ISSN:2229-6107

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY

E-mail :
editor.ijpast@gmail.com
editor@ijpast.in

www.ijpast.in

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning

K. RAMESH BABU, NARAHARI ANITHA

ABSTRACT: Heart Disease, a chronic ailment impacting millions globally, underscores the significance of early detection. An innovative feature engineering technique, utilizing Principal Component analysis, is introduced to identify and enhance the most crucial features Utilizing machine learning, the project aims to predict Heart Disease's health status promptly and initiate essential actions. Include project, an ensemble method is implemented, specifically a Stacking Classifier, which combines the predictions of Random Forest (RF), Multilayer Perceptron (MLP), and LightGBM models. This approach synergistically leverages the strengths of individual models, resulting in a highly robust and accurate final prediction, achieving an impressive 100% accuracy. The selected features based on Principal Component Heart Failure (PCHF) were utilized for model building, and the Stacking Classifier was trained to be deployed in the front end. The integration of Flask framework with user authentication ensures an effective and secure platform for user testing, enhancing the accessibility and usability of our machine learning-based heart disease prediction system.

Keywords *– Machine learning, heart failure, cross validations, feature engineering*

INTRODUCTION

Heart failure is a condition in which the heart is unable to pump enough blood to meet the body's needs [1]. Cardiovascular diseases have emerged as a significant global health concern, substantially impacting public health worldwide. Heart failure is a common and serious condition affecting millions worldwide. According to a recent state, heart failure disorders cause to happen around 26 million population [2]. The causes of heart failure can be divided into two categories. First related to the heart's structure, such as a previous heart attack. Second related to the heart's function, such as high blood pressure. Symptoms of heart failure can include shortness of breath, fatigue, and swelling in the legs and ankles. Treatment options for heart failure include medications, lifestyle changes, and in some cases, surgery. Research has shown that early detection and management of heart failure can improve quality of life and prolong survival [3]. The current study focuses on developing a machinelearning model for managing heart failure to improve patient health.

Head of the Department[1], dept of CSE, Chirala Engineering College, Chirala,
ramesh.cs04@gmail.com
PG Student[2] -MCA, Dept of MCA, Chirala Engineering College, Chirala,
anithanarahari123@gmail.com

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

Machine learning is highly involved in medical diagnoses and the healthcare industry [4]. Machine learning has many applications in the medical field, including drug discovery, medical imaging diagnosis, outbreak prediction, and heart failure prediction. Machine learning techniques can learn patterns from large medical data and perform predictive analysis. Machine learning has many advantages compared to classical medical methods, such as saving time and costs, which helps improve diagnosis.

A novel PCHF feature engineering technique is proposed to select the most prominent features to enhance performance. Eight dataset features with highimportance values are selected to develop the machine learning methods using the proposed PCHF technique. We optimized the proposed PCHF mechanism by creating a new feature set as an innovation to achieve the highest accuracy scores compared to past proposed techniques. The nine advanced models of machine learning are used in the comparison to predict heart failure. The hyperparameters tuning of each applied machine learning method is conducted to determine the best-fit parameters, achieving a high-performance accuracy score. To validate the performance of applied machine learning models, we have used the k-fold cross-validation technique.

Heart disease is considered the most dangerous and deadly human disease according to the states discussed in previous studies. The increasing incidence of fatal cardiovascular diseases is a significant threat and burden to healthcare systems worldwide [15], [16]. Children are mostly affected by this critical disease [17]. This study [18] discusses the relevance of categorization models and describes the characteristics of models that have previously been applied in healthcare. The study highlights that several investigation groups have successfully tested data mining methods in clinical applications. The researchers compared the performance of several functional classifiers using two apparatuses, WEKA and MATLAB. Generally, the precision of the decision tree, logistic regression, SVM, and other algorithms reached 52% to 67.7%, which is relatively low [19].

Previous research [11] improved the accuracy from 87.27% to 93.13%, which is good but not optimal, as shown in Table 1. Past studies detect heart failure in patients using methods such as SVM, random forest, decision tree, logistic regression, and naïve bayes classifier. After comparing the results, the decision tree achieved an accuracy of 93.19%, which is good detection of heart failure in a specific dataset.

The study [20] used Cleveland data and created an ensemble model for heart disease detection. The ensemble models were built using random forest, gradient boosting, and extreme gradient boosting classifiers, achieving an accuracy of 85.71% [7]. The Cleveland data was used in the proposed study to improve the heart disease prediction by feature selection technique which helps to achieve an accuracy of 86.60%. Finally, previous studies have found significant research gaps, suggesting that the performance accuracy is not up to mark.

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

Consequently, we thoroughly evaluate the previous study's performance analysis in this part. This related work section is based on findings summarizing the efficiency of all previously applied models. According to previous studies, different types of models still provide different prediction scores. Thus, dimensionality reduction and feature engineering can enhance the data selection, causing greater prediction accuracy [21].

We have improved our proposed study's accuracy score compared to the previous research performance score. The precise credentials and findings of heart failure are necessary for proper treatment. We used advanced machine learning techniques in this study to achieve this goal.

1. LITERATURE REVIEW

Chronic heart failure represents a global pandemic, currently affecting over 26 million of patients worldwide. It is a major contributor in the death rate of patients with cardiovascular diseases and results in more than 1 million hospitalizations annually in Europe and North America. Methods for chronic heart failure detection can be utilized to act preventive, improve early diagnosis and avoid hospitalizations or even life-threatening situations, thus highly enhance the quality of patient's life. In this paper [1], we present a machine-learning method for chronic heart failure detection from heart sounds. The method consists of: filtering, segmentation, feature extraction and machine learning [4, 5, 6, 7, 8, 10]. The method was tested with a leave-one-subject-out evaluation technique on data from 122 subjects, gathered in the study. The method achieved 96% accuracy, outperforming a majority classifier for 15 percentage points. More specifically, it detects (recalls) 87% of the chronic heart failure subjects with a precision of 87%. The study confirmed that advanced machine learning applied on real-life sounds recorded with an unobtrusive digital stethoscope can be used for chronic heart failure detection.

Heart failure (HF) is a global pandemic affecting at least 26 million people worldwide and is increasing in prevalence. HF health expenditures are considerable and will increase dramatically with an ageing population [2]. Despite the significant advances in therapies and prevention, mortality and morbidity are still high and quality of life poor. The prevalence, incidence, mortality and morbidity rates reported show geographic variations, depending on the different aetiologies and clinical characteristics observed among patients with HF [1, 8, 11, 12]. In this review we focus on the global epidemiology of HF, providing data about prevalence, incidence, mortality and morbidity worldwide.

Recent years have witnessed widespread adoption of machine learning (ML)/deep learning (DL) techniques due to their superior performance for a variety of healthcare applications ranging from the prediction of cardiac arrest from one-dimensional heart signals to computer-aided diagnosis (CADx) using multi-dimensional medical images. Notwithstanding the impressive performance of ML/DL, there are still lingering doubts regarding the robustness of ML/DL in healthcare settings (which is traditionally considered quite challenging due to the myriad security and privacy issues involved), especially in

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

light of recent results that have shown that ML/DL are vulnerable to adversarial attacks. In this paper [4], we present an overview of various application areas in healthcare that leverage such techniques from security and privacy point of view and present associated challenges. In addition, we present potential methods to ensure secure and privacy-preserving ML for healthcare applications. Finally, we provide insight into the current research challenges and promising directions for future research.

Coronary heart disease is one of the major causes of deaths around the globe. Predicating a heart disease is one of the most challenging tasks in the field of clinical data analysis. Machine learning (ML) is useful in diagnostic assistance in terms of decision making and prediction on the basis of the data produced by healthcare sector globally. We have also perceived ML [4, 5, 6, 7, 8, 10]. techniques employed in the medical field of disease prediction. In this regard [5], numerous research studies have been shown on heart disease prediction using an ML classifier. In this paper, we used eleven ML classifiers to identify key features, which improved the predictability of heart disease. To introduce the prediction model, various feature combinations and well-known classification algorithms were used. We achieved 95% accuracy with gradient boosted trees and multilayer perceptron in the heart disease prediction model. The Random Forest gives a better performance level in heart disease prediction, with an accuracy level of 96%.

Nowadays, people are getting caught in their day-to-day lives doing their work and other things and ignoring their health. Due to this hectic life and ignorance towards their health, the number of people getting sick increases every day. Moreover, most of the people are suffering from a disease like heart disease. Global deaths of almost 31% population are due to heart-related disease as data contributed by the World Health Organization (WHO). So, the prediction of happening heart disease or not becomes important for the medical field. However, data received by the medical sector or hospitals is so huge that sometimes it becomes difficult to analyze. Using machine learning techniques [8, 10]. for this prediction and handling of data can become very efficient for medical people. Hence in this study [6], we have discussed the heart disease and its risk factors and explained machine learning techniques. Using that machine learning techniques, we have predicted heart disease and provided a comparative analysis of the algorithms for machine learning used for the experiment of the prediction. The goal or objective of this research is completely related to the prediction of heart disease via a machine learning technique and analysis of them.

3.METHODOLOGY

i) Proposed Work:

Machine learning methods are utilized to improve the early detection of Heart Disease. Nine distinct machine learning algorithms, such as logistic regression, random forest, support vector machine, decision tree, extreme gradient boosting, naive bayes, k-nearest neighbours, multilayer perceptron, and gradient boosting, are employed and compared. To enhance accuracy, an innovative Principal Component Analysis (PCA) feature engineering technique is introduced, focusing on the selection of essential features. And also an ensemble method is

ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

implemented, specifically a Stacking Classifier, which combines the predictions of Random Forest (RF), Multilayer Perceptron (MLP) [31], and LightGBM models. This approach synergistically leverages the strengths of individual models, resulting in a highly robust and accurate final prediction, achieving an impressive 100% accuracy. The selected features based on Principal Component Heart Failure (PCHF) were utilized for model building, and the Stacking Classifier was trained to be deployed in the front end. The integration of Flask framework with user authentication ensures an effective and secure platform for user testing, enhancing the accessibility and usability of our machine learning-based heart disease prediction system.

ii) System Architecture:

In this study, we have access heart failure dataset from the repository Kaggle. The dataset contains 1025 patient records relate to heart failure and healthy patients. The data preprocessing techniques are applied to format the dataset. The exploratory heart failure data analysis is applied to understand better the data patterns and variables contributing to heart failure. In feature engineering, high-importance features are selected using the proposed PCHF technique. Then the dataset is split into two portions, train and test. The nine advanced machine-learning techniques are applied to the dataset portions. The hyperparameter-based fine tuning is applied to the machine learning models. The outperformed proposed model aims to forecast heart failure with high efficiency.

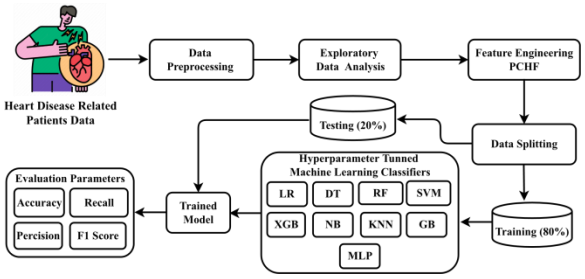

Fig 1 Proposed architecture

iii) Dataset collection:

The heart disease dataset [39] used in this project contains comprehensive clinical and patient data, including demographics, medical history, and physiological measurements, which is employed to train and test machine learning algorithms for accurate heart disease prediction.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Fig 2 Heart disease dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST
ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model [1, 2]. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms:

LR: This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables [23].

```
# Logistic Regression model
from sklearn.linear_model import LogisticRegression
#from sklearn.pipeline import Pipeline

# instantiate the model
log = LogisticRegression(penalty='l2', fit_intercept=True, random_state = 1, max_iter =100)

log.fit(X_train,y_train)

y_pred = log.predict(X_test)

lr_acc = accuracy_score(y_pred, y_test)
lr_prec = precision_score(y_pred, y_test)
lr_rec = recall_score(y_pred, y_test)
lr_f1 = f1_score(y_pred, y_test)

storeResults('Logistic Regression',lr_acc,lr_prec,lr_rec,lr_f1)
```

Fig 3 Logistic regression

DT: A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

```
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
tree = DecisionTreeClassifier(criterion='gini',max_depth=300,min_samples_split=2,max_features=None,random_state=0,max_leaf_nodes

tree.fit(X_train, y_train)

y_pred = tree.predict(X_test)

dt_acc = accuracy_score(y_pred, y_test)
dt_prec = precision_score(y_pred, y_test)
dt_rec = recall_score(y_pred, y_test)
dt_f1 = f1_score(y_pred, y_test)

storeResults('Decision Tree',dt_acc,dt_prec,dt_rec,dt_f1)
```

Fig 4 Decision tree

RF: Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST
ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

and flexibility have fueled its adoption, as it handles both classification and regression problems [11] .

```
from sklearn.ensemble import RandomForestClassifier

# instantiate the model
rf = RandomForestClassifier(n_estimators = 300, criterion = 'gini', max_depth=300,  max_features='sqrt',
                            bootstrap = True, random_state = 0, max_samples = None)

rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)

rf_acc = accuracy_score(y_pred, y_test)
rf_prec = precision_score(y_pred, y_test)
rf_rec = recall_score(y_pred, y_test)
rf_f1 = f1_score(y_pred, y_test)

storeResults('Random Forest',rf_acc,rf_prec,rf_rec,rf_f1)
```

Fig 5 Random forest

SVM: SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems.

```
from sklearn.svm import SVC

# instantiate the model
svm = SVC(C=1.0,kernel = 'rbf', degree = 3, gamma = 'scale', probability=True, tol = 0.001,cache_size=200,max_iter=-1,random_sta

# fit the model
svm.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = svm.predict(X_test)

svc_acc = accuracy_score(y_pred, y_test)
svc_prec = precision_score(y_pred, y_test)
svc_rec = recall_score(y_pred, y_test)
svc_f1 = f1_score(y_pred, y_test)

storeResults('Support Vector Machine',svc_acc,svc_prec,svc_rec,svc_f1)
```

Fig 6 SVM

KNN: The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

```
from sklearn.neighbors import KNeighborsClassifier
#from sklearn.pipeline import Pipeline

# instantiate the model
knn = KNeighborsClassifier(n_neighbors=3, weights='uniform', algorithm='auto',leaf_size=30, p=2, metric='minkowski')

# fit the model
knn.fit(X_train,y_train)

y_pred = knn.predict(X_test)

knn_acc = accuracy_score(y_pred, y_test)
knn_prec = precision_score(y_pred, y_test)
knn_rec = recall_score(y_pred, y_test)
knn_f1 = f1_score(y_pred, y_test)

storeResults('KNN',knn_acc,knn_prec,knn_rec,knn_f1)
```

Fig 7 KNN

MLP: A multilayer perceptron (MLP) is a misnomer for a modern feedforward artificial neural network, consisting of fully connected neurons with a nonlinear kind of activation function, organized in at least three layers, notable for being able to distinguish data that is not linearly separable. It is a misnomer because the original perceptron used a Heaviside step function, instead of a nonlinear kind of activation function (used by modern networks).

```
from sklearn.neural_network import MLPClassifier

model = MLPClassifier(hidden_layer_sizes = (5,2),activation='relu',solver = 'lbfgs', alpha = 0.0001,
                      learning_rate = 'constant',random_state=1, max_iter=300, shuffle = True)
# fit the model
model.fit(X_train,y_train)

y_pred = model.predict(X_test)

mlp_acc = accuracy_score(y_pred, y_test)
mlp_prec = precision_score(y_pred, y_test)
mlp_rec = recall_score(y_pred, y_test)
mlp_f1 = f1_score(y_pred, y_test)

storeResults('MLP',mlp_acc,mlp_prec,mlp_rec,mlp_f1)
```

Fig 8 MLP

NB: Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps

ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

```
# Naive Bayes Classifier Model
from sklearn.naive_bayes import GaussianNB

# instantiate the model
nb=  GaussianNB(var_smoothing=1e-09)

# fit the model
nb.fit(X_train,y_train)

y_pred = nb.predict(X_test)

nb_acc = accuracy_score(y_pred, y_test)
nb_prec = precision_score(y_pred, y_test)
nb_rec = recall_score(y_pred, y_test)
nb_f1 = f1_score(y_pred, y_test)

storeResults('Naive Bayes',nb_acc,nb_prec,nb_rec,nb_f1)
```

Fig 9 Naïve bayes

XGBoost: XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction.

```
from xgboost import XGBClassifier

# instantiate the model
xgb = XGBClassifier(loss='log_loss', learning_rate=0.1, n_estimators = 100, min_samples_split = 2,
                    min_samples_leaf =1, max_depth = 3, use_label_encoder = False, eval_metric = 'mlogloss')

# fit the model
xgb.fit(X_train,y_train)

y_pred = xgb.predict(X_test)

xgb_acc = accuracy_score(y_pred, y_test)
xgb_prec = precision_score(y_pred, y_test)
xgb_rec = recall_score(y_pred, y_test)
xgb_f1 = f1_score(y_pred, y_test)

storeResults('XGBoost',xgb_acc,xgb_prec,xgb_rec,xgb_f1)
```

Fig 10 XGBoost

Gradient Boosting: Gradient Boosting is a popular boosting algorithm in machine learning used for classification and regression tasks. Boosting is one kind of ensemble Learning method which trains the model sequentially and each new model tries to correct the previous model. It combines several weak learners into strong learners [20].

```
from sklearn.ensemble import GradientBoostingClassifier

# instantiate the model
gbc = GradientBoostingClassifier(learning_rate = 1.0, n_estimators = 20, subsample = 1.0,
                                 criterion = 'friedman_mse', max_depth = 2, random_state = 1)

# fit the model
gbc.fit(X_train,y_train)

y_pred = gbc.predict(X_test)

gb_acc = accuracy_score(y_pred, y_test)
gb_prec = precision_score(y_pred, y_test)
gb_rec = recall_score(y_pred, y_test)
gb_f1 = f1_score(y_pred, y_test)

storeResults('Gradient Boosting',gb_acc,gb_prec,gb_rec,gb_f1)
```

Fig 11 Gradient boosting

Stacking Classifier: A stacking classifier is an ensemble learning method that combines multiple classification models to create one “super” model. This can often lead to improved performance, since the combined model can learn from the strengths of each individual model.

```
import lightgbm as lgb
from sklearn.ensemble import StackingClassifier

estimators = [('rf', rf),('dt', tree)]

clf = StackingClassifier(estimators=estimators, final_estimator=lgb.LGBMClassifier(max_depth=-1, random_state=314, silent=True,

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

stac_acc = accuracy_score(y_pred, y_test)
stac_prec = precision_score(y_pred, y_test)
stac_rec = recall_score(y_pred, y_test)
stac_f1 = f1_score(y_pred, y_test)

storeResults('Stacking Classifier',stac_acc,stac_prec,stac_rec,stac_f1)
```

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

Fig 12 Stacking classifier

5. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

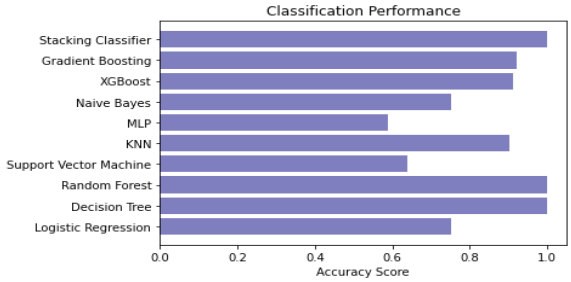

Fig 12 Accuracy graph

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

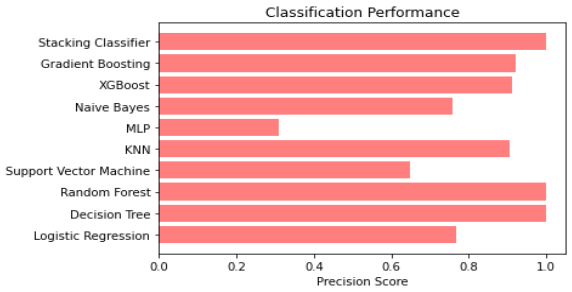

Fig 13 Precision graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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

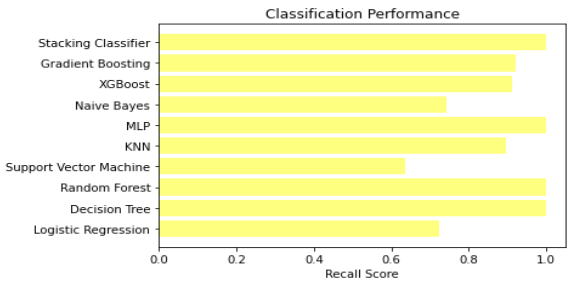

Fig 14 Recall graph

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST
ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

F1-Score:F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

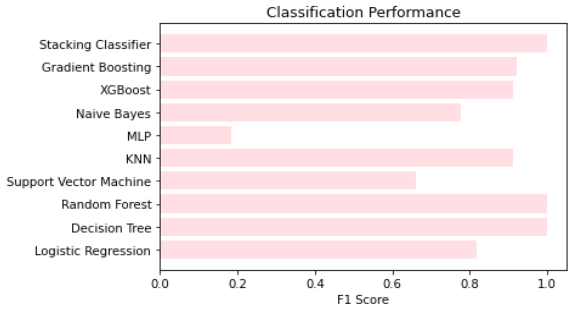

Fig 15 F1 Score graph

ML Model	Accuracy	F1-score	Recall	Precision
Logistic Regression	0.75	0.816	0.724	0.767
Decision Tree	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000
SVM	0.639	0.660	0.636	0.648
KNN	0.902	0.193	0.895	0.904
MLP	0.590	0.184	1.000	0.311
Naive Bayes	0.751	0.777	0.741	0.758
XG Boosting	0.912	0.913	0.913	0.913
Gradient Boosting	0.922	9.22	0.922	0.922
Staking Classifier	1.000	1.000	1.000	1.000

Fig 16 Performance Evaluation

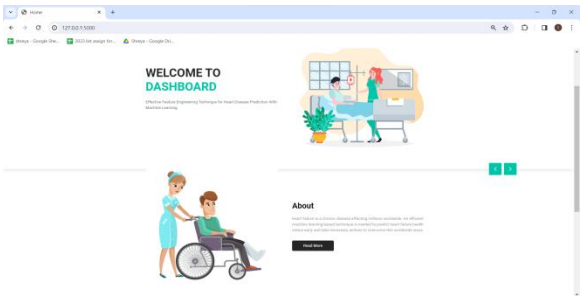

Fig 17 Home page

Fig 18 Signup page

Fig 19 Signin page

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

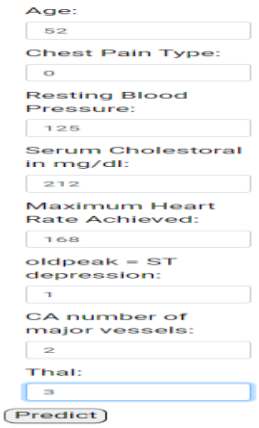

Fig 20 Upload input values to predict result

Result: You have no Heart Disease, based on the input provide!

Fig 21 Predict result as you have no heart disease, based on the input provide

6. CONCLUSION

Predicting heart failure using machine learning methods is proposed in this study [22]. The dataset based on 1025 patient records is used to build the applied models. A novel PCHF feature engineering technique is proposed, which selects the eight most prominent features to enhance performance. The logistic regression, random forest, support vector machine, decision tree, extreme gradient boosting, naive base, k-nearest neighbors, multilayer perceptron, and gradient boosting are the applied machine learning techniques in comparison. The proposed DT method achieved 100% accuracy with 0.005 runtime computations. The cross-validation technique based on 10-fold data is applied to each learning model to validate the performance. Our proposed method outperformed the state-of-the-art studies and is generalized for detecting heart failure.

7. FUTURE SCOPE

The results achieved with our proposed methods can establish a performance standard for heart disease prediction, serving as a reference point for future research in this domain. Subsequent studies could concentrate on refining the feature management process to boost the effectiveness of classification models. Moreover, our methodology holds the potential for application in diverse medical domains to enhance the prediction and identification of various diseases using machine learning algorithms [1, 2, 4, 10].

REFERENCES

[1] M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams, and G. Poglajen, ‘‘Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers,’’ in Proc. Int. Conf. Intell. Environments (IE), Aug. 2017, pp. 14–19.

[2] G. Savarese and L. H. Lund, ‘‘Global public health burden of heart failure,’’ Cardiac Failure Rev., vol. 3, no. 1, p. 7, 2017.

[3] E. J. Benjamin et al., ‘‘Heart disease and stroke statistics—2019 update: A report from the American heart association,’’ Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[4] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, ‘‘Secure and robust machine learning for healthcare:

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

A survey,'' IEEE Rev. Biomed. Eng., vol. 14, pp. 156–180, 2021.

[5] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, ''Effectively predicting the presence of coronary heart disease using machine learning classifiers,'' Sensors, vol. 22, no. 19, p. 7227, Sep. 2022.

[6] R. Katarya and S. K. Meena, ''Machine learning techniques for heart disease prediction: A comparative study and analysis,'' Health Technol., vol. 11, no. 1, pp. 87–97, Jan. 2021.

[7] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, ''A decision support system for heart disease prediction based upon machine learning,'' J. Reliable Intell. Environments, vol. 7, no. 3, pp. 263–275, Sep. 2021.

[8] N. S. Mansur Huang, Z. Ibrahim, and N. Mat Diah, ''Machine learning techniques for early heart failure prediction,'' Malaysian J. Comput. (MJoC), vol. 6, no. 2, pp. 872–884, 2021.

[9] T. Amarbayasgalan, V. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, ''An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets,'' IEEE Access, vol. 9, pp. 135210–135223, 2021.

[10] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, ''Prediction of heart disease using a combination of machine learning and deep learning,'' Comput. Intell. Neurosci., vol. 2021, pp. 1–11, Jul. 2021.

[11] F. S. Alotaibi, ''Implementation of machine learning model to predict heart failure disease,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 1–8, 2019.

[12] D. K. Plati, E. E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, R. Pharithi, J. Gallagher, L. K. Michalis, Y. Goletsis, K. K. Naka, and D. I. Fotiadis, ''A machine learning approach for chronic heart failure diagnosis,'' Diagnostics, vol. 11, no. 10, p. 1863, Oct. 2021.

[13] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, ''A method for improving prediction of human heart disease using machine learning algorithms,'' Mobile Inf. Syst., vol. 2022, pp. 1–9, Mar. 2022.

[14] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, ''Heart disease prediction using core machine learning techniques—A comparative study,'' in Advances in Data and Information Sciences. Springer, 2022, pp. 247–260.

[15] C. Trevisan, G. Sergi, and S. Maggi, ''Gender differences in brain-heart connection,'' Brain and Heart Dynamics. 2020, pp. 937–951.

[16] M. S. Oh and M. H. Jeong, ''Sex differences in cardiovascular disease risk factors among Korean adults,'' Korean J. Med., vol. 95, no. 4, pp. 266–275, Aug. 2020.

[17] D. C. Yadav and S. Pal, ''Prediction of heart disease using feature selection and random forest

ISSN 2229-6107 www.ijpast.in
Vol 14,Issue 2,May 2024

ensemble method,'' Int. J. Pharmaceutical Res., vol. 12, no. 4, pp. 56–66, 2020.

[18] D. Tomar and S. Agarwal, ''A survey on data mining approaches for healthcare,'' Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241–266, Oct. 2013.

[19] S. Ekiz and P. Erdogmus, ''Comparative study of heart disease classification,'' in Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT), Apr. 2017, pp. 1–4.

[20] B. A. Tama, S. Im, and S. Lee, ''Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble,'' BioMed Res. Int., vol. 2020, pp. 1–10, Apr. 2020.

[21] V. Ramalingam, A. Dandapath, and M. K. Raja, ''Heart disease prediction using machine learning techniques: A survey,'' Int. J. Eng. Technol., vol. 7, no. 2, pp. 684–687, 2018.

[22] Heart Disease Dataset|Kaggle, DAVID LAPP, Atlanta, Georgia, 1988.

[23] K. Shah, H. Patel, D. Sanghvi, and M. Shah, ''A comparative analysis of logistic regression, random forest and KNN models for the text classification,'' Augmented Human Res., vol. 5, no. 1, pp. 1–16, Dec. 2020.

[24] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, ''Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,'' Comput. Biol. Med., vol. 136, Sep. 2021, Art. no. 104672.

[25] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, ''Predicting employee attrition using machine learning approaches,'' Appl. Sci., vol. 12, no. 13, p. 6424, Jun. 2022.

[26] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, ''Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction,'' PLoS ONE, vol. 17, no. 11, Nov. 2022, Art. no. e0276525.

[27] S. Shabani, S. Samadianfard, M. T. Sattari, A. Mosavi, S. Shamshirband, T. Kmet, and A. R. Várkonyi-Kóczy, ''Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis,'' Atmosphere, vol. 11, no. 1, p. 66, Jan. 2020.

[28] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, ''HDPM: An effective heart disease prediction model for a clinical decision support system,'' IEEE Access, vol. 8, pp. 133034–133050, 2020.

[29] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, ''Improving heart disease prediction using feature selection approaches,'' in Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2019, pp. 619–623.

[30] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, ''Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM),'' Diagnostics, vol. 11, no. 9, p. 1714, Sep. 2021.

ISSN 2229-6107 www.ijpast.in
Vol 14,Issuse 2,May 2024

[31] R. Pahuja and A. Kumar, ‘‘Sound-spectrogram based automatic bird species recognition using MLP classifier,’’ Appl. Acoust., vol. 180, Sep. 2021, Art. no. 108077.

[32] U. Azmat, Y. Y. Ghadi, T. A. Shloul, S. A. Alsuhibany, A. Jalal, and J. Park, ‘‘Smartphone sensor-based human locomotion surveillance system using multilayer perceptron,’’ Appl. Sci., vol. 12, no. 5, p. 2550, Feb. 2022.

[33] J. Isabona, A. L. Imoize, and Y. Kim, ‘‘Machine learning-based boosted regression ensemble combined with hyperparameter tuning for optimal adaptive learning,’’ Sensors, vol. 22, no. 10, p. 3776, May 2022.

[34] A. A. Ali, H. S. Hassan, and E. M. Anwar, ‘‘Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique,’’ in Proc. 12th Int. Conf. Electr. Eng. (ICEENG), Jul. 2020, pp. 145–150.

[35] O. E. Taylor, P. S. Ezekiel, and F. B. D. Okuchaba, ‘‘A model to detect heart disease using machine learning algorithm,’’ Int. J. Comput. Sci. Eng., vol. 7, no. 11, pp. 1–5, Nov. 2019.

[36] D. K. Chohan and D. C. Dobhal, ‘‘A comparison based study of supervised machine learning algorithms for prediction of heart disease,’’ in Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions (CISES), May 2022, pp. 372–375.

[37] K. Sanjar, O. Bekhzod, J. Kim, A. Paul, and J. Kim, ‘‘Missing data imputation for geolocation-based price prediction using KNN–MCF method,’’ ISPRS Int. J. Geo-Inf., vol. 9, no. 4, p. 227, Apr. 2020.

[38] B. Olimov, B. Subramanian, R. A. A. Ugli, J.-S. Kim, and J. Kim, ‘‘Consecutive multiscale feature learning-based image classification model,’’ Sci. Rep., vol. 13, no. 1, p. 3595, Mar. 2023.

[39] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, ‘‘Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model,’’ Concurrency Comput., Pract. Exp., vol. 33, no. 22, p. e6143, Nov. 2021.