JES
Journal of Engineering Sciences

Vol 13, Issue 01, Jan /2022
ISSN NO:0377-9254

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

Name : G. Indu
Mail ID: ganjiindu666@gmail.com
MTech CSE, Vaagdevi College of Engineering, bollikunta, warangal

Guide:
Email ID: rajmatrix2000@gmail.com
Name : Prof P. Rajkumar
Associate Professor of CSE Department,
Vaagdevi College of Engineering, bollikunta, warangal

Abstract:

Heart disease is one of the complex diseases and globally many people suffered 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 diagnosis 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 neighbor, 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 novel fast conditional mutual information feature selection algorithm to solve 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. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyperparameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms. The experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. The suggested diagnosis system (FCMIM-SVM) achieved good accuracy as compared to previously proposed methods. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

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 diagnosisand 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. 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 analyze. Thus, to develop a noninvasive 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 aresult, 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. 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.

II. Related work:

in this paper Larry A. Allen, MD, MHS, Co-Chair; Lynne W. Stevenson, MD, Co-Chair proposed by Providers have an ethical and legal mandate to involve patients in medical decisions. Shared decision making recognizes that there are complex trade-offs in the choice of medical care.1 Shared decision making also addresses the ethical need to fully inform patients about the risks and benefits of treatments.2 In the setting of multiple reasonable options for medical care, shared decision making involves clinicians working with patients to ensure that patients' values, goals, and preferences guide informed decisions that are right for each individual patient. Grounded in the ethical principle of autonomy,3 judicial decisions (eg, Cruzan v Missouri Department of Health4) and legislative actions (eg, the Patient Self-Determination Act5) have repeatedly affirmed the rights of patients or duly appointed surrogates to choose their medical therapy from among reasonable options.

In this paper M. Durairaj* and Nandhakumar Ramasamy** proposed by Medical diagnostics systems are evaluated by employing large information databases, but it endures many failures to extract data from Database. There is no sufficient tool available to discover the major relationship between the data. In such case, the core knowledge of healthcare data is extracted by applying the data mining methods. The extracted knowledge can be used for the perfect diagnosis and further treatment. Infertility is an emotional cause of fertility in all over the world over past years. Treatment for infertility includes set of

procedures like IUI, IVF, ICSI, and GIFT to cure the disease. Predicting the success rate for the infertility treatment can be done by using the pre-processed data from the database. This paper explains the existing Pre-processing method and analyzes the accuracy of prediction rate after pre-processing. It is evident that the accuracy is increased up to 90% after pre-processing the raw data using the existing techniques. Hybridge different techniques together will provide a better result which is taken as the future direction of this work.

In this paper Anh L. Bui, Tamara B. Horwich, and Gregg C. Fonarow proposed by Heart failure (HF) is a major public health issue, with a prevalence of over 5.8 million in the USA, and over 23 million worldwide, and rising. The lifetime risk of developing HF is one in five. Although promising evidence shows that the age-adjusted incidence of HF may have plateaued, HF still carries substantial morbidity and mortality, with 5-year mortality that rival those of many cancers. HF represents a considerable burden to the health-care system, responsible for costs of more than $39 billion annually in the USA alone, and high rates of hospitalizations, readmissions, and outpatient visits. HF is not a single entity, but a clinical syndrome that may have different characteristics depending on age, sex, race or ethnicity, left ventricular ejection fraction (LVEF) status, and HF etiology. Furthermore, pathophysiological differences are observed among patients diagnosed with HF and reduced LVEF compared with HF and preserved LVEF, which are beginning to be better appreciated in epidemiological studies. A number of risk factors, such as ischemic heart disease, hypertension, smoking, obesity, and diabetes, among others, have been identified that both predict the incidence of HF as well as its severity. In this Review, we discuss key features of the epidemiology and risk profile of HF.

III. MATERIALS AND METHOD:

All the research materials and techniques background are discussed in the following subsections.

A. DATA SET:

Cleveland Heart Disease 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 extracted features is formed.

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. STANDARD STATE OF THE ART FEATURES SELECTION ALGORITHMS:

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
[11]	HD diagnosis using ML classifiers	The Proposed method accuracy is very low.	Computationally less complex.
[22]	MLP+SVM	Computationally complex.	The performance of the propose method is high in terms of pre diction accuracy.
[23]	ANN+Fuzzy Logic	More execution time required to generate results.	Accuracy is high.
[19]	ANN ensemble based diagnosis system	Computationally complex.	High accuracy.
[17]	HD diagnosis system based on NB, DT and ANN	The NB and DT performance are low.	ANN achieved high performanc in term of accuracy
[18]	Three phase technique based on ANN	High computation time.	High accuracy.
[20]	ANN-FUZZY-AHP	Computationally complex.	Achieved high accuracy.
[25]	Relief-Rough set based method for HD detection	Computation time is high.	High accuracy due to selectio of appropriate feature for trainin and testing of the model.
[27]	Hybrid ML method	Low accuracy.	Low computation time.

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

IV. ANALYSIS AND DISCUSSIONS OF EXPERIMENTAL RESULTS:

A. EXPERIMENTAL DESIGN SETUP:

Supervised classification experiments have been conducted in order to evaluate the classification performance of classifiers. 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 performances were evaluated on selected features. Furthermore, LOSO CV method is applied with each classifier. To test the performances of the classifiers, various performance evaluation metrics are computed.

B. EXPERIMENTAL RESULTS:

1) RESULTS OF DATA PRE-PROCESSING TECHNIQUES The different statistical operations such as removing attributes missing values, Standard Scalar (SS), Min-Max Scalar, means, standard division have been applied to the dataset. The results of these operations are reported in Table 5. The processed dataset has 297 instances and 13 inputs attribute with one output Label. Data Visualization is the presentation of data in graphical format. It helps people understand the significance of data by summarizing and presenting huge amount of data in a simple and easyto-understand format and helps communicate information

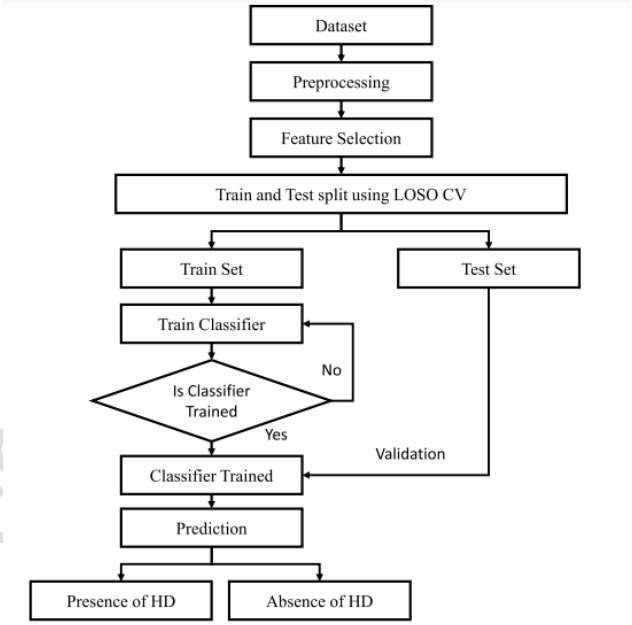

FIGURE 1. Proposed heart disease identification system.

clearly and effectively. Figure 2 is the histogram of the data set represents the frequency of occurrence of specific phenomena which lie within a specific range of values and arranged in consecutive and fixed intervals and Figure 3 describes the co-relation among the features of the dataset using heat map. The heat map, which is a two-dimensional representation of data in which colors represent values. A single heat map provides a quick visual summary of information. More elaborate heat maps allow the viewer to understand complex datasets. Furthermore, Heatmap can be super useful when we want to see which intersections of the categorical values have higher concentration of the data compared to the others.

RESULTS LOSO CV FOR CLASSIFIERS PERFORMANCE ON FULL FEATURES SET:

In this section, on full features set the classifiers performances are measured with the LOSO validation method. Classifiers

TABLE 6. Selected features by Relief, MRMR, LASSO, and LLBFS:

FS Algorithm	Order	Feature	Feature Code	Score
Relief	1	13	THA	0.247
	2	9	EIA	0.227
	3	3	CPT	0.217
	4	11	PES	0.131
MRMR	5	12	VCA	0.128
	6	8	MHR	0.123
	1	3	CPT	0.59
	2	5	SCH	0.575
	3	11	PES	0.574
	4	12	VCA	0.542
	5	2	SEX	0.523
	6	13	THA	0.486
LASSO	1	2	SEX	0.15
	2	12	VCA	0.14
	3	9	EIA	0.13
	4	3	CPT	0.1
	5	11	PES	0.08
	6	13	THA	0.08
LLBFS	1	13	THA	0.596
	2	12	VCA	0.592
	3	3	CPT	0.59
	4	2	SEX	0.579
	5	11	PES	0.574
	6	10	OPK	0.561

TABLE 7. Features selected by FCMIM FS algorithm:

S.no	Feature code	score
1	SEX	0.523
2	CPT	0.217
3	RBP	0.165
4	SCH	0.575
5	RES	0.696
6	MHR	0.123
7	EIA	0.298
8	OPK	0.561
9	PES	0.574
10	THA	0.486

other parameters values also passed during the training process. Table 8 represents the performance evaluation of classifiers with LOSO CV. According to Table 8, the classifier logistic regression has good performance that obtained 84% accuracy, 93% specificity, and 75% sensitivity and MCC was 84%, and processing time was 0.003 seconds at C = 10 as compared with others values of parameter C. The K-NN, different experiments conducted with different values of k.

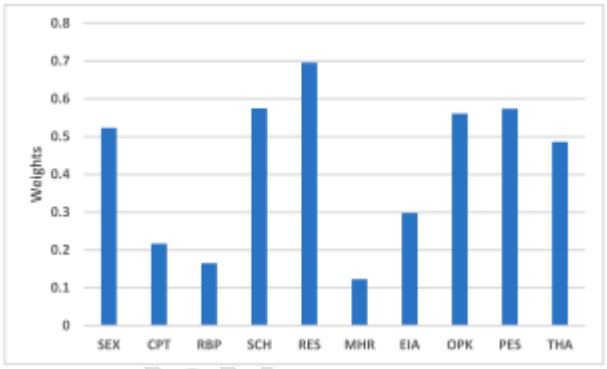

FIGURE 2. Features selected by FMIM FS algorithm.

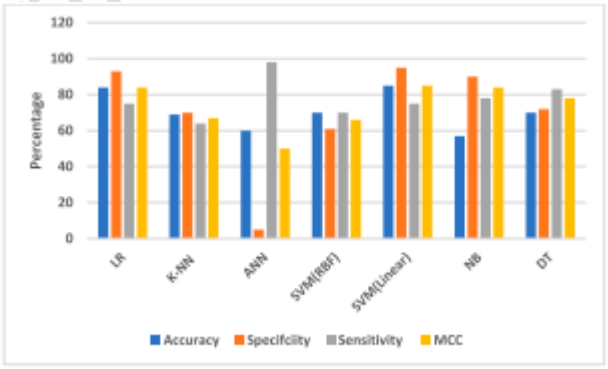

FIGURE 3. Classifiers performance with LOSO CV on set of full features.

ANN was trained with hidden neurons but at 10 hidden neurons give better performance result with accuracy 60%, specificity 100%, and sensitivity 0%. SVM (RBF) with C = 100, g = 0.001 has 61% specificity, 70% sensitivity and 70% accuracy. The SVM linear kernel has 95% specificity, 75% sensitivity, and 85% accuracy. The NB was third good classifiers which have 90% specificity, 78% sensitivity and 80%accuracy. DT has 72% specificity, 83% sensitivity, and 70% accuracy. Figure 7 shows that the SVM outperformed as

compared to the other five classifiers. The accuracy of SVM (linear) is 85%, sensitivity 77%, and specificity 95%, and 85% accuracy. Logistic regression is second good classifier has 84% accuracy. The third important classifier is NB and its specificity is 90%, sensitivity is 78%, and classification accuracy is 80%. The worst classifiers were K-NN at k = 1 with LOSO cross-validation. The MCC of SVM is 85% pretty good and SVM is good classifier for heart disease prediction. In Figure 11, we have been shown the execution time of each algorithm in which classifier Svm (linear) on C = 100 and g = 0.009 processing time is 30.145 seconds and logistic regression at C = 10 is 0.003 seconds very fast exaction time as compared to others classifiers with LOSO cross-validation method. Table 8 shows the LOSO cross validation classifiers performance with full features.

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. Furthermore, 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 as shown in Table 17. Further, the performance of machine learning based method FCMIMSVM 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.

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.