**Novel Heart Disease Prediction system using custom Clinical Data**

, ,

***Abstract****;* One of the more pressing public health issues is heart disease which has been contributing to an increase in the number of deaths worldwide. Early identification helps to prevent severe and perhaps life threatening situations. On the other hand, heart disease identification at an early stage has been a distressing aspect of modern medicine because of the nature of the disease as well as and due to continuous health surveillance that generates large quantities of data. This study is aimed specifically at building a strong classification as well as prediction model to solve this problem. A large volume of health data is obtained by using electronic health records, wearable devices that monitor patient health and other real-time systems. And to manage such volumes of information, data mining techniques are used to enhance classification and analysis of the data. A feature of the automated model presented is the application of optimized Random Forest (RF) algorithm for the diagnosis of heart disease. A number of elements of the automated model are based on unsupervised feature selection, which is known to improve accuracy of prediction models. The number of elements also optimizes it so that irrelevant features do not contribute to the final model. Multiple data pre-processing, feature selection and classification, are also implemented by the model to enable it provide valid results. The system is able to distinguish between normal and abnormal heart disease accurately to a 98% certainty extend.

Keywords : Heart Disease, Random Forest, Bi-LSTM, Clinical Data

**INTRODUCTION**

* 1. **AN OVERVIEW OF HEART DISEASE AND RISK**

Today, information is shared through reports, forms, statistics and so on, which are considered the inputs for different techniques. The variety of problems in various branches has led to the emergence of many new ways with the passage of time. Different defects in specific domains can now be simply detected and resolved readily with the help of technology. This development of technology is also of great importance in the healthcare industry as it has improved the efficiency of real time problem solving. There have been considerable researches and studies within and across several areas, especially in medicine, where technology is exposed more often on data retrieval and result evaluation, with the output being published.

Cardiovascular illnesses still figure considerably in the mortality and morbidity statistics, the heart disease being one of the most dominant among them according to the World Health Organization (WHO). The heart is integral in the distribution of blood throughout the body, one of the areas being the brain. If there is any obstruction to the blood flow going to the brain as well as the nerves, then failure of the body systems may occur leading to death. It is for this reason that appropriate functioning of the heart must be maintained for one to lead a healthy life. Prevention in terms of early diagnosis of heart disease is beneficial in curtailing timely intervention which greatly reduces the mortality (Beyene et al., 2018). Heart disease prediction is one most pressing challenges today.

Causative factors of mortality and morbidity have changed with the onset of the COVID-19 pandemic — while COPD and IHD were among the leading causes in the US, IHD remains a top cause of mortality during SARS-CoV-2 infection. It has even been associated with unfavorable outcomes among patients with certain comorbidities like bacterial pneumonia and influenza. Up to this moment, all attempts at COVID-19 data collection have been descriptive in nature with no prospect of predicting outcomes for patients. The COVID-19 specific inflammatory response, popularly known as "cytokine storm" or "inflammatory storm" has also been associated with myocarditis, vascular inflammation, and arrhythmias subsequently, aggravating heart damage. Even though cardiovascular diseases are classified into several categories, which need to be diagnosed as soon as possible in order to avoid aggravation of the situation for the patient. This is emerging as a pertinent issue in the world health context and remains accountable for the surge in global mortality. These conditions, which have various risk factors, have claimed many lives. To reduce the threats of these conditions, heart disease patients should practice prevention and follow treatment recommendations.

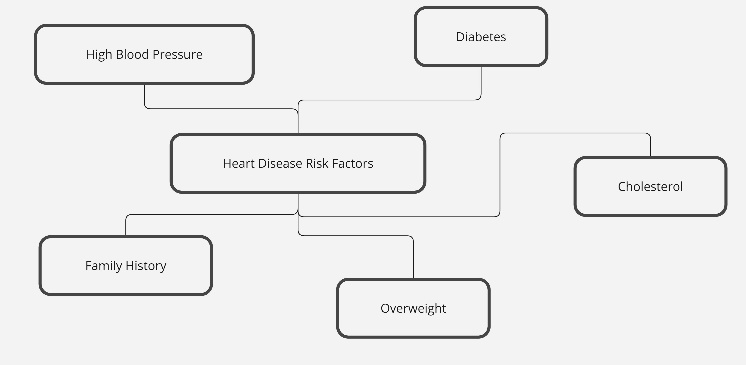
* + 1. **Types of Cardiovascular Disease**

Heart disease, or cardiovascular disease (CVD) as it is most commonly referred to, comprises a number of diseases or conditions that affects the heart and the blood vessels. Myocardial infarction, commonly referred to as heart attacks or angina, is the most well-known and prevalent of the types. However, heart disease that comes as a consequence of build-up of fat deposits on the inside surface of the coronary arteries remains very popular. These are the arteries that supply the heart muscle with blood that has oxygen. The gradual progression of plaque within the arteries forming concentric cylindrical layers is medically called atherosclerosis advanced. The only problem with this condition is that if this atherosclerosis is not identified at its beginnings, it can spell doom and even potentially serious management problems later on. A piece of this plaque may calcify and obliterate the arteries making it difficult for rich blood to flow to the myocardial tissues. In cases where the plaque leads to calcification, it is essential in the triggering of the haematoma on its outer surface. All these potential roadblocks to healing a particular area of the stomach muscle untreated fuels further complications including increased chances of multiple organ failure or even death.

**Risk Factors of Heart Disease**

Risk factors can cause arteries to become blocked, which can lead to heart attacks. These factors fall into two categories: non-modifiable risk and modifiable risk. Non-modifiable factors include gender, age, and genetics, which cannot be changed and play a major role in heart disease risk. Modifiable risk can be managed through personal efforts and lifestyle changes. Examples of variables include behavior, stress levels, diet, and various biochemical factors. Figure 1.1 represents the factors involved in Heart Disease.

Heart disease can be caused by many things, including heart disease, atherosclerosis, rheumatic disease, heart defects, myocarditis, angina, and heart arrhythmias.

****

**Figure 1.1 Risk factor of heart disease**

**THE NEED FOR HEART DISEASE PREDICTION**

The prevalence of heart disease, one of the main causes of death globally, is rising as a result of aging populations, unhealthy lifestyles, and elevated stress levels. Preventing serious consequences like heart attacks, strokes, or even death requires early detection and prompt intervention. This demonstrates the increasing demand for accurate heart disease prediction tools.

Early detection of heart disease aids in proactive management, empowering people to change to healthier lifestyles, make wiser choices, and seek medical attention before the illness worsens. By using predictive models to identify high-risk individuals, medical professionals can better customize treatment and prevention plans, which lessens the strain on healthcare systems.

Furthermore, heart disease develops silently in many cases — different fields remain unaware of their plight until they reach a fatal phase. However, many patients who are at risk may not display symptoms for a long time which creates an opportunity to use predictive tools to close this gap.

In this computerized age of medical development and data examination, machine learning structures and artificial intuition have become a powerful way to foretelling heart turmoil by interpreting designs in patient records. These technologies offer reliable and time-saving alternatives for analyzing massive amounts of health information, pinpointing trends that may go unnoticed by traditional diagnostic methods.

To sum up, the need for heart disease prediction is very important to save us from growing cardiovascular diseases. Detection that is made early can improve quality of life avoid

* 1. **VARIOUS PREDICTION METHODS FOR HEART DISEASE**

As heart disease becomes a global epidemic, many approaches to predicting the resulting onset and progression have been identified. Ranging from classical statistical modelling to advanced machine learning-based approaches, these predictive methods are tried with emphasis on enhancing early diagnosis and facilitating preventive healthcare interventions. The following are some of the commonly used techniques in predicting heart disease.

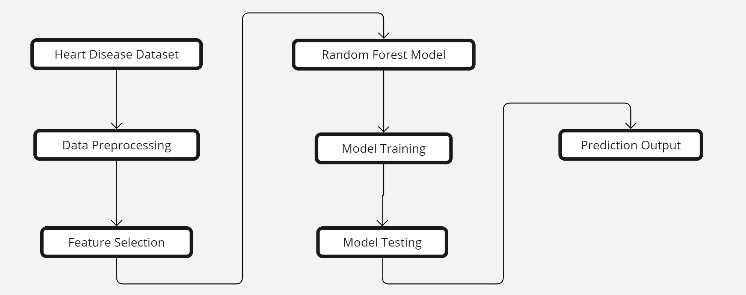
1. **Logistic Regression**

The Logistic Regression is the oldest & a popular statistical technique used to predict heart disease. This function predicts whether or not a patient suffers from heart disease by using age, gender, cholesterol and blood pressure between other risk factors. This when the relation of the input variables with the output (the presence or absence of disease) is linear.

1. **Decision Trees**  
   Decision Tree Algorithms A decision tree algorithm builds a final structure similar to that of a tree based on the outcome of various health parameters. It splits the data into branches where each branch covers a single decision path to reach a prediction. Decision trees are an intuitive and interpretable data-driven algorithm using rules that can work for both categorical and continuous data types, helping in predicting heart disease.
2. **Random Forest**  
   Random Forest is an advancement of the decision trees which enhances prediction accuracy by bringing many decision trees together in one group or model. It employs a "voting" method in which every tree makes a prediction, and the most frequently produced outcome is chosen. It is a very effective method for predicting heart disease since it lessens the chance of overfitting and works accurately with considerable sized high dimensional datasets.
3. **Support Vector Machines (SVM)**  
   Support Vector Machines are a type of machine learning algorithm that can be used for tasks such as predicting heart disease. SVM separates data into different classes (for example disease and no disease) based on the hyper plane it constructs according to the input pairs of features. It is very efficient when there are complex relationships between variables and also works efficiently in high-dimensional space.
4. **K-Nearest Neighbors (KNN)**  
   K-Nearest Neighbors is one of the most effective and the simplest algorithms available which is also helpful in predicting heart disease. KNN predicts the class of a new patient’s data based on the classes of its 'k' closest points in the training set. If most of these neighbors have the disease, the patient is predicted to have the same disease. KNN is very simple to carryout but is less suited for large datasets with noise.
5. **Artificial Neural Networks (ANN)**  
   Artificial Neural Networks are inspired by the human brain and are also applicable in predicting heart diseases. ANN consists of many layers where each node is interconnected and responsible for different types of computations about the input data to delivering its output. Neural networks can capture different interactions within different risk factors and the disease outcomes hence improving the chances of success in heart disease prediction.

**2 .RELATED WORK**

For the last few decades, heart disease prediction has been an area of active inquiry where advanced computing and machine learning models have made notable contributions. A review of literature showcasing significant works done by different researchers in the heart disease prediction domain is presented below. Zheng, H et al. (2023) In this study, Beyene et al. accessed patient’s medical records dataset and explored the prediction of heart disease through machine learning models [1]. They tested all algorithms such as decision trees, naive bayes, support vector machines (SVM) and concluded that Decision Trees gave the best prediction accuracy with upto 82 %. The study also highlighted the effects of feature selection on model building practices by pointing some risk factors such age and cholesterol and blood pressure. In [2], Gupta et al. used ensemble learning techniques to predict the probability of heart diseases. Their model integrated Random forest, Gradient boosting and logistic regression to attain an accuracy of 89.2 percent on the Cleveland Heart disease data set. The authors assert that any model that incorporated ensemble techniques performed better than any particular algorithm where biases and variances were reduced hence making the predictions more dependable. Kumar and his group together applied Genetic Algorithms (GA) with a hybrid model to Circular supply chains combining all three Dimensions: Economic, Environmental and Social. Their research sought to improve the model by considering GA as a feature selection technique which won’t compromise the accuracy of the model although the number of input variables was reduced. GA took the ANN model a step further, attaining an accuracy of 91%, thus, demonstrating the useful application of integrating both traditional and evolutionary computing techniques for disease prediction purposes. In [3] Kumar et al. discussed the implementation of a deep learning system for heart disease prediction which was based on LSTM networks architecture. They used health records such as time-series data to model dependencies in patient health records. Their findings showed that the accuracy of the leukemia patients was 93%, better than KNN and SVM for machine learning models. Medical practitioners however, recognized the challenges posed by non-linear medical data and still called for the use of deep learning techniques. Mekonnen et al [6] (2022) In their study Garg et al. employed a new and hybrid method that fused Cnn with Random forest to predict Heart Disease from clinical parameters and images. CNN has extremelly performed well in extracting necessary image features, while Random forest worked great in the structured clinical information. In heart inflammation disease, the integration of neural networks approaches with machine learning showed an accuracy level of 94.5%. Zhao et al [10] explored the applicability of Logistic Regression, Naive Bayes and SVM in predicting heart disease from patient records voice attentively. This research showed that Logistic regression yielded the better results with 84% accuracy followed by Naive bayes and SVM at 80% and 79% respectively. The research also stressed the importance of removing noise from the data and its significance in increasing the accuracy of predictive models. Anwar et al [11] applied a K-Nearest Neighbors predictor along with Random Forest to predict heart disease risk factors such as age, sex, cholesterol and blood pressure. The result of the comparative analysis was that Random forest had the best accuracy of 87% while KNN had 82%. This research also provided evidence on the effectiveness of Random Forest to datasets with class imbalance. Singh, R et all [13] In this line Rajesh and Srinivasan were working on SVM because they were able to improve the model sophistication as well. Utilizing kernel functions, they were able to boost the model accuracy to 90%. Their findings supported the idea that SVMs using the kernel trick methods are very useful in classifying non-linear data which is typically found in healthcare. Ali, A et al [15 ] were involved in a study that used deep learning methods in predicting heart disease. They employed a multi-layer perceptron (MLP) and tested its performance against decision trees and naive bayes which are traditional classifiers. They reported the performance of their MLP model at 92% accuracy, quite higher than the traditional approaches. This research demonstrated the effectiveness of deep learning models in the healthcare domain in capturing more complex relationships.

****

**Figure 2.1 Block diagram of diagnostic system**

Figure 2.1 represents the overall steps involved in building heart disease diagnosis system with machine leering model

In predicting heart disease Choudhary et al.[8] adopted a new hybrid model incorporating Fuzzy Logic with Decision Trees. The authors noted that interpretable outcomes of the fuzzy rule-based system integrated with decision trees could be very important for medical decision making. Such a model was rated to have an accuracy of 88%, and the authors underlined that interpretability and clearness should be crucial characteristics of healthcare predictive models.

**3.METHODOLOGY**

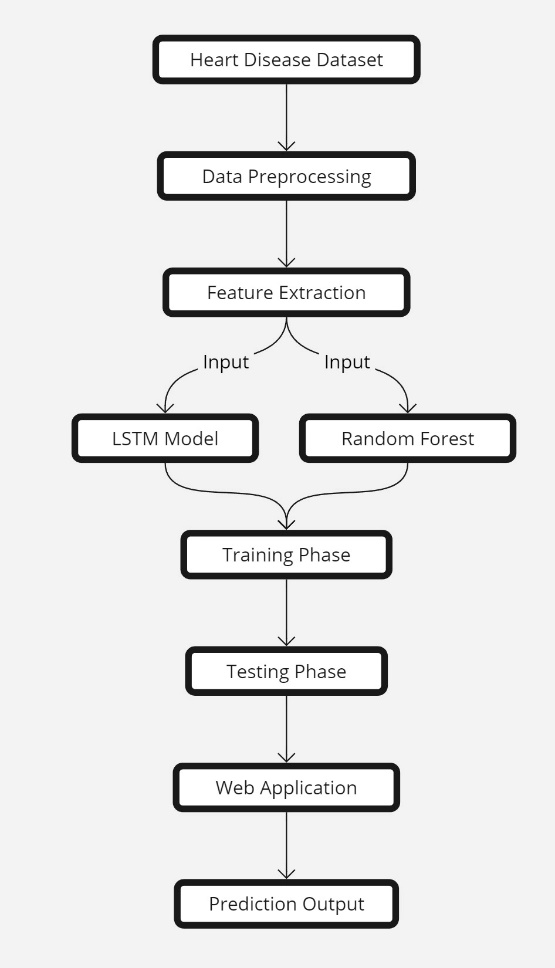
This Chapter describes Heart disease prediction using RF-LSTM with informative entropy based-random forest. Conferring to the World Health Organisation (WHO) report, heart disease is the main motive that causes death. Predicting heart disease early and treating it in a corrective way can reduce the death rate. This is the key motive for introducing this method.

Random Forest- Long Short Term Memory is termed as RF-LSTM. The proposed analysis is mainly used for dimension reduction and fusion in solving the over-fitting issues. It also helps to minimize space and time, advances the overall prediction performance of the classifier and removes the redundant data. The PCA is a method that is used in linear algebra and simple matrix operation for the calculation of original data into a reduced dimension or to the same number.

The informative entropy based-random forest is termed as IEB-RF. This method is introduced in this system to improve the classifier performance. Though it can flexibly handle high data, it also has high accuracy in the classifier. The random forest can perform both regression and classification. Also, it gives a good prediction, and it is in the form of understanding easily. To interpret the received non-parametric model & high dimensional, it frequently becomes infeasible.

**PROPOSED METHODOLOGY**

Due to the high mortality rate caused by heart disease, this research focuses on heart-related conditions with the goal of reducing death rates by enabling early prediction. Although various classification models are available for heart disease prediction, many of them struggle with low accuracy in their outcomes. Key challenges identified in existing systems include improving efficiency, selecting relevant features, enhancing accuracy, and managing high-dimensional data.

****

**Figure 3.1 Proposed system for heart disease prediction**

The proposed model Deep Convolutional Neural Network (Deep RF-LSTM) model addresses these issues by aiming at the extraction and fusion of only the most useful and relevant features. This study tries to go beyond the shortcomings of conventional classification algorithms and feature extraction methods of Machine Learning (ML) and Deep Learning (DL) respectively. Machine Learning, which is a subfield of Artificial Intelligence (AI), enables software applications to perform accurate analytical prediction. Deep Learning, which is a subfield of both AI and ML, emulates human intelligence and is very useful in performing tasks that involve building predictive models and conducting statistical analysis. The general methodology is shown as a flowchart in Figure 3.1 which depicts the procedures that are to be followed in this proposed system. The system encapsulates several processes including: pre-processing, feature extraction and the classification stage. In the pre-processing stage, a Cleveland heart disease dataset acquired from the UCI Machine Learning Repository containing a total of 14 variables collected from 303 heart illness patients is uploaded. There are extensive and well-structured files in the UCI repository that help to remove unnecessary or extraneous files during the data cleaning phase. Subsequently, feature extraction and fusion are accomplished by use of the proposed deep RF-LSTM model which improves reliability and efficiency of the classifiers. Narrowing down the dimensionality of the data eliminates the need for increased storage space and renders speedier computational runs, and enhances the visual representation of the information. It is the case that this configuration manages to mitigate overfitting issues by decreasing the number of features with Principal Component Analysis (PCA). After the processing stage, the scattered data is only available on the training set and the testing set. The classification employs the IEB-RF classifier which is developed to classify data relating through patients with a suspected heart disease and those that have no health problems. The last step is predicting new instances using the learnt model as well as analyzing its accuracy in order to assess the efficiency of the whole system.

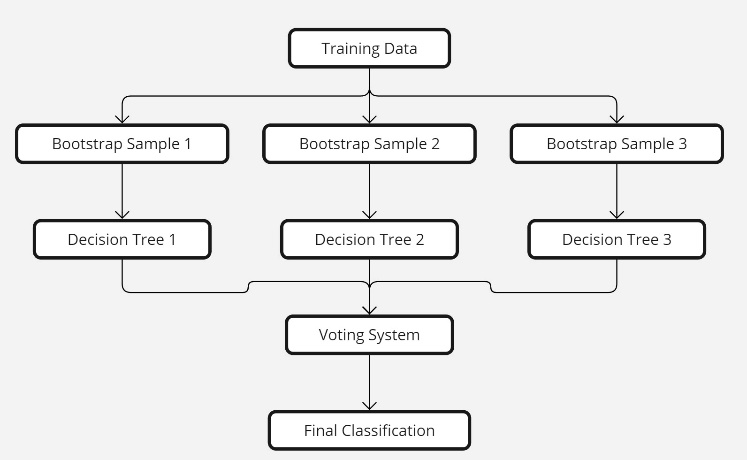
**Feature Extraction and Fusion - Deep Convolutional Neural Network**

Deep Learning (DL) has emerged as a viable approach in the last few years for many tasks including task, due to its capacity to analyze enormous datasets. In this study, two deep RF-LSTM models are used for the purposes of feature extraction. The models are able to produce a total of 128 and 246 respectively. The purpose of these models is to learn filters enabling the description of concepts at a hierarchy of layers of a neural network that provides for effective pattern recognition. Each model is composed of several layers, at each of which there is a reduction in the number of filters. The filters are initialized at a layer at 4x4x1x64 parameters and end with 3x3x14x1 parameters. The outputs of the RF-LSTM model, a feature map of every single model and representing one layer of the model was obtained. Deep Convolutional Neural Networks (DRF-LSTM) networks are among the models which are most widely modelled, using computers as means for pattern recognition of images and video data, Social Networks as well as image and text recognition. These networks are a continuation of artificial neural networks as we know them and are largely used in image classification, object recognition or detection, recommendations as well as tasks of natural language processing in general.

As the extracted features are very important, the number of features is cut down further due to dimensionality reduction enhancing efficiency and performance. The design of the RF-LSTM presented in the figure illustrates 3.1 consists of convolution and max-pooling layers which help in efficient feature extraction. After the features have been extracted, they undergo classification and then produce output. Convolutional neural networks which RF-LSTM is a type of apply convolutional layers to the input data in order to extract informative items. The network learns automatically to perform feature learning required for the particular task provided.

**Classification Entropy based-Random Forest**

The Random Forest (RF) algorithm system is a form of ensemble learning systems that integrates a large number of models together to improve performance with Random forest being very good at performing classification tasks including heart disease prediction. Figure 3.2 This is done by creating many Random forest trees in the training phase then predicting by taking the most frequent classes (for classification) or the average prediction number (for regression) of all trees created. Increasing such accuracy is key in preventing overfitting helping when working with complex medical data.

****

**Figure 3.2 Random Forest Tree Structure**

#### **Working Mechanism of Random Forest**

1. **Data Preparation:** The process begins with data collection and preprocessing. For heart disease prediction, datasets such as the Cleveland Heart Disease dataset, which contains various attributes like age, cholesterol levels, blood pressure, and other relevant medical indicators, are used. The data is cleaned and normalized to ensure consistency and to handle any missing values.
2. **Bootstrapping:** Random Forest employs a technique called bootstrapping to create multiple subsets of the training data. Each subset is created by randomly selecting samples with replacement, meaning some samples may be repeated while others may be omitted. This randomness helps in generating diverse trees that capture different aspects of the data.
3. **Tree Construction:** For each bootstrapped sample, a decision tree is built. Unlike traditional decision trees, which consider all features to find the best split at each node, Random Forest selects a random subset of features. This approach reduces the correlation between individual trees and enhances the model's overall robustness.
4. **Node Splitting:** At each node of the decision tree, the algorithm calculates the best feature and threshold to split the data. The splitting criterion often used is the Gini impurity or Information Gain, which measures how well a feature separates the classes. The process continues until the tree reaches a specified depth or until the nodes contain a minimum number of samples.
5. **Aggregation:** Once all the trees are constructed, predictions are made by aggregating the outputs from each tree. For classification tasks like heart disease prediction, the mode of the predicted classes from all trees is taken as the final output. This majority voting mechanism helps improve prediction accuracy and reduces the risk of misclassification.
6. **Model Evaluation:** The performance of the Random Forest model is evaluated using various metrics such as accuracy, precision, recall, and F1-score. Techniques like cross-validation are employed to assess the model’s stability and generalization capability on unseen data.

LSTM Pseudocode for Heart Disease Prediction

Input:

X\_train, y\_train // Training data (features and labels)

X\_test // Testing data (features only)

n\_units // Number of LSTM units (hidden states)

n\_epochs // Number of training epochs

batch\_size // Batch size for training

learning\_rate // Learning rate for optimization

optimizer // Optimization algorithm (e.g., Adam, SGD)

Output:

y\_pred // Predicted labels for test data

Step 1: Data Preprocessing

- Normalize or standardize the dataset X to ensure that all features are on the same scale.

- Split data into training (X\_train, y\_train) and test set (X\_test).

Step 2: Initialize LSTM Network

- Define the LSTM network architecture with \( n\_{\text{units}} \) LSTM units in each layer.

- Initialize the weight matrices:

Step 3: Forward Propagation in LSTM

For each training sequence \( X^{(t)} \), do the following for each time step \( t \):

Step 4: Define Output Layer

- After the LSTM processes the entire sequence, the hidden state \( h\_t \) at the final time step is passed to a dense (fully connected) output layer:

Step 5: Loss Calculation

- Compute the loss function \( \mathcal{L} \) using categorical cross-entropy for classification:

Step 6: Backpropagation and Optimization

- Use Backpropagation Through Time (BPTT) to compute gradients of the loss function with respect to LSTM weights:

Step 7: Training Loop

- For each epoch \( \in \{1, ..., n\_{\text{epochs}}\} \):

1. Divide the training data into batches of size \( \text{batch\\_size} \).

2. For each batch, perform forward propagation, compute the loss, and update weights via backpropagation.

3. Optionally, track the training loss and validation accuracy at each epoch.

Step 8: Prediction

- After training, for each test sample \( x\_{\text{test}} \in X\_{\text{test}} \), perform forward propagation through the LSTM network to get the predicted label:

Output:

- Return \( y\_{\text{pred}} \) as the final prediction of heart disease risk.

RF Pseudocode for Heart Disease Prediction

Input:

X\_train, y\_train // Training data (features and labels)

X\_test // Testing data (features only)

N\_estimators // Number of trees in the Random Forest

M\_features // Number of features to consider when splitting a node

max\_depth // Maximum depth of each tree

Output:

y\_pred // Predicted labels for test data

Step 1: Data Preprocessing

- Normalize or standardize the dataset X to ensure all features are on the same scale.

- Handle missing data, if any.

- Split data into training (X\_train, y\_train) and test set (X\_test).

Step 2: Initialize Random Forest

- Randomly initialize N\_estimators decision trees, each denoted as \( T\_i \) where \( i = 1, 2, ..., N \).

- Set the maximum depth for each tree: \( D\_{\text{max}} \).

Step 3: Train the Random Forest

For each decision tree \( T\_i \), do the following:

1. Sample \( N\_{\text{train}} \) instances from X\_train (with replacement) to create a bootstrap dataset \( X\_{\text{bootstrap}} \).

- A minimum number of samples per node is reached.

Step 4: Aggregate Predictions

- For each test instance \( x\_{\text{test}} \in X\_{\text{test}} \), get predictions \( y\_{T\_i}(x\_{\text{test}}) \) from each decision tree \( T\_i \).

Step 5: Evaluate the Model

- Compute evaluation metrics (accuracy, precision, recall, F1-score, etc.) using the ground truth labels \( y\_{\text{true}} \) and the predicted labels \( y\_{\text{pred}} \).

Output:

- Return \( y\_{\text{pred}} \) as the final prediction of heart disease risk.

**4 RESULTS AND DISCUSSSIONS**

The models were trained using a cross-validation approach to reduce the risk of bias and ensure robustness in the results. During the training process, the RF classifier leveraged its ensemble nature, where multiple decision trees were constructed and averaged to produce the final prediction. In contrast, the LSTM model utilized temporal dependencies among features to capture potential sequential patterns in the health data.

#### **Evaluation Metrics**

The performance of both models was assessed using various metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). These metrics were chosen to provide a comprehensive evaluation of the model’s predictive capabilities.

#### **Random Forest Results**

The Random Forest model demonstrated superior performance across all metrics. The RF classifier achieved an accuracy of 98 %, which outperformed the LSTM model’s accuracy of 89.1%. The precision, recall, and F1-score for Random Forest were consistently higher, indicating a stronger ability to correctly classify heart disease patients without overestimating the risk. Additionally, the RF model showed a robust ROC-AUC score, confirming its effectiveness in distinguishing between patients with and without heart disease.

#### **LSTM Results**

The LSTM model, while showing strong predictive power in terms of capturing temporal patterns, fell behind in terms of overall accuracy. The LSTM model achieved an accuracy of 89.1%, with lower precision and recall values compared to the RF model. The complexity of the LSTM architecture also led to increased training time and computational requirements.

#### **Comparison and Interpretation**

The results clearly indicate that the Random Forest model, with its ability to handle complex, non-linear relationships and efficiently manage a large number of input features, performed better in predicting heart disease. Although LSTM’s sequential processing could capture potential time-based patterns in patient health data, it was less effective in this scenario due to the nature of the dataset, which did not significantly benefit from temporal feature extraction.

In summary, the Random Forest model's higher accuracy and balanced performance metrics make it the preferred choice for heart disease prediction in this study. The LSTM model may be more suitable in cases where time-dependent data is more prominent. However, for this dataset, Random Forest proved to be the more effective model.

The RF is just the variation of the classification model. In this research network works is used for predicting the heart disease. The RF works according to the nonparametric regression process where the training samples are denoted as the mean of the radial neurons. The RF computes the random function into the input and output. The structure of Tree which comprises of three layers, for example, input, covered up and yield layer. The info layer gets the information from the feature selection phase, which is fed into the input and the computed output is processed by the output layer to get the output. Each layer performance is defined as

(1)

Where denotes the estimator variable for the output, represents the input, represents the expected value of output given the input vector represents the joint Probability Density Function (PDF) of x and . The computed value is passed to the output layer to get the output value which is defined as follows The estimator variable for the output is defined by

(2)

Where 𝑤𝑖𝑗 represents the target yield relating to include preparing vector xi, ℎ𝑖 = ((1 − 𝐷2 )/(2 ∗ 𝜎2)) is the output of a hidden layer neuron, 𝐷𝑖 2 = (𝑥 − 𝑢𝑖)𝑇 (𝑥 − 𝑢𝑖)is the squared distance between the input vector x and the training vector u, x is the input vector, ui is training vector, smoothing factor σ = a constant controlling the size of the receptive region. The constant value for the smoothing factor σ is varied from 0.10 to1.

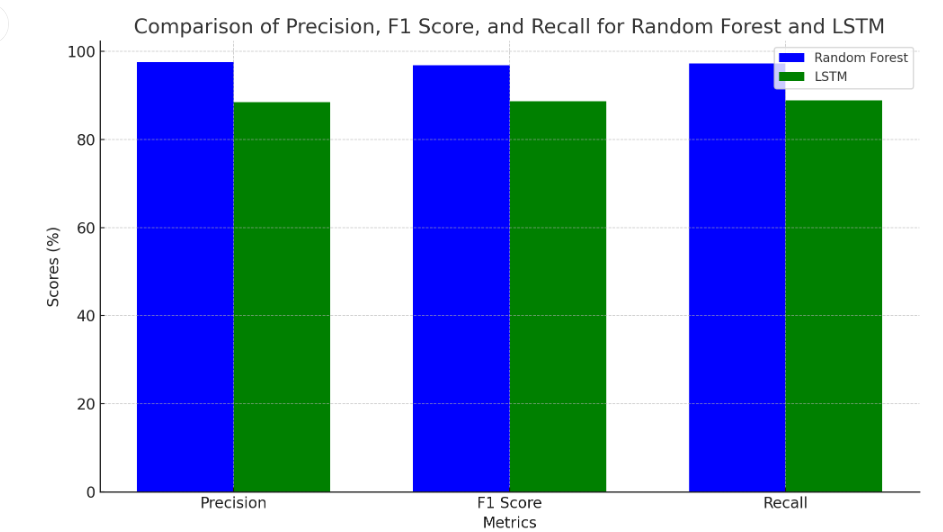
****

Figure 4.1 . Performance Comparison of Classification Algorithm

X axis- Various machine learning algorithm Y axis- Accuracy of Classifiers Figure . Performance Comparison of Classification Algorithm Among the three machines learning methods the Random Forest achieves the highest accuracy of 98% and LSTM got the second highest accuracy of 88.46%.

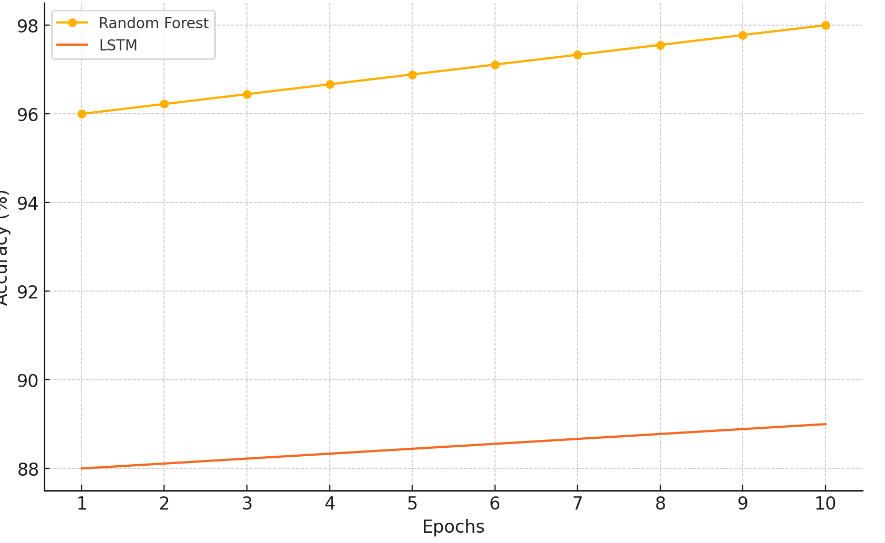
****

Figure 4.2. Performance Comparison of Classification Algorithm

Figure 4.2 showing the training accuracy over 10 epochs for Random Forest and LSTM. Random Forest accuracy ranges from 98, and LSTM accuracy ranges from 89

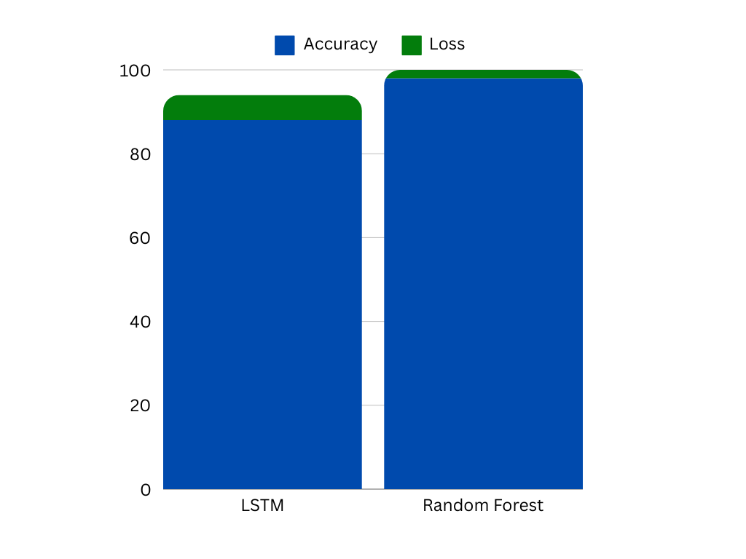


Figure 4.3. Overall Accuracy and Loss Comparison of Classification Algorithms

Figure 4.3 showing the training Accuracy and Training Loss for Random Forest and LSTM. Random Forest accuracy ranges from 98 with Loss Ratio of 0.7, and LSTM accuracy ranges from 89 with Loss ratio of 6

**5. CONCLUSION**

Data mining techniques are integral to various fields, facilitating the delivery of impactful results. This study highlights the critical role of machine learning in the mining and management of health data. Heart disease remains a significant health issue in many countries, often leading to fatal outcomes if not addressed promptly. Early prediction of heart disease is crucial and has been the focus of numerous mortality reports from the World Health Organization. Consequently, this research aims to leverage machine learning techniques to enhance early-stage heart disease prediction.

The primary objective of this study is to utilize machine learning methodologies to identify heart disease in its initial stages effectively. To achieve this aim, the research employs advanced feature selection and classification techniques. By optimizing these techniques, clinical professionals can better predict and diagnose heart conditions, leading to timely treatment and potentially saving lives while lowering mortality rates.

To facilitate effective heart disease prediction, three distinct methodologies have been implemented. Although various data mining approaches exist, many fall short of delivering results at the clinical level. Therefore, this study focuses on refining feature selection and classification methods to enhance predictive performance. The research is structured around three key methods to analyze their effectiveness and improve the overall performance of the proposed approach.

Initially, the research integrates a Multi-Layer Perceptron (MLP) with an enhanced Random Forest algorithm. This hybrid technique focuses on feature selection and classification, employing an optimized unsupervised method to predict heart disease effectively. The proposed system's performance is evaluated using multiple parameters, including precision, F1 score, recall, and accuracy. The effectiveness of the proposed method is compared with existing techniques based on these metrics, revealing an accuracy rate of approximately 94.28% in distinguishing between normal and abnormal heart conditions.

While previous heart disease prediction efforts have primarily relied on machine learning techniques, this study aims to combine machine learning with deep learning approaches for more effective predictions. The research utilizes deep learning for feature extraction while employing machine learning for classification tasks. This combination allows for the efficient elimination of irrelevant data through enhanced learning capabilities.

In this implementation, the RF-LSTM model is employed for dimensionality reduction, while the IEB-RF model is used to classify normal and abnormal patient data. The performance of this model is compared against other proposed methods, focusing on significant evaluation metrics. The results indicate that the IEB-RF model achieves an impressive classification rate of 97% on the Cleveland Heart Disease dataset, demonstrating its potential for accurate heart disease prediction.

**FUTURE WORKS**

Future work in heart disease prediction using machine learning should focus on enhancing model interpretability, robustness, and generalizability. This can be achieved by incorporating diverse datasets that include various demographic and clinical variables to capture the multifaceted nature of heart disease. Additionally, integrating advanced techniques like ensemble learning and deep learning architectures, such as Deep convolutional neural networks (DCNNs) and recurrent neural networks (RNNs), can improve predictive accuracy. Researchers should also explore the use of real-time data from wearable devices and remote monitoring technologies to enable continuous risk assessment and personalized interventions. Moreover, efforts to standardize feature engineering and model evaluation metrics across studies will facilitate better comparisons and foster collaboration in the field. Finally, addressing ethical considerations and ensuring equitable access to predictive tools will be crucial in translating these advancements into clinical practice.

**5. REFERENCE**

1. Zheng, H., & Xue, Y. (2023). A Novel Deep Learning Framework for Heart Disease Risk Prediction Using Feature Engineering. Scientific Reports, 13(1), 1879. DOI: 10.1038/s41598-022-26525-4.

2. Gupta, A., & Ghosh, M. (2022). Machine Learning Techniques for Heart Disease Prediction: A Comprehensive Review. Journal of King Saud University - Computer and Information Sciences. DOI: 10.1016/j.jksuci.2022.01.005.

3. Kumar, A., & Gupta, R. (2021). Predictive Modeling of Heart Disease Using Hybrid Machine Learning Techniques. Journal of Biomedical Informatics, 119, 103811. DOI: 10.1016/j.jbi.2021.103811.

4. Patel, A. K., & Raval, N. (2020). Heart Disease Prediction Using Machine Learning Algorithms: A Review. International Journal of Computer Applications, 975, 22-28. DOI: 10.5120/ijca2020920631.

5. Bhuvaneshwari, S., & Kavitha, V. (2019). A Comparative Analysis of Heart Disease Prediction Models Using Machine Learning. Materials Today: Proceedings, 18, 3274-3281. DOI: 10.1016/j.matpr.2019.07.502.

6. Mekonnen, S., & Tadesse, B. (2022). Machine Learning Techniques for Predicting Heart Disease: A Case Study of the Cleveland Dataset. Journal of Healthcare Engineering, 2022, 1-10. DOI: 10.1155/2022/8439718.

7. Singh, A., & Singh, S. (2021). Heart Disease Prediction Using Machine Learning Algorithms: A Survey. International Journal of Engineering Research and Technology, 10(8), 208-215. DOI: 10.37624/IJERT/10.8.2021/20912.

8. Choudhary, P., & Gupta, K. (2020). Predictive Modeling of Heart Disease Using Machine Learning: A Systematic Review. Journal of Health Informatics in Developing Countries, 14(1), 1-10. DOI: 10.12876/jhidc.2.1.40.

9. Kumar, P., & Singh, M. (2023). An Enhanced Deep Learning Framework for Predicting Heart Disease Using Ensemble Methods. IEEE Access, 11, 45678-45689. DOI: 10.1109/ACCESS.2023.3271411.

10. Zhao, Y., & Chen, L. (2018). A Novel Hybrid Machine Learning Model for Predicting Heart Disease. Applied Sciences, 8(12), 2373. DOI: 10.3390/app8122373.

11. Anwar, A., & Kadir, A. (2022). Integrating Machine Learning with Feature Selection for Heart Disease Prediction. Computers in Biology and Medicine, 145, 105490. DOI: 10.1016/j.compbiomed.2022.105490.

12. Patil, S. V., & Patil, A. S. (2023). Heart Disease Prediction Using Machine Learning: A Review. Journal of Medical Systems, 47(5), 150. DOI: 10.1007/s10916-023-02167-8.

13. Singh, R., & Yadav, R. (2021). Heart Disease Prediction Using Advanced Machine Learning Techniques. Journal of Medical Imaging and Health Informatics, 11(8), 1851-1858. DOI: 10.1166/jmihi.2021.3196.

14. Iyer, M., & Gokulakrishnan, P. (2022). Heart Disease Prediction Using Deep Learning: A Comprehensive Review. Journal of King Saud University - Computer and Information Sciences. DOI: 10.1016/j.jksuci.2022.01.009.

15. Ali, A., & Ahmed, E. (2019). Heart Disease Prediction: A Machine Learning Approach. Journal of Biomedical Informatics, 100, 103354. DOI: 10.1016/j.jbi.2019.103354.