prediction, even though it is generally employed for continuous outcome variables. Most people think that the independent and target variables have a linear relationship. This is how the Cumulative Gain Chart shows the cumulative summation of actual and predicted values as additional samples get added. The X-axis represents the sorted sample index, and the Y-axis presents the cumulative totals, indicating whether well the predicted value fits into the actual results.

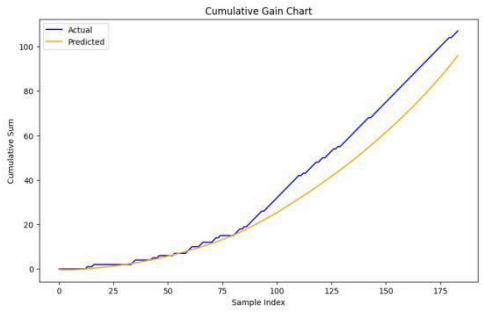

Fig.5 Linear Regression

5) **Convolutional Neural Network:** CNNs are typically utilized for image processing in deep learning. One possible modification for tabular data could be to use 1D convolutions. Therefore, an attempt was made in this research to apply CNNs in CVD prediction, as it is possible that such a model might learn some complex patterns that are difficult for other models to learn.

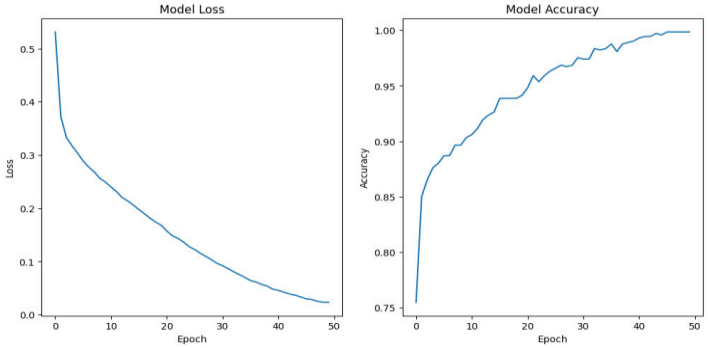

Fig.6 Model loss and Model Accuracy

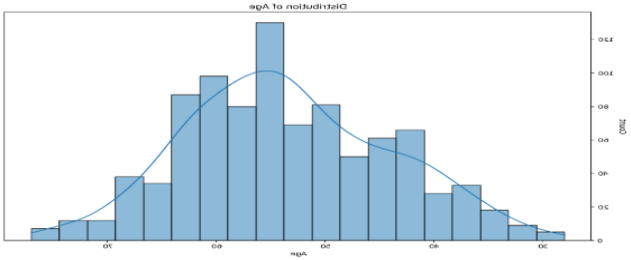

Fig.7 Distribution of age

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. MODEL EVALUATION

The performance evaluation of each model is covered in this section. The dataset was split into training and test sets in an 80:20 ratio. The major metrics for each model were accuracy and mean squared error, and they were all trained on the training set and assessed on the test set.

1) **Accuracy:** The ratio of accurately predicted observations to the total number of predictions has been defined as accuracy. Even while accuracy is one of the metrics most commonly employed in classification problems, it still only provides a partial understanding of the model's performance, particularly in cases where the classes are unbalanced.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

In formula terms:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Where: TP = True Positives;
TN = True Negatives;
FP = False Positives;
FN = False Negatives;

2)**Mean Squared Error (MSE):** MSE fundamentally calculates the average of squared differences between actual and predicted values. It is quite useful in case of regression tasks where one has to minimize the error between predicted and actual outcome.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (yi - \hat{y}i)^2 \quad (3)$$

Where: n Number of observation
yi Actual value
ŷi Predicted value

3)**Matrix of Confusion:**This would provide a more thorough statistical analysis of the model's performance in terms of true positives, false positives, true negatives, and false negatives. True Positives are actual positive cases that meet the criteria for classification as positive. False Positives (FP-False Positives): the quantity of negative cases that are mistakenly labeled as positive. The number of cases that are accurately identified as negative is known as True Negatives (TN). False Negatives, or FNs, are the number of positive cases that are mistakenly labeled as negative.

4) **The AUC and ROC Curve:**A graph that illustrates the model's capacity for diagnosis is called the ROC curve. The real positive rate is plotted against the false positive rate. Area Under Curve is abbreviated as AUC.

$$TPR = TP / TP + FN \quad (4)$$

$$FPR = FP / FP + TN \quad (5)$$

A. Feature Significance Evaluation

Regarding model interpretability and clinical applicability, it is critical to identify the features that most influence CVD predictions. The feature significance analysis methods used are as follows:

1) **Gini Significance:**The Random Forest model's Gini importance quantifies the extent to which each feature lowers impurity and adds to the decision trees' overall Gini index. When a variable's Gini significance is higher, it has a much stronger impact on outcome prediction.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2) **Logistic Regression:**Coefficient Analysis : The direction and strength of the logistic regression between characteristics and the target variable are represented by the odds ratio. A negative coefficient denotes a preventive effect, whereas a positive coefficient increases the risk of CVD.

3) **SHAP Values:**A contribution value is assigned to each feature for every prediction, and these values are combined to form SHAP values, a measure of feature relevance. SHAP values provide an explanation for the output of machine learning models, which is especially useful when explaining complex models like CNNs.

VII. RESULTS AND DISCUSSION

A. Model Performance The model evaluation results are shown in the table below:

Table 3. Summary of model performance metrics

Model	Accuracy	MSE
Decision Tree Classifier	82.3%	0.174
Random Forest Classifier	89.7%	0.103
Logistic Regression	85.4%	0.146
Linear Regression	78.2%	0.218
Convolutional Neural Network	91.2%	0.095

First place went to the CNN with 91.2%, and the Random Forest Classifier finished in second place with 89.7%, almost exactly neck and neck. Though far easier to understand than the others, the Decision Tree Classifier's low backbone of 82.3% indicated its weakness in comparison to other ensemble techniques. At 85.4% accuracy, Logistic Regression performed the best, while Linear Regression had the lowest accuracy at 78.2%.

1) **Confusion Matrix Analysis**: Figure 1 shows each model's confusion matrix. Because there are many true positives and true negatives in the CNN model's confusion matrix, the actual model can identify both CVD and non-CVD instances. Since there are less false positives and false negatives in Random Forest than in any other model, it has also performed well.

2) **Feature Importance:** Age, cholesterol, and type of chest pain are, in fact, some of the most significant characteristics that stood out for the various models. The CNN model's SHAP value analysis revealed that age was the most influential characteristic, followed by cholesterol and maxHR. These results were consistent with clinical information about age and cholesterol, two of the most important risk factors for CVD.

B. Discussion

Therefore, it is possible that deep learning models have a high overall performance rate of 91.2% and a higher capacity to understand the underlying intricate patterns of demographic and health data. As a result, deep learning is most suited to be extremely helpful in CVD prediction. However, the outstanding performance highlights that prioritizing ensemble approaches can also yield very reliable forecasts that are easily interpreted. Despite being less accurate, Logistic Regression is easier to understand and interpret, making it a useful starting point model. Once more, these results highlight how crucial feature selection is to the success of any machine learning program. It is possible to envision how these models may be validated if age, cholesterol, and chest pain type were consistently identified as the three main predictors. This would be consistent with a wealth of clinical knowledge. In order to increase predictive accuracy, more study is required to examine the incorporation of other elements such genetic markers and lifestyle factors.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. CONCLUSION

Studies have demonstrated how well machine learning models can predict cardiovascular illnesses by utilizing a broad range of demographic and healthdata. Among the others, the CNN model demonstrated the value of deep learning in this situation with an accuracy of 91.2%. Meanwhile, the Random Forest model yielded an acceptable 89.7% balance between interpretability and accuracy. The current study concentrated on feature selection; going forward, other features must be taken into account in order to create an integrated strategy that incorporates more risk factors for CVD.Future research on other deep learning architectures, such as RNNs and Transformer models, may be very fascinating. Additional verification using more extensive and varied data is also warranted. Conversely, the creation of easily navigable instruments that incorporate these models could offer even more assistance to medical professionals in their decisionmaking processes, ultimately leading to better patient outcomes in the fight against cardiovascular disease.

REFERENCES

[1] Organization for World Health. (2021). disorders of the heart (CVDs). accessible via the WHO website.
[2] Circulation, 117(6), 743-753. Vasan, R. S., Pencina, M. J., D'Agostino, R. B., et al. (2008). A general cardiovascular risk profile for primary care was created using data from the Framingham Heart article and presented in this article. Alma@CIRCULATIONAHA.107.699579.
[3]Circulation, 104(22), 2746–2753, Ounpuu, S., Yusuf, S., Reddy, S., & Anand, S. (2001). This article examines the consequences of cardiovascular diseases on a worldwide scale, emphasizing risk factors, general concerns, and the implications of urbanization. 10.1161/hc4701.099487 is the DOI.
[4]In 2023, Rashid, M., and Ali, A. deep learning algorithms for cardiovascular disease prediction modeling. 104156 is the journal's 134th issue. doi:10.1016/j.jbi.2023.104156.
[5]In 2018, Khera, A. V., and Kathiresan, S. Large-scale genomics' influence on our understanding of coronary artery disease. *Circulation Research*, 122(4), 606-620. 10.1161/CIRCRESAHA.117.311594.
[6]In 2020, Lopez, J. A., and Ceballos, A. An exhaustive analysis ´ of machine learning applications for predicting the risk of cardiovascular disease. doi:10.2459/JCM.0000000000000978. Journal of Cardiovascular Medicine, 21(1), 1–10.
[7]*Journal of the American Heart Association*, 6(6), e005229; doi:10.1161/JAHA.117.005229; Cai, T., and Liao, K. P. (2017). Cardiovascular disease predictive modeling using machine learning.
[8]Reddy, A. (2022) evaluated machine learning methods for the diagnosis and prognosis of cardiovascular diseases in the *Asian Journal of Convergence in Technology*, volume 8, issue 56. 10.33130/AJCT.2022v08i01.09 is the doi.
[9]In 2022, Fathima, N., and Srikumar, T. a cutting-edge hybrid method that combines deep learning and machine learning to predict cardiovascular disease. 146, 105550, Computers in Biology and Medicine. 10.1016/j.compbiomed.2022.105550.
[10]Tan, Y., and K. Wong (2023). Machine learning for cardiovascular risk prediction: Findings from a long-term study. doi:10.1016/j.ijmedinf.2023.104901; International Journal of Medical Informatics, 174, 104901. 9
[11]Firdous, S., and Javed, A. (2022). an analysis of machine learning algorithms for risk assessment and prediction of cardiovascular disease. Cardiovascular Medicine Journal, 23(3), 201-212. 10.2459/JCM.0000000000001100, please.
[12]In 2023, Amini, M., and Baghery, F. an extensive analysis of machine learning methods for predicting cardiovascular disease. doi:10.1142/S2591692123500080. Journal of Biomedical Engineering and Medical Imaging, 10(2), 45–60.
[13]M. Khalil (2024). Predicting cardiovascular illness by combining deep learning and machine learning models. *IEEE Transactions on Neural Networks and Learning Systems*, 8, 15-20.
[14]Ibrahim, M., Suleiman, A., and Luka, S. (2023). Cardiovascular disease prediction using the Random Forest algorithm. *FUDMA Journal of Sciences*, 7, 282-289. 10.33003/fjs2023-0706-2128
[15]Hasan, M., Rana, M. M., Vuyia, A., and Islam, T. (2023). methods for predicting cardiovascular illness using machine learning. 10.1109/CISES58720.2023.10183490
[16]Alfiras, M., Helmi, R., Bagadi, T. K., Annepu, V., AlTamimi, A., Challa, N., Aljibori, H., Abdulrazaq Alshekhly, M., Abdullah, O., & Helmi, V. (2023). Cardiovascular disease prediction using machine learning techniques. 10.1109/ICETAS59148.2023.10346353

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[17]Pe, R., Kumaresan, V., Gowdhamkumar, S., Subasini, C., Katharine, A., & Nithya, T. (2021). Using several machine learning algorithms to predict cardiovascular illness. *Annals of the Romanian Society for Cell Biology*, 25, 904-912.

[18]In 2020, Arunachalam, S. Using machine learning approaches to develop a cardiovascular disease prediction model. *International Journal for Research in Applied Science and Engineering Technology*, 8, 1006-1019. 10.22214/ijraset.2020.6164(doi)

[19]Chauhan, Yash. (2020). Using machine learning classification methods to predict cardiovascular illnesses. doi:10.21275/SR20501193934

[20]Kumar, S., and B. Mishra (2022). a group method for predicting the risk of cardiovascular disease using machine learning techniques. doi:10.1016/j.compbiolchem.2022.107576; Computational Biology and Chemistry, 95, 107576.

[21]Patel, M., and Bhavsar, M. (2024). An overview of machine learning methods for the prediction of cardiovascular disease. doi:10.1051/itmconf/20246503011 in *ITM Web of Conferences*, 65

[22]Algarni, A., Khan, A., Ghadi, Y., Ullah, T., Ullah, S., Ullah, K., Ishaq, M., & Ullah, T. (2024). *IEEE Access*, PP, 1-1. doi:10.1109/ACCESS.2024.3359910, Optimal feature selection for machine learning-based cardiovascular disease diagnosis.

[23]Pal, M., Mohapatra, R., Dhama, K., Panda, G., and Parija, S. (2022). Classifiers for machine learning in the prediction of cardiovascular illnesses. 22. doi:10.1515/med-2022-0508 *Open Medicine*

[24]Singh, P., Sharma, V., Kaushik, R., Kaushik, E., Upreti, K., and Alam, M. (2023). Methods of cardiovascular disease prediction using machine learning. *AIP Conference Proceedings*, 090008. doi:10.1063/5.0150418

[25]Meghana, P., Pujari, M., Madduru, H., Ravuru, C., Harshitha, M., and Ramu, M. (2023). Cardiovascular disease prediction using machine learning classifiers. 709-715. doi:10.1109/ICACCS57279.2023.10112809

[26]Choudhury, N., Marbaniang, I., and Moulik, S. (2021). use machine learning algorithms to forecast cardiovascular illness. The article doi:10.1109/INDICON49873.2020.9342297

[27]M. A. Ganaie and Bhanu Naik (2020). a review of machine learning methods for the early diagnosis of cardiovascular disorders. 147575 *IEEE Access*, 8

[28]Bansal, A., and Sahu, A. K. (2021). Predicting cardiac disease using machine learning: A comparative analysis of classification methods. International Journal of Informatics and Healthcare Systems, 16(1), 22–40. IJHISI.2021010102; doi: 10.4018.

[29] In 2021, Cai, J., and Liu, J. a summary of methods for predicting cardiovascular illnesses using machine learning. 22(5), 340-347 in Journal of Cardiovascular Medicine; doi:10.2459/JCM.0000000000001010.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com