International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Heart Disease Prediction Using DNN

Aradhya Sharma[1], Simranjeet Kaur[2], Mitali Joshi[3], Palak Rathour[4], Ayushi Chand[5], Ayush Bagri[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering, School of Science and Technology, Swami Rama Himalayan University

Abstract

Heart disorders is still considered a severe cause of death in the world, and timely identification and risk forecast is really crucial in the modern health-care system. As more medical information becomes available, machine learning algorithms are exhibiting to be advantageous in spotting patterns and aiding in clinical decision-making. This work is on forecasting the risk of Atherosclerotic Heart Disease (AHD) from a structured data source with multiple machine learning models. The proposed mechanism involves data refinement, feature construction, and model training, all of which aims for the betterment of the risk assessment performance. Multiple models were tested, such as Logistic Regression, Decision Tree, Random Forest and a Deep Neural Network (DNN) were tested. The results prove that DNN is more accurate and has higher F1 and recall scores as compared to others, which indicates that DNN has the capability to detect the patients with high risks effectively without the presence of false-negative cases. The findings have proved that feature engineering is quite important and its combination with complex learning models has helped in improving prediction accuracy. Further studies will be conducted for incorporating real-time data from patients along with explaining the deep learning algorithms.

Keywords: Machine Learning, Heart Disease Prediction, Random Forest, Logistic Regression, Deep Neural Network, Healthcare Analytics, Classification Models, Feature Engineering, Model Evaluation, AHD, Risk Prediction.

INTRODUCTION

As per what the WHO suggests [1], in a vicinity of 17.9 million deaths are caused by cardiovascular diseases (CVDs) worldwide annually or 32% of all deaths. Coronary artery disease (CAD) is the most common type of CVD, narrowing of the coronary arteries which results in myocardial infarction (MI) and heart failure. Quick as well as accurate diagnosis of the heart disease is essential for prompt treatment and better prognosis of the patients. The traditional methods of diagnosis involve physical examination, electrocardiograms and echocardiograms, but these methods are often expensive and need a specialist to perform them. As machine learning techniques evolved, patterns and risk factors within medical data have been identified and utilized to aid clinical decision making and improve predictive analytics in recent years [3, 4].

There have been multiple studies that examined the use of data insights extraction and ML algorithms in assessing cardiac diseases. For example, Palaniappan and Awang [3] have designed an intelligent heart disease assessment structure using classification techniques and Amin et al. [4] were able to find the

significant features that will facilitate better forecasting system. Other methods such as ensemble and advanced learning methods have also been suggested to boost the accuracy and robustness of diagnosis [5].

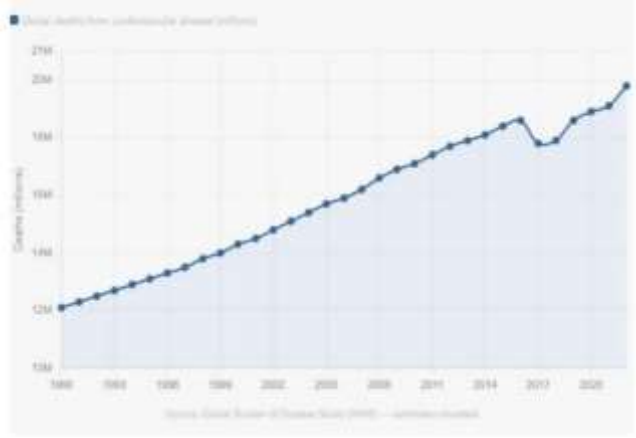


Fig. 1. Global deaths from cardiovascular disease (millions).

We explore four popular classifiers: Logistic Regression (LR),
Random Forest (RF), Decision Trees (DT) and a Deep Neural Network (DNN) for a heart disease dataset from the UCI Machine Learning Repository [2, 7] which is a combination of heart disease datasets from Cleveland, Hungarian and Long Beach. The main aims of this research are:

- To preprocess and engineer features from the data.
- To train and tune each classifier.
- To undertake a thorough comparative review based on several performance indicators.
- To discover the best algorithm to be used in clinical setting.

RELATED WORK

There are multiple works that have investigated the employment of ML and DL for forecasting the risk of heart disease in a patient. And also got to know that CVDs are the prominent contributor of fatality [1] and the need for the development of early and accurate diagnosis systems is emphasized.
Melillo et al. [12] have previously performed some early work using the CART algorithm to detect reduced cardiac function and reported a recall of 93.3% and specificity of 63.5%. Rahhal et al. [13] proposed further improvements by using deep neural networks (DNNs) for the analysis of electrocardiogram (ECG) signals to extract features and classify with greater accuracy.
Guidi et al. [14] used a medical support system that evaluated several algorithms for example Support Vector Machine (SVM), Random Forest, and neural networks. They found that Random Forest and CART had an overall correctness of about 87.6% as compared to others.
Another method was to use NLP with rule-based techniques to process the unstructured clinical notes, yielding an accuracy of 93.37% [15]. Likewise, Parthiban and Srivatsa [16] used SVM for predicting heart disease in diabetic patients with an accuracy of 94.60% with the features like age, blood pressure, blood sugar level etc.
The curse of dimensionality, which occurs with high-dimensional datasets, is another difficulty when predicting heart disease. In response to this, dimensionality reduction techniques have been widely adopted in the research. Rajagopal and Ranganathan [17] proved through study that feature selection plays a vital role to increase the efficiency the of model by choosing the most relevant features.

Yaghoubi et al. [18] and Asl et al. [19] obtained better prediction accuracy by applying machine learning techniques with good feature reduction. Other ensemble and hybrid models like the one proposed by Bashir et al. [5] have also proved to be very useful in improving the classification accuracy.

The datasets typically employed in the vast majority of studies are contained in the UCI Machine Learning Repository [7] and in particular the Cleveland dataset originally introduced by Detrano et al. [2] which provides some important clinical features that are pertinent to heart disease. Traditional machine learning models like Logistic Regression, Decision Trees [9] and Random Forest have been performing well, and deep learning models [10] have been found promising in improving performance as well as good feature engineering.

Overall, the results from the existing studies show that the feature selection, dimensionality reduction and ensemble techniques are important in the enhancement of prediction accuracy. Yet, there still exists a number of issues like overfitting and generalizing the model, which would require more research on this area.

DATASET AND PROCESSING

A. Dataset Description

The UCI Cleveland Heart Disease Dataset [7] is the most commonly used dataset for predicting heart disease.It contains 919 records, each with 13 different features.These features include age, sex, type of chest pain (cp), resting blood pressure (trestbps), cholesterol levels (Chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of the ST segment, number of major vessels (ca), thalassemia (thal), and the target variable (num).The target variable is a binary number, where 0 means no disease and 1 means the person has heart disease.There are 410 cases of no disease and 508 cases of heart disease in the dataset.

B. Preprocessing Steps

One of the most important step in improving the quality of the input data which can further help developers to improve the performance of the model is called data preparation. This study systematically processed the data with several steps such as filling the missing values, detecting the outliers, encoding the categorical data, and feature scaling.

1. **Dealing Missing Values:** No incomplete records or null values were detected in the dataset and it was first checked. The data set was relatively clean and no significant missing values were found. If minor inconsistencies were detected, the approaches taken were deemed to be appropriate and included removal or imputation of data to ensure data integrity.
2. **Outlier Detection and Treatment**: Outliers were determined in order to discover values that had a marked difference from the rest of the data set. This is because outliers have an impact on the effectiveness of machine learning techniques since they lead to biases. each of the numerical features and the IQR was found as follows:

$$IQR = Q3 - Q1$$
$$Lower\ Bound = Q1 - 1.5 \times IQR$$
$$Upper\ Bound = Q3 + 1.5 \times IQR \qquad (1)$$

Instances that deviate from these bounds were considered potential outliers. Additionally, boxplots were used for visual inspection to validate the presence of extreme values.

Instead of directly removing all outliers (to avoid loss of important medical information), their impact was minimized using scaling techniques and careful evaluation. This approach ensured that rare but clinically significant cases were not discarded.

3. Encoding of Categorical Variables: Machine learning algorithms need numbered entry features; therefore, all categorical variables in the dataset were transformed into numerical formats using suitable encoding techniques. Based on the nature of each variable, three types of encodings were applied: Binary Encoding, Ordinal Encoding, and One-Hot Encoding.

Binary Encoding: Attributes with only 2 categories were binary encoded. Gender (F/M) and Exercise Induced Angina (N/Y) were binary attributes in this data set. The labels were associated with numerical observations as follows: Gender: F → 0, M → 1; Exercise Induced Angina: N → 0, Y → 1.

Ordinal Encoding: Variables that have a natural order were encoded using ordinal encoding. The attribute RestingECG and ST_Slope were ranked attributes: RestingECG: Normal → 0, ST → 1, LVH → 2; ST_Slope: Up → 0, Flat → 1, Down → 2.

One-Hot Encoding: ChestPainType is a nominal variable, where the categories (TA, ATA, ASY, NAP) have no natural ordering. One-Hot Encoding was applied. The first category was dropped to avoid multicollinearity with drop_first=True.

4. Attributes Construction: Baseline UCI data gives important clinical markers but raw variables do not always fully reflect the complex and non-linear physiological interactions that underlie cardiovascular risk. To overcome this, 5 domain specific features were engineered to give the models representation at a higher level of patient health:

- Cardiac Workload Index: This is an index derived as a substitute for the Rate-Pressure Product. From a clinical perspective, a high workload at rest is indicative of myocardial stress and/or elevated oxygen consumption, both of which are important clues for the presence of HF.
- Cholesterol-to-Age Ratio: Developed to adjust metabolic data according to the clinical risk for cholesterol levels in different age groups. This will enable the model to differentiate between the "expected" age-related levels and early onset hypercholesterolemia.
- In the original dataset, the severity of ST depression (Oldpeak) is a continuous variable; however, clinical guidelines often suggest that an ST depression severity of ≥2 mm is a high-risk sign of significant ischemia. To make these critical cases explicit, a binary flag was used.
- Angina-Slope Interaction: Cardiovascular risk is more likely multiplicative than additive. The model captures the high-risk
- synergy of the combination of Exercise Induced Angina and ST_Slope.
- Blood Pressure Categorization: Continuous RestingBP observations were grouped into classification: Normal, Prehypertension, Hypertension.

5. Feature Scaling: Our dataset consists of multiple clinical features such as age, cholesterol, MaxHR etc. All these features are measured on different scales and units, which can negatively impact the performance of machine learning model if not properly handled. In this project we have used Standardization technique of feature scaling which is also termed as Z-score Normalization. This is done to make sure the features have an average of 0 and a spread of 1.

$$zi = xi - \mu / \sigma \qquad (2)$$

Here, xi is the value of the feature, μ is the mean, and σ is the standard deviation.

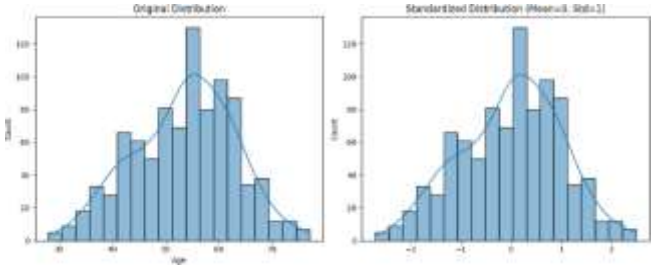


Fig. 2. Comparison of feature distribution before and after standardization.

Feature standardization was applied to the features to transform the data to zero mean and unit variance. In Fig. 2, we have considered age as our example feature to illustrate the actions the transformation performs on the feature space, transforming the shape of the distribution without affecting the original shape, thus ensuring that all the variables are equally important during model training. This led to a quicker converging and more stable training process.

C. Train-Test Split

The data set was divided into two parts: one for training the model and one for testing its performance. The training part used 80% of the data, and the testing part used 20%.This division was done randomly to ensure the model's results were reliable. get an unbiased representation of samples from both subsets.

A constant random state was used in splitting, for reproducibility. This will allow the same results to be obtained in several runs. Using a separate test set helps to ascertain the generalization efficiency of the mechanism, and prevents overfitting. Partitioning of the datasets was done to ensure that the model was evaluated and optimized efficiently. A split of 80:20 was used initially where 80% of the data would be used in training while 20% remained as an independent test set. In the training set, part of the data was reserved as the validation set in order to track the progress of different epochs and also enable early stopping of overfitting. The random split is controlled by a fixed value of the random state so that the same data split could be achieved later in case the need arises. Evaluating the effectiveness of the final constructed model using only the test set gives us an unbiased estimation of its generalization ability on unseen clinical data.

D. Dataset Summary

1) Numerical Variables:

a) **Age:** 28-77
b) **Resting Blood Pressure:** 0-200mm Hg
c) **Cholesterol:** 0-603mg/dl
d) **Maximum heart rate:** 60-202bpm

2) Categorical Variables:

a) **Gender:** Male / Female
b) **Chest Pain Type:** ASY, NAP, ATA, TA
c) **Resting ECG:** Normal, LVH, ST

3) Binary Variables:

a) **Fasting Blood Sugar**: Presence/Absence
b) **Exercise Angina:** Yes/No
c) **Obesity**: Present/Absent
d) **Diabetes:** Present/Absent

4) Target Variable:

Risk of Atherosclerotic Heart Disease: Presence=1, Absence=0

PROPOSED METHODOLOGY

A. Logistic Regression

The logistic function is used in logistic regression, which applies a sigmoid function to convert a linear mix of inputs into a probability that ranges from 0 to 1.
It does not fit a line rather it fits an S-shaped curve.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (3)$$

$$P(y = 1|x) = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

where $x_1, x_2, \ldots, x_n$ are the input features, β_0 is the intercept, and $\beta_1, \ldots, \beta_n$ are the model coefficients.
Logistic Regression is used in this study to predict whether a person has heart disease or not based on their clinical features. The model uses a sigmoid function on the linear combination of input variables (see the equation (3) and (4)) to predict the probability of a positive outcome.
This data set has been split into training and testing sets. Logistic Regression with L2 regularization was used to train the model. Optimization of the model parameters was done using a suitable solver to minimize binary cross-entropy loss.There are certain strengths that are present in the Logistic Regression model, making it a reliable algorithm when applied in medical diagnosis. Interpretability is one of such strengths because the coefficients for each variable reveal whether the feature increases or decreases the value of the prediction. It means that the algorithm can be applied to solve health-related tasks since it reveals what risk factors affect patients. Thus, age, cholesterol levels, resting blood pressure, maximum heart rate achieved, and type of chest pain may significantly affect predictions made by the model.

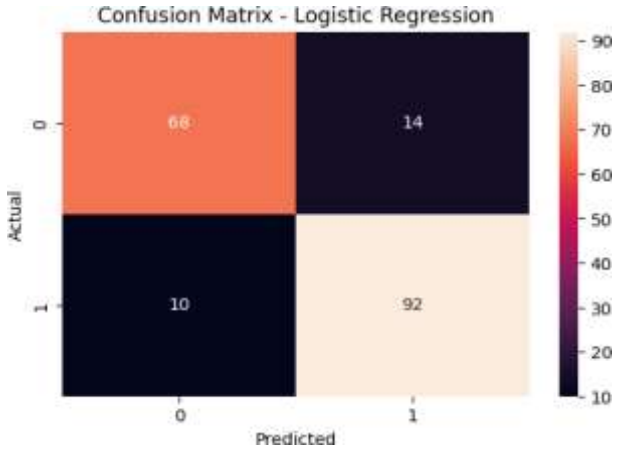


Fig. 3. Confusion Matrix of Logistic Regression.

The confusion matrix is a representation of the distribution of true positives, true negatives, false positives and false negatives of the Logistic Regression model and is shown in Fig. 3. These values are applied to calculate the evaluation metrics like accuracy, precision, recall, and F1 score etc.
This confusion matrix was used to assess the performance of the Logistic Regression model. The overall accuracy of the model is 87%, which demonstrates a good prediction ability for the test set. The model achieved a precision, recall, and F1 score of 0.87, 0.83, and 0.85 respectively for the negative class (no heart disease).
This means that although the model works well in identifying non-disease cases, there are also a few cases which are wrongly classified as non-disease cases. The precision, recall, and F1 score for the positive class

(heart disease) were 0.87, 0.90, and 0.88 respectively. A higher recall value means that the model is performing well in correctly identifying the patients who suffer from heart diseases and reducing the chances of false positives.

The macro and weighted average precisions, recalls, and F1-scores are all around 0.87, indicating a balance between the classes.

B. Random Forest

Random Forest is based on ensemble learning principle that constructs multiple decision trees are combined to improve predictive performance. Each tree is trained on a random subset of the data (bootstrapping) and randomness is also added at node splitting to obtain diversity among trees.

In classification problems, the final prediction takes the majority votes of all the decision trees, which is defined as in (5):

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_n(x)\} \quad (5)$$

Where $h_i(x)$ represents the prediction of the i-th decision tree.

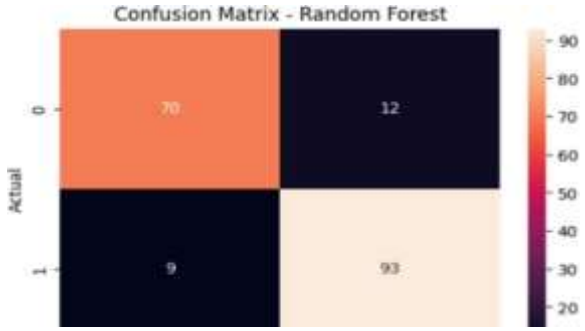


Fig. 4. Confusion Matrix of Random Forest.

The performance of the Random Forest model was assessed based on accuracy, precision, recall and the F1-score. The confusion matrix in Fig. 4 shows the distribution of the true positive, true negative, false positive and false negative. Upon this basis, the model was found to have an overall accuracy of around 88.6% on the test set, which shows that the model has good predictive ability. The model had a precision of 0.89, a recall of 0.85 and an F1-score of 0.87 for the negative class (no heart disease) meaning that the model is good at identifying non-disease cases, but some of them are incorrectly classified.

The model showed a precision of 0.89, recall of 0.91 and an F1-score of 0.90 for the positive class (presence of heart disease). The greater the recall, the better the model was able to accurately identify patients who suffer from heart disease with very few false negatives.

The macro and weighted averages of precision and recall as well as the F1-score are all close to 0.88–0.89 reflecting the balanced performance on both classes.

C. Deep Neural Network (DNN)

To predict heart disease, a complex relationship between clinical features is modeled using a Deep Neural Network (DNN) that is a non-linear classifier. The key difference between traditional machine learning algorithms and DNNs is the ability of DNNs to capture complex patterns and relationships in high-dimensional data, which is achieved by their hierarchical feature representation using multiple hidden layers.

The transformation performed by each neuron in the network is mathematically expressed as shown in (6):

$$y^{(l)} = f\left(\sum_{j=1}^{n} w^{(l)} x^{(l-1)} + b^{(l)}\right) \quad (6)$$

Where w_{ij}^{l} denotes the synaptic weight, x_j^{l-1} represents the input from the previous layer, b_i^{l} is the bias term, and $f(\cdot)$ is a nonlinear activation function.

A sequential neural network model is built in this study, with the input layer having as many neurons as there are features in the dataset, and the hidden layers having 16 and 8 neurons, respectively. The hidden layers use the Rectified Linear Unit (ReLU) activation function to provide non-linearity and overcome the vanishing gradient problem. The output layer is composed of one neuron with a sigmoid activation function, for a binary classification of whether or not a person has a heart disease.

The model is optimized with Adaptive Moment Estimation (Adam) optimizer, which is designed to efficiently handle gradient-based optimization at adaptive learning rates. It is used in binary classification tasks due to the use of the binary cross-entropy loss function which measures the difference between the predictions and true values.

Training Performance and Convergence Analysis

To train the DNN model, 20% of the training data was used to validate the model and 100 epochs were used with 32 batch size. The validation loss was used for an early stopping mechanism to avoid overfitting. The training was carried out continuously and the model's accuracy was improved continuously during the training process. The accuracy of training rose from about 57.9% in the first epoch to about 89% for subsequent epochs, showing good learning of feature representations. At the same time, the training loss consistently decreased, which was the evidence of appropriate optimization of model parameters.

The validation accuracy also increased from 63.2% to ~83%, indicating that the model had good generalization performance on unseen data. The validation loss was observed to decrease initially followed by a steady state after a few epochs, which is a sign of model convergence.

It was noticed that the improvement in the validation loss was quite small after around 30–35 epochs and the validation loss was not much varying from epoch to epoch. The early stopping mechanism successfully stopped the training at an optimal point, maintaining the best model weight and avoiding overfitting.

As a whole, the DNN model reached a stable convergence with a balance between the bias and variance, and it was the most reliable classifier for the prediction of heart disease in the comparison with rest of the algorithms

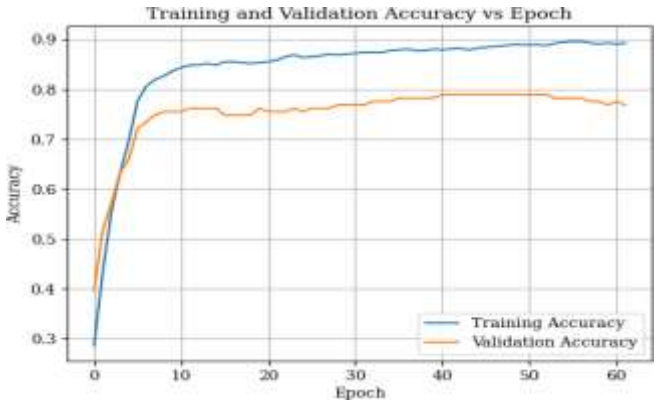


Fig. 5. Variation of training accuracy over epochs for the DNN model.

The graph of training accuracy in Fig. 5 demonstrates a gradual rise in the accuracy of the model during training, suggesting the model has learned the patterns in the data well. The validation accuracy starts to converge at about 78%, which is considered good generalization performance. It is observed that there is a small difference in the accuracies of the training and the validation sets, indicating mild over fitting, and this is prevented by setting the number of epochs for early stopping.

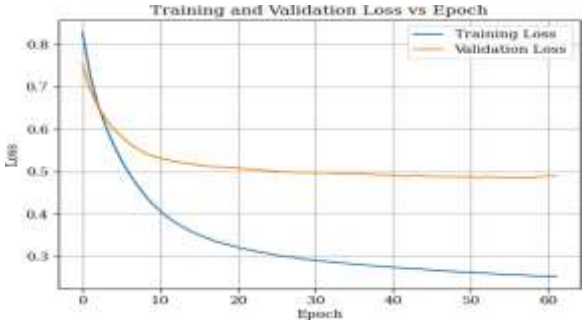


Fig. 6. Variation of training loss over epochs for the DNN model.

The loss curves of the proposed DNN model for training and validation losses are shown in Fig. 6. The training loss gradually gets reduced during the training period, showing that the parameters of the model are being continually optimized. The validation loss drops off sharply in the first epochs and then levels off after ~30 epochs, indicating that the model is converging. This slight overfitting is reflected in the gap between training and validation loss in later epochs but is suppressed by early stopping.

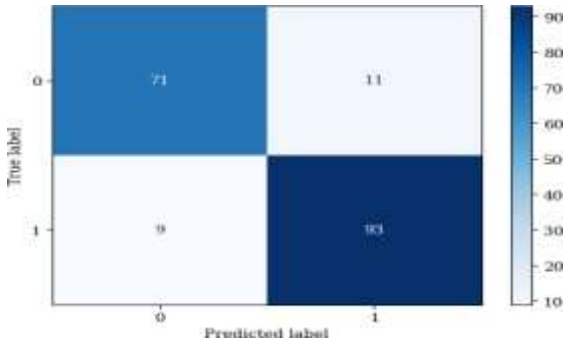


Fig. 7. Confusion Matrix of DNN

The proposed DNN model is further tested by using the confusion matrix. The model is able to correctly predict 71 times that the subject was negative (True Negatives) and 93 times it was positive (True Positives) as seen in Fig. 7. There are 11 false Positives and 9 false Negatives, however.

The results show that the model has a high accuracy and relatively low misclassification rate. The lower false negative rate showed that the model was effective in diagnosing patients with heart disease, which is a vital role in medical diagnosis.

D. Decision Tree

In this study the Decision Tree classifier is trained using parameters from the heart disease dataset that have been cleaned up beforehand. The Decision Tree classifier learns how to make decisions based on rules to put patients into two groups: those with heart disease and those without it.

The results of the experiment show that the Decision Tree classifier is correct about 78.26% of the time. For patients with heart disease the model is precise about 79% of the time remembers about 83% of the time. Has an F1-score of 0.81. The fact that the recall is higher means the model is good at finding patients with heart disease, which's very important for healthcare systems that try to predict things.

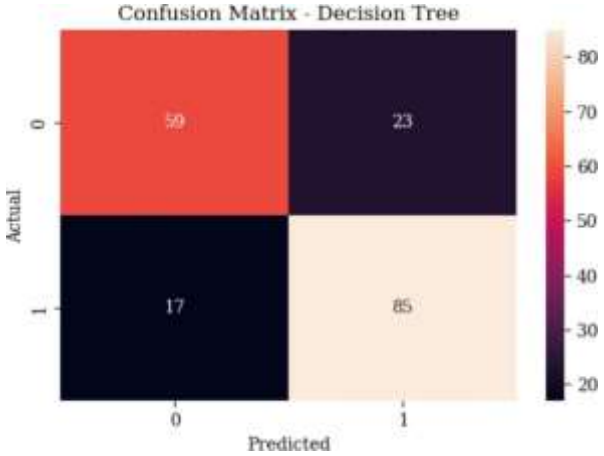


Fig.8. Confusion Matrix of Decision Tree

The results of the Decision Tree classifier are shown in the confusion matrix in Fig. 8. The model correctly said that 59 people without heart disease were healthy and that 85 people with heart disease actually had the disease.. The model made mistakes when it said 23 healthy people had heart disease and when it said 17 people with heart disease were healthy.

The chart shows that the Decision Tree model is good at finding people with heart disease and it remembers 83% of them.. The fact that there are more false positives and false negatives compared to the proposed DNN model means it is not as reliable and not as good at making general predictions.

When doctors are trying to diagnose people it is very important to avoid missing people with heart disease because if they do not get treated they might get sicker. Even though the Decision Tree classifier is easy to understand and makes decisions based on rules it does not do as well as other models because it has trouble understanding complicated relationships in the heart disease dataset. The Decision Tree classifier has trouble with this, which means it is not as good, at predicting heart disease.

EVALUATION METRICS

A comparative analysis of the performance of different classification algorithms is done by comparing Accuracy, Precision, Recall and F1-score. In this study the models taken into consideration are Logistic Regression, Random Forest, Decision Tree, and Deep Neural Network (DNN).

A. Discriminative Performance: ROC-AUC

The Receiver Operating Characteristic (ROC) curve is used to test the classifiers if they have threshold independent diagnostic performance. The curve depicts the association between sensitivity and specificity and is created with respect to True Positive Rate (TPR) against False Positive Rate (FPR). The area under the curve (AUC) provides a scalar quantity corresponding to the probability that the model ranks a random patient with CHD higher than a healthy one as stated in (7). This metric is especially useful in a clinical setting because it provides a measurement of the performance that is independent of both class distribution and misclassification costs.

$$AUC = \int TPR(FPR)\, d(FPR) \qquad (7)$$

B. Predictive Accuracy and Robustness

Accuracy is the simplest measure of accuracy but is not useful when there are an imbalance of prevalence across classes. Additionally, to get a well-balanced assessment, we also use F1-Score, which is the harmonic mean between Precision and Recall. This ensures that the model's predictive capability is not

skewed towards the majority class (typically healthy controls) and is robust in predicting the minority class (heart disease patients).

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \quad (8)$$

C. Clinical Sensitivity and Precision

As far as cardiac diagnostics are concerned, the price of a False Negative (Type II Error) – when a diseased patient is (Sensitivity) is kept in mind, so that the diseased population is identified at maximum level. Correspondingly, Precision (Positive Predictive Value) is tracked to mitigate the number of False Positives (Type I Error) to avoid unnecessary clinical actions and patients' psychological suffering.

Sensitivity (Recall): Lowest risk of a Type II error (missed heart disease).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision (PPV): Measures the chance of a positive prediction being correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

RESULTS AND DISCUSSION

TABLE 1. COMPARATIVE PERFORMANCE OF CLASSIFIERS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.87	0.87	0.87	0.87
Random Forest	0.886	0.89	0.89	0.89
Decision Tree	0.78	0.78	0.78	0.78
DNN	0.89	0.89	0.91	0.90

A. Comparative Performance Analysis

The results of the investigative studies of the four classification algorithms are reported in Table II. To measure each model's predictive ability in a holistic manner, the performance was judged using the following measures—Accuracy, Precision, Recall, and F1-Score. Among the evaluated approaches, the proposed model achieved the highest predictive accuracy on all of the metrics with a peak Accuracy of 0.89 and an F1-Score of 0.90. The performance hierarchy that is observed in the model comparison bar chart is evident, with DNN outperforming all other models and RF close behind, while LR is performing at a high enough level, and the Decision Tree is the least effective with an overall correctness of 0.78.

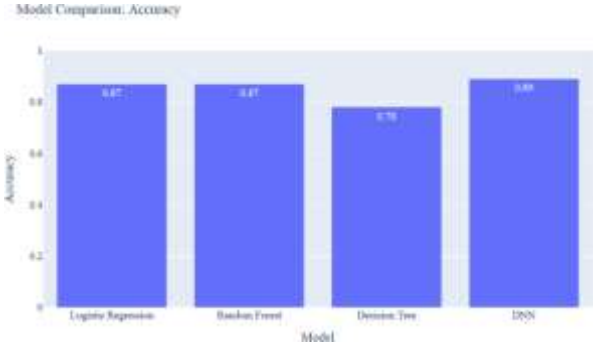


Fig. 9. Model Comparison based on Accuracy

The difference in performances is further displayed in the bar chart of accuracy of the models as shown in Fig. 9. The visualization shows that the Decision Tree is very far behind the other models in terms of margin, the DNN and Random Forest model by a huge margin and the Logistic Regression model also by a huge margin. The large recall values (between 0.90 and 0.91) from the top three models reflect that they are very effective at reducing false negatives, which is key for correct diagnosis of the illness.

B. Classification Robustness and ROC Analysis

In order to further examine the diagnostic ability of the classifiers, an analysis of Receiver Operating Characteristic (ROC) curves was done, which is depicted in Fig. 10. One of the most important measures of the models' strength in distinguishing between classes is called the Area Under the Curve (AUC). As in the numerical results, the Logistic Regression and Random Forest models both had an AUC of 0.93, and the DNN was very close with +0.92. The curve of Decision Tree (AUC = 0.77) is much steeper, with less stability in the classification threshold than the smoother curves of the ensemble and deep learning methods. In the ROC AUC plot, the Logistic Regression, Random Forest and DNN curves are close to the top left corner which shows that these models have low false-positive rate and high sensitivity.

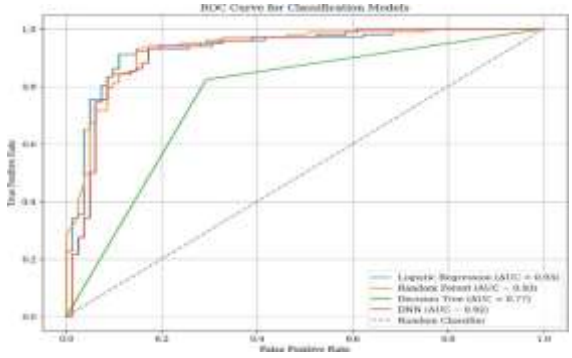


Fig. 10. Comparative ROC analysis of models.

C. Feature Importance and Impact of Engineering

A feature importance analysis was performed side-by-side to test the medical interpretability of the models and effectiveness of the preprocessing pipeline. This comparison is shown graphically in Fig. 11, and indicates that the change in priorities has been brought about by the introduction of engineered variables.

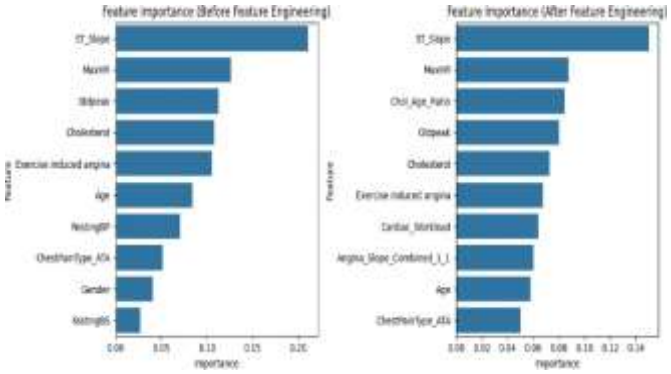


Fig. 11. Correlation heatmap illustrating relationships between features and the target variable.

In order to investigate the predictive capability of the selected feature set, a Pearson correlation analysis was performed. The correlation heatmap is shown in Fig. 10, displaying the direct associations between the engineered features, the baseline physiological markers and the target variable Risk of AHD. The heatmap shows that the newly derived metrics exhibit a stronger correlation with the target variable. Indeed, the Cardiac_Workload and Chol_Age_Ratio have shown a clear positive correlation with the AHD risk, when compared with their raw counterparts. Additionally, the visualization is used to detect and reduce possible

multicollinearity between features, and the engineered interactions add new and non-redundant information to the models, which are used to enhance the models' ability to classify. The above correlation pattern is very clear, and it is the basic basis for the change in predictive importance that is seen in the subsequent model-based analysis.

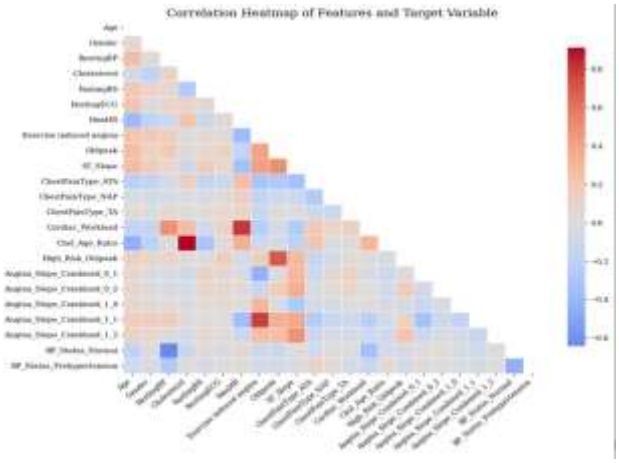


Fig. 12. Feature Importance Analysis Pre and Post Engineering

Baseline Analysis (Pre-Engineering): In the first state, the model used a large amount of physiological markers. The most important predictor was identified as the ST_Slope with an influence score greater than 0.20 followed by MaxHR and Oldpeak. These are known medical markers but the model's use of raw Cholesterol and Age suggests a more generic approach to risk assessment which may not address individual patient factors.

Post-Engineering Refinement: Following the implementation of domain-specific features the internal logic of the classifier became more nuanced. The engineered Chol_Age_Ratio became the 3rd most important feature, justifying normalization of cholesterol by age. Both the Cardiac_Workload and Angina_Slope interaction term were also important, effectively capturing the synergistic risk of cardiac stress and exertion-induced pain. Interestingly, ST_Slope continued to be the most predominant predictor in both scenarios, highlighting its core importance to facilitate the detection of atherosclerotic heart disease

Discussion of Findings: Based on the results, one can conclude that for the current dataset, more complex models such as DNN and ensemble models such as Random Forest models are more powerful in predicting the results. Interestingly, Logistic Regression and Random Forest both perform identically which may indicate that the relationship between the features is fairly linear or that the RF had hit its performance plateau with the given features.The slight performance gap between the Decision Tree and the other models could be attributed to the fact that the Decision Tree tends to overfit or may not be able to learn as effectively as the deeper or ensemble-based models. The combined evidence of the accuracy metrics and the ROC analysis indicates that the DNN should be the preferred model to deploy, with Logistic Regression being a very competitive and interpretable model.

CONCLUSION

This study was successful in developing and evaluating several ML algorithms to quantify the likelihood of Atherosclerotic Heart Disease (AHD). The dataset was carefully preprocessed using methods such as outlier detection, encoding categorical variables as one-hot vectors, and normalizing data using Z-scores to prepare it for efficient classification. One major aspect of this work was the use of domain specific feature engineering, which involved creating various domain-specific indices that were shown to improve the

models' capacity to capture complex clinical patterns, including the Cardiac Workload and Cholesterol-to-Age Ratio.

Based on the comparative analysis, it was noted that a definite performance hierarchy among the algorithms exists. The correctness of Logistic Regression and Random Forest was high at 0.87 and were also interpretable, whereas the Decision Tree was not. The model that produced the best result was the Deep Neural Network (DNN) with a high degree of correctness of 0.89 and an F1-Score of 0.90. Most significantly for medical diagnosis, the DNN showed the best recall of 0.91, meaning that it had the highest sensitivity in determining which patients were at risk, and avoided the terrible negative consequence of false negatives. The outcomes are the proof that although old models are still competitive and easily interpretable, deep learning models paired with clinically relevant feature engineering are the most accurate method for cardiac risk stratification. Going forward, future research will explore the use of longitudinal patient information and integrating Explainable AI (XAI) to give clinicians clear, personalized information about what the model is deciding. At the end, the role of predictive systems within clinical practice is poised to be a transformative force, particularly in the context of timely intervention for cardiac ailment.

REFERENCES

1. World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheet, 2021 [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. R. Detrano, A. Janosi, W. Steinbrunn et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.
3. S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Proc. IEEE/ACS Int. Conf. Computer Systems and Applications, Doha, 2008, pp. 108–115.
4. M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82–93, 2019.
5. S. Bashir, U. Qamar, and F. H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting," Australasian Physical and Engineering Sciences in Medicine, vol. 38, no. 2, pp. 305–323, 2015.
6. P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn,andA.
7. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," arXiv preprint arXiv:1707.01836, 2017.
8. D. Dua and C. Graff, "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
9. H. D. K. H. Bharti and M. A. Ansari, "Prediction of heart disease using machine learning algorithms," in Proc. Int. Conf. Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 132–137.
10. J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
11. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
12. K. Srinivas, G. Raghavendra Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in Proc. Int. Conf. Computer Science and Education, 2010, pp. 1344–1349.
13. A. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients

suffering from congestive heart failure via long-term heart rate variability," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 727–733, 2013.

14. A. Rahhal, S. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," Information Sciences, vol. 345, pp. 340–354, 2016.
15. G. Guidi, M. Pettenati, and P. Melillo, "A machine learning system to improve heart disease diagnosis using clinical data," Procedia Computer Science, vol. 112, pp. 160–169, 2017.
16. Y. Zhang, M. Zhang, and D. Shen, "Combining natural language processing and structured data for improved clinical prediction," Journal of Biomedical Informatics, vol. 76, pp. 1–10, 2017.
17. G. Parthiban and S. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012.
18. S. Rajagopal and S. Ranganathan, "Feature selection and classification in heart disease prediction using machine learning techniques," International Journal of Computer Applications, vol. 97, no. 10, pp. 1–5, 2014.
19. M. Yaghoubi, M. S. Ayatollahi, and H. Bahrami, "Heart disease prediction using data mining techniques with feature reduction," International Journal of Advanced Computer Science and Applications, vol. 8, no. 4, pp. 1–6, 2017.