Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 1 (2025)

A Multilayer Perceptron Technique Combined with Correlation-Based Feature Selection is Used in the Heart Disease Prediction System

[1]Chikka krishnappa T. K. , [2]Mohammed Tajuddin,

[1]Research scholar, DSCE, Affiliated to VTU, India,

[2]Professor, DSCE, Dept. of CSE (Cyber Security),

Abstract: Cardiac disease prediction helps physicians to make accurate recommendations on the treatment of the patients. The use of machine learning (ML) is one of the solution for recognising heart disease-related symptoms. The goal of this study is to suggest a methodology for identifying the most relevant features of cardiac disease characteristics by applying a feature selection technique. The data set used in this study was Framingham heart disease dataset (FHS). It was collected from KAGGLE Machine Learning repository. There are 16 attributes and a mark in the dataset that has been validated by four ML classifiers. There are two feature selection methods, Correlation Based Feature selection (CBFS) and Principle Component Analysis (PCA) was used for the comparison in the study. By using CBFS Method five highly correlated features are selected for the study, and by using PCA thirteen features are selected. The experimental result shows that Correlation Based Feature Selection with Multilayer perceptron (CBFS with MLP) obtained the highest accuracy for this dataset

Keywords: LBP, Hybrid PSO optimization, Treatment etc

I. INTRODUCTION

The research concentrates on the two feature selection methods for data reduction before building the predictive models by classification algorithms. These reduced features are then passed into the classification algorithms to design the models for the heart disease prediction. These models are used for the comparison of accuracy of the classifier. Principle Component Analysis and Correlation Based feature selection methods are used for finding out the reduced features. The selected features are inputted to four different classifiers such as Navie Bayes, ADABOOST, MLP and SMO. The accuracy of each model is compared with the other.

Devansh Shah studied various attributes related to heart disease[1]. The study was conducted with Na¨ıve Bayes, decision tree, K-nearest neighbor, and random forest algorithms[1]. The experimental result proves that K-nearest neighbor algorithm exhibits the highest accuracy. Hamidreza Ashrafi Esfahani[2] formulated a model to predict cardiovascular disease. The model includes decision trees, Neural Networks, Rough set, Na¨ıve Bayes and SVM for implementation. On comparing the results achieved, it was revealed that the hybrid model of Rough Set, Na¨ıve Bayes and Neural Network obtained the highest accuracy.

II. PROPOSED SYSTEM

Framingham Heart Study (FHS) from Kaggle Machine Learning repository is used for the study. Two feature selection methods along with four classification algorithms are used for the study. CBFS and PCA are the methods used for the dimensionality reduction. MLP, Navie Bayes, Sequential Minimum Optimiser (SMO) and ADABOOST algorithms are used for classification. The reduced feature set from both the feature selection methods are inputted to different classifiers. Eight different Machine Learning Models were created for Heart Disease Prediction. Accuracy of these Models are compared with each other.

Description of the Data Set: The Framingham Heart Study (FHS) dataset was collected from Kaggle. The dataset consists of 4241 records. It contain sixteen features, as shown in figure 1,including AGE, PREVALENT HYP, SYSBP, DIABP, GLUCOSE, SEX, EDUCATION, CURRENT SMOKER, CIGSPERDAY, BPMEDS, PREVALENT STROKE, BMI, HEART RATE, DIABETES, TOTCHOL and PREDICTOR VARIABLE.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure1: Dataset

Classification Algorithms: In Machine Learning various forms of classification techniques are available. Classification techniques used for this study was described below.

III. METHODOLOGY

The methodology generates artifcial instances of underrepresented categories, achieving a balanced distribution across both types. The process of classification involves the use of chosen characteristics and using various classifiers such as support vector machine (SVM)21, PCA22, linear discriminant analysis (LDA)23, naïve Bayes (NB)24, decision tree (DT)25, and random forest (RF)26. Ultimately, the classifier predicts an individual's presence or absence of heart disease. The mechanism used in the suggested approach for heart disease forecasting is shown in Fig.2.

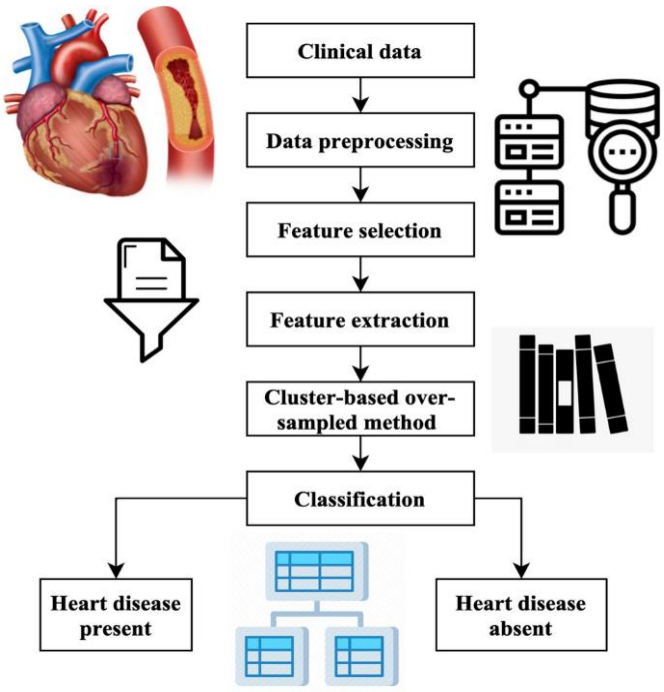


Figure2: Architecture of Heart disease detection process

Multilayer Perceptron (MLP): MLP is a subset of Artificial Neural Network. MLP comprises one or more than one hidden layers aside from one input and one output plate. The Perceptron is made of an input layer and a totally

linked output layer. MLPs have the same levels of input and output, but could have several levels concealed within them.

Adaboost In machine learning, AdaBoost(Adaptive Boosting) is a supervised learning algorithm. It is used for combining several weak classifiers together to generate a strong classifier.

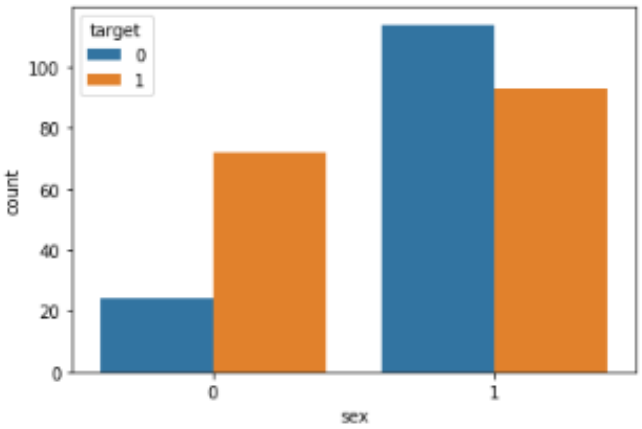
Naive Bayes Naive Bayes is a Machine Learning algorithm based on Probability theory in statistics. The term naive suggests that the elements that go through the software are autonomous of each other, That is, the value of one characteristic, does not explicitly influence or alter the value of any of the other characteristics used in the algorithm. The Bayes theorem tells us how we can compute the conditional probability.

IV. RESULTS AND DISCUSSIONS

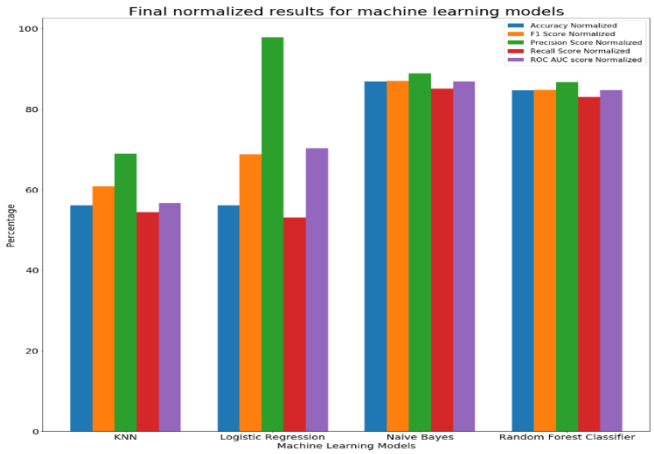
The performance of the suggested strategy is evaluated via thorough testing in the simulation part. Multiple metrics, including accuracy, sensitivity, specificity, and precision, are calculated to assess the predictive efficacy of the methodology. The effectiveness of the suggested strategy is validated by a comparative study conducted against current models.

Sequential Minimal Optimization (SMO): The sequential minimal optimization is more effective to solve the SVM problem compared to traditional Quadratic Programming algorithms such as the interior-point method. The SMO algorithm can be viewed as a method of decomposition by which a problem of optimization of multiple variables is decomposed into a set of sub problems, each optimizing an objective feature of a limited number of variables, usually only one, whereas all other variables are treated as constants which remain unchanged in the sub problem.

Feature Selection: During feature selection the most relevant features are extracted from the data set. Redundancy can be avoided using this method. Since irrelevant features are excluded from the input data, feature selection can increase the accuracy of prediction. In this study Correlation Based Feature Selection (CBFS) and Principle Component Analysis (PCA) is used for feature selection. After feature selection the reduced data set is applied to four different classification Algorithm.



Gender Plot

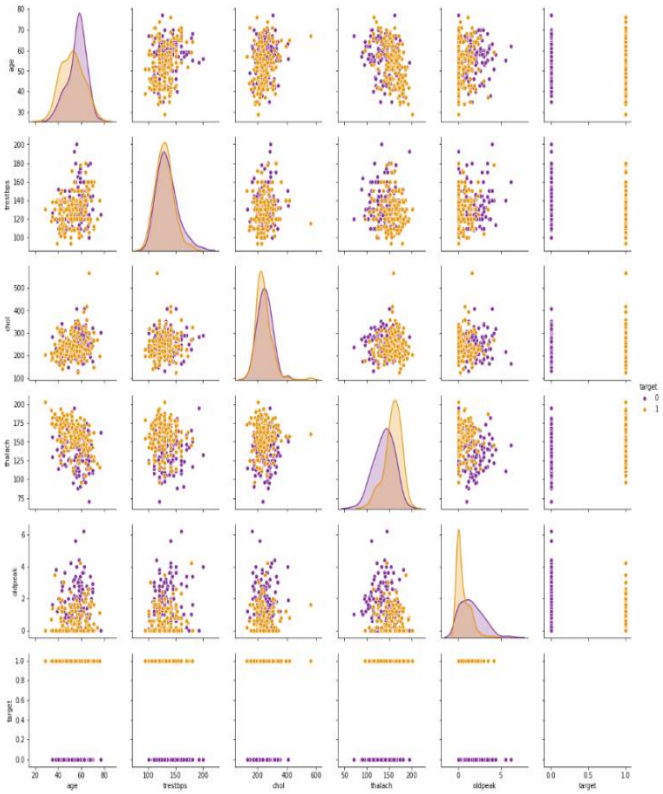


Principal Component Analysis (PCA) Thirteen features were selected by PCA during feature selection. The features selected by the PCA algorithm are AGE, PREVALENTHYP, SYSBP, DIABP, DIABETES, SEX, BPMEDS, TOTCHOL, PREVALENTSTROKE, CIGSPERDAY, EDUCATION, BMI, CURRENT SMOKER.

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	0.0	0.0	1.0	2.0	3.0
trestbps	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	0.0	0.0	1.0	1.0	2.0
thalach	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
exang	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slope	303.0	1.399340	0.616226	0.0	1.0	1.0	2.0	2.0
ca	303.0	0.729373	1.022606	0.0	0.0	0.0	1.0	4.0
thal	303.0	2.313531	0.612277	0.0	2.0	2.0	3.0	3.0
target	303.0	0.544554	0.498835	0.0	0.0	1.0	1.0	1.0

V. CONCLUSION

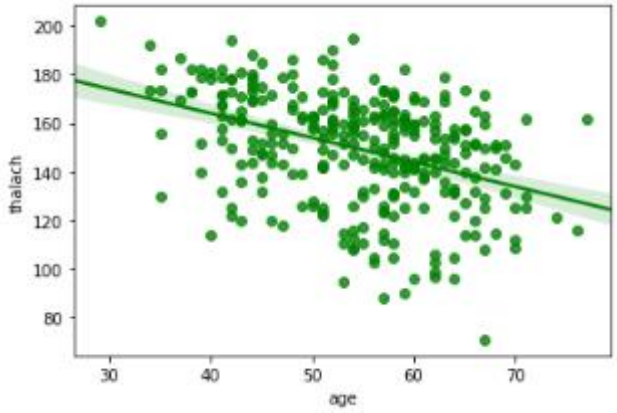
During the study the performance of two different feature selection methods CBFS and PCA are evaluated. Eight different classifier models are developed by combining the feature selection and classification algorithms. The performance of each model was evaluated. Performance measures such as Accuracy, Precision, Recall, F Measure and ROC are evaluated for finding out the best classifier. From the result it is proven that the model CBFS with MLP Classifier shows the maximum performance for FHS dataset.



Scatter plot

Heatmap

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```



Thalach v/s Age

REFERENCES

[1] Devansh Shah, Samir Patel and Santosh Kumar Bharti 2020 Heart Disease Prediction using Machine Learning Techniques Springer.

[2] Hamidreza and Ashrafi Esfahani 2017 Cardiovascular disease detection using a new ensemble classifier IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI).

[3] V.V.Ramalingam 2018 Prediction of Heart Diseases Using Machine Learning International Journal of Engineering and Technology.

[4] Amin Ul Haq 2019 A Hybrid Intelligence System Frame Work for Prediction of Heart Disease Using ML, Mobile Information Systems.

[5] Shadman Nashif Heart Disease Detection by Using Machine Learning Algorithm and a Real time Cardiovascular Health Monitoring System World journal of Engineering and Technology,vol.06 no.04,2018,article id 88650.

[6] Poornima Singh 2018 Effective heart disease prediction system using data mining techniques International journal of Nano medicine.

[7] M. Gunay and T. Ensarı 2019 Predictive churn analysis with machine learning methods IEEE 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1- 4.

[8] R. Suguna, M. Shyamala Devi and Rincy Merlin Mathew 2019 Customer Churn Predictive Analysis by Component Minimization using Machine Learning International Journal of Innovative Technology and Exploring Engineering.

[9] G. Battista, C. Sassi, M. Zompatori, D. Palmarini, and R.Canini, "Ground-glass opacity: Interpretation of high resolution CT findings,"La Radiolo-gia Medica , vol. 106, pp. 425–442, 2003.

[10] Z. G. Yang, S. Song, and S. Talcashima, "High- resolution CT analysis of small Tumor region adenocarcinoma revealed on screening helical CT," Amer. J.Roentgenol. , vol. 176, no. 6, pp. 1399–1407, 2001.

[11] T. Aoki, Y. Tomoda, H. Watanabe, H. Nakata, T. Kasai,H. Hashimoto, M. Kodate, T. Osaki, and K. Yasumoto, "Peripheral Tumor region adenocarcinoma: Correlation of thin-section findings with histologic factors and survival," Radiology, vol. 220, pp. 803–809, 2001.

[12] J. J. T. Owen, D. E. McLoughlin, R. K. Suniara, and E.J. Jenkinson, "The role of mesenchyme in thymus development," Current Topics Microbiol. Immunol. , vol. 251, pp. 133–137, 2000.

[13] M. R. Melamed, B. J. Flehinger, M. B. Zaman, R. T. Heelan, W. A. Perchick, and N. Martini, "Screening for Tumor region cancer: Results of the memo-rial sloan- kttering study in New York", Chest, vol. 86, no. 1, pp. 44– 53, 1984.

[14] C. V. Zwirewich, S. Vedal, R. R. Miller, and N. L. M¨uller, "Solitary pulmonary nodule: High-resolution CT andradiologic-pathologic corre-lation," Radiology, vol. 179, no.2, pp, 469–476, 1991.

[15] S. F. Huang, R. F. Chang, D. R. Chen, and W. K. Moon,"Characterization of spiculation on ultrasound lesions," IEEE Trans. Med. Imag. , vol. 23, no. 1, pp. 111–121, Jan. 2004.

[16] M. Noguchi and Y. Shimosato, "The development and progression of adenocarcinoma of the Tumor region," Cancer Treatment Res., vol. 72, pp. 131– 142, 1995.

[17] T. V. Colby and C. Lombard. "Histiocytosis X in the Tumor region," Human Pathol. , vol. 14, no. 10, pp. 847– 856, 1983.

[18] V. J. Lowe, J. W. Fletcher, L. Gobar, M. Lawson, P. Kirchner, P. Valk, J. Karis, K. Hubner, D. Delbeke, E. V. Heiberg, E. F. Patz, and R. E. Coleman, "Prospective investigation of positron emission tomography in Tumorregion nodules," J. Clin. Oncol. , vol. 16, no. 3, pp. 1075– 1084, 1998.

[19] K. S. Lee, Y. Kim, and S. L. Primack, "Imaging of pulmonary lymphomas," Amer. J. Roentgenol., vol. 168, no.2, pp. 339–345, 1997.

[20] J. W. Gurney, "Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis: Part 1. Theory," Radiology, vol. 186, no. 2, pp. 405–413, 1993.

[21] J. J. Erasmus, H. I. McAdama, and J. H. Connolly, "Solitary pulmonary nodules: Part II. Evaluation of the indeterminate nodule," Radiographics , vol. 20, no. 1, pp. 59–66, 2000.

[22] L. Song, X. Liu, L. Ma, C. Zhou, X. Zhao, and Y. Zhao,"Using HOG-LBP features and MMP learning to recognize imaging signs of lesions," in Proc. Comput.-Based Med. Syst. , 2012, pp. 1–4.

[23] X. Ye, X. Lin, G. Beddoe, and J. Dehmeshki. "Efficientcomputer-aided detection of ground-glass opacity nodules inthoracic CT images," in Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2007, pp. 4449–4452.

[24] T. W. Way, B. Sahiner, H. P. Chan, L. Hadjiiski, P. N. Cascade, A. Chughtai, N. Bogot, and E. Kazerooni, "Computer-aided diagnosis of pul-monary nodules on CT scans: Improvement of classification performance with nodule surface features," Med. Phys., vol. 36, no. 7, pp. 3086–3098, 2009.

[25] H. Chen, Y. Xu, Y. Ma, and B. Ma, “Neural network ensemble-based computer-aided diagnosis for differentiationof Tumor region nodules on CT im-ages clinical evaluation,”Acad. Radiol. , vol. 17, no. 5, pp. 595–602, 2010.

[26] H. U. Kauczor, K. Heitmann, C. P. Heussel, D. Marwede, T. Uthmann, and M. Thelen, “Automatic detection and quantification of ground-glass opacities on high-resolution CT using multiple neural networks: Comparison with a density mask,”Amer. J. Roentgenol., vol. 175,no. 5, pp. 1329–1334, Nov. 2000.

[27] K. G. Kim, J. M. Goo, J. H. Kim, H. J. Lee, B. G. Min,K. T. Bae, and J. G. Im, “Computer-aided diagnosis oflocalized ground-glass opacity in the Tumor region at CT: initial experience,” Radiology, vol. 237, no. 2, pp. 657–661,2005.