International Journal for Modern Trends in Science and Technology
Volume 10, Issue 04, pages 275-282.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol10issue04.html
DOI: https://doi.org/10.46501/IJMTST1004040

Earlier Heart Diseases Detection and Identification using KNN Models

Dr D J Samatha Naidu | G.Gayathri

Department of MCA, Annamacharya PG College of Computer Studies, Rajampet Andhra Pradesh, India.

To Cite this Article
Dr D J Samatha Naidu and G.Gayathri, Earlier Heart Diseases Detection and Identification using KNN Models, International Journal for Modern Trends in Science and Technology, 2024, 10(04), pages. 275-282. https://doi.org/10.46501/IJMTST1004040

Article Info
Received: 04 April 2024; Accepted: 24 April 2024; Published: 25 April 2024.

ABSTRACT

Heart diseases (HD)is the critical health issue and numerous people have been suffered by this disease around the world. The HD occurs with common symptoms of breath shortness. In existing works researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Machine learning predictive models include ANN, LR, K-NN, SVM, DT, and NB are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, mRMR, LASSO and Local learning-based-features-selection (LLBFS) have been used to select the features. We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection.

Keywords: Heart disease classification, features selection, disease diagnosis, intelligent system, medical data analytics.

1. INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis and proper treatment can save the lives of many people. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a result, the ratio of death decreases. The Cleveland heart disease data set was used by various researchers for the identification problem of HD. The machine learning predictive models need

proper data for training and testing. The performance of machine learning model can be increased if balanced dataset is use for training and testing of the model.

In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator(LASSO),Relief, Minimal-Redundancy-Maximal-Relevance (MRMR), Local-learning-based-features-selection (LLBFS), Principle component Analysis (PCA), Greedy Algorithm (GA), and optimization methods, such as Anty Conley Optimization (ACO), fruit fly optimization (FFO), Bacterial Foraging Opti- mization (BFO) etc. Similarly Yun *et al.* Feature selection has great influence in numerous applications such as building simpler, increasing learning performance, creating clean and understandable data. The feature selection from big data is challenging job and create big problems because big data has many dimensions. Further, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues. For big data analytics challenges .

Feature selection is very important to resolved in designed unsupervised hashing scheme, called topic hyper graph hashing, to report the limitations. Topic hypergraph hashing effectively mitigates the semantic shortage of hashing codes by exploiting auxiliary texts around images. The proposed Topic hyper graph hashing can achieve superior performance equaled with numerous state-of-the- art approaches, and it is more appropriate for mobile image retrieval. The feature selection algorithms are classified into three type such as filter based, wrapper based and embedded based. All these feature selection mechanisms have some advantages and limitations in certain cases. The filter based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measure the usefulness of a subset of features by actually training the classifier on it. The filter method is less computationally complex than wrapper method. To evaluate all classifiers on data and find that they get, on average, 50% of the cases right. Furthermore, appropriate cross validation techniques and performance evaluation metrics are critical necessary for a model when model is train and test on dataset. We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include ANN, LR, K-NN, SVM, DT, and NB are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, mRMR, LASSO and Local-learning-based-features-selection (LLBFS) have been used to select the features. We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection. Leave-one-subject-out cross-validation (LOSO) technique has been applied to select the best hyper-parameters for best model selection Apart from this, different performance . The art existing methods in the literature, such as NB,Three phase ANN (Artificial neural Network) diagnosis system, Neural network ensembles (NNE), ANN-Fuzzy-AHP diagnosis system (AFP) [20], Adaptive-weighted-Fuzzy-system-ensemble (AWFSE) [21].

2. LITERATURE REVIEW

[1] Q S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6S,pp. 1009–1015, 2019.

Cardiovascular diseases are challenging to predict and diagnose due to the underlying dysfunctions associated with reflex mechanisms. Considering the mortality ratio and economy burden by the cardiovascular disorder, various researchers seek to diagnose this pernicious disease at its earliest by analysing the healthcare data. In recent times, researchers made seminal contributions however, the unavailability of an extensive and

fundamental article motivated us to prepare a literature review on a cardiovascular disease. We conduct a comprehensive database search between the years 2000 and 2017 using different keyword combinations to get distinguished articles about the disease. We provide descriptive insights to fill the uncovered research gaps. This paper attempts to uncover the state-of-the-art data mining approaches and tools that can be used to diagnose the cardiovascular disease at its initial. To our knowledge until now there is no competent and comprehensive article on cardiovascular disorder prognosis and identification using knowledge mining and machine learning approaches. The topic is diverse as well progressive hence demands additional research to understand newly identified discoveries about the disease.

[2] Q S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," *Int. Arab J. Inf. Technol.*, vol. 15, no. 2, pp. 224–231, 2018.

Software components are imperative parts of a system which play a fundamental role in the overall function of a system. A component is said to be secure if it has a towering scope of security. Security is a shield for unauthorized use as unauthorized users may informally access and modify components within a system. Such accessing and modifications ultimately affect the functionality and efficiency of a system. With an increase in software development activities security of software components is becoming an important issue. In this study, a fuzzy logic based model is presented to handle ISO/IEC 18028-2 security attributes for component security evaluation. For this purpose an eight input, single output model based on the Mamdani fuzzy inference system has been proposed. This component security evaluation model helps software engineers during component selection in conditions of uncertainty and ambiguity

[3] Q S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*. [Online]. Available: http://arxiv.org/abs/1811.12808

The correct use of model evaluation, model selection, and algorithm selection techniques is vital in academic machine learning research as well as in many industrial settings. This article

reviews different techniques that can be used for each of these three subtasks and discusses the main advantages and disadvantages of each technique with references to theoretical and empirical studies. Further, recommendations are given to encourage best yet feasible practices in research and applications of machine learning. Common methods such as the holdout method for model evaluation and selection are covered, which are not recommended when working with small datasets. Different flavours of the bootstrap technique are introduced for estimating the uncertainty of performance estimates, as an alternative to confidence intervals via normal approximation if bootstrapping is computationally feasible. Common cross-validation techniques such as leave-one out-cross-validation and k-fold cross-validation are reviewed, the bias-variance trade-off for choosing k is discussed, and practical tips for the optimal choice of k are given based on empirical evidence. Different statistical tests for algorithm comparisons are presented, and strategies for dealing with multiple comparisons such as omnibus tests and multiple-comparison corrections are discussed. Finally, alternative methods for algorithm selection, such as the combined F-test 5x2 cross validation and nested cross-validation, are recommended for comparing machine learning algorithms when datasets are small.

[4] Q A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT)*, Mar. 2019, pp. 1–4.

Detection of Heart Disease (HD) by using models of machine learning (ML) is very effective in early stages. The HD treatment and recovery is effective if detected the disease at initial stages. HD identification by machine learning (ML) techniques has been developed to assist the physicians. In this study we proposed an Identification system by using ML models to classify the HD and healthy subjects. Sequential backward selection of feature algorithm was used to select more appropriate features to increase the classification accuracy and reduced the computational time of predictive system.

Cleveland heart disease dataset was for evaluation of the system. The dataset 70% used for training and remaining for validation. The proposed system performances have been measured by using evaluation metrics. The experimental results shows that Sequential Backward Selection (SBS) algorithms choose appropriate features and these features increase the accuracy using K-Nearest Neighbour supervised machine learning classifier. The good accuracy of this study suggests that the proposed model will effectively identify the HD and healthy subjects

[5] Q S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," ***IEEE Access*****, vol. 7,pp. 81542–81554, 2019.**

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

3. PROPOSED METHOD

A. DATA SET

Clevel and Heart Disease [29] dataset is considered for test- ing purpose in this study. During the designing of this data set there were 303 instances and 75 attributes, however all published experiments refer to using a subset of 14 of them. In this work, we performed pre-processing on the data set, and 6 samples have been eliminated due to missing values. The remaining samples of 297 and 13 features dataset is left and with 1 output label. The output label has two classes to describe the absence of HD and the presence of HD. Hence features matrix 297*13 of extracted features is formed. The dataset matrix information's are given in Table 2

B. PRE-PROCESSING OF DATA SET

The pre-processing of dataset required for good representation. Techniques of pre-processing such as removing attribute missing values, Standard Scalar (SS), Min-Max Scalar have been applied to the dataset.

C. STANDARAD STATE OF THE FEATURES SELECTION ALOGORITHMS

After data pre-processing, the selection of feature is required for the process. In general, FS is a significant step in constructing a classification model. It works by reducing the number of input features in a classifier, to have good

TABLE 1. Summary of the previous methods

Ref	Technique	Limitations	Advantages	Acc(%)
[11]	HD diagnosis using ML classifiers	The Proposed method accuracy is very low.	Computationally less complex.	77
[22]	MLP+SVM	Computationally complex.	The performance of the proposed method is high in terms of prediction accuracy.	80.41
[23]	ANN+Fuzzy Logic	More execution time required to generate results.	Accuracy is high.	87.4
[19]	ANN ensemble based diagnosis system	Computationally complex.	High accuracy.	89.01
[17]	HD diagnosis system based on NB, DT and ANN	The NB and DT performance are low.	ANN achieved high performance in term of accuracy	88.12
[18]	Three phase technique based on ANN	High computation time.	High accuracy.	88.89
[20]	ANN-FUZZY-AHP	Computationally complex.	Achieved high accuracy.	91.1
[25]	Relief-Rough set based method for HD detection	Computation time is high.	High accuracy due to selection of appropriate feature for training and testing of the model.	92.32
[27]	Hybrid ML method	Low accuracy.	Low computation time.	88.07

predictive and short computationally complex models [30]. We have been used four standard state of the art FS algorithms and one our proposed FS algorithm in this study.

1)RELIEF

Relief Algorithm assigns weights to each data set features and updated weights automatically. The features having

high weight values should be selected and low weight will be discarded. Relief and K-NN algorithm process to determine the weights of features are the same [32].

2)MINIMAL-REDUNDANCY-MAXIMAL-RELEVANCE

MRMR algorithm chooses features that are suitable for the prediction and selected features that are non redundant. It does not take care of the combination of features [32]. The MRMR pseudo code is given in algorithm 2 [34].

3)LEAST-ABSOLUTE-SHRINKAGE-SELECTION-OPERATOR ALGORITHMS

LASSO choose feature based on modifying the absolute coefficient value of the features. Then these features coefficient values set to zero and finally zero coefficient features are eliminated from the features set. In the selected features set

Algorithm 1 Pseudo-Code for Relief FS Algorithm

Input: S: Training data (feature vectors with class labels), Parameter m: number of random training samples out of total samples used to W.

Output: W: weights for each feature

1: $n \leftarrow$ total number of training samples

2: $d \leftarrow$ number of features (dimensions)

3: $W[A] \leftarrow 0.0$;Feature weights set

4: **for K** $\leftarrow$ 1 to m do

5: Randomly choose a 'Target' sample R_k

6: Find a nearest hit H and nearest miss M

7: for A$\leftarrow$1 to a do

8: W[A] $\leftarrow$ W[A] - $diff(A, R_k, H)/m + diff(A, R_k, M)/m$

9: **end for**

10: **end for**

11: **Return** W ;weight vector of features that calculate the quality of feature

those features to include who coefficient have a high value. Sometime LASSO selects irrelevant features and includes in the subset of feature [35].

4) LOCAL LEARNING BASED FEATURES SELECTION ALGORITHMS

LLBFS assigns weights to features and reduced the complexity of non-linear problems into linear. Features having large

TABLE 2. Cleveland heart disease dataset 2016

S.no	Feature Name	Feature Code	Description
1	Age	AGE	Age in years
2	sex	SEX	Male=1,Female=0
3	chest pain	CPT	Atypical angina=1 Typical angina=2 Asymptomatic=3 Non-anginal pain=4
4	resting blood pressure	RBP	mm hg, hospitalized
5	serum cholesterol	SCH	In mg/dl
6	$fastingbloodsugar > 120mg/dl$	FBS	$fastingbloodsugar > 120mg/dl$ (T =1) (F=0)
7	resting electrocardiographic	RES	Normal=0 ST T=1 Hypertrophy=2
8	maximum heart rate	MHR	—
9	exercise induced angina	EIA	yes=1 no=0
10	old peak=ST depression induced by exercise relative to rest	OPK	—
11	The slope of the Peak Exercise ST Segment	PES	Up Sloping=1 Flat=2 Down Sloping=3
12	number of major vessels (0–3) Colored by fluoroscopy	VCA	
13	thallium scan	THA	Normal=3 Fixed defect=6 Reversible defect=7
14	label	LB	Heart disease patient=1 Healthy=0

Algorithm 2 Pseudo code for MRMR

Input: CF : Set of initial candidate features, $num\ R$: number of reduced features wanted.

Output: SF: Selected features

1: for each feature fi ∈ CF do

2: relevance ← mutual Info (fI, class)

3: redundancy ← 0

4: for each feature fj ∈ CF do

5: redundancy ← redundancy + mutual Info (fi, fj)

6: end for

7: mrmr Values[fi] ← relevance − redundancy

8: end for F

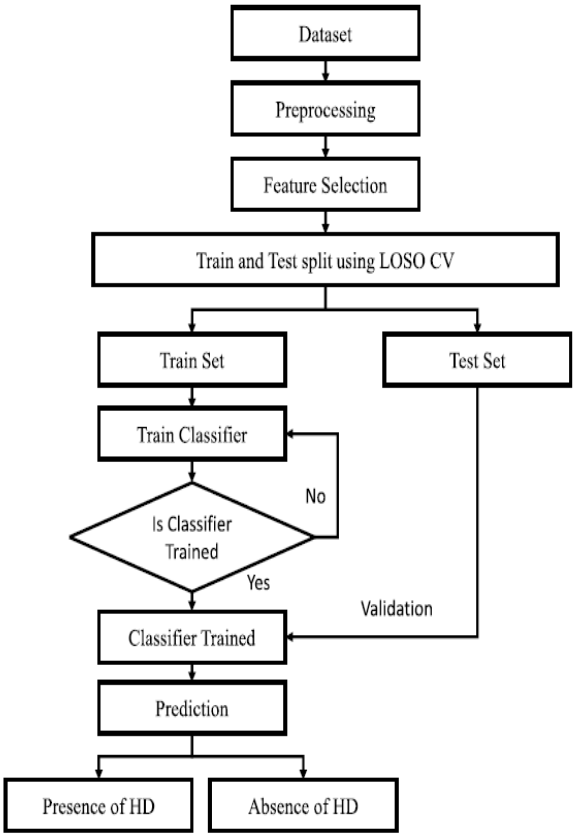
9: SF ← sort (mrmr Values). take(num R)

10: return SF

F The set of selected feature weighted values are selected and features weights are small discarded from a subset of features [36].

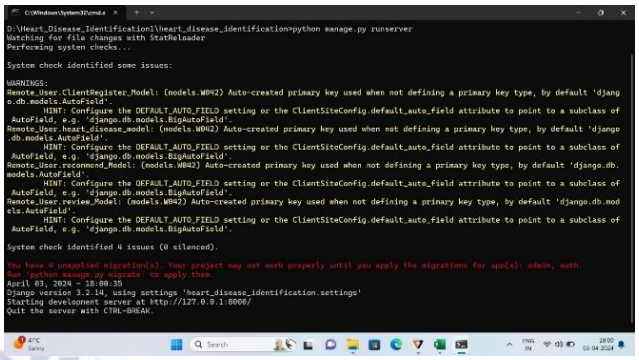
4. RESEARCH METHODOLOGY

SYSYTEM ARCHITECTURE

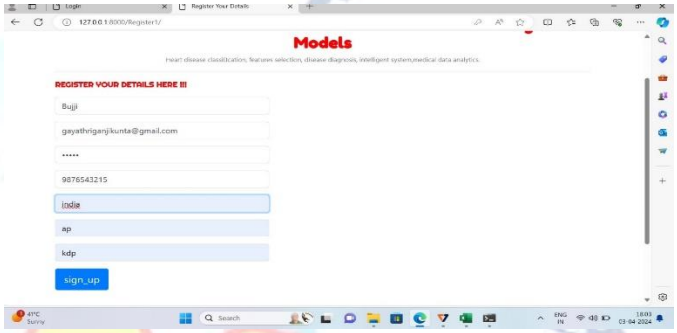


RESULT AND ANALYSIS

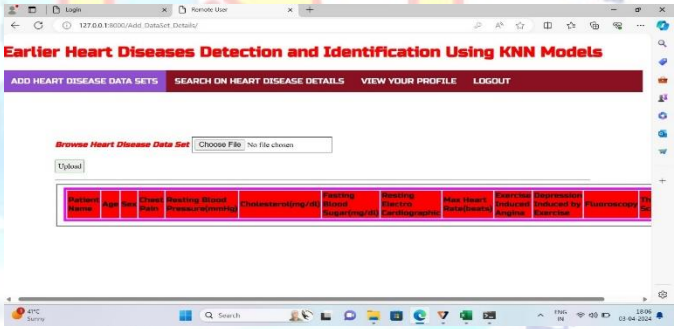
Supervised classification experiments have been conducted in order to evaluate the classification performance of classi- fiers. In the first phase, standard features selection algorithms are applied such as Relief, MRMR, LASSO and LLBFS for selection of appropriate features. Then in the second phase of experiments, the proposed FS algorithm was used for features selection. Then the classifiers performance were evaluated on selected features. Furthermore, LOSO method is applied with each classifier. To test the performances of the classifiers, various performance evaluation metrics are com- puted. All the experiments have been performed in a python environment using different machine learning libraries on an Intel(R) C i7-2400 CPU @3.10 GHz system.



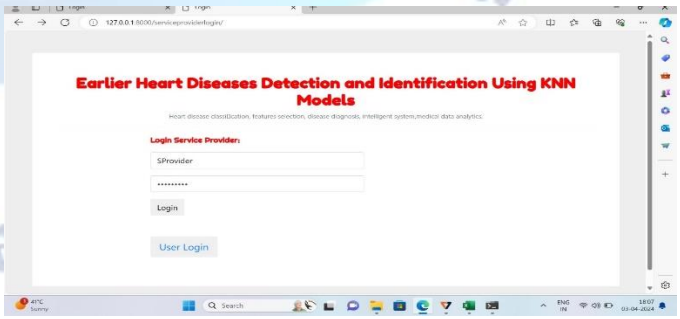
Screen1: CMD Running Process



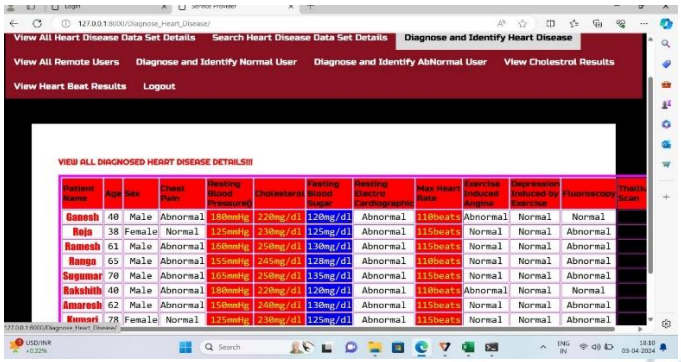
Screen2: Enter user details



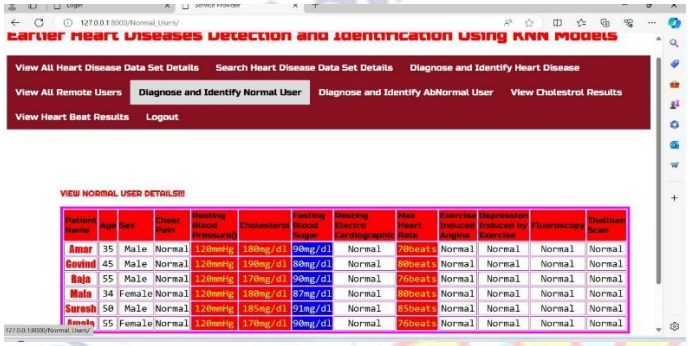
Screen3: Add Heart disease Data Set



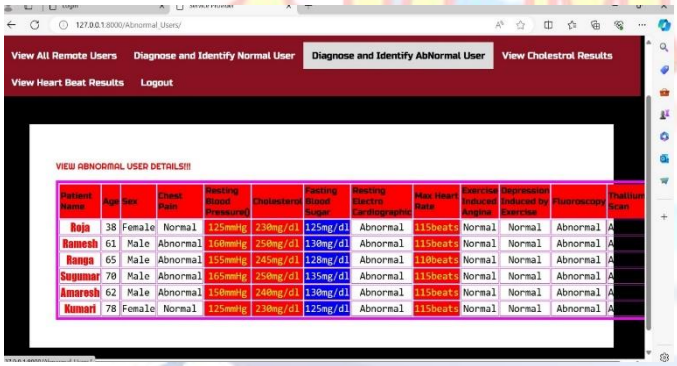
Sreen4: login Service Provider



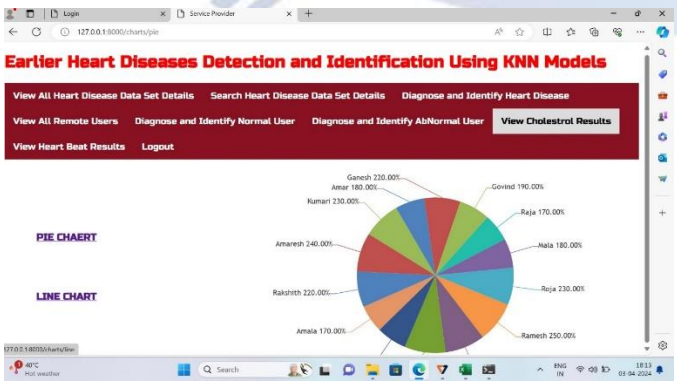
Screen5: Diagnose and identify heart disease



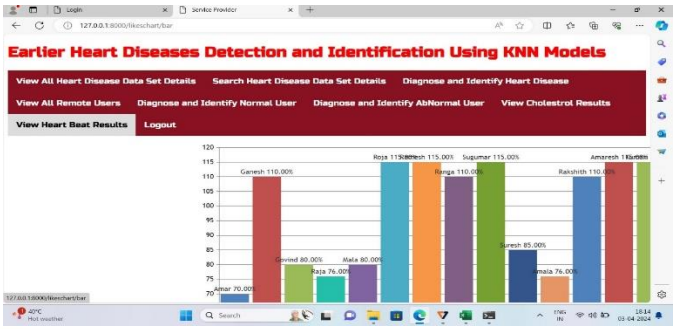
Screen6: Diagnose and identify Normal user



Screen7: Diagnose and Identify Abnormal User



Screen8: View Cholestrol Results



Screen9: View Heart Beat Result

5. CONCLUSION

In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include LR, K-NN, ANN, SVM, NB, and DT are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyperparameters selection. The system is tested on Cleveland heart disease dataset. Further more, performance evaluation metrics are used to check the performance of the identification system. According to Table 15 the specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. Therefore for ANN with relief is the best predictive system for detection of healthy people. The sensitivity of classifier NB on selected features set by LASSO FS algorithm also gives the best result as compared to the sensitivity values of Relief FS algorithm with classifier SVM (linear). The classifier Logistic Regression MCC is 91% on selected features selected by FCMIM FS algorithm. The processing time of Logistic Regression with Relief, LASSO, FCMIM and LLBFS FS algorithm best as compared to MRMR FS algorithms, and others classifiers. Thus the experimental results show that the proposed features selection algorithm select features that are more effective and obtains high classification accuracy than the standard feature selection algorithms. According to feature selection algorithms, the most important and suitable features are Thallium Scan type chest pain and Exercise-induced Angina. All FS algorithms results show that the feature Fasting blood sugar (FBS) is not a suitable heart disease

diagnosis. The accuracy of SVM with the proposed feature selection algorithm (FCMIM) is 92.37% which is very good as compared previously proposed methods.

6.ACKNOWLEDGEMENT

We thankful to all the referred journal authors for their elaborative study helps me to write this paper.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ''Epidemiology and risk profile of heart failure,'' Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[2] M. Durairaj and N. Ramasamy, ''A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,'' Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock,R. M. Arnold, N. R. Cook, G. M. Felker,

[4] G. S. Francis, P. J. Hauptman,

[5] E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ''Decision making in advanced heart failure: A scientific statement from the American heart association,'' Circulation, vol. 125, no. 15,pp. 1928–1952, 2012.

[6] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ''Innovative artificial neural networks-based decision support system for heart diseases diagno- sis,'' J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.

[7] Q. K. Al-Shayea, ''Artificial neural networks in medical diagnosis,'' Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[8] J. Lopez-Sendon, ''The heart failure epidemic,'' Medicographia, vol. 33, no. 4, pp. 363–369, 2011.

[9] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup,M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera,D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson, and Y. J. Woo, ''Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association,'' Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[10] S. I. Ansarullah and P. Kumar, ''A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,'' Int. J. Recent Technol. Eng., vol. 7, no. 6S,pp. 1009–1015, 2019.