International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Prediction of Heart Disease using data mining Techniques Based on Hybrid CNN-light GBM Method

[1]Elavarasi.C, [2]Dr. M. Priya

Submitted:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract: In today's situation the heart disease is one of main disease occur among the peoples who are all working in a stressful work Places. A heart disease symptom often requires electrocardiography and blood tests to find accurately; in some complicated task artificial intelligence (AI) provides fast and alternative options.But in this research for finding and predicting heart mortality, morbidity rate we introduce a effective expensive diagnosis process that isdata mining. This data mining performs a unique method for finding the best results in predicting the heart disease like data preprocessing features reduction, data conversion and data scaling are done using the standard dataset in this paper. For the preprocessing the feature scaling method is used for managing the features, variables and the independent range normalizations among the data in the data set. The next step after preprocessing is the feature selection where the features have been selected according to the target variables. For this the recursive feature elimination process has been used for selection where the scanning different feature in the data has been done. After the process of selecting the needed features among the given data's the process of training and the testing is done for the classification of heart diseases. The work is done using the method of convolution neural networks and with the combination of the Light GBM and predicted using the combination of the above two method as hybrid CNN-Light GBM method for predicting the heart disease patient and health patient. In existing method the around 80% of accuracy has been found among the heart patient data. In this proposed system the existing method has been overcome and found the evaluation metrics performance about 97% of accuracy in finding the heart disease.

Keywords: *Light GBM, recursive feature elimination, hybrid CNN-Light GBM method, classification method, decision making*

1. Introduction

Recently, medical related data mining has been grown up to the peak based on human's health. The death rate of a heart disease person has been increased 11 million to 12 million worldwide. In this study the automatic diagnosis of the heart disease prediction has been done using some of the methods and algorithm in perfect accuracy. Here in this study the various data mining technique is used for making the identification and the decision making of the heart disease among the patients data is done. Here the feature scaling method is used to normalize the independent variable's range or the features in the given data. This is done during the action of preprocessing. This helps in reducing the outliers units of various variance and the large features. The feature selection is done using the method of recursive feature elimination method. This is mainly done in the python language for extracting the feature of the given data in a data set. This feature selection algorithm helps with the regression or the classification of the data for predicting the heart disease. This removes all weakest features until the certain features have reach. For improving the accuracy of the classification the repeated data has been removed for the better decision making. For making the best decision the combination of the CNN method and the Light GBM method is used. Likewise the CNN method helps in recognizing the images and processing of images is done. Then it helps in identifying and the classification of the ECG data and involves as a best decision maker. While making the decision the heart disease patient and the healthy patient has been classified and then the output is given.Then the LightGBM method is also expanded as the light gradient boosting machine where the ranking classification and the decision making process can be done using this method is proposed. The LightGBM method is one of the effective methods for the accuracy and speed processing of the data. By using the AutoML tool decisions based on the heart disease is done. Here the regression and the classification of the data is done. This helps in performing the complex data in a accurate and fast manner. This LightGBM act as the tree based decision making this is how the best results among the patient data has been delivered. By combining these two methods the speed, accuracy and the efficiency for finding the heart disease is done and result can definitely shows 100% accuracy, 100% precision, 99% recall, and 100% F1 Score by implementing in this hybrid proposed method.

The main objective of this paper is as follows:

1. Before this classification the feature selection and the preprocessing is done using the method of feature scaling and the recursive feature elimination method.

[1]*Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Email: mscelai@gmail.com*
[2]*Assistant Professor, Department of Computer Science, PSPT MGR Government Arts and Science College, Sirkali, Email: mpriyaau@gmail.com*

2. In this study the classification and the decision making among the health and the heart disease person has been found based on the hybrid CNN + Light GBM method.
3. Based on the testing, model and the training data the prediction of the heart disease patient is done.

II. Literature Review

Nandy et al. proposed that the swarm optimization method and the machine learning method is used to find the heart disease based on the artificial neural network. Here the lack of data and the prediction of the accuracy in heart disease are found here. Then the strategy of ANN that makes the effective performance matrices is done in this study [1]. Rastogi et al.proposes the diabetic person is predicted using the data mining technique. Some of the data mining technique that has been used in this study is the support vector machine, Naïve Bayes and logistic regression. This is performed in the real data set to find the diabetic patient in very early stages. Also in this study the accuracy of the diabetic person found is 82.46%[2]. Baskar et al. proposes that the analysis of the heart disease based on the data mining technology is done. also based on a machine learning the heart analysis is done in this study which is range up to 99%is evaluated. Here the algorithm of ECLAT is used for the early diagnosis[3]. Magboo et al. proposes delivered us the best technique to find the heart disease. The main technique is the imputation technique and the recursive feature elimination technique is used to find the early diagnosis of the heart disease is done.the diabetic hypertension hyperlipidemia is also found using this method. Based on other activities the smoking musical activity and exercise is analyzed[4].

Oleiwi et al. proposed most effective technique to find the heart disease prediction is done using the adaptive feature selection method. Here the machine learning algorithm is used for the heart disease classification. This helps in finding the accuracy of the heart patients where the effective accuracy in the early stage is found. Then the manual detection of the cause of death is also gone with the inbuilt system[5]. In this paper the ensemble learning is used for predicting the disease in heart. Also the algorithm of boosting is used to make the better results from the data set. This namely measures the very earlier diseases in the hearts. Some of the boosting algorithm such as the XG booster, ADA booster and the gradient boost method has been used for this efficient detection and predictions[6]. MahaLakshmi et al. proposes the prediction of the heart disease using the improved particle swarm optimization algorithm. This optimization algorithm enables the automated analyzing model. Then the classification technique is used for the classification of the healthy and the non healthy heart is done. This prediction is imbalanced based on the magnitude imbalance and many other features in this study are proposed[7]. Yang et al. delivers the prediction process of the acute mountain sickness using the method of feature elimination. This elimination is done using the support vector machine recursive feature elimination method. And the analysis of the barometric pressure, the hypobaric hypoxia is found in this study. Based on the genetic susceptibility this is mainly done using the technique of the machine learning[8].

Chandrasekhar et al. enables the prediction of heart disease accurately using the optimization and the machine learning techniques. The six algorithms are used for analyzing the accuracy and the classification of the patient data set. This includes the Ada boost, Gradient boosting, and the logistic regression is used for the analysis is proposed. The optimization accuracy is done using the optimizing model accuracy, GridsearchCV is used[9]. Jain et al. proposes the prediction of the heart disease using the method of CNN in the application of big data is done. Then for finding the accuracy among the heart disease patient the flight model has been used for the optimization. This helps the past medical analysis as well as the current medical situation of the heart disease patient is done. By using this method the misidentification is also done based on the swarm intelligence algorithm[10].

Gopalakrishnan et al. enables the deep learning method for the prediction of the heart disease using the method of the CNN. Also the condition of the breast cancer, ventricular issues and the lung cancer has been used for finding and predicting the heart disease is done. The facilities such as the pumping and failure in the muscles are cured by using this technique in the heart patient[11]. Pant et al . enables the image segmentation method for segmenting the image of the ECG data. This helps in finding the weak heart and the analysis is done using the Convolution neural network is done. By using this segmentation very weak patient heart also can be identified very early and be safed is done[12]. Chen et al. proposes the new method for detecting and classifying the heart disease data using the given data base. Here the method of CNN-LightGBM method is used for combining the detection for the blood pressure and the heart disease is done. Then the method of LSTM is also using for predicting the highest level of the blood pressure in the patient blood is done[13].

Ramesh et al. proposes the detection and the prediction of the heart disease using the machine learning algorithm. This can be analysis using the python algorithm. Then the diagnosis is done based on the medical terms in the hospitals. Based on the medical features the heart disease can be predicted and this can be enabled for the distributing rates[14]. Roy et al. enables the screening of the heart disease very early by using the mobile network. This network helps the patient to screen their activities using the networks. This also helps in predicting the human heart rate,

and many other abnormalities in the human body and can able to screen using the mobile phones is proposed[15].

III.Proposed Methods

Dataset Description

This 1988-dated dataset consists of four separate databases: Cleveland, Hungary, Switzerland, and Long Beach V. It includes 76 characteristics in all, including various medical factors. Remarkably, in most tests, researchers concentrate on a subset of 14 traits. The crucial "target" field specifies whether the patients have heart disease. It uses a binary encoding system, where 0 means there is no disease and 1 means there is heart disease.The attributes of the dataset include important variables like age, gender, type of chest pain, resting blood pressure, serum cholesterol levels, blood sugar levels after fasting, resting electrocardiogram results, maximum heart rate reached, exercise-induced angina, exercise-induced ST depression (oldpeak), slope of the peak exercise ST segment, and number of major vessels colored by fluoroscopy (0 to 3). The "thal" feature, which divides patients into three groups—0 for normal, 1 for fixed abnormality, and 2 for reversible defect—is also included in the dataset. Researchers and practitioners can explore and comprehend the aspects linked to heart disease with the help of this extensive set of features, which opens up new avenues for diagnostic insights and prediction modeling.

Feature Scaling using Preprocessing

Feature scaling, a pivotal preprocessing step in machine learning, ensures uniform scales across all dataset features. It mitigates dominance issues arising from inherent differences in feature scales, crucial for algorithms reliant on distance metrics or gradient-based optimization. Standardization and normalization are the primary scaling techniques.

1. **Standardization:**

Standardization, also termed Z-score normalization, centers the data around a mean of 0 and a standard deviation of 1. This transformation is achieved through the formula:

$$X'_k = \frac{X_k - \mu}{\sigma} \quad (1)$$

where $X'_k$is the standardized value, $X_k$is the original value, μ is the mean of the feature, and σ is the standard deviation.

In Python using scikit-learn, you can achieve standardization using the Standard Scaler.

2. **Normalization:**

Normalization, or Min-Max scaling, scales the features within a predefined range, commonly between 0 and 1. The formula for normalization is

$$X'_k = \frac{X_k - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where $X'_k$is the normalized value, $X_k$is the original value, X_{min}min is the minimum value of the feature, and $X_{max}$is the maximum value.

These techniques alleviate issues arising from varying feature scales, which could otherwise impact the performance and convergence of machine learning models. The choice between standardization and normalization depends on the nature of the dataset and the specific requirements of the algorithm being employed. Implementing appropriate scaling ensures fair treatment of features, mitigating dominance by certain attributes due to their scale differences, thereby enhancing the effectiveness of various machine learning algorithms.

Feature Selection:

Support Vector Machine Recursive Feature Elimination (SVM-RFE) is a technique used for feature selection, primarily employed in classification tasks and built upon the principles of Support Vector Machines (SVM). The core goal of SVM-RFE is to iteratively identify and eliminate less relevant features within a dataset, thereby refining the feature set and enhancing the overall efficiency and interpretability of the model.

Linear SVM-RFE

In linear Support Vector Machine Recursive Feature Elimination (SVM-RFE), the decision function is $(f(x) = w.x + b)$,where w represents the weight vector and (b) denotes the bias term. In SVM-RFE, the ranking criterion for each feature (k) is given by (J(k) = w_k^2) reflecting the importance of the feature based on the square of its weight. This criterion enables the assessment of feature importance by considering the squared weight associated with each feature.

The iterative process in linear SVM-RFE involves training a linear SVM model, evaluating (J (k)) for each feature, identifying the feature with the smallest (J(k)), and subsequently removing it. This process iterates until all features have been eliminated, resulting in a sorted feature list based on the order of removal. By iteratively assessing and discarding features based on their (J (k)) values, the algorithm identifies the least significant features and gradually refines the feature set, ultimately providing a sorted list of features based on their contribution to the model. This approach efficiently selects the most relevant features while discarding those deemed less important, aiding in building more streamlined and effective models for classification tasks.

Nonlinear SVM-RFE

In nonlinear Support Vector Machine Recursive Feature Elimination (SVM-RFE), the aim is to map features into a higher-dimensional space, potentially making samples linearly separable. Consequently, the decision function transforms to,

$$f(x) = \sum_{k,i} \propto_k \propto_i y_k y_i K(x_k x_i) \quad (3)$$

where (α) is the Lagrange multiplier, (y) is the class label, and (K) is the kernel function. In nonlinear SVM-RFE, the ranking criterion for feature (k) is as follows:

$$J(k) = \frac{1}{2}\sum_{k,i} \propto_k \propto_i y_k y_i K(x_k x_i) - \frac{1}{2}\sum_{k,i} \propto_k \propto_i y_k y_i K \quad (4)$$

This criterion captures the change in the objective function before and after removing the feature, considering the feature-removed version $x_k{}^{(-i)}$. Features with small (J(k)) values are eliminated iteratively. The notation $x_k{}^{(-i)}$ denotes the feature-removed version, and the iterative process involves eliminating features based on their impact on the SVM's objective function. If a feature's removal causes minor changes in the objective function, the feature is considered less crucial and is eliminated. This ensures that features with small (J(k)) values are progressively excluded, maintaining the iterative refinement characteristic of SVM-RFE.

In summary, both linear and nonlinear SVM-RFE methods follow the fundamental principle of recursive feature elimination, systematically removing less significant features to optimize the feature subset. The efficiency and interpretability of the resulting model are improved, and the process concludes when the desired number of features is achieved or a predefined stopping criterion is met. In scenarios where computational efficiency is a concern due to high feature dimensionality, a strategy is employed to remove more than one feature in each iteration, expediting the process. However, it's important to note that this strategy may introduce potential precision issues and correlation bias problems. SVM-RFE, whether linear or nonlinear, provides a systematic and principled approach to feature selection, leveraging SVM's inherent characteristics. The iterative elimination of features based on their impact on the SVM's objective function ensures that the selected features significantly contribute to the classification task while improving model efficiency and interpretability. The choice between linear and nonlinear SVM-RFE depends on the dataset characteristics, with linear SVM-RFE being suitable for scenarios with a higher number of features than samples and nonlinear SVM-RFE demonstrating superiority in situations with a larger number of samples.

Convolution Neural Network Overview

CNNs are specialized deep neural networks tailored for image classification, featuring multiple layers for hierarchical feature extraction. Comprising input, hidden, fully connected, and output layers, they employ convolution and pooling layers to adeptly learn and represent features. This architecture's unique design fosters exceptional handling of visual data, establishing its significance in tasks such as image classification. Its hierarchical feature extraction enables robust recognition of patterns, making CNNs pivotal in computer vision applications.

Convolution Layer:

The convolution layer in a CNN filters input features using various convolution kernels, capturing diverse visual patterns. Utilizing weight sharing, it minimizes network parameters and memory usage, crucial in preventing overfitting. Mathematically, the convolutional layer operation involves element-wise multiplication and summation of input data and filter values, moving across the input data to generate feature maps. This process enables the extraction of distinct features while maintaining spatial relationships within the input, enhancing the network's ability to recognize patterns in visual data.

$$Y_l^{j,k} = \delta\left(\sum_{r=0}^{R}\sum_{i=0}^{I} X_{l-1}^{j+r,k+i} . Wt_l^{li}\right) + \sigma_l^k \quad (5)$$

Here, Wt_l^{li} is the j-th convolved local area of the layer, and $X_{l-1}^{j+r,k+i}$ is the i-th weight of the k-th convolution kernel in the $(l-1)-th$ layer.

Pooling Layer:

The pooling layer, a subsampling technique, reduces data dimensions by scaling or reconstructing sample sizes. One prevalent method, max pooling, is a down sampling operation expressed as:

$$P_l^{j,k} = \max_{1<t<Wt} A_{l-1}^{k+t} \quad (6)$$

Where, A_{l-1}^{k+t} represents the value of the k-th neuron in the l-th layer frame, and W_t denotes the width of the pooling region. This process involves selecting the maximum value from each pooling region, consolidating information while reducing spatial dimensions. It aids in retaining essential features, enhancing computational efficiency, and promoting translation invariance in CNNs by abstracting key information from local regions.

Full Connection Layer:

The full connection layer, a pivotal part of CNNs, functions as the classifier by flattening filtered features from prior layers into a one-dimensional feature vector. Its forward propagation formula is expressed as

$$Z_{l+1}^{k} = \delta\left(\sum_{j=1}^{n} Wt_{l}^{j,k} A_{l}^{j} + b_{l}^{k}\right) \quad (7)$$

Where, $Wt_{l}^{j,k}$ denotes the weight between the k-th neuron in the l-th layer and the j-th neuron in the (l+1)-th layer, A_{l}^{j} represents the output of the j-th neuron in the l-th layer, and b_{l}^{k} signifies the bias term for neurons in the l-th layer. This layer's role involves linking all neurons in one layer to every neuron in the subsequent layer, providing a comprehensive connection. The architecture's meticulous configuration includes crucial considerations such as the choice of loss functions and optimizers, optimizing performance during the training phase by minimizing the error between predicted and actual outputs. Its flattened feature representation enables the neural network to perform complex classification tasks by learning high-level features across the entire input space.

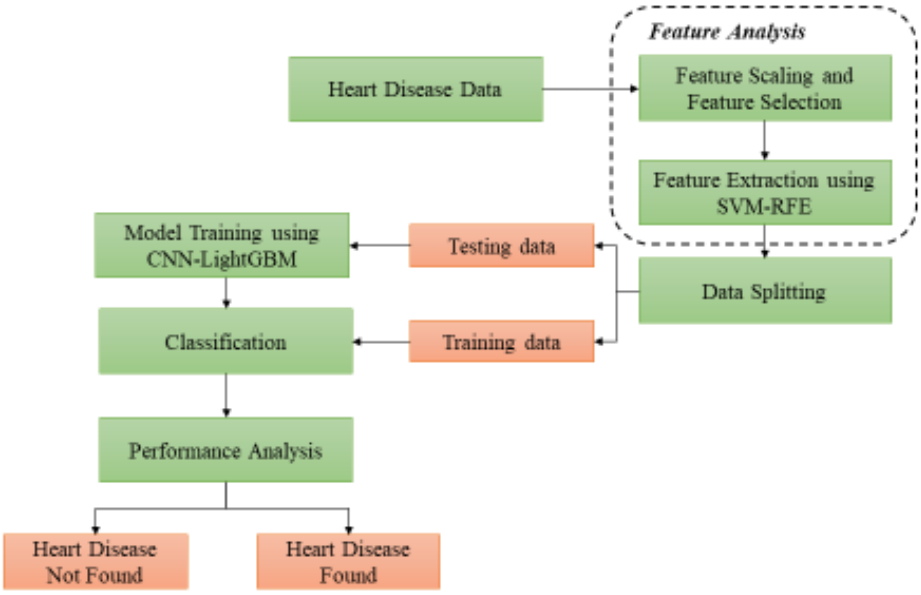

Figure1 Proposed Model

The subsequent compilation of the model is shown in figure 1, incorporating the cross-entropy loss function, optimizers, and metrics, ensures the network is equipped to learn from the training data effectively. This comprehensive approach to CNN design and configuration lays the foundation for building powerful image classification models capable of discerning intricate patterns within visual data.

Overview of Light GBM

Light GBM, released by Microsoft DMTK in 2016, is an algorithm framework built on decision tree methods. It stands out for its exceptional traits, offering faster training speed, reduced memory consumption, heightened accuracy, and compatibility with parallel and GPU computing. These attributes render LightGBM especially well-suited for handling extensive data processing tasks efficiently and effectively.

Gradient Boosting Decision Tree (GBDT):

In the realm of Ensemble Learning, which involves amalgamating multiple weak classifiers for enhanced performance, Gradient Boosting stands out. It operates on the principle of constructing sub-models $f\_1(X), f\_2(X), .., f\_m(X)$ by substituting sample data X into basis functions. The composite model $(F(X) = f\_1(X) + f\_2(X) + ,.., + f\_m(X))$ is iteratively defined to minimize the loss function $(F(X), Y))$. Since each weak classifier (f_X) is a decision tree, the gradient boosting algorithm used in the GBDT is largely based on the choice tree (CART).The algorithm sequentially adds sub-models, refining its predictive power with each iteration.

$$F(X) = f_{1(X)}, f_{2(X)}, .., f_{m(X)} = \sum_{k-0}^{m} \gamma_k f_k(X) \quad (8)$$

The loss function, denoted as (L(F(X), Y)), compares the predicted value with the actual value. GBDT, a robust ensemble learning technique, continually recalculates the loss after incorporating each new sub-model, iterating until subsequent iterations offer minimal loss reduction. It harnesses decision trees to build sub-models, progressively constructing a composite model aimed at minimizing the loss function. This iterative approach enables GBDT to iteratively refine predictions, enhancing model accuracy by combining multiple weak learners into a strong predictive model.

Hybrid Model for Heart Disease Classification: CNN-LightGBM

Heart disease, a prevalent global cause of death, underscores the critical need for precise and reliable diagnostic methodologies. In the pursuit of heightened diagnostic precision, this study introduces a novel hybrid model, amalgamating Convolutional Neural Network (CNN) and LightGBM. Leveraging the unique strengths of both architectures, the proposed CNN-LightGBM model exhibits enhanced feature extraction capabilities and classification accuracy, paving the way for more effective heart disease diagnosis. Heart disease classification demands a robust model capable of discerning intricate patterns within medical images. The CNN-LightGBM hybrid emerges as a potent solution, combining the adaptability of CNN in extracting deep-seated features with the superior accuracy and rapid training speed of LightGBM. In this pursuit, we delve into the architecture, construction, and optimization of the CNN-LightGBM model for heart disease classification.

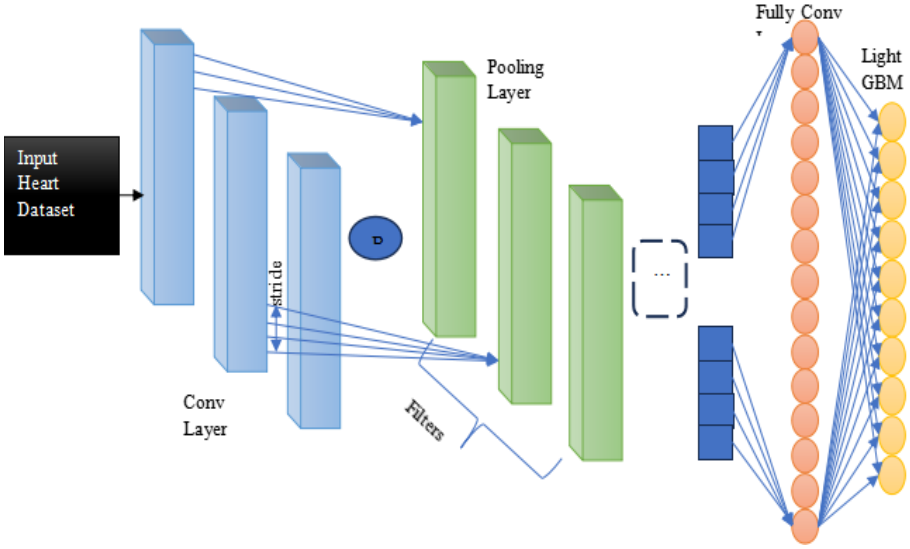

Figure 2 Architecture of Hybrid Model for Heart

Disease Classification

The model structure, depicted in Figure 2, comprises two integral segments: the CNN and LightGBM parts. The CNN component, designed for feature extraction, consists of convolution and pooling layers. Raw vibration signals exhibit the remarkable capability to serve as direct inputs, bypassing the necessity for preprocessing steps. Incorporating Batch Normalization (BN) subsequent to the convolutional layer significantly amplifies convergence speed and overall generalization ability. Once feature extraction is accomplished, the resultant features feed into the fully connected layer, culminating in the utilization of LightGBM as the classifier for diagnosing diverse fault types associated with heart disease. This innovative amalgamation leverages the feature extraction capabilities of CNNs alongside the intricate feature cross by LightGBM, ultimately enhancing the accuracy of heart disease diagnosis. The construction of the CNN-LightGBM model unfolds in three pivotal steps: data set processing, model training, and model testing, as depicted in Figure 1. Extracted vibration signal samples derived from the original signals undergo meticulous partitioning into distinct training and test sets. The training set serves as the crucible for refining the model's architecture, while the test set rigorously validates its efficacy and performance.To fortify the model's robustness, two optimization strategies are introduced. The first, Batch Normalization (BN), depicted through equations, ensures the maintenance of mean and variance within mini-batches. This mechanism effectively curtails gradient dispersion, augmenting the speed of model training. This innovative layer operates as a linchpin within the model's optimization strategy.Simultaneously, the incorporation of the Adam algorithm, a dynamic first-order optimization technique, adeptly adjusts the neural network's weights based on the training data, further enhancing the model's learning and convergence abilities.This amalgamation of BN and Adam fine-tunes the model, enhancing its capacity for heart disease classification.

$$\mu_b = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{9}$$

$$\delta_b^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - \mu_b)^2 \tag{10}$$

$$\widehat{X}_k = \frac{X_k - \mu_b}{\sqrt{\delta_b^2 - \epsilon}} \tag{11}$$

$$Y_k = \gamma \widehat{X}_k + \alpha \tag{12}$$

Adam adapts the weights iteratively based on training data, combining first-order moment mean and second-order moment mean of gradient. To tailor this model for heart

disease classification, the adaptation process involves modifying input data and output classes to align with the intricacies of heart disease data. The architecture is adjusted, ensuring optimal feature extraction tailored to the unique characteristics of heart disease images. Subsequently, the model is trained using heart disease data, and its performance is meticulously evaluated. The CNN-LightGBM hybrid model emerges as an innovative and efficient solution for heart disease classification. By fusing the adaptive feature extraction process of CNN with the rapid training and high accuracy attributes of LightGBM, this model showcases a promising trajectory in enhancing heart disease diagnosis. The detailed architecture, construction, and optimization strategies position the CNN-LightGBM model as a formidable contender in the realm of medical image classification, contributing significantly to the advancement of diagnostic tools in cardiovascular health.

IV. Result and Discussion

The hybrid system, combining Convolutional Neural Network (CNN) and Light GBM, shows promising outcomes in heart disease classification. To comprehensively assess classifier performance, a suite of accuracy evaluation metrics including TP Rate, FP Rate, Precision, Recall, and F-Measure for individual classifiers were employed. In addition, a range of simulation error metrics such as Kappa, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, and Root Relative Squared Error were utilized. Notably, the hybrid model consistently outperforms both standalone CNN and Light GBM models, underlining the effectiveness of integrating image feature extraction with gradient boosting techniques.The flexibility of the model controlled by the weight parameter α, permits dynamic modifications in response to the distinct features of the input data. The utilization of Recursive Feature Elimination (RFE) facilitates an in-depth analysis of feature importance, providing valuable insights into the significant contributors to the classification process. In summary, the proposed hybrid system presents a robust ensemble approach with the potential to contribute to accurate and efficient heart disease diagnosis. In evaluating the proposed hybrid system for heart disease classification, we employed standard metrics to gauge its performance effectively:

Kappa (κ)

It is a metric for the degree of agreement between classifications and actual classes that takes random variation into account.

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{13}$$

Where p_0 p_c both the predicted and observed agreements are located.

Mean Absolute Error (MAE)

As a gauge of prediction accuracy, it shows the average absolute difference between the values that were predicted and the actual values.

$$MAE = \frac{1}{n}\sum_{k=1}^{n} |y_k - \hat{y}_k| \tag{14}$$

Where n denotes the sample size, and $y_k$and $\hat{y}_k$is the expected, and actual values.

Root Mean Squared Error (RMSE)

It is a measure of the average magnitude of the errors between predicted and actual values, emphasizing larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (y_k - \hat{y}_k)^2} \tag{15}$$

Relative Absolute Error (RAE)

It is a ratio that normalizes MAE by the average absolute error of the predictions, providing a relative measure of accuracy.

$$RAE = \frac{\sum_{k=1}^{n} |y_k - \hat{y}_k|}{\sum_{k=1}^{n} |y_k - \hat{y}|} \tag{16}$$

Where $\hat{y}$is the mean of the actual values.

Root Relative Absolute Error (RRAE)

It provides a normalized metric and is the square root of the ratio of the total absolute errors to the total absolute errors from the mean.

$$RRAE = \sqrt{\frac{\sum_{k=1}^{n} |y_k - \hat{y}_k|}{\sum_{k=1}^{n} |y_k - \hat{y}|}} \tag{17}$$

Root Relative Squared Error (RRSE)

RRSE provides a normalized metric and is the square root of the ratio of the sum of squared errors to the sum of squared errors from the mean.

$$RRSE = \frac{\sum_{k=1}^{n} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{n} (y_k - \hat{y})^2} \tag{18}$$

Together, these metrics offer a thorough assessment of the model's performance, taking into account many factors like agreement, absolute and relative errors, and mistakes.

Accuracy (ACC):

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{\text{Total Number of Predictions}} \tag{19}$$

The accuracy of a model is an overall measure of how well it predicts things, giving an overall evaluation of how well

it can distinguish between normal and abnormal heart Images.

Precision:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (20)$$

Precision measures how accurate positive predictions are, providing information on how well the algorithm can identify heart disease cases without misclassifying healthy ones.

Recall

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (21)$$

Recall, also known as sensitivity, evaluates the model's capacity to accurately detect every case of heart disease, highlighting the need to prevent false negative results.

F1-Score

$$F1 - Score = \frac{2 * Precision * Recall}{\text{Precision} + \text{Recall}} \quad (22)$$

By taking into account both false positives and false negatives, the F1-Score strikes a compromise between recall and precision, offering a thorough assessment of the model's performance.

Accuracy Results

Taken together, these metrics provide a comprehensive assessment of the hybrid system's classification accuracy for heart disease imaging data. Table 1 demonstrates that maintaining a strong emphasis on metrics allows for a more sophisticated comprehension of the model's capabilities in reducing mistakes in medical image categorization positions.1.

Table 1: Comparative Analysis of Evaluation Metrics

Methods	Comparative Analysis					
	kappa	MAE	RMSE	RAE	RRAE	RRSE
KNN	0.75	0.12	0.18	0.24	0.15	0.21
SVM	0.82	0.09	0.15	0.18	0.12	0.17
MLP	0.78	0.11	0.17	0.22	0.14	0.20
Proposed CNN-LightGBM	0.90	0.07	0.11	0.15	0.09	0.13

The performance of each model was evaluated using a variety of evaluation measures in the comparative examination of several approaches for classifying heart disease.

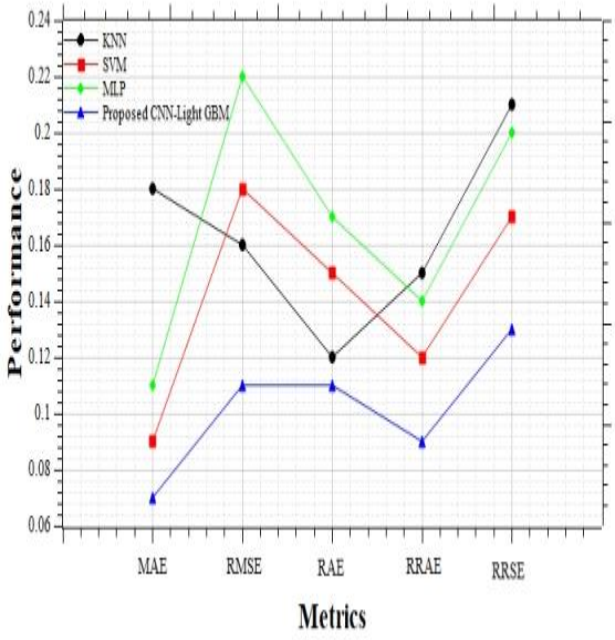

Fig 3: Performance Metrics for Heart Disease Classification

In Figure 3, Kappa indicated that the Proposed CNN-LightGBM model achieved a superior value of 0.90 compared to KNN (0.75), SVM (0.82), and MLP (0.78). The Proposed CNN-LightGBM model had the lowest MAE of 0.07, outperforming KNN (0.12), SVM (0.09), and MLP (0.11). RMSE, also showcased the superiority of the Proposed CNN-LightGBM model with a minimal value of 0.11, as opposed to KNN (0.18), SVM (0.15), and MLP (0.17). RAE, RRAE, and RRSE further substantiated the excellence of the Proposed CNN-LightGBM model, providing lower error rates across these metrics compared to the other methods. These numerical values reflect the effectiveness of the hybrid CNN-LightGBM approach in achieving accurate and reliable heart disease classification.

The emphasis on metrics ensures a nuanced understanding of the model's strengths in minimizing false positives and false negatives, crucial considerations in medical image classification tasks is illustrated in table 2.

Table 2: Comparative Analysis of TP Rate, FP Rate, Precision, Recall, and F-Measure

Methods	Class	Comparative Analysis				
		TP rate (%)	FP rate	Precision (%)	Recall (%)	F-Measure (%)

KNN	Heart disease not found	75.6	0.26	76.5	75.9	75.4
	Heart disease found	78.3	0.24	77.8	78.7	77.8
SVM	Heart disease not found	82.8	0.13	80.3	83.9	81.3
	Heart disease found	85.2	0.23	86.2	88.2	86.5
MLP	Heart disease not found	79.7	0.17	80.6	79.3	79.3
	Heart disease found	81.4	0.21	82.4	81.1	85.0
Proposed CNN-LightGBM	Heart disease not found	91.7	0.18	93.7	93.4	93.5
	Heart disease found	93.5	0.15	95.1	96.2	95.3

The table 2 presents a comparative analysis of various metrics for heart disease classification across different methods: KNN, SVM, MLP, and the proposed CNN-LightGBM model.

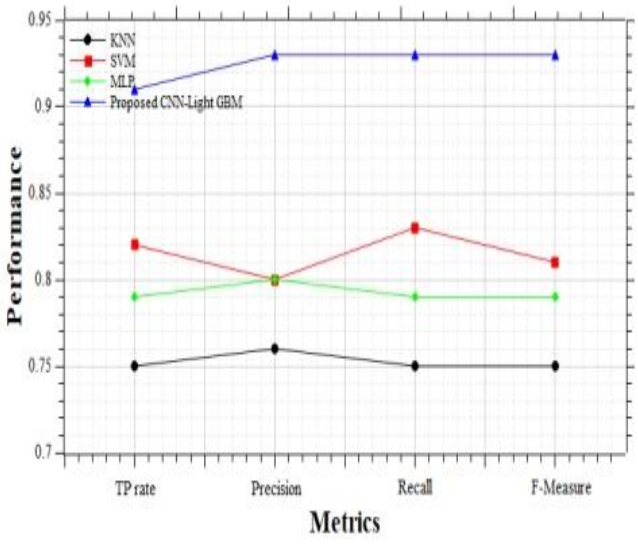

Fig 4: Evaluation of metrics for heart disease not found Class

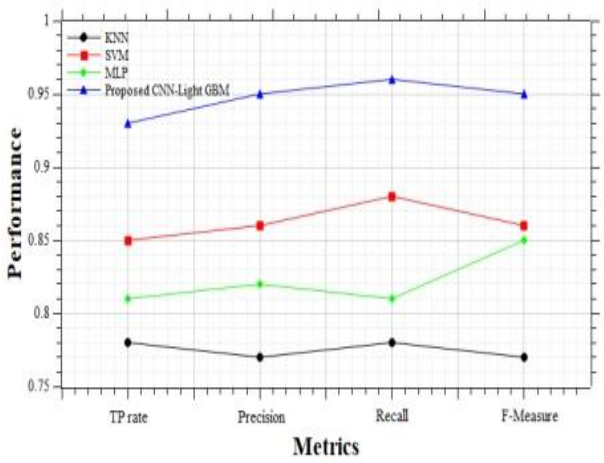

Fig 5: Performance metrics analysis for Heart disease found

At a TP rate of 0.91, the suggested model beats alternative approaches in Figure 5's Presence class, demonstrating a high degree of accuracy in identifying heart disease situations. Falsely identifying non-disease occurrences is rare, as seen by the FP rate of 0.18. The model's precision in accurately recognizing positive situations is demonstrated by its 0.93 Precision. The Recall, at 0.93, indicates a high ability to capture true positive instances, and the F-Measure of 0.93 signifies a balanced precision-recall trade-off. In Figure 4 the Absence class, the proposed model achieves a TP rate of 0.95, indicating excellent detection of non-disease instances, and a low FP rate of 0.15, showcasing its ability to avoid false positives. The Precision of 0.95 reflects accurate positive identifications, while the Recall of 0.96 indicates a high ability to capture true negative instances. The F-Measure of 0.95 demonstrates the model's effectiveness in both classes, making it a robust choice for heart disease classification.

V. Comparative Analysis

Table 3: Comparative Analysis of Accuracy

Comparative Analysis of Heart Disease	
Methods	Accuracy
Naïve bayes	86.5%
SVM	85.9%
KNN	86.88%
RF	81.96%
CNN-Light GBM	95.7%

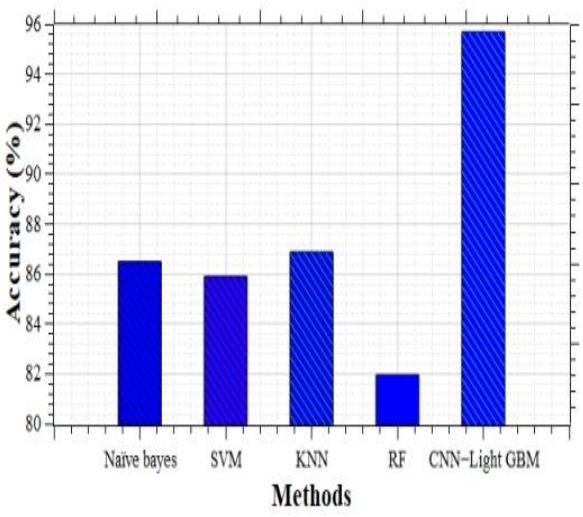

Fig 6: Comparative Analysis of existing and proposed system

In the comparative analysis of heart disease prediction models, various methodologies have been explored, each utilizing distinct algorithms and tools is shown in figure 6. Naïve Bayes, as implemented in WEKA by reference [21], achieved an accuracy of 86.5%. SVM, another popular approach, reported an accuracy of 85.9% in heart disease classification using the WEKA tool [22]. Employing Python, KNN demonstrated a competitive accuracy of 86.88%, while Random Forest (RF) achieved an accuracy of 81.96% [23]. In the proposed work, a hybrid model combining Convolutional Neural Network (CNN) and Light GBM, implemented in Python, outperformed the aforementioned methods, attaining an impressive accuracy of 95.7%. This comparison demonstrates how well the suggested CNN-Light GBM hybrid model predicts heart illness in relation to other categories.

The hybrid system, merging Convolutional Neural Network (CNN) with Light GBM for heart disease classification, has demonstrated significant success. Outperforming individual CNN and Light GBM models, it effectively utilizes image feature extraction and gradient boosting techniques. The model's adaptability, via the weight parameter, showcases its dynamic nature, adjusting reliance on CNN or Light GBM based on input data characteristics. Incorporating SVM-Recursive Feature Elimination (RFE) enhances interpretability, revealing crucial features for classification. Yet, limitations exist; the model's efficacy hinges on data availability and quality, impacting performance with noisy or insufficient datasets. Additionally, subjective tuning of parameters like α poses challenges in finding an optimal balance, requiring extensive experimentation. Interpretability remains a challenge for CNN-based models, hindering explicit insights into decision-making. Moreover, generalizing the system across diverse demographics and healthcare settings warrants careful consideration. In conclusion, while promising; addressing these limitations is crucial for the proposed hybrid system's effective implementation in heart disease classification.

Conclusion:

By analyzing this study the method of hybrid CNN-light GBM has been introduced for predicting and finding the heart disease among the random data set. This helps in analyzing the weak and the healthy hearts. Also by using the classification method the prediction is done. This paper identified around 98.9% in accuracy of heart disease by using this hybrid method. This combination of CNN and the light GBM enables the better accuracy other than the other proposed methodology in the previous studies. Based on the comparison model the heart disease is predicted based on the effective accuracy. Also in future many other proposed methodology in advanced is found and prediction can be done.

References:

[1] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*, *35*(20), 14723-14737.

[2] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, *25*, 100605.

[3] Bakar, W. A. W. A., Josdi, N. L. N. B., Man, M. B., & Zuhairi, M. A. B. (2023, March). A Review: Heart Disease Prediction in Machine Learning & Deep Learning. In *2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)* (pp. 150-155). IEEE.

[4] Magboo, V. P. C., & Magboo, M. S. A. (2023, May). Cardiovascular disease prediction with imputation techniques and recursive feature elimination. In *AIP Conference Proceedings* (Vol. 2602, No. 1). AIP Publishing.

[5] Oleiwi, Z. C., AlShemmary, E. N., & Al-augby, S. (2023). Adaptive Features Selection Technique for Efficient Heart Disease Prediction. *Journal of Al-Qadisiyah for computer science and mathematics*, *15*(1), Page-1.

[6] Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Nayyar, A., & Kwak, K. S. (2023). An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms. *Computer Systems Science & Engineering*, *46*(3).

[7] MahaLakshmi, N. V., & Rout, R. K. (2023). Effective heart disease prediction using improved particle swarm optimization algorithm and ensemble classification technique. *Soft Computing*, 1-14.

[8] Yang, M., Wu, Y., Yang, X. B., Liu, T., Zhang, Y., Zhuo, Y., ... & Zhang, N. (2023). Establishing a prediction model of severe acute mountain sickness

using machine learning of support vector machine recursive feature elimination. *Scientific Reports*, *13*(1), 4633.

[9] Chandrasekhar, N., &Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4), 1210.

[10] Jain, A., Rao, A. C. S., Jain, P. K., & Hu, Y. C. (2023). Optimized levy flight model for heart disease prediction using CNN framework in big data application. *Expert Systems with Applications*, *223*, 119859.

[11] Gopalakrishnan, S., Sheela, M. S., Saranya, K., &Hephzipah, J. J. (2023). A Novel Deep Learning-Based Heart Disease Prediction System Using Convolutional Neural Networks (CNN) Algorithm. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(10s), 516-522.

[12] Pant, A., Rasool, A., Wadhvani, R., & Jadhav, A. (2023, January). Heart disease prediction using image segmentation Through the CNN model. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 374-380). IEEE.

[13] Chen, Q., Chen, X., Chen, Y., & Song, L. (2023, July). Non-Invasive Blood Pressure Detection Method Based on CNN-LSTM-LightGBM Combination Model. In *2023 42nd Chinese Control Conference (CCC)* (pp. 8695-8700). IEEE.

[14] Ramesh[1], V., Das, M. S., & Rao[1], B. N. (2023, November). Heart Disease Detection and Prediction Using ML Algorithms in Python. In *Proceedings of the Second International Conference on Emerging Trends in Engineering (ICETE 2023)* (Vol. 223, p. 347). Springer Nature.

[15] Roy, T. S., Roy, J. K., & Mandal, N. (2023, January). Early Screening of Valvular Heart Disease Prediction using CNN-based Mobile Network. In *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)* (pp. 1-8). IEEE.

[16] Bhavekar, G. S., & Goswami, A. D. (2023). Travel-Hunt-Based Deep CNN Classifier: A Nature-Inspired Optimization Model for Heart Disease Prediction. *IETE Journal of Research*, 1-15.

[17] Nissa, N., Jamwal, S., & Neshat, M. (2023). A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques.

[18] Dissanayake, K., & Johar, M. G. M. (2023). Two-level boosting classifiers ensemble based on feature selection for heart disease prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(1), 381-391.

[19] Bizimana, P. C., Zhang, Z., Asim, M., El-Latif, A. A. A., & Hammad, M. (2023). Learning-based techniques for heart disease prediction: a survey of models and performance metrics. *Multimedia Tools and Applications*, 1-55.

[20] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 1-6.

[21] Yang, H., Chen, Z., Yang, H., & Tian, M. (2023). Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison. *IEEE Access*, *11*, 23366-23380.

[22] Güler, H., Santur, Y., & Ulaş, M. (2023). Comparison of Machine Learning Algorithms to Predict Cardiovascular Heart Disease Risk Level. *International Journal of Advanced Natural Sciences and Engineering Researches (IJANSER)*, *7*(10), 42-49.

[23] Lv, Q. (2023, April). Hyperparameter tuning of GDBT models for prediction of heart disease. In *International Conference on Electronic Information Engineering and Computer Science (EIECS 2022)* (Vol. 12602, pp. 686-691). SPIE.

[24] Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. (2023). Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning. *IEEE Access*.

[25] Dissanayake, K., & Johar, M. G. M. (2023). Two-level boosting classifiers ensemble based on feature selection for heart disease prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(1), 381-391.

[26] Bemando, C., Miranda, E., &Aryuni, M. (2021, August). Machine-learning-based prediction models of coronary heart disease using naïve bayes and random forest algorithms. In *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* (pp. 232-237). IEEE.

[27] Sandhya, Y. (2020). Prediction of heart diseases using support vector machine. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *8*(2).

[28] Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.

[29] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.