IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE IDENTIFICATION METHOD BY USING MACHINE LEARNING CLASSIFICATION

V. NARESH KUMAR
Associate Professor
Department of CSE
CMR TECHNICAL CAMPUS, HYDERABAD

K. RAKSHITHA
B.Tech student
Department of CSE
CMR TECHNICAL CAMPUS, HYDERABAD

V. BHARGAVI
B.Tech student
Department of CSE
CMR TECHNICAL CAMPUS,HYDERABAD

B.Tech student
Department of CSE
CMR TECHNICAL CAMPUS, HYDERABAD

G. GOPI KRISHNA
B.Tech student
Department of CSE
CMR TECHNICAL CAMPUS,HYDERABAD

***ABSTRACT:** Heart disease is one of the complex diseases and globally many people suffer from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. In this article, we proposed an efficient and accurate system to diagnose heart disease and the system is based on machine learning techniques. The system is developed based on classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbour, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed a novel fast conditional mutual information feature selection algorithm to solve the feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. In this Classification the accuracy obtained by comparing the algorithms KNN, SVM, ANN is*

74.19 percent.

***Key words:** Heart disease classification, features selection, K-nearest neighbour(KNN), Support Vector machine(SVM), Artificial neural networks(ANN)*

I. 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 The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam. and proper treatment can save the lives of many people [5]. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually [6]. Most of the people in the United States are suffering from heart disease [7]. 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. Moreover, it is expensive and computationally difficult to analyse .

Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a 107562 This work is licensed under a Creative Commons Attribution 4.0 License. For more information,see-https://creativecommons.org/licenses/by/4.0/ VOLUME 8, 2020 J. P. Li et al.: HD Identification Method Using ML

Classification in E-Healthcare result, the ratio of death decreases [9] and [10]. The Cleveland heart disease data set was used by various researchers [11] and [12] 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. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. In literature various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD. In order to improve the predictive capability of machine learning model data pre-processing is important for data standardization. Various Pre-processing 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 (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 Optimization (BFO) etc. Similarly Yun et al. [13] presented different techniques for different type of feature selection, such as feature selection for high-dimensional small sample size data, large-scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi-view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Jundong et al. [14] discussed the challenges of feature selection (FS) for big data. It is necessary to decrease the dimensionality of data for various learning tasks due to the curse of dimensionality. 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 of feature selection is very important to resolved.

II. LITERATURE SURVEY

In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnosis HD. This research study present some existing machine learning based diagnosis techniques in order to explain the important of the proposed work. Detrano et al. [11] developed HD classification system by using machine learning classification techniques and the performance of the system was 77% in terms of accuracy. Cleveland dataset was utilized with the method of global evolutionary and with features selection method. In another study Gudadhe et al. [17] developed a diagnosis system using multi-layer Perceptron and support vector machine (SVM) algorithms for HD classification and achieved accuracy 80.41%. Humar et al. designed HD classification system by utilizing a neural network with the integration of Fuzzy logic. The classification system achieved 87.4% accuracy. Resul et al. developed an ANN ensemble based diagnosis system for HD along with statistical measuring system enterprise miner (5.2) and obtained the accuracy of 89.01%, sensitivity 80.09%, and specificity 95.91%. Akil et al. designed a ML based HD diagnosis system. ANN-DBP algorithm along with FS algorithm and performance was good. Palaniappan et al. proposed an expert medical diagnosis system for HD identification. In development of the system the predictive model of machine learning, such as navies bays (NB), Decision Tree (DT), and Artificial Neural Network were used. The 86.12% accuracy was achieved by NB, ANN accuracy 88.12% and DT classifier achieved 80.4% accuracy. Olaniyi et al. developed a three-phase technique based on the artificial neural network technique for HD prediction in angina and achieved 88.89% accuracy. Samuel et al. developed an integrated medical decision support system based on artificial neural network and Fuzzy AHP for diagnosis of HD. The performance of the proposed method in terms of accuracy was

91.10%. Liu et al. proposed a HD classification system using relief and rough set techniques.

The proposed method achieved 92.32% classification accuracy. In [15] proposed a HD identification method using feature selection and classification algorithms. Sequential Backward Selection Algorithm (SBS FS) for Features Selection. The classifier K-Nearest Neighbour (K-NN) performance has been checked on full and on selected features set. The proposed method obtained high accuracy. In another study MOHAN et al. designed a HD prediction method by using hybrid machine learning techniques. He also proposed a new method for significant feature selection from the data for effective training and testing of machine learning classifier.

They have been recorded 88.07% classification accuracy. Geweid et al. designed HD identification techniques by using improved SVM based duality optimization technique. In the above literature the proposed

HD diagnosis methods limitation and advantages have been summarized in for better understanding the important of our proposed approach. All these existing techniques used numerous methods to identify the HD at early stages. However, all these techniques have lack of prediction accuracy and high computation time for prediction of HD. According to the prediction accuracy of HD detection method need further improvement for efficient and accurate detection at early stages for better treatment and recovery. Thus, the major issues in these previous approaches are low accuracy and high computation time and these might be due the use of irrelevant features in dataset. In order to tackle these problems new methods are needed to detect HD correctly. The improvement in prediction accuracy is a big challenge and research gap.

More than 5.8 million adults in the USA are living with HF.[1] This syndrome affects more men than women, and its prevalence greatly increases with advancing age.

Studies estimate the overall prevalence of HF in the population to be about 2–3%. From self-reported data obtained by the National Health and Nutrition Examination Survey, the prevalence in the USA was 2.6% in

2006.[1] Studies with validated diagnoses of HF include cohort studies, such as the Rochester Epidemiologic Project in Olmsted County, MN, where the prevalence of HF was found to be 2.2%.

Here, prevalence increased with age, reaching

8.4% in those aged ≥75 years compared with

0.7% in those 45 to 54 years of age. The Rotterdam cohort showed similar trends, with a HF prevalence of 1% in those aged 55 to 64 years, compared with over 10% in those aged

≥85 years.

The worldwide prevalence of HF seems to have been increasing over the past decades.This trajectory may reflect growing awareness and diagnosis of HF, an aging population, increasing incidence of HF, improvement in the treatment and management of cardiovascular disease, or a combination of some or all of these potential explanations. Curtis and colleagues concluded from a cohort study of more than 600,000 US

Medicare beneficiaries that the prevalence of

HF increased from 90 to 121 per 1,000 between 1994 and 2003, although the rate of increase has slowed in the past few years, possibly reflecting stabilized incidence and mortality. By contrast, prevalence as measured by HF admission rates declined in Canada between 1994 and 2004; this difference may in part have resulted from higher admission thresholds in Canada during this time period, owing to limited hospital bed availability.

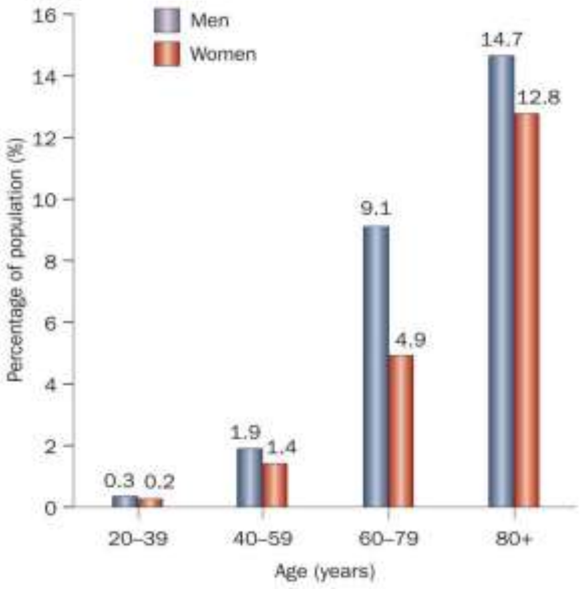


fig1: graph of increase in HD for respective ages

III. PROPOSED SYSTEM

The system has been designed for the identification of heart disease. The performances of various machine learning classifiers for HD identification have been checked on selected features. We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include

ANN, K-NN and SVM are used for the identification of HD. One of 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 assessment metrics have been used for classifiers performances evaluation. The proposed method has been tested on Cleveland HD dataset. Furthermore, the performance of the proposed technique has been compared with state of the art existing methods in the literature, such as Three phase ANN (Artificial neural Network) diagnosis system , Neural network ensembles (NNE).

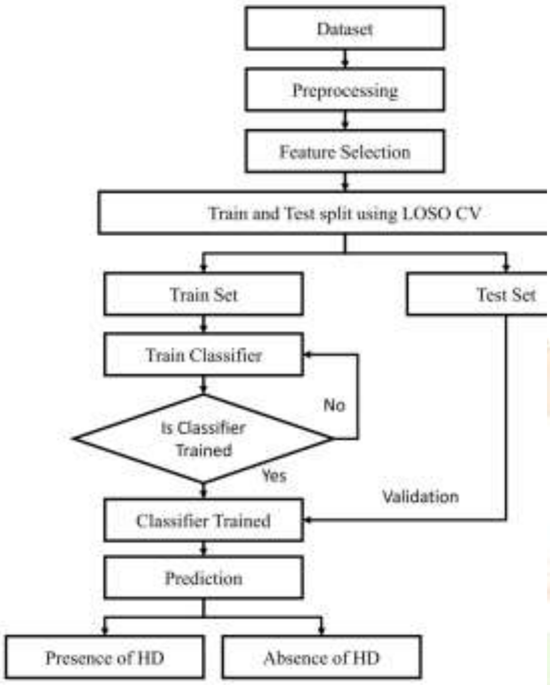


fig.2 : Proposed system architecture

Description: Firstly we need to upload the dataset in the database then we need to import the dataset. Now the data will be preprocessed and a feature selection algorithm LOSO is applied to train and test the dataset. After training and testing the dataset we obtain the shape of both trained and tested data. We need to apply KNN, SVM, ANN algorithms to the trained and tested data , now we obtain the accuracy of the three algorithms and the algorithm with best accuracy is used for

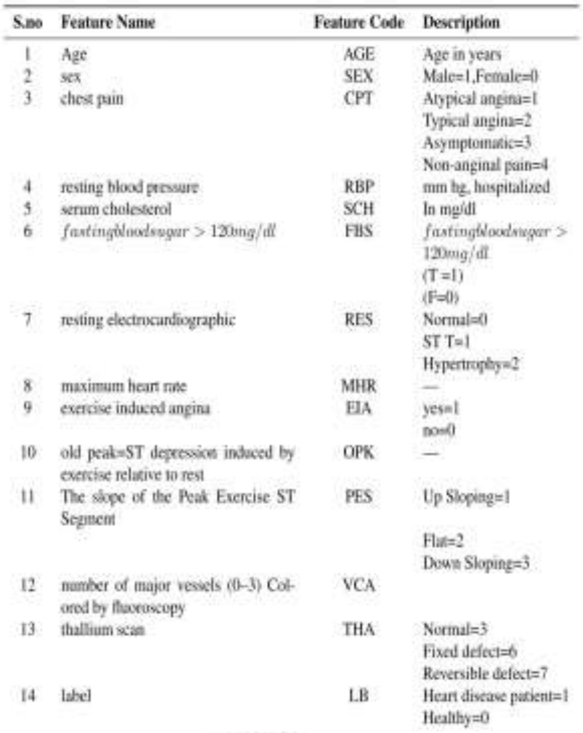
Prediction.

A. Data Set

Cleveland Heart Disease [16] dataset is considered for testing 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 the extracted features are formed. The dataset matrix information is given in Table 1. **B. Pre-processing of the Data set** The pre-processing of the 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. Standard state of art features selection algorithms :

After data pre-processing, the selection of features 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 predictive and short computationally complex models. We have used the LOSO FS algorithm in this study.



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

Table-1 Cleveland Heart disease dataset

V. IMPLEMENTATION

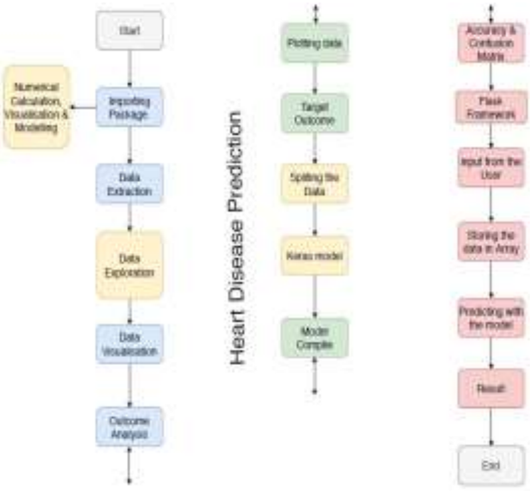


fig 3: Implementation flowchart

IMPLEMENTATION STEPS-

- Gathering the dataset from database
- Pre-processing the dataset and analysis dataset
- Splitting the datasets into training and testing in the ration80 % SVM, ANN,
  KNN, models to analysis the data
- Obtain the accuracy in prediction

Attributes :-

Age, gender, chest pain(CPT), blood pressure, tobacco, adiposity, chd, family history, sbp, alcohol

LEAVE-ONE-SUBJECT-OUT(LOSO) CROSS VALIDATION TECHNIQUE :

In this LOSO validation strategy, one sample is separated as test data and remaining subjects to train the model. The test subject is predicted as HD otherwise, the subject is classified as healthy.

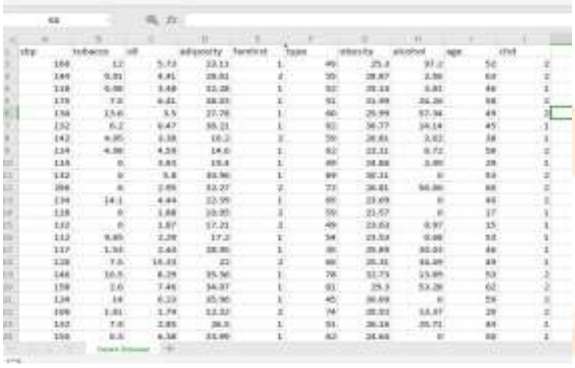


fig 4: attributes identification

PreProcessing the data:

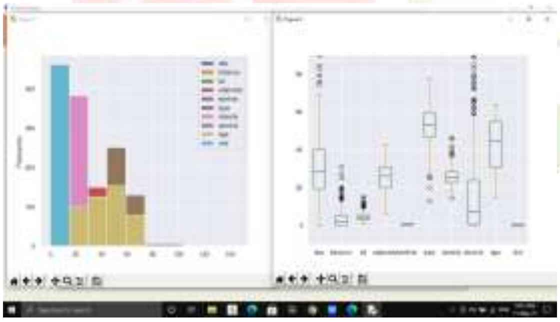


fig 5: Preprocessing the data

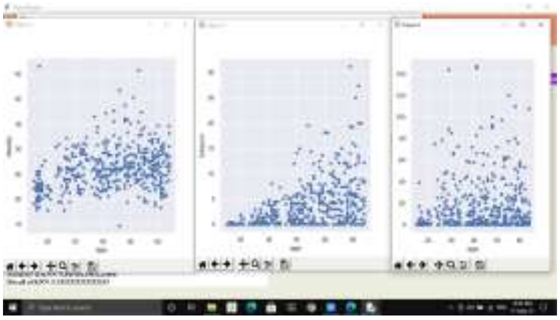


fig 6: Levels of alcohol, obesity and tobacco

After preprocessing the data we need to train and test the models. we get shape of trained and tested data in this step and it is also represented in form of graph Train and Test Data :

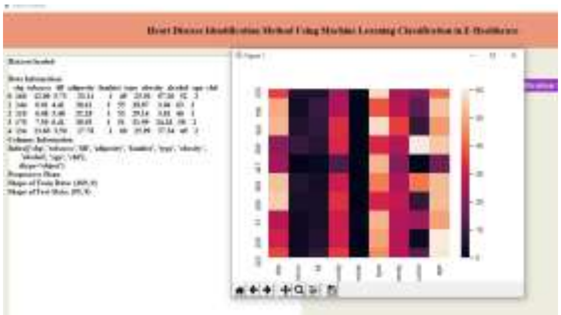

Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare

fig 7: shape of trained and tested data

Now we run the algorithms KNN, ANN and SVM to get the Accuracy, Recall and Precision values of each algorithm. We also represent them in the form of graphs.

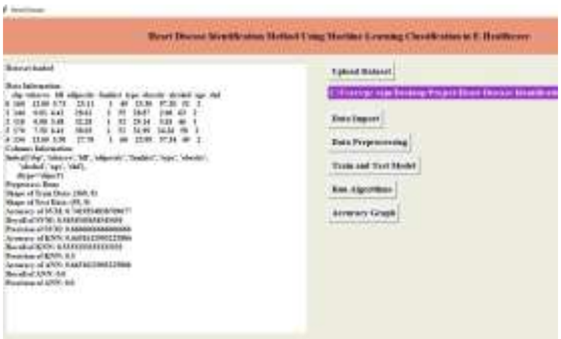


fig 8:accuracy levels of algorithms

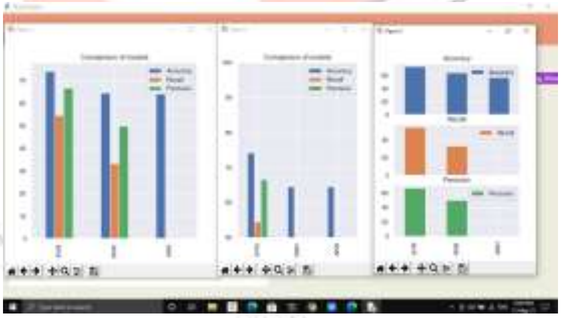


fig 9: comparison of accuracy levels

After comparing the algorithms, the best algorithm with high accuracy is given as output. In this Project we obtained SVM as the best one which shows high accuracy compared to KNN and ANN.

VI. CONCLUSION AND FUTURE SCOPE

In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include K-NN, ANN and SVM 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. Furthermore, performance evaluation metrics are used to check the performance of the identification system. According to 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. Further, the performance of machine learning based method FCMIM SVM is high then Deep neural network for detection of HD. A little improvement in prediction accuracy have great influence in diagnosis of critical diseases. The novelty of the study is developing a diagnosis system for identification of heart disease. In this study, four standard feature selection algorithms along with one proposed feature selection algorithm is used for features selection. LOSO CV method and performance measuring metrics are used. The Cleveland heart disease dataset is used for testing purpose. As we think that developing a decision support system through machine learning algorithms it will be more suitable for the diagnosis of heart disease. Furthermore, we know that irrelevant features also degrade the performance of the diagnosis system and increased computation time. Thus another innovative touch of our study to used features selection algorithms to selects the appropriate features that improve the classification accuracy as well as reduce the processing time of the diagnosis system. In the future, we will use other features selection algorithms, optimization methods to further increase the performance of a predictive system for HD diagnosis. The controlling and treatment of disease is significance after diagnosis, therefore,i will work on treatment and recovery of diseases in future also for critical disease such as heart, breast, Parkinson, diabetes.

VII. ACKNOWLEDGEMENT

Apart from the efforts of us, the success of paper depends largely on encouragement and guidelines of many others. We take this opportunity to express my profound gratitude to CMR Technical Campus College Management for motivating us and for providing all the facilities required for this work. We are deeply indebted to Chairman
Shri C.Gopal Reddy , Secretary Smt. C.Vasantha Latha, Director Dr.A.Raji Reddy,
HOD CSE Dr.K.Srujan Raju, PRC Co-ordinator Dr.Suvarna Gothane, Project Guide V.Naresh Kumar who always has been a constant source of inspiration for me.

VIII. 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, G. S. Francis, P. J.
Hauptman, 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.

[4] S. Ghwanmeh, A. Mohammad, and A.
Al-Ibrahim, ‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’ *J. Intell. Learn.*
Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.

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

[6] J. Lopez-Sendon, ‘‘The heart failure epidemic,’’ *Medicographia*, vol. 33, no. 4, pp. 363–369, 2011.

[7] 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.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ''Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity,'' *J. Roy. Soc. Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[9] 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.

[10] 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.

[11] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfifisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ''International application of a new probability algorithm for the diagnosis of coronary artery disease,'' *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, Aug. 1989.

[12] J. H. Gennari, P. Langley, and D. Fisher, ''Models of incremental concept formation,'' *Artif. Intell.*, vol. 40, nos. 1–3, pp. 11–61, Sep. 1989.

[13] Y. Li, T. Li, and H. Liu, ''Recent advances in feature selection and its applications,'' *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 551–577, Dec. 2017.

[14] J. Li and H. Liu, ''Challenges of feature selection for big data analytics,'' *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 9–15, Mar. 2017.

[15] H. Kahramanli and N. Allahverdi, ''Design of a hybrid system for diabetes and heart diseases,'' *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 82–89, Jul. 2008.

[16] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ''A hybrid classification system for heart disease diagnosis based on the RFRS method,'' *Comput. Math. Methods Med.*, vol. 2017, pp. 1–11, Jan. 2017.

[17] M. Gudadhe, K. Wankhade, and S. Dongre, ''Decision support system for heart disease based on support vector machine and artificial neural network,'' in Proc. Int. Conf. Comput. Commun. Technol. (ICCCT), Sep. 2010,pp. 741–745.