JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org **Vol.2, Iss.2 Pages:105-122**

AN INTELLIGENT SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION USING IMPROVED DENSITY BASED SPATIAL CLUSTERING

G Priyadarshini

Assistant professor,
Department of B.C.A, KG College of Arts and Science, Coimbatore, Tamilnadu, India

Article Received: April 2021 Published: July 2021

Abstract

One leading cause of global deaths in humans is the CHD (Coronary Heart Disease) which affects arteries running in heart. An early diagnosis can help clinicians in taking precautionary measures and treat patients accurately for avoiding cardiac arrests. Diagnostically intelligent systems have been a great source of support in healthcare, but unfortunately the amount of data needed for these systems is unbalanced and may not result in expected precisions in detections. Misclassifications of diseases are a major issue in existing systems. Hence, this research work proposes an intelligent diagnosis of CHDs considering information imbalance. The selected dataset is pre-processed for eliminating noise, data inconsistencies and equalizing missing values. The samples are classified using distribution frequencies. This paper proposes an intelligent system for CHD diagnosis while emphasizing on data imbalances which need to be normalized. A multitude of techniques are used in this work to achieve desired objectives. The evaluations of this proposed work in terms of sensitivity, specificity, accuracy and F-measure show that this proposed approach can detect CHDs by ensuring optimality in heart disease predictions.

Keywords: Heart disease, synthetic minority over-sampling technique (SMOTE), Enhanced Density-Based Spatial Clustering of Applications with Noise (EDBSCAN) Clustering, Data partitioning, Ensemble learning, weighted aging ensemble classifier (WAE)

1 INTRODUCTION

One prominent dysfunction in the human body is CHD which affects humans in their middle or old age ages and has high mortality rates [1]. Men are most likely to be affected by CHDs when compared to women. According to WHO, twenty four percent of deaths in non-communicable diseases is attributed to CHDs in India. One third of the deaths in a global scale and above fifty percent in deaths in US and developed countries are due to heart ailments. At least seventeen million die every year due to CVDs (Cardiovascular Diseases) where Asia tops the ranks in mortality due CHDs [2]. CHDD (Cleveland Heart Disease Database) has required information for identifying CHDs including patient's Age, gender, family history, smoking/ alcohol intakes, blood cholesterol level, dieting habits, blood pressure level, obesity and information about physical activities. Excessive intake of vices and hereditary risk factors like diabetes and high blood pressure also culminate in CHDs [3]. Major risk factors like eating habits, obesity and physical inactiveness are controllable. CHDs or CVDs can be classified into several types namely angina pectoris, cardiomyopathy, congestive heart failure, congenital heart disease, myocarditis and arrhythmias. Predicting these types based only on risk factors is a complex issue.

Overcoming imbalances, imprecision and uncertainties in data is significant to deployment of MDSS (Medical Decision Support Systems) for their clarity in decisions [4]. The main benefit of using these systems is in their ability to detect diseases in reduced times and cost. Opinion of medical experts is very important to assess patient's condition. Medical practitioners have exploited MDSSs in their diagnostic processes where intelligent systems using MLTs (Machine Learning Techniques) have been involved for improvements inferring diseased conditions from medical modalities like X-rays, CT (Computer Tomography) scans and surgical imagery [5]. Physician normally uses their experience of similar cases to identify the root causes of a disease and the start treating patients. Various tests are also advised to confirm the line of treatment for a diseased patient. A MDSS can assist physicians in their informed decisions. For example, an undiagnosed patient's disease inferences combined with physician's expertise can result is proper justifications as these computerized systems infer from large numbers of electronically stored patient records [6].

Healthcare is one area where volumes of information are generated like clinical history and disease symptoms [7]. DMTs (Data Mining Techniques) have a proven track record in identifying hidden patterns in voluminous data and hence are used in processing medical datasets for discovering hidden patterns. However, medical datasets are widely heterogeneous by nature [8]. Thus formatting and organizing them to obtain required information becomes a necessity. MLTs have been applied in medical processing where examples can be SVMs (Support Vector Machines), DTs (Decision Trees), NNs (Neural Networks), NB (Naïve Bayes) and LRs (Logistic Regressions). These techniques have classified medical data. Analytical studies on DMTs reveal that NB, NNs, DTs and their associative classifications have been

predicting CHDs effectively [9]. Associative classifications produce higher levels of accuracy when compared to traditional classifiers. The study in [10] opined a rank for classifiers for medical information namely NB, NNs and DTs in the listed order [10]. ANNs (Artificial Neural Networks) and supervised networks have also been employed in disease predictions.

Various techniques have predicted risks in CHDs with accuracy where hybrid classifying methods have shown improved accuracies in by overcoming weakness of one algorithm in the combination [11]. Hybrid systems can result in improved classification efficiencies. This research work proposes an intelligent diagnostic technique for identifying CHDs and also considers data imbalances in its processing. The proposed scheme uses mean to divide the dataset into smaller subsets and application of CART (Classification And Regression Trees) for modelling the divided partitions. The following section is a study of related literature, while section three details on the proposed scheme. Section four is obtained results of this work while section five concludes this paper with future scope.

2 LITERATURE REVIEW

This section reviews literature related to the study. MLTs have been consistently been used for classification in various fields by researchers. New techniques have been proposed again and again for improving classification performances of MLTs.

Kurt et al [12] in their study compared a multitude of techniques for judging their performances which included LR, CART, MLPs (Multi-Layer Perceptrons), RBF (Radial Basis Function), and SOFMs (Self-Organizing Feature Maps). The predictor variables used in the study were age, gender, smoking habits, family history on diabetes, hypercholesterolemia, hypertension, and BMI (Body Mass Index). The techniques were evaluated for accuracy based on HCA (Hierarchical Cluster Analysis), ROC curves and MDS (Multi-Dimensional Scaling (MDS). The study found ROC curve values of 0.753, 0.783, 0.721, 0.745, and 0.675, for LRs, MLPs, RBF, CART and SOFMs respectively. They also found that MLPs, CART, LRs, and RBF performed better than SOFMs in predicting CHDs based on MDS and HCA values.

Abdar et al [13] proposed improvisations for SVM variants in their study. A data pre-processing method with normalization improved the performances of these variants. The syudy also used, GA (Genetic Algorithm) and PSO (Particle Swarm Optimization) combined with a 10-fold cross-validation executed twice to optimized feature selections and classifier's parameters. Their scheme enhanced traditional MLT performances. Their novel optimization technique called N2Genetic optimizer classified accurately. Their experiments showed that their proposed N2Genetic-nuSVM classifier achieved 93.08 % accuracy with a F1-score of 91.51% while predicting CHDs from the Z-Alizadeh Sani dataset and thus proving the competitiveness of their proposed scheme.

Motwani et al [14] used MLT for an automatic selection of features where the feature's information gain was ranked for selections. Their scheme was a boosted ensemble algorithm with ten old stratified cross-validation. Their data was formed from patient details collected over a period of 5 years where their study used 25 clinical and 44 CCTA parameters for evaluations. SSSs (Segment Stenosis Scores), SISs (Segment Involvement Scores), modified DIs (Duke indexes) and regular CHD factors of age, gender, risk factors and FRSs (Framingham risk score) were used to identify segments features for identifying non-calcified/mixed/ calcified plaques. Their MLT classification exhibited a higher ROCs with FRS or CCTA values while SSS, SIS, DI predicted the cause of deaths in 745 cases. They listed their values thus, ML: 0.79 vs. FRS: 0.61, SSS: 0.64, SIS: 0.64, DI: 0.62; $P< 0.001$.

Neelima et al [15] in their scheme selected optimal features using a bi-directional pooled variance strategy. Traditional classification constraints of unstable accuracy while using k-fold cross validations was overcome in this study by the use of CS (Cuckoo Search), a swarm intelligence technique. Their experimental selection of optimal features when compared with other contemporary models using forward feature selections or SVMs (FFS&SVM). The experimental results confirmed the fact that their proposed scheme of bidirectional pooled variance estimation based classifications built on CS (BPVE&CS) outperformed the other model namely (FFS&SVM).

Nasarian et al [16] used a hybrid technique as their feature selection algorithm. They named it 2HFS and tested it on the Nasarian CAD dataset with clinical, environmental and work place features. They nullified data imbalances in the dataset using SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (ADAptive SYNthetic) techniques. DTs, GNB (Gaussian NB), RF (Random Forest) and XGBoost classifiers were used for predictions with 2HFS-selected features as inputs. The experimental tests showed that their scheme's selected features yielded a classification accuracy of 81.23% with SMOTE and XGBoost. They further extended their tests to Hungarian, Long-beach-va and Z-Alizadeh Sani datasets where accuracy of 83.94%, 81.58% and 92.58% were obtained. Thus, their scheme's results confirmed the effectiveness and superiority of their feature selections algorithm 2HFS.

CHDs detections using MLTs were evaluated by Qin et al [17] in their study. Their proposal adopted multiple criteria to evaluate feature's importance. They combined a heuristic search strategy with 7 classification algorithms to stress on the importance of feature selections in the Z-Alizadeh Sani CHD dataset.

Their proposed EA-MFS (Ensemble Algorithm Based on Multiple Feature Selection,) integrated multiple feature selection techniques to an ensemble technique. Their algorithm also adopted Bagging approach for increasing data diversity and used voting for decisions. Their integrations were selective based on classifiers in their ensemble process. Functional perturbation for multiple selections of their proposed EA-MFS algorithm could describe relationships between features and improve classifications while being robust. EA-MFS

algorithm reduced total dependency on dataset samples thus proving the significance of the scheme's application to clinical evaluations of CHDs.

Tuncer et al [18] in their study used DWT (Discrete Wavelet Transform). They combined wavelets with 1D-HLP (1 Dimensional Hexadecimal Local Pattern) technique for detecting arrhythmia. ECG (Electro Cardio Gram) signals of 10 s time slots were decomposed by DWT into 5 levels. Their proposed scheme extracted 512 features in each decomposed level using low pass filters. A feature set (Dimension 3072)was constructed by concatenating extracted features. The feature set's dimensionality was then reduced by using NCA (Neighborhood Component Analysis) to obtain 64, 128 and 256 features. 1NN (1Nearest Neighborhood) was used to classify samples based on four distance metrics namely city block, Euclidean, spearman and cosine. Their scheme achieved a classifying accuracy of 95 % in identifying seventeen classes of arrhythmia from the MIT-BIH Arrhythmia ECG dataset. Their results were far superior in their comparative analysis when compared with other methods used in arrhythmia detections from ECG signals.

Morphological operations figured in the study by Kandala et al [19]. Their proposal classified heartbeats with non-linear morphological features that were found suitable by their voting operation. Their proposal was tested on computers was targeted to be run on FPGAs (Field-Programmable Gate Arrays). They evaluated their proposed scheme on the MIT-BIH (Massachusetts Institute of Technology- Beth Israel Hospital) database following AAMI (Association for the Advancement of Medical Instrumentations). Their experimental simulations demonstrated the scheme's superiority specific to minority predictions where fusion and unknown classes with 90.4% and 100% were classified.

Wozniak et al [20] added weights to classifiers to improve their accuracies. The proposed a WAE (Weighted Aging Ensemble) Classifier named AB-WAE (Accuracy Based WAE) which could easily adapt to characteristic changes caused by concept drifts. Their model did not assess drifts but instead changed the order of ensemble classifications by assigning weights. Each classifier was trained based on incoming data chunks where AB-WAE chose the best ensemble considering the chunk's fixed ensemble size, previously trains and new trains. Their discussed WAE modification used ensembles of homogeneous classifiers only as sophisticated combination rule based on support functions could be applied resulting in enhanced classification accuracy as confirmed in their experimentations.

Cermak et al [21] presented a stream-based IP flow data processing application for real-time attack detection using similarity search techniques. This approach extends capabilities of traditional detection systems and allows to detect not only anomalies and attacks that match exactly to predefined patterns but also their variations. The approach is demonstrated on detection of SSH authentication attacks. And describe a process of patterns definition and illustrate their usage in a real-world deployment. Here show that this approach provides sufficient performance of IP flow data processing for real-time detection while maintaining

versatility and ability to detect network attacks that have not been recognized by traditional approaches.

Mienye et al [22] developed an improved machine learning method is proposed for the prediction of heart disease risk. The technique involves randomly partitioning the dataset into smaller subsets using a mean based splitting approach. The various partitions are then modelled using classification and regression tree (CART). A homogeneous ensemble is then created from the different CART models using an accuracy based weighted aging classifier ensemble, which is a modification of the weighted aging classifier ensemble (WAE). The approach ensures optimal performance is achieved. The experimental results on the Cleveland and Framingham datasets achieved classification accuracies of 93% and 91%, respectively, which outperformed other machine learning algorithms and similar scholarly works. The receiver operating characteristic curves further validates the improved performance of the proposed ensemble learning approach. The results show that heart disease risk can be predicted effectively by the proposed ensemble.

This paper has reviewed MLTs including methods that combine classifiers or partition samples based on majority vote and other techniques. Certain ensemble learning methods also combine classifiers. Ensemble methods can also be used on partitioned datasets. This work focuses on developing an ensemble learning model that will improve accuracy of CHD predictions.

3 PROPOSED METHODOLOGY

The proposed methodology detects heart ailments from healthcare data. Initially data is pre-processed which includes re-sampling by non-stratified random sampling. SMOTE, COR (Cleaning data attributes Out of Range) and RD (Remove Duplicates). The pre-processed data is then applied with EDBSCAN (Enhanced Density-Based Spatial Clustering of Applications with Noise) for increasing efficiency of predictions. The final part of the proposed methodology uses different CART models using a modified WAE for improved accuracy. The proposed intelligent system for diagnosing CHDs is depicted in Figure 1.

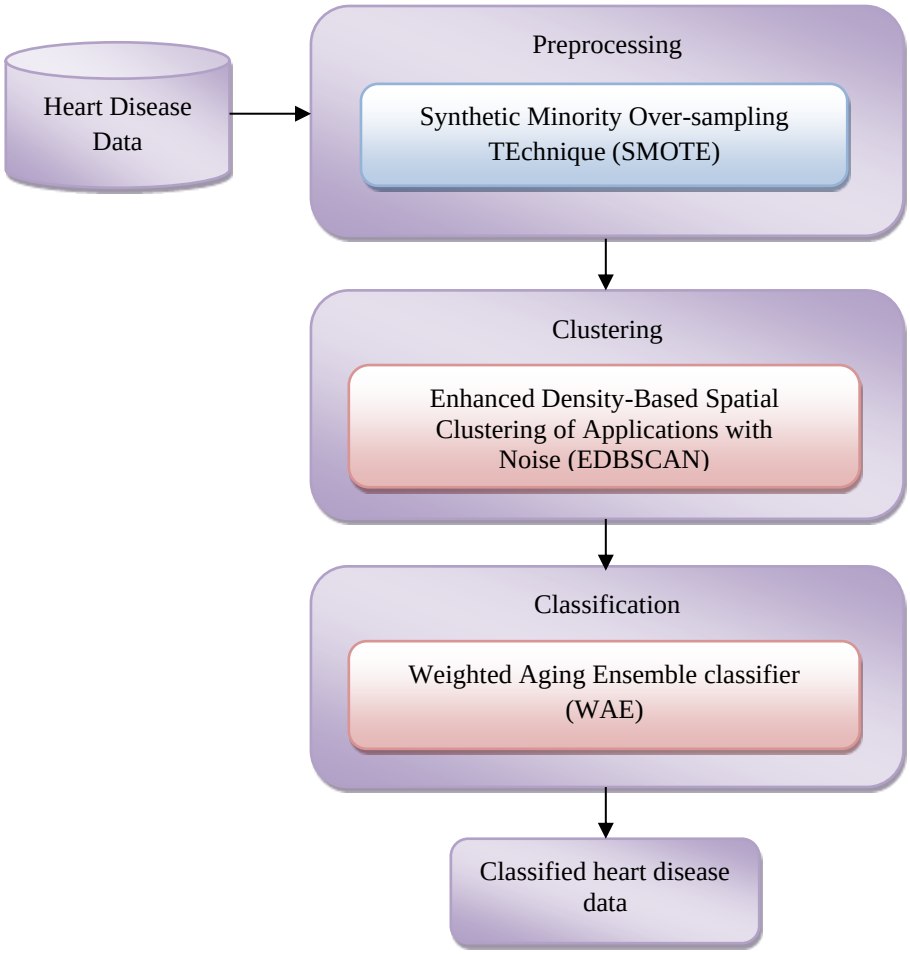

Fig 1 Overall flow of the CHD prediction model

3.1. Data collection

The Cleveland dataset used in this study was downloaded from UCI (University of California, Irvine) repository [22]. Samples had 76 attributes with missing values. The dataset samples include demographic and health record information like age, gender, cholesterol level, blood pressure, alcohol intake, diabetes, etc.

3.2. Pre-processing using SMOTE

CHD levels in the data was found to be imbalanced and solved in this work by using the oversampling approach of SMOTE [23]. The first phase stage of this work is pre-processing executed using re-sampling of non-stratified random sampling, SMOTE, COR, and elimination of duplicates. The steps used result in maximizing minority class of data by creating mock data to extend decision-making. The first step is re-sampling to estimate sample accuracy of statistically or assign random replacements from a subset of the samples. Subsequently, SMOTE balances each minority class with $k = 5$ (nearest neighbors). The over-sampling rates were adjusted by referencing amount of data at each level thus normalizing data to the same

healthy level and indicated factors against CHDs. Re-sampling was used for a better distributions, but resulted in duplicate samples. SMOTE identified duplicates thus balancing data samples. SMOTE processes data produced beyond range attributes which was overcome by removing these overshoots and duplicates.

3.3. Clustering using EDBSCAN

The pre-processed data is further clustered in this phase. Density based clustering can discover arbitrarily shaped clusters in spatial databases which have noise where density is number of points within a specified radius. DBSCAN algorithm creates clusters regions with high densities based on density connectivity [24]. The algorithm is explained with definitions below:

Definition 1: Any point that is not a core point is a noise point and discarded.

Definition 2: A ε-neighbourhood *are* objects within the radius of ε from an object and shown in figure.2.(a).

Definition 3: If an object is in the ε- neighbourhood, it contains at least a minimum number of objects (MinPts) then the object is called a core object.

Definition 4: Any object q is directly density-reachable to p if q is within the ε-neighbourhood of p and p is a core object.

Definition 5: An object p is density-reachable from q w.r.t ε and MinPts if there exists a chain of objects *p1,…,pn,* with *p1=q, pn=p* such that *pi+1* is directly density-reachable from *pi* w.r.t ε and *MinPts* for all 1<= *i* <= *n* as shown in fig. 2.(c).

DBSCAN Process is depicted in Figure 2. DBSCAN can also be summarized as steps and are detailed below:

- Choose a point ***p.***
- Get all density-reachable points from ***p*** w.r.t ε and MinPts.
- If p is a core-point create a cluster
- Else move to the next point
- Continue the process until all the points have been processed.

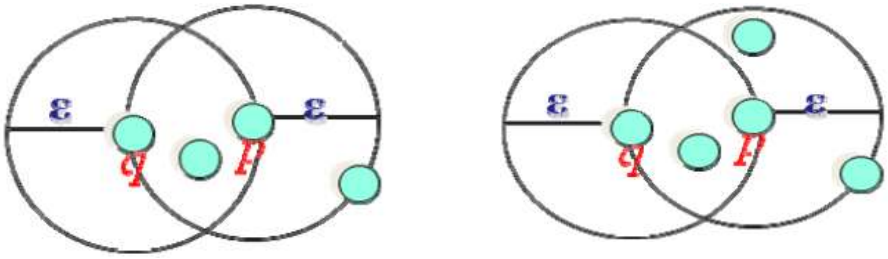

Figure 2.(a) Figure 2.(b)

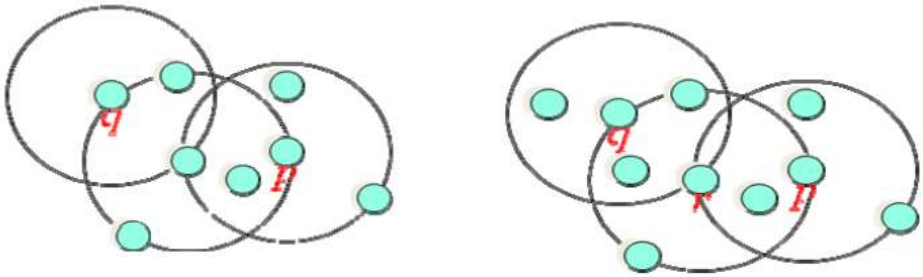

Figure 2.(c) Figure 2.(d)

Fig 2.(a). ε- neighbourhood of p, q. p is core object with MinPts=4. Fig 2.(b). q is directly density reachable from p. Fig 2.(c). q is density-reachable from p. Fig 2.(d). p and q are density-connected to each other by r.

3.3.1. DBSCAN's Threshold based Splitting

Disadvantages of DBSCAN are: It not discriminate clusters within clusters; sensitive to MinPts and Eps parameters; does not identify clusters when densities vary and clustering fails in sparse dat. This work overcomes these disadvantages and forms Clusters with varied densities using the following steps:

1. Data is split into smaller regions until homogeneity criteria based on a threshold value
2. Small regions are formed for obtaining final regions of interest using DBSCAN using similarity measures namely Euclidian and Manhattan and by changing values the of Eps and MinPts.
3. Performing Post processing or dealing with noise points.

The dataset is split into four parts as dividing it into odd number of parts may not result in crisp operations or clusters may overlap. Splitting is done until homogeneity criteria is met using:

Step 1: Splitting starts from the highest level (Dataset)and each possible child regions represented with

a supposed mean point.

Step 2: Compute Euclidian distances between the main region and its probable child regions.

Step 3: IF computed distance is greater than the threshold value THEN

Split the region based on splitting the condition.

Else

Do not make any changes in the region

Step 4: Repeat the steps until all homogeneous are found OR are too small to split.

Step 5: Decide on the splitting threshold value. Higher value decreases accuracy of segmentation while lower value results in very small regions.

Hence, this work chooses a median point for the division of data sets into sub regions and the average distances of all samples is taken as the threshold value.

3.3.2. EDBSCAN Algorithm

EDBSCAN takes two inputs (Eps and MinPts). The process is started with lower of Eps and MinPts which is then increased gradually for iterations. The Euclidian distance is defined in Equation (1) while Manhattan distance is defined in Equation (2)

$$d(i,j) = \sqrt{(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots .. + \left|x_{ip} - x_{jp}\right|^2)} \quad (1)$$

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots . + \left|x_{ip} - x_{jp}\right| \quad (2)$$

In Equations (1) and (2) $i = (x_{i1}, x_{i2}, \ldots\ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots\ldots, x_{jn})$ are n-dimensional data objects. On forming clusters are formed, points that are left out and points that are considered as noise are post processed in this work.

3.3.3. Post-Processing

Clusters which are very small and do not fall into clusters are post processed. Noisy points are merged into its nearest distanced cluster as noisy regions do not get connected with other clusters and cannot be merged until their surrounding noisy regions get merged. This is iterated

to merge all noisy points and thus these regions get merged into clusters which are nearest to it.

3.3.4. DBSCAN Iteration

Clusters within sub-regions are identified by increasing Eps and MinPts maginally for each iteration as the initial clusters generated by EDBSCAN has less significance and more repetitive splits are needed.

3.4. WAE Classification

Homogeneous ensemble learning is used in this work for classifications. Mean based splitting randomly divides the dataset into smaller subsets which are then modelled using CART. Accuracy based AB-WAE. A modified version of WAE classifies the processed samples. The resultant ensemble model is then used for predicting CHDs and explained below:

Assume the dataset is D = {(*xi* + *yi*), *i* = 1, 2,…., *N*} where the independent variable is depicted as Equation (3)

$$x_i = [x_{i1}, x_{i2}, x_{i3}, \dots\dots\dots, x_{ip}] \qquad (3)$$

If the dependent variable of D is y_i where $y_i \in \{0, 1\}$, the independent variables from multiple instances or rows can be depicted as Equation (4)

$$x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \\ \vdots \\ x_{Nj} \end{bmatrix} \qquad (4)$$

And the weighted mean of x_j can be computed using Equation (5)

$$\bar{x_j} = \frac{1}{N}\sum_{i=l}^{N} w_i x_{ij}, w_i \geq 0, \ \sum_{i=l}^{N} w_i = 1 \qquad (5)$$

Data points with greater weights contribute to mean more than data points with lower weights. As per Equation (5), weights can be 0 but cannot have negative value. This work uses randomizations for partitioning and in each at partition the variable set, {1,2, …,p} is drawn randomly as it can produce more number of trees with minimal variances and improve performances. Considering *D* as the root node, it has to be split into different homogeneous sets by randomly selecting a data point from the variable set, {1,2,..,p} or *j*. This data point in the set is replaced during further selections. Splitting *D* into two partitions using mean-based partitioning is represented as Equation (6)

$$D = \begin{cases} D_{11}, & if\ x_{ij} < \overline{x_j} \\ D_{12}, & if\ x_{ij} \geq \overline{x_j} \end{cases} \quad (6)$$

If D_{11} and D_{12} are root nodes, when Equation (6) is applied D_{11} results in $D_{21,}$ D_{22} and D_{12} generates D_{23} and D_{24}. Thus, partitioning is repeated until termination condition is met. The termination is based on two rules which ensure the dataset is not over partitioned. The first condition for stop is the maximum tree height, H_{max} to control continuous growth of the tree or over-partitioning. The tree growth is stopped when $H\ reaches\ H_{max}$. The root node's height is zero and one for r D_{11} and D_{12} etc. The second rule for stopping tree growth is checking is partitions are very small or count of instances in very low. Assuming $N(D_{kl})$ is the instances count of D_{kl}, the tree growth stops when $N(D_{kl}) \leq N_{min}$ (preferred minimum instances count) in the partition. The stoppage of tree growths is depicted in Equation (7)

$$H = H_{max}\ or\ N(D_{kl}) \leq N_{min} \quad (7)$$

After partitioning of D, CART is applied to each partition as DTs are highly interpretable and robust. CART identifies misclassification probability of randomly chosen instance using Gini Impurity. The Gini measure can be computed using Equation (8)

$$Gini = 1 - \sum_{i=1}^{J} p_i^2 \quad (8)$$

where p_i is the probability of an object being classified into a particular class, J set of items with classes and $i \in \{1, 2,\ldots, J\}$. Thus, the forest has T_{max} trees due to CART which are then applied with AB-WAE for accuracy. The AB-WAE is a modified WAE. Assuming Ψi *is the classier*, $Pa(\Psi i)$ its predictive accuracy frequency and $itter(\psi i)$ is the count of iterations. Its weight $\omega(\Psi i)$ can be represented as Equations given below

$$P_a(\Psi_i) > P_a^{\pi}\ then\ \omega(\Psi_i) = P_a(Psi_i) \quad (9)$$

else

$$\omega(\Psi_i) = \frac{P_a(\Psi_i)}{\sqrt{itter(\Psi_i)}} \quad (10)$$

where P_a^{π} is classifier's average accuracy in the ensemble π. The final prediction of the ensemble Ψ can be obtained using

$$\Psi(x) = i \quad if$$

$$\sum_{t=1}^{T_{max}} \omega(\Psi_t) F_t^{(i)}(x) = \max_{j \epsilon \{1,,2,\ldots\ldots J\}} \sum_{t=1}^{T_{max}} \omega(\Psi_t) F_t^{(i)}(x) \quad (11)$$

The proposed ensemble assigns weights to classifiers based on their accuracy and the time taken for execution. Classifier weights below a specified threshold level are removed from the

ensemble. Using accuracy as optimization criteria ensures the ensemble achieves optimality of results.

4 RESULTS AND DISCUSSION

This section displays the results obtained in this work and are depicted as figures or tables as and when necessary. The proposed work was tested and evaluated on the Cleveland heart dataset from UCI machine learning repository. It has seventy six 76 attributes, but experiments used fourteen attributes. Experiments were conducted for identifying CHD cases from the dataset. Computations used ranged from 0 (no presence) to 4. The proposed method was evaluated with other MLTs based on binary classification criteria. The metrics used for evaluations were TPs (True Positives), FPs (False Positives) (FP), TNs (True Negatives) and FNs (False Negatives) and these metrics were used to compute other related metrics. Precision is no of relevant instances retrieved. Recall is the proportion of retrieved relevant instances. Sine, these measures are often conflicting, they can be combined with equal weights to obtain F-measure. Accuracy is defined as the proportion of correctly predicted instances relative to all predicted instances. Precision is defined as the ratio of correctly found positive observations to all of the expected positive observations.

$$\text{Precision} = TP/TP+FP \qquad (12)$$

Sensitivity or Recall is defined the ratio of correctly identified positive observations to the over-all observations.

$$\text{Recall} = TP/TP+FN \qquad (13)$$

F - measure is defined as the weighted average of Precision as well as Recall. As a result, it takes false positives and false negatives.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \qquad (14)$$

Accuracy is calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = (TP+FP)/(TP+TN+FP+FN) \qquad (15)$$

Table.1. illustrate the performance comparison results between the proposed and existing method for two dataset as frahmingam and cleveland. In this table it observed that the proposed method effectively predict the heart disease data for frahmingam dataset compare to the cleveland dataset.

Table1 Performance comparison results between the proposed and existing method for two dataset as frahmingam and cleveland.

Datsets	**FRAHMINGAM**			**CLEVELAND**		
Metrics	**LR**	**WAE**	**EDBSCAN**	**LR**	**WAE**	**EDBSCAN**
Accuracy	79.4811	80.8962	95.6107	50.5495	59.3407	93.3333
Precision	56.6833	58.6253	95.7768	10.8998	11.8681	95.7746
Recall	55.1163	57.9131	95.7720	16.9143	20	88
F-measure	55.5861	58.2311	95.6107	12.4052	14.8966	90.9759

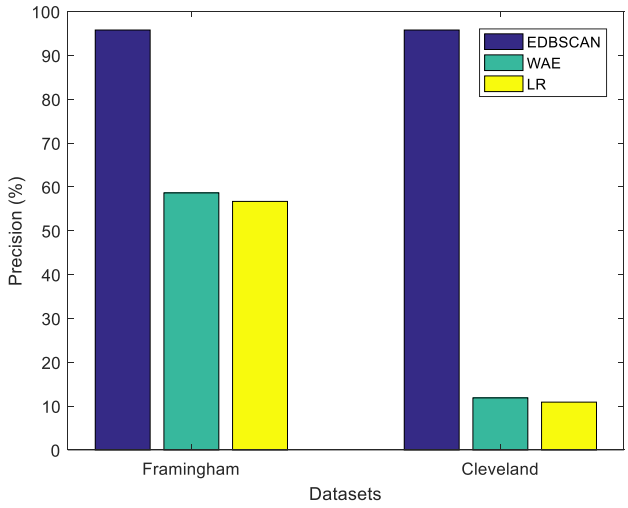

Fig 3 Precision comparison results between the proposed and existing method for classifying the heart disease data

The fig.3 illustrates that the precision comparison results between the proposed and existing method for classifying the heart disease data. From the results it concludes that the proposed EDBSCAN technique has high precision results compare to the existing classification techniques.

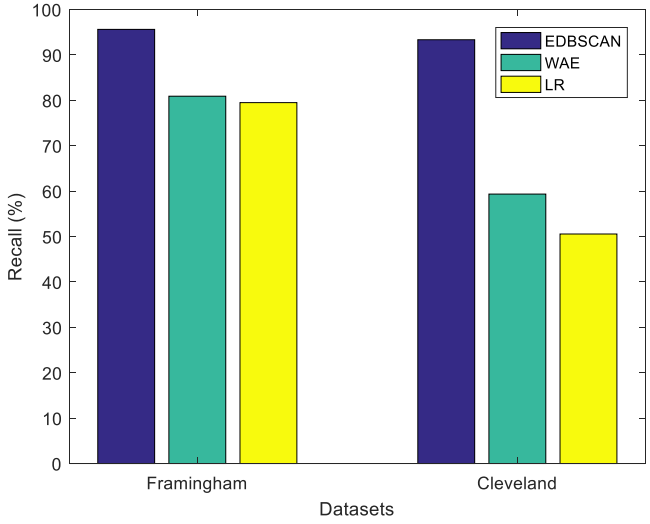

Fig 4 Recall comparison results between the proposed and existing method for classifying the heart disease data

The fig.4. illustrates that the recall comparison results between the proposed and existing method for classifying the heart disease data. From the results it concludes that the proposed EDBSCAN technique has high recall results compare to the existing classification techniques.

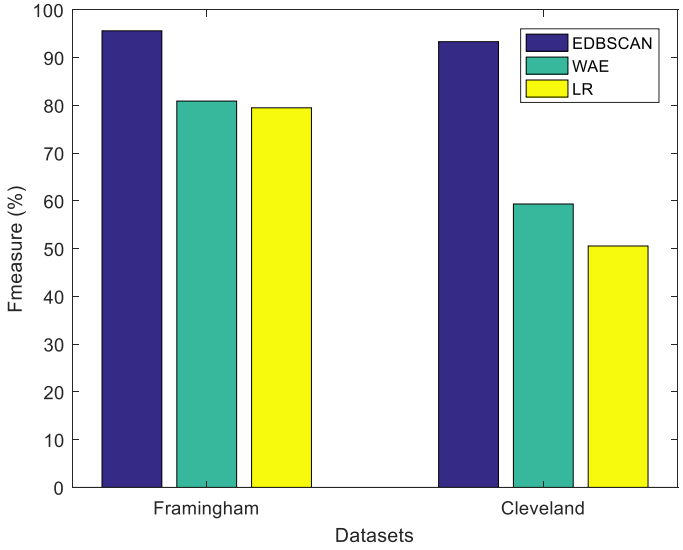

Fig 5 F-measure comparison results between the proposed and existing method for classifying the heart disease data

The fig.5. illustrates that the F-measure comparison results between the proposed and existing method for classifying the heart disease data. From the results it concludes that the proposed

EDBSCAN technique has high F-measure results compare to the existing classification techniques.

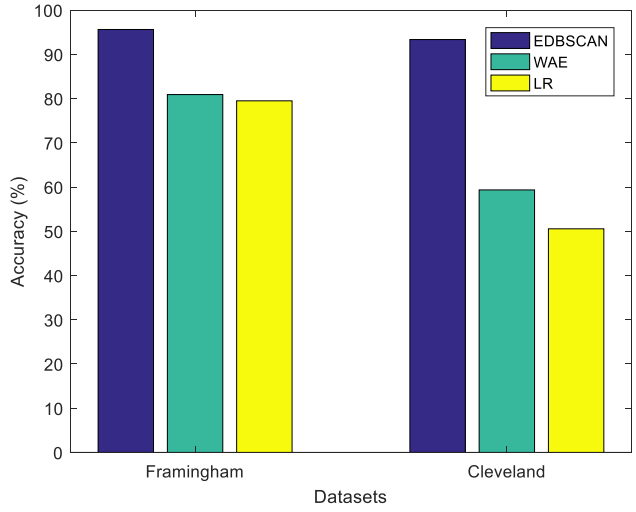

Fig 6 Accuracy comparison results between the proposed and existing method for classifying the heart disease data

The fig.6. illustrates that the accuracy comparison results between the proposed and existing method for classifying the heart disease data. From the results it concludes that the proposed EDBSCAN technique has high accuracy results compare to the existing classification techniques.

4 CONCLUSIONS

One leading cause of global deaths in humans is the CHDs which affects arteries running in heart. An early diagnosis can help clinicians in taking precautionary measures and treat patients accurately for avoiding cardiac arrests. Diagnostically intelligent systems have been a great source of support in healthcare, but unfortunately the amount of data needed for these systems is unbalanced and may not result in expected precisions in detections. Misclassifications of diseases are a major issue in existing systems. Hence, this research work has proposed intelligent diagnosis of CHDs. Efficient detection oh CHDs can benefit healthcare This research work uses a multitude of techniques for identifying CHDs. The study solves the pronlem of data imbalances in datasets by its pre-processing steps where SMOTE, COR, duplicate removals are executed. This cleaned data is then clustered by CART and finally an ensemble classifier detects CHDs. The results of the proposed ensemble method show it is a good strategy for improving the accuracy of classifiers. This work has proposed, implemented and demonstrated the utility of the proposed scheme. It can be concluded from evaluation that the proposed scheme can be implemented for identifying CHDs from dataset information. In future directions, this paper proposes to use improved feature selection approaches for

maximizing accuracy of CHD predictions and thus save more human lives. As a future work, additional feature selection techniques will be used to improve the accuracy of ensemble algorithms and build an model that can predict specific heart disease types.

REFERNCES

1. Jackson, G., Boon, N., Eardley, I., Kirby, M., Dean, J., Hackett, G., ... & Miner, M. (2010). Erectile dysfunction and coronary artery disease prediction: Evidence-based guidance and consensus. *International journal of clinical practice*, *64*(7), 848-857.
2. Wilson, P. W., & Evans, J. C. (1993). Coronary artery disease prediction. *American journal of hypertension*, *6*(11_Pt_2), 309S-313S.
3. Little, W. C., Constantinescu, M., Applegate, R. J., Kutcher, M. A., Burrows, M. T., Kahl, F. R., & Santamore, W. P. (1988). Can coronary angiography predict the site of a subsequent myocardial infarction in patients with mild-to-moderate coronary artery disease?. *Circulation*, *78*(5), 1157-1166.
4. Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., ... & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, *111*, 103346.
5. Hampe, N., Wolterink, J. M., Van Velzen, S. G., Leiner, T., & Išgum, I. (2019). Machine learning for assessment of coronary artery disease in cardiac ct: a survey. *Frontiers in cardiovascular medicine*, *6*, 172.
6. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.
7. Taneja, A. (2013). Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, *6*(4), 457-466.
8. Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In *2016 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1-5). IEEE.
9. Sa, S. (2013). Intelligent heart disease prediction system using data mining techniques. *International Journal of healthcare & biomedical Research*, *1*, 94-101.
10. Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). Heart disease prediction using data mining techniques. In *2017 International Conference on Intelligent Computing and Control (I2C2)* (pp. 1-8). IEEE.
11. Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. *International journal on recent and innovation trends in computing and communication*, *2*(10), 3003-3008.
12. Qin, C. J., Guan, Q., & Wang, X. P. (2017). Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection. *Biomedical Engineering: Applications, Basis and Communications*, *29*(06), 1750043.
13. Tuncer, T., Dogan, S., Pławiak, P., & Acharya, U. R. (2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. *Knowledge-Based Systems*, *186*, 104923.
14. Kandala, R. N., Dhuli, R., Pławiak, P., Naik, G. R., Moeinzadeh, H., Gargiulo, G. D., & Gunnam, S. (2019). Towards real-time heartbeat classification: evaluation of nonlinear morphological features and voting method. *Sensors*, *19*(23), 5079.
15. Wozniak, M. (2017, November). Accuracy based weighted aging ensemble (ab-wae)—algorithm for data stream classification. In *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)* (pp. 21-24). IEEE.
16. Cermak, M., Laštovička, M., & Jirsik, T. (2019, April). Real-time Pattern Detection in IP Flow Data using Apache Spark. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)* (pp. 521-526). IEEE.

Cite this article:

G Priyadarshini, “An Intelligent System for Coronary Artery Disease Prediction Using Improved Density Based Spatial Clustering”, Journal of Multidimensional Research and Review (JMRR), Vol.2, Iss.2, pp.105-122, 2021.