Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-190143
IOS Press

Intelligence – based decision support system for diagnosing the incidence of hypertensive type

M. Ambika[a], G. Raghuraman[a], L. SaiRamesh[b] and A. Ayyasamy[c,*]

[a]*Department of Computer Science and Engineering, SSN College of Engineering Kalavakkam, Chennai, Tamil Nadu, India*

[b]*Department of Information Science and Technology, CEG, Anna University Chennai, Tamil Nadu, India*

[c]*Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu, India*

Abstract. Hypertension is a major non-communicable disease, a silent killer that serves as a root cause for many entangled maladies. Early analysis and detection will play vital roles in reducing the prevalence of hypertension and its associated risk factors. As medicine moves forward, there is a need for sophisticated decision support systems to make real-time predictions. Since most medical applications need to deal with multi-class problems, high diagnostic prediction accuracy is extremely important. The quality of data also significantly affects the learning model's performance. These issues induce the need for proper exploration and investigation of the multi-class medical dataset. This research intends to present an intelligent learning model that can explore medical data and offer decision support for domain experts and individuals. As clinical data tend to be, grimy appropriate pre-processing techniques are essential to ensure high data quality. This paper deals with the poor-quality data using computational statistical techniques. The prominent features are obtained by employing diverse feature selection techniques and provide a competitive report. We evolved a supervised learning model that can handle multi-class issues in diagnosing medical data categories. This model will learn from the data samples by using a multi-class support vector machine technique to generate precise predictions. We evaluated our learning model by using a real-time hypertension dataset obtained from primary health centres. The proposed approach improves predictive accuracy, precision and recall for handling the multi-class dataset above that of existing techniques. The outcome positively reveals that the proposed intelligent model is effective in undertaking medical decision-making task.

Keywords: Medical decision support system, medical diagnosis, pre-processing, feature selection, machine learning, multi-class classifier, hypertension

1. Introduction

Hypertension a prime non-communicable disease is the elevated level of blood pressure. It is commonly termed as a silent killer, because individuals can have it for a long time without signs and symptoms [1]. Uncontrolled hypertension can lead the peril of health problems like coronary heart disease, heart or renal failure and stroke [2, 3]. The growing pattern of prevalence of hypertension (stage I & II) will surge from 118.2 million in 2000 to 213.5 million by 2025 [4]. Health-damaging personal behaviours can contribute

*Corresponding author. A. Ayyasamy, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu, India. E-mail: samy7771@yahoo.co.in.

to the high burden of this chronic disease [1]. These determinants are modifiable and controllable. Hence, early detection, identification and awareness will play a stepping stone to lessen the occurrence of hypertension [5]. Thus, a precise diagnostic method which could help a physician to propose an appropriate therapy is still a very desirable tool. The prime objective of this paper is to promote decision-making through the early analysis and diagnosis of hypertension to evoke the quality of treatment.

Medical decision support systems have been the focus of intense research for years [6]. It assists medical professionals and other health expert with the clinical decision-making task [7]. The element of computer science which is effectively pertained in medical science is Artificial Intelligence. This system utilizes a style of artificial intelligence known as machine learning, which gives rise to an intelligent model [8, 9]. Machine learning methods will empower the user to learn from the given data to make appropriate predictions. Such model offers support for the decision-making process in many areas of health care like diagnosis, prognosis, and monitoring and hospital management [10, 11].

1.1. Motivation

The medical information captured through the laboratory and clinical processes suffers from poor quality due to erroneous instruments measuring some property or human error while enlisting it [12, 13]. These data contain noise, outliers, missing values, redundant attribute, inconsistency and in an unsuitable form. Before commencing an automated learning model to draw useful information, the data must be renewed to a suitable form [12].The problem of hypertensive type diagnosis is the crucial stage as it entails different classes. According to the literature reviews, most related work deals with the binary-class classification problem. The condition is either the patient has the disease or not. Consequently, there is a need for a model which can learn from such data using multi-class learning techniques, to make accurate real-time predictions [14]. The performance of the most learning algorithms relies on the class distribution. Medical data commonly own an imbalance class distribution. The reason is that in case of the medical dataset the normal cases arise abundantly and the abnormal cases appear infrequent. This problem degrades the eminence of the learning system, as it learns better from the popular class [15, 16]. Hence, it is necessary to equalize the dataset as it can lead for misclassification.

In this regard, this paper puts forward a novel approach to support decision making by examining the medical data aiding to make accurate real - time prediction. As the information quality influences the learning model's performance, there is a need for proper exploration and investigation on the medical dataset [2]. Thus choosing the appropriate combination of pre-processing techniques considerably affects the classification potential [17]. The proposed model addresses these challenges using statistical value computation. This model eradicates the irrelevant and redundant attributes to reinforce the quality of prediction using the appropriate feature selection technique. Sampling based approach is used to balance the class.

The fundamental motivation of this exploration is to assemble a learning model by adopting supervised learning techniques which can manage multi-class problems. There has been extensive investigation on the diligence of supervised machine learning algorithms for medical data [18]. Most research evidences that Support Vector Machine (SVM) furnishes a high level of accuracy when equated to other techniques [19]. Here, the One-versus-All Multiclass Least Square Twin Support Vector Machine technique is proposed for categorizing the hypertensive patients. At that point, the execution of the proposed work is assessed in terms of accuracy, precision and recall.

The main contributions of this paper can be sketched as follows. To propose an intelligent learning model with the following practical implications. First, to ensure the quality of the data to improvise the efficiency, as the accuracy is very crucial in medical application. Second, to evacuate unnecessary features and provides a reduced set of features to enhance the fineness of prediction. Next level is to train the learning model using supervised machine learning technique to attain eminent levels of prediction accuracy and finally to evaluate the model using various learning techniques.

The rest of the paper is structured as follows. Section 2 throws light on hypertension and its prevalence, section 3 details the work related to the model of the medical application. The proposed work and its methodology are dilated section 4. Section 5 presents the experiment and the discussion of evaluation results. Finally, conclusion and the suggestion for the future direction in this work are illustrated in section 6.

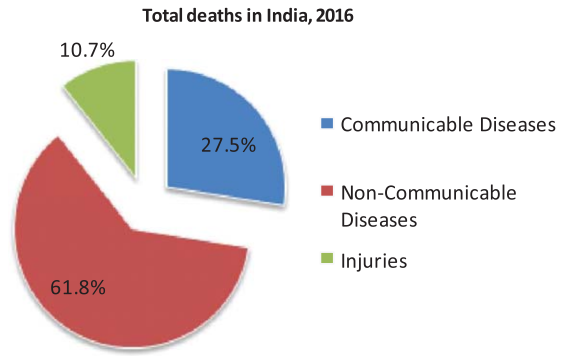

Fig. 1. Contribution of major disease groups to total deaths in india, 2016.

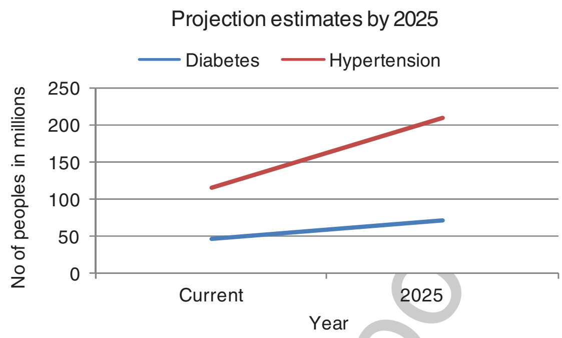

Fig. 2. Pervasiveness estimation by 2025.

2. Background

2.1. Non – communicable diseases (NCDs)

Non - Communicable Diseases (NCDs) grew into the prevailing public health dispute of the modern world [20]. In 2016, it accounted for 70% of the preeminent cause of fatalities worldwide [21]. Disorders like cardiovascular diseases, diabetes, cancer, pulmonary diseases, kidney disease, trauma, disasters, and emergency medical services, mental health, diseases of bones and joints come under this category. Such chronic diseases are not contagious or transferable, and common among individuals who are younger than 45 years [22].

India was facing a rapid health changeover with the rising predominance of most preeminent NCDs from 1990 to 2016. More than six million individuals died due to non-communicable diseases which were twice that of communicable diseases in 2016 [5, 23]. Figure 1 proves that the NCD's report for high demise rates in India and substitutes infectious diseases for the high morbidity and mortality [22]. NCD would not just seriously affect the human health, but also on economic growth [24].

The Cardiovascular diseases were the driving motive of death, followed by chronic respiratory diseases, diabetes, cancers and urogenital disorders. High blood pressure is a noteworthy among them. Hypertension is the raised level of blood pressure. Uncontrolled hypertension can lead to the various health issues [2, 3]. Figure 2 shows the pervasiveness estimation of diabetes and hypertension in 2025 [4, 21]. Hypertension can be discerned among various categories [25] as depicted in the Table 1.

Table 1
Categories of hypertension

Classification	Systolic Pressure (mmHg)	Diastolic Pressure (mmHg)
Normal	90–119	60–79
Pre-hypertension	120–139	80–89
Hypertension – Stage 1	140–159	90–99
Hypertension – Stage 2	≥160	≥100

Health damaging personal behaviors are the primary drivers for these persistent diseases. However, the social conditions in which people live and work also play a crucial role. Risk factors comprise behavior like tobacco utilization, smoking, alcohol consumption, unhealthy diet, and obesity, lack of physical activities, stress and environmental factors leads to the major disease burden of NCDs [1]. These determinants are modifiable and controllable, with the goal that the occurrence of NCDs can be diminished. Kaberi, et. al [26] elaborated the various criteria for examination of hypertension. Early analysis and detection will play vital roles in lowering the prevalence of hypertension and its associated risk factors [5]. This article acquaints a way to predict and diagnose hypertension by employing a machine learning techniques in order to reinforce the standard of treatment.

2.2. Electronic health records and decision support system

Electronic health records (EHR) also known as Electronic Medical Record (EMR) is an electronic version of the patient medical record. It generally records all basic and medial related information collected during the clinical process to support decision making [27]. It provides a remarkable opportunity to pertain informatics techniques in healthcare [28]. Clinical decision support technologies have manifested the ability to enhance patient care across diverse health care settings [6, 29]. Numerous studies proposed that clinical decision support interventions are effective in assisting physician and health professionals, with respect to process outcomes such as guideline adherence [7, 30, 31]. Moreira et. Al [6] offers a comparative analysis of the most meaningful solutions for intelligent systems in health care. This investigation is imperative for understanding the diverse approaches used in recent studies, to support and justify the best technology for the development of further research on the topic.

3. Related work

3.1. Hypertension and its related work

The healthcare is the dynamic research area which requires the usage of several fields like data mining, web mining, machine learning, natural language processing, web personalization, recommender system and security.This section reveals some prevailing works associated with Hypertension (HT) prediction and classification in data mining environment. Bartosz [32] et al. proposed a computer support system that can identify the type of hypertension using hierarchical one-class classifier which can handle imbalanced medical data. Vilaplana et al. [33] suggested a cloud computing tool for supervising the hypertensive patients.

Omboni and Ferrai [34] developed a Home Blood Pressure Tele-monitoring technique for managing the hypertensive patients. From this paper, it was analyzed that the HBPT was an appealing tool for amending blood pressure control of hypertensive patients. Abdullah, et al. [35] designed a fuzzy expert system to analyze the hypertension risk for patients aged between 20s, 30s and 40s. Fard et al. [36, 51–54] proposed a home health care by using a hybrid version of recent and old metaheuristics approaches. Melillo et al. [37] introduced a cloud centered remote processing for automatic risk investigation of hypertensive patients via data mining functionalities. This paper portrait the m-health problem based on a cloud with versatility, extensible and transparent. Inbarani et al. [10] suggested a supervised hybrid feature selection based PSO and rough set method for medical diagnosis. In this work, a novel supervised feature selection technique hybridization of PSO, PSO based Relative Reduct (RR) and PSO based Quick Reduct (QR) were presented for disease diagnosis.

Jiang et al. [38] suggested a strategy for predicting acute hypotensive episodes, using Hilbert-Huang transforms and multiple genetic programming (GP) classifiers. Fard et al. [39, 40, 55] introduces a new mathematical formulation considering new suppositions in this research area and obtained solution approach based on Lagrangian relaxation theory. Wu, et al. [41] projected a novel way for analyzing the function of the human right ventricle by extracting the kinematic features of the cardiac. This approach was anatomically consistent, and it offers new metrics to enhance the prognosis and understanding of cardiovascular diseases such as Pulmonary Hypertension (PH). Wu, et al. [42] suggested a kinematic feature extraction technique for hypertension classification. The article details about the potential shape analysis approach for the intervention spectrum. From this paper, it was noticed that the PH was diagnosed by analyzing the shape of the right side of the interventricular septum.

Sarah et al. [43] elaborated a review over the implication of personalized medicine for treating hypertension. Fernando, et al. [44] offered a prediction model using logistic regression, to assess the relationship among the factors like age, gender, race, smoking habit, BMI, kidney disease and diabetes. Muhammad et al. [45] presented a Hybrid Prediction Model, which can offer an early prediction of type 2 diabetes and hypertension based on risk-factors for individuals. This approach used an integrated combination of DBSCAN-based outlier detection, SMOTE, and Random Forest techniques for diabetes and hypertension forecast. Goli et al. [46] briefed the advantages and constraints of machine learning models in traditional medicine. According to the literature review, among the various data mining techniques Support Vector Machines (SVMs), Bayesian Networks (BNs), Artificial Neural Networks (ANNs), and decision trees were identified as most eminent methods for knowledge discovery.

Chayakrit et al. [47] showed recent innovations in the field of computer science and medicine, illustrating the impressive approach of AI for the prediction of early stages of hypertension. This article portrays comparative assessment of different machine learning techniques and their pros and cons. LaFreniere et al. [48] identified the important risk factors and presented a neural network model for predicting hypertension with about 82% accuracy. George et al. [49] describes the state-of-the-art in artificial intelligence and machine learning methods for cardiovascular disease diagnosis and prognosis. The appropriate literature was dissected and contrasted with respect to the factor such as, dataset, feature space, the employed predictive modelling strategies and their discriminating or predictive capacity.

3.2. Research gap and objectives

Recent research shows that machine learning and data mining approaches have a vital role in achieving the mission of disease analysis and predictions. But every technique outperforms on specific data and has its limitations. There is bounded work on artificial intelligence at the current stage in the context of hypertension treatment because of limitations in the amount and quality of data and computational power. In general, the typical performance of machine learning techniques for hypertension diagnosis ranges from about 80 to 90% accuracy in regard to the data and the model [47].

Most related investigate in hypertension has been focusing on the prediction of hypertension with binary categories. In case of hypertensive type diagnosis, it is an essential need for a model which can handle multiple categories. Second, ensuring the quality of the data captured during clinical trials. The study shows that a high amount of data preparation research was carried out in order to improve the performance of the learning system. Especially most research works focus on the data reduction task, and particularly in feature selection.

Class distribution degrades the eminence of the learning system. Hence, it is needed to balance the class with appropriate technique. Very few research works have been conducted in class balancing, which can deal with binary categories. Moreover, while capturing clinical trials, there is no integration of personal behaviour and past medical history information. Ensuring high data quality is more credible to bring prominent results. According to the statistical report, in the entire process of data processing, the workload of the pre-processing stage is more than 60% [50]. Consequently, there is a necessity for combination of diverse pre-processing techniques to yield better results. Thus, the proposed intelligent model can deal with raw clinical trials in order to promote early analysis and diagnosis of hypertension with high prediction accuracy.

4. Material and methods

The aim of this study is to broaden an intelligent learning model for the analysis of medical data and also to study from the data samples using multi-class techniques, aiding to make accurate real - time prediction. Thereby, we particularly focus on all aspects of hypertension-related issues. The overall flow of the proposed model is illustrated in Fig. 3.

Initially, the user medical information are collected and converted into an Electronic Health Record (EHR). Then, the normalization process is applied to the data. The dirty data is pre-processed to eliminate missing values, outlier and imbalance data. The pre-processed output is given to the feature selec-

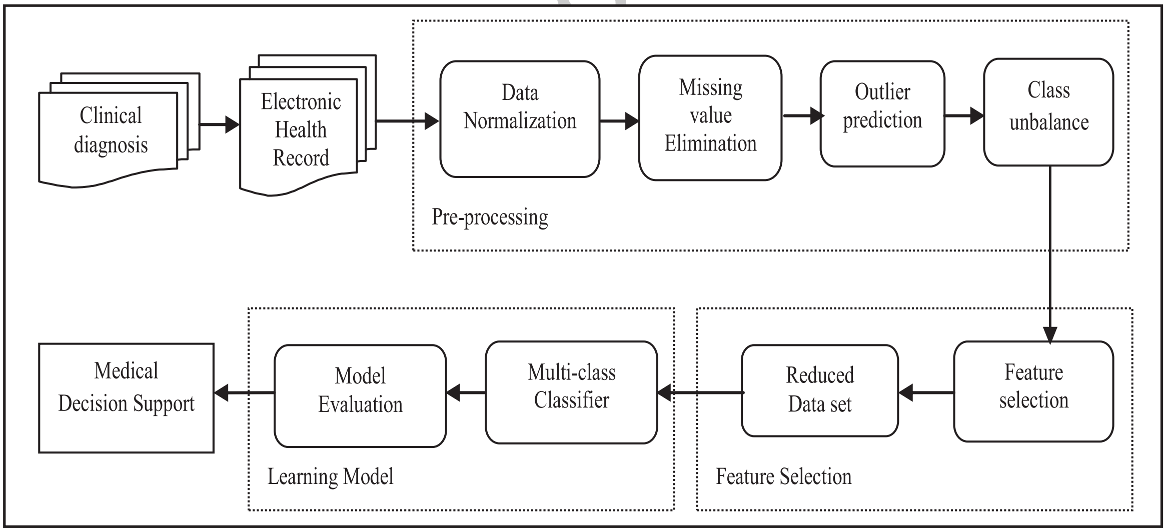

Fig. 3. Intelligence – based decision support system.

tion process, an important step in learning process for selecting the reduced set of features. Finally, multi-class SVM technique is invoked for classifying the hypertensive patients. After that, using testing data the performance of the intended system is evaluated.

4.1. Data collection

The initial stage is collecting the medical dataset. The sample data used for training and testing the proposed framework are gathered from the Primary Health Centers with the support from Tamil Nadu Health System Project - Non Communicable Disease wing. With the assistance from the health experts and trained professionals, data are obtained through various forms like, interviewing with the patients, lab investigation and doctors report. The clinical data comprise the patient medical details including,

- Demographic information
- Past medical history
- Personal behavioural information,
- Anthropometric Measurement
- Blood Pressure Measurement
- Lab investigation and Doctors' report.

Totally, there are 225 samples in the dataset and each sample has 44 features, inclusion of the class attribute. The samples are categorized into four classes, namely, normal hypertension (normal), pre-hypertension (preHT), newly diagnosed hypertension (newHT), known hypertension (KHT). We converted these clinical data into Electronic Health Record, an electronic version of the patient's clinical history, which would help in further processing like diagnosis, prediction and treatment of the patient. Through the presence of digital data, the decision support process can be earned easier. Necessary pre-processing is enforced to renew data into an appropriate form using normalization and cleaning technique.

4.2. Data pre-processing

The medical data captured during the laboratory, and clinical process suffer from poor quality with noisy and missing values, redundant attributes, inconsistent representation, imprecision and unbalanced, and in an unsuitable format. This may be because of erroneous instruments measuring some property or human error while enlisting it. Before we begin an automated learning model to extract useful information, the data must be renewed into a suitable form. Initially, the medical records are normalized to nominal values by performing the statistical value computation.

4.2.1. Missing feature values

Incomplete records are an inevitable problem in dealing with most of the real-world dataset, especially in medical application [56, 57]. There are numerous causes for the missing values of the attributes.

- Attribute value is missing because it was forgotten or lost.
- Certain features is not appropriate for the given instance,
- Human error while enrolling the data

Different strategies are available to handle these kinds of missing data. This work follows mean substitution method. The absent values are supplanted by computing mode and mean values of the attribute from the available cases, for the nominal and numeric attribute respectively.

4.2.2. Noise elimination

The noise in the data is considered as outlier or errors. Generally, in the medical dataset the outlier may occur due to faulty data, variability in the measurement, instrument error, clinical error, human error or simply through natural deviations. The machine learning algorithm may have poor performance in the presence of these noisy data. Hence, it is needed to recognize and remove these outliers.

This proposed work uses statistical methods to eliminate the outlier. Interquartile range (IQR) is used to find outliers in the dataset [60]. It is the measures of dispersion that can be computed as the difference of upper quartile and lower quartile as in Equation 1 [61]. The middle 50% of a distribution is the range of IQR. The values that lie beyond the IQR are treated as outlier. Since, an outlier is a perception which diverts a lot from the other.

$$IQR = Q_3 - Q_1 \tag{1}$$

$$Q_1 = \frac{n+1}{4} \tag{2}$$

$$Q_3 = \frac{3(n+1)}{4} \tag{3}$$

4.2.3. Class imbalance

Most of the machine learning techniques works well when the numbers of instances of each class are generally equivalent. The medical dataset commonly

Table 2
Procedure for class imbalance problem

Step 1: M_{min} *is a set of minority class from the dataset M,*
Step 2: For each instance $x_i \in M_{min}$ *Find its K - nearest neighbors by using Euclidean distance.*
Step 3: To create a synthetic sample, Randomly select one of the K-nearest neighbors, Calculate a feature difference between x_i *and its neighbor. Multiply this feature vector difference by a random number rand* $\in [0, 1]$ *Finally add this vector to* x_i *to get the synthetic sample* x_{new} *as in Equation 4*

possesses an unbalanced class distribution, where one class comprises large numbers of samples while the others constituted with the small numbers. The reason is that in case of the medical dataset the normal cases occur abundantly and the abnormal cases occur infrequently. This problem degrades the quality of the learning system, as it learns better from the majority class. Hence, it is necessary to equalize the dataset as it can lead for misclassification.

There are different ways available to deal with this problem. This model mainly focused on sampling based approach and uses SMOTE (Synthetic Minority Over-sampling Technique) technique [62]. The method of over-sampling the minority class is done by taking each minority class instance and creating synthetic samples using K-nearest neighbor technique as in Equation 4 [63]. The numeric results indicate that the suggested approach can help classifiers to achieve better performance.

$$x_{new} = x_i + (\hat{x}_i - x_i) \times rand \qquad (4)$$

Where,$x_i \in M_{min}$ is an instance of minority class in the original dataset,$\hat{x}_l$ is one of the K - nearest neighbors of x_i, $rand \in [0, 1]$ is a real random number, x_{new} is the new synthetic sample.

4.3. Applying feature selection

In general, the number of attributes and their dependency rely an impact on the accuracy of the learning models. This issue can be dealt by preferring appropriate features from the basic feature set. Feature selection as a pre-processing step in machine learning, used to elite most enlightening features from the given medical datasets [10]. It supports in reducing the dimensions, eliminating irrelevant and redundant features, in order to increase learning accuracy and reducing the computation cost for diagnosis of the disease [65].

This work employs various well-known feature selection techniques such as GainRatio, InfoGain, Chi-squared, and Principal Component Analysis. Based on the performance efficiency of the various techniques the best strategy is chosen and employed.

4.4. Supervised machine learning classification

The leading purpose of the supervised machine learning is to build a classification model by considering the target class of the training instances [8]. The ensuing classifier is employed to assign class labels for the testing instances, whose attribute values are known, yet the target class is unknown. Common machine learning algorithms are k-Nearest neighbour, Naïve bayes, decision tree, neural network, genetic algorithm and support vector machine (SVM). The choice of selecting the precise learning algorithm is a crucial step. Most literature reports proves that SVM performs better in accuracy when compared with other learning algorithms, as accuracy plays an imperative role when dealing with medical datasets [68].

As the most medical application needs to deal with multi-class classification, the recent researchers extend the conceptualization of binary SVM to the multi-class problem scenario [69]. The binary LSTSVM is extended to multi-class series using "One-versus-All" (OVA) principle [14]. One-versus-All concept belongs to the "decomposition - reconstruction" strategy, where the data points in a specific class are trained with the data points of the rest of all the other classes.

In general for K-class classification problem, this approach produces K-non parallel hyper-planes, by elucidating K-linear equations one for each and every class. The classifier designates the class to the test data point by finding the minimum distanced hyper-plane. For K-class issues, it will build K-binary LSTSVM classifiers, such that the ith classifier considers the data points of ith class with positive class label and data points of residual classes with negative class labels.

The OVA - MLSTSVM classifier with three classes is graphically presented in Fig. 4 [14]. This classifier constructs three non parallel hyper-planes, mutually for every class. The distinctive shapes denote the statistics points of different classes. Hyper-planes are acquired in the manner that the data points of each class lie closer to their own hyper-plane and lie far apart from other hyper-planes.

The proposed model use Linear One-versus-All Multi-class Least Squares Twin Support Vector Machine (Linear OVA-MLSTSVM) to train the

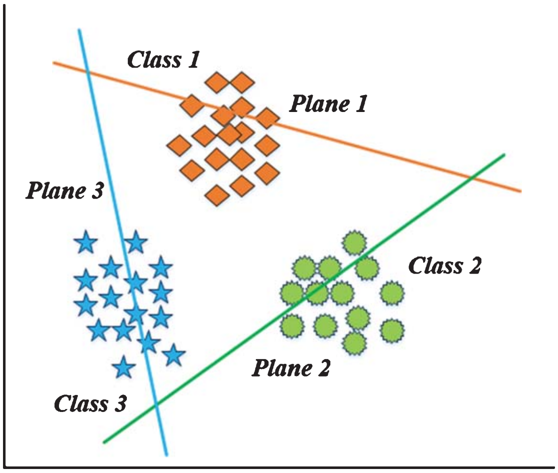

Fig. 4. Multi-class least square twin support vector machine classifier with three classes.

classifier. The training dataset for multi-class classification is denoted as in Equation (5).

$$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots, (x_1, y_1)\} \tag{5}$$

Where, $l \to$ *size of training dataset*
$x_i \in R^n, \to$ *input datapo* int s *in* $n -$ *dimentional real space* R, *such that* $i = 1, 2, \ldots..l.$
$y_i \in \{1, 2, \ldots, K\} \to$ *corresponds to* $K -$ *class lables*

The data points are represented as a matrix. The matrix X_i denote the data points ith class and the matrix Y_i denotes the data points of other classes, such that $X_i \in R^{l_i \times n}$, $Y_i \in R^{(l-l_i)\times n}$ and

$$Y_i = [(X_1)^T, (X_2)^T, \ldots, (X_K)^T,]^T \tag{6}$$

The ith hyper-plane is derived using the Equation (7) by solving the the ith linear equation as in Equation (8).

$$f_i = (w_i.x) + b_i = 0 \tag{7}$$

$$\begin{aligned} &\min(w_i, b_i, \varepsilon_i)\tfrac{1}{2}\,\|X_i w_i + e_{i1} b_i\|^2 + \tfrac{c_i}{2}\varepsilon_i^T \varepsilon_i \\ &s.t\,(Y_i w_i + e_{i2} b_i) + \varepsilon_i = e_{i2} \end{aligned} \tag{8}$$

Where, $c_i \to$ *non - negagive penality parameter*
$\varepsilon_i \to$ *slack variable*
w_i & $b_i \to$ *parameters of the ith hyper - plane*
$w_i \to$ *denotes the normal vector to the hyperplane*
$b_i \to$ *refers to the bias term.*
$e_{i1} \in R^{l_i}$ *and* $e_{i1} \in R^{l-l_i} \to$ *are vectors of one's*

The solution of the Equation (9) determines the ith hyper-plane parameters.

Table 3
Algorithm for multi-class support vector machine

Training Phase:
step 1: *Define two matrices* A_i *and* B_i
step 2: *Select the penalty parameter* $c_i > 0$.
step 3: *Solve and evaluate the hyperplane parameters according to Equation* (9)
step 4: *Determine the hyperplane using the Equation* (7)
Testing Phase:
step 5: *Assign class to new data point by using decision function in Equation* (10)

$$\begin{bmatrix} w_i \\ b_i \end{bmatrix} = u_i = (B_i^T B_i + \frac{1}{c_i} A_i^T A_i)^{-1} B_i^T e_{i2} \tag{9}$$

In this way hyper-plane for each and every class is obtained. Finally, the classifier predicts the class for the data points of the test data, based upon the premise of its distance from the hyper-plane. The class, corresponding to the hyper-plane which lies nearest is doled out to it. The decision function is given in the Equation (10).

$$d(x) = \arg \min_{i=1,\ldots,K} \frac{|w_i.x + b_i|}{\|w_i\|} \tag{10}$$

Where, $x \to$ *test data point*

The algorithm for Linear OVA-MLSTSVM consists of two phases as in Table 3.

- Training Phase – train the classifier with training dataset by constructing hyper-planes.
- Testing Phase – assign class of the testing dataset by using distance factor.

5. Implementation and discussion

5.1. *Performance Analysis of Multi-Class Learning Model*

In assessing multi-class classification problem, classifier overall performance is generally described in step with the confusion matrix related to the each classifier. As the proposed approach utilizes “One-versus-All” (OVA) principle, the performance measure for a multi-class model can be comprehended with a set of several binary classification models. For K-class issues, it will build K-binary classifiers, in a way that the ith classifier considers the data items of ith class as positive label and data items of residual classes as negative labels. Then the confusion matrix for one of the categories may have the structure as with Table 4.

Table 4
Confusion Matrix for binary classification problems

Actual/predicted	Positive class	Negative class
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

The row of the table signifies the predicted class, as the column stands for the actual class. From this confusion matrix, TP and TN symbolize the number of positive and negative cases that are correctly categorized. On the other hand, FP and FN indicate the quantity of misclassified negative and positive instances, respectively [72]. Based on the values in the matrix, it is conceivable to compute varied performance metrics. Precision is the ratio of quantity of correct prediction to the total predictions made. In other terms it is number of items correctly defined as positive out of the total items distinguished as positive. Recall also termed as sensitivity or True Positive Rate (TPR) is the number of appropriate predictions divided by the total elements within that category. F-measure is a harmonic mean of precision and recall. False Positive Rate (FPR) is certainly number of items wrongly recognized as positive out of all true negatives.

Accuracy is the most valuable assessment metric for both binary and multi-class classification problems. It is denoted as the number of items correctly discerned as either truly positive or truly negative from the total number of items. Graphically, it is the sum of the value in the diagonal of the confusion matrix divided by the total number of predictions made. Precision, recall, F-measure and accuracy are calculated using Equations (11–14) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{15}$$

Misclassification error or error rate is the measures, the ratio of incorrect predictions over the total quantity of instances evaluated as in Equation (16).

$$Error\ rate = \frac{FP + FN}{TP + FN + FP + TN} \tag{16}$$

To deal with multiclass problems, some of the performance metrics have been prolonged for multi-class classification evaluations [73]. Equations (17–20) denote average precision, recall, accuracy and error rate.

$$Averaged\ Accuracy = \frac{\sum_i^k \frac{TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i}}{K} \tag{17}$$

$$Averaged\ Precision = \frac{\sum_i^k \frac{TP_i}{TP_i+FP_i}}{K} \tag{18}$$

$$Averaged\ Recall = \frac{\sum_i^k \frac{TP_i}{TP_i+FN_i}}{K} \tag{19}$$

$$Averaged\ Error\ Rate = \frac{\sum_i^k \frac{FP_i+FN_i}{TP_i+FN_i+FP_i+TN_i}}{K} \tag{20}$$

Where, $K \rightarrow$ No of classes; for class i, TP_i – true positive for class i, FP_i – false positive for class i, FN_i – false negative for class i, TN_i – true nrgative for class i.

5.2. Performance result of the multi-class learning model with hypertension data

This segment portrays the illustration of the proposed model. The sample data used for training and testing the proposed system are collected from the Primary Health Centers with the support from Tamil Nadu Health System Project - Non Communicable

Table 5
Comparitive Analysis of performance of various multi-class classifiers with imbalanced data

Learning models	TPR (%)	FPR (%)	Average precision (%)	Average Recall (%)	Average F-measure (%)	Average accuracy (%)
SVM	91.6	07.3	88.3	91.6	89.9	91.6
ANN	91.1	07.4	88.2	91.1	89.4	91.1
IBK	84	01.29	83.4	84	83.1	84
NaiveBayes	83.1	01.2	85	83.1	83	83.1

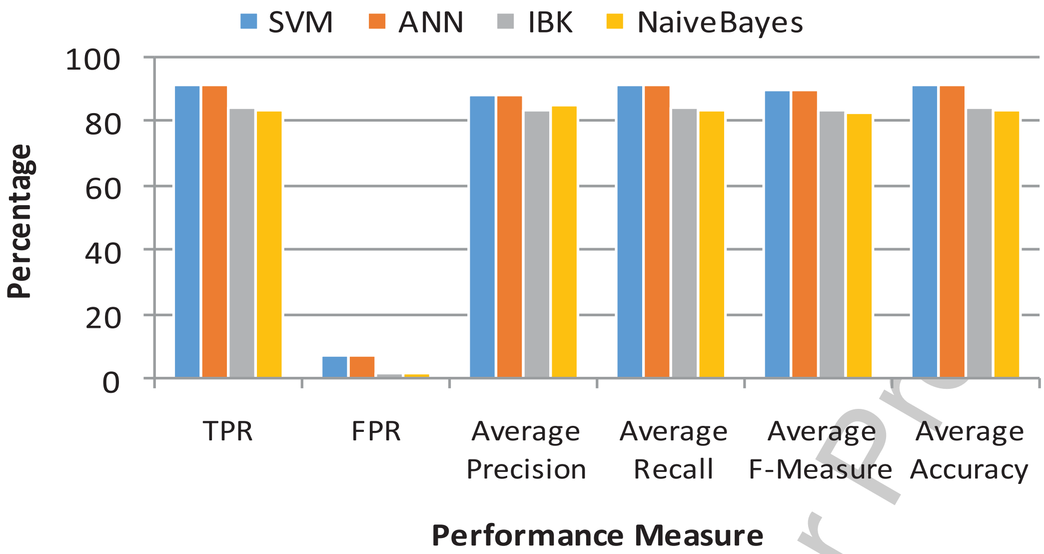

Fig. 5. Performance measure of the hypertension dataset with imbalanced data.

Disease wing. Dataset comprises 225 samples with 44 attributes, inclusion of the class attribute. The samples are categorized into four classes, with 109 normal, 100 preHT, 8 newHT, 8 KHT features.

As hypertensive type diagnosis involves different classes, it is desirable to use multi-class learning model. The effectuation of the suggested learning model using Multi-class Least Square Twin Support Vector Machine classifier are extensively studied and compared with other learning techniques like Naive-Bayes, Instance Based K-neighbors (IBK), Artificial Neural Network (ANN). 10-fold cross-validation is used for the validation. The performance comparisons of the diverse classifier, with the imbalanced dataset are elaborated in Table 5 and Fig. 5. The first set of experiment is with the unbalanced dataset, in which the suggested model shows better performance over different strategies.

5.3. Performance Result with Class balancing and Feature Selection

The performance of the most learning algorithms relies on the class distribution. Medical data commonly have an imbalanced class distribution. In this work the minority classes are re-sampled used SMOTE technique. The class distribution and class wise experiment results are recapitulated in the Tables 6 and 7. It is interesting to note an enhancement in overall performance is recorded after class balancing. Here, the comparisons between the existing and proposed methods are also discussed based on the performance metrics.

Table 6
Class Distribution Before and After Class Balancing

Dataset	Total No. of samples	Class	No. of samples
Original	225	Normal	109
		preHT	100
		KHT	8
		newHT	8
Balanced	401	Normal	109
		preHT	100
		KHT	96
		newHT	96

The next set of experiment is to investigate the performance with balanced data and reduced attributes. The prominent reduced attributes are elected using diverse well-known feature selection techniques such as GainRatio, InfoGain, Chi-squared, and Principal Component Analysis. As per evaluation metric revealed in Fig. 6, it leads us to conclude that the salient feature selected using information gain owns high computational efficiency over other methods.

Accordingly, after implementing class balancing and feature selection, we briefly observed that the opted methodology performs better and gains the highest accuracy over other. The experiment results are furnished in Table 8 and Fig. 7, portray the overall average accuracy of the various learning models and its improvisation after class balancing and feature selection. Thus, accuracy of the proposed learning model improves after the balanced data distribution. Despite the fact that the performance of ANN is more

Table 7
Class-wise Comparative Analysis of Precision – Recall

Learning models	Class	Before balancing		After balancing	
		Precision (%)	*Recall (%)*	*Precision (%)*	*Recall (%)*
SVM	Normal	93	97.2	93.8	97.2
	preHT	89.3	92	86	86
	KHT	100	100	100	100
	newHT	0	0	92.4	88.5
ANN	Normal	95.5	96.3	97.2	97.2
	preHT	86.4	95	94.7	90
	KHT	100	62.5	100	100
	newHT	0	0	93.1	97.9
IBK	Normal	88.8	87.2	90.3	85.3
	preHT	80.2	89	79.1	87
	KHT	100	50	100	100
	newHT	33.3	12.5	94.6	90.6
NavieBayes	Normal	96.6	77.1	98.7	67.9
	preHT	75.8	94	69.7	92
	KHT	66.7	75	97.9	99
	newHT	60	37.5	91.8	92.7

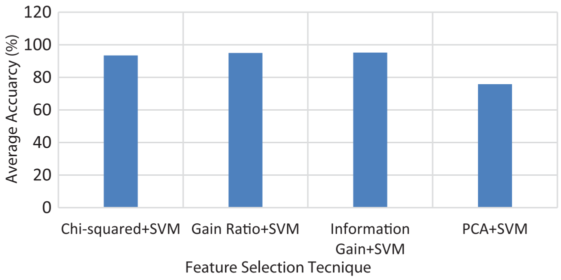

Fig. 6. Classification accuracy of the hypertension dataset after feature selection.

similar to the proposed model, it has degraded in the elapsed time need for execution is exceptionally high (Fig. 8). Hence, the proposed work affords support for a medical decision support system with high efficiency.

As clinical data tends to be dirty, the machine learning classifiers suffer from misclassification error. The misclassification error refers to the quantity of the individual wrongly categorized to a different category. The proposed system eliminates the error rate with appropriate data cleaning process, class balancing and employing salient feature selection techniques. Figure 9, portraits the average error rate in the learning process. From the perception, it is inferred that the error rate of the balanced data is reduced compared to the imbalanced data. Furthermore, it can also evoke that our system expels highly possible misclassification rather than eradicating all feasible misclassification and still suffer from minor error rate. Hence, this system needed further improvement in handling misclassification errors.

6. Conclusion and future work

The prime intent of this paper is to promote an intelligent learning model to endorse decision making by probing the medical data. Most related investigation in hypertension has been focusing on the prediction with binary categories. In this regard, this model handled multiple categories for predictions. As the quality of data affects the learning model's performance, this paper puts forward a combined approach of diverse pre-processing techniques. Thus enhancing the nature of the data to own high computational efficiency. The skewed class distribution degrades the

Table 8
Comparitive performance analysis of learning models with balanced data and Feature Selection

Learning models	Balanced data: Average precision	Average recall	Average F-measure	Average accuracy	Balanced data+feature selection: Average precision	Average recall	Average F-measure	Average accuracy
SVM	93	93	93	93.0175	95.2	95.3	95.2	95.2618
ANN	93.5	93	93.2	93.2594	95.8	95.8	95.8	95.7581
IBK	90.8	90.5	90.6	90.5237	91.1	90.8	90.9	90.7731
NaiveBayes	89.6	87.3	87.3	87.2818	95.1	95	94.9	95.0125

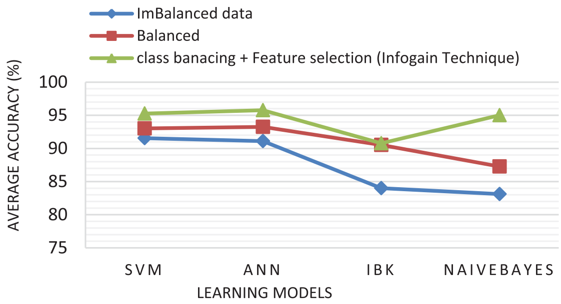

Fig. 7. Comparison of learning model accuracy for Hypertension dataset.

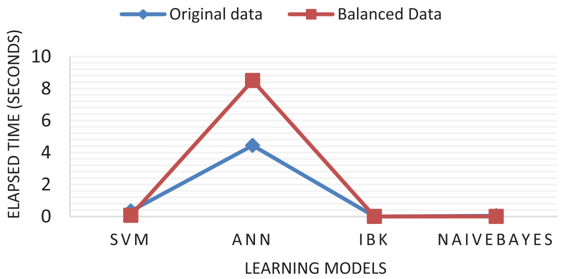

Fig. 8. Time efficiency of the Hypertension dataset.

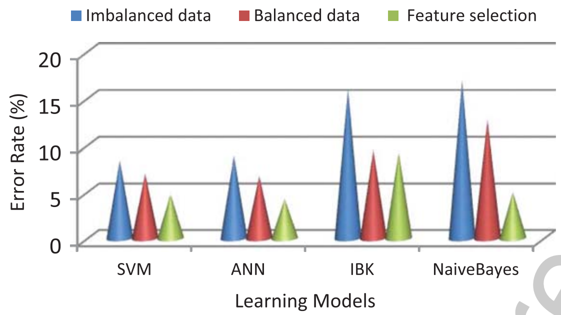

Fig. 9. Comparative analysis of the error rate of the various learning model for Hypertension dataset.

eminence of the learning system with less efficiency in dealing with minority class. This model balanced the class using sampling based approach (SMOTE). The comparative analysis report of imbalanced and balanced class proves an improved accuracy with less error rate. The disclosure of significant features using information gain, contributes to the better comprehensibility of the modeled classifier and a better inferring of the learned concept. The proposed work trains the learning model using supervised machine learning technique and attained eminent levels of prediction accuracy when equated to other techniques. The experiment outcome shows that the proposed strategy provides improved performance in medical diagnosis with very less error rate. As this system deals with medical data, it needs enhancement in handling misclassification errors. As a future work, this learning model can be streamlined by focusing misclassification patterns using optimization techniques. Moreover, personal behavioral information and past medical history can be taken into concern for prediction, so that right decision can be made at the right time as prevention plays a vital role to reduce the incidence of hypertension.

References

[1] National Institute of Medical Statistics, Indian Council of Medical Research (ICMR), "IDSP Non-Communicable Disease Risk Factors Survey, Phase-I States of India, 2007–08". National Institute of Medical Statistics and Division of Non-Communicable Diseases, Indian Council of Medical Research, New Delhi, India, 2009. [Online]. Available: http://www.icmr.nic.in/final/IDSP-NCD%20Reports/Phase-1%20States%20of%20India.pdf

[2] I. Kadi, A. Idri and J.L. Fernandez-Aleman, "Systematic mapping study of data mining–based empirical studies in cardiology". Health Informatics Journal. pp. 1–30, 2017. [Online]. Available: https://doi.org/10.1177/8081460458217717636

[3] H. Benhar, A. Idri and J.L. Fernandez-Aleman, "A Systematic Mapping Study of Data Preparation in Heart Disease Knowledge Discovery", *Journal of Medical Systems* **43**(17), 2019. [Online]. Available: https://doi.org/10.1007/s10916-018-1134-z

[4] Ministry of Home Affairs, Report on causes of deaths in India 2001–2003, Office of the Registrar General of India, Govt. of India, 2010. [Online]. Available: http://www.cghr.org/wordpress/wp-content/uploads/Causes_of_death_2001-03.pdf

[5] T.A. Gaziano, K.S. Reddy, F. Paccaud and S. Horton, "Cardiovascular Disease. Disease Control Priorities in Developing Countries". 2nd edition. Washington (DC): World Bank; Chapter 33, 2006. [Online]. Available: https://doi.org/10.1596/978-0-8213-6179-5

[6] M.W.L. Moreira, J.J.P.C. Rodrigues, V. Korotaev, J. Al-Muhtadi and N. Kumar, "A Comprehensive Review on Smart Decision Support Systems for Health Care," in IEEE Systems Journal. 2019, pp. 1–10 [Online]. Available: doi: 10.1109/JSYST.2018.289012

[7] M. Pereboom, I.J. Mulder, S.L. Verweij, R.T.M. van der Hoeven and M.L. Becker, "A clinical decision support system to improve adequate dosing of gentamicin and vancomycin", *International Journal of Medical Informatics* (2019), [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2019.01.002

[8] Florin Gorunescu, Smaranda Belciug. "Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stabilization", J. BIOMED. INFORM, vol. 49, pp. 112–118, 2014. [Online]. Available: http://dx.doi.org/10.4316/ieee.1959.3422

[9] Foster et al., "Machine learning, medical diagnosis, and biomedical engineering research – commentary", BIOMED ENG ONLINE, vol. 13(94), 2014. [Online]. Available: http://dx.doi.org/10.1186/1475-925X-13-94

[10] H. Hannah Inbarani, Ahmad Taher Azar and G. Jothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, *COMPUT METH PROG BIO* **113** (2014), 175–185. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2013.10.007

[11] J. Horsky, J. Aarts, L. Verheul, D.L. Seger, H. Van Der Sijs and D.W. Bates, "Clinical Reasoning in the Context of Active Decision Support During Medication Prescribing," *Int. J. Med. Inform.* **97** (2017), 1–11. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2016.09.004

[12] S.L. Ting, C.C. Shum, S.K. Kwok, A.H.C. Tsang and W.B. Lee, "Data mining in biomedicine: current applications and further directions for research". *J. Softw. Eng. Appl.* **2** (2009), 150–159. [Online]. Available: https://doi.org/10.4236/jsea.2009.23022

[13] S. Almuhaideb and M.E.B. Menai, "Impact of Preprocessing on Medical Data Classification". *Front. Comput. Sci.* **10** (2016), 1082–1102. [Online]. Available: https://doi.org/10.1007/s11704-016-5203-5

[14] D. Tomar and S. Agarwal, A comparison on multi-class classification methods based on least squares twin support vector machine, *KNOWL-BASED SYST* **81** (2015), 131–147. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2015.02.009

[15] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue and G. Bing, "Learning from class-imbalanced data: Review of methods and applications". *Expert Syst. Appl.* **73** (2017), 220–239. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.12.035

[16] H. He and E.A. Garcia, "Learning from imbalanced data". *IEEE Trans. Knowl. Data Eng.* **21** (2009), 1263–1284. [Online]. Available: https://doi.org/10.1109/TKDE.2008.239

[17] Stamatios-Aggelos N. Alexandropoulos, Sotiris B. Kotsiantis and Michael N. Vrahatis, "Data Preprocessing in Predictive Data Mining", *The Knowledge Engineering Review* **34** (2019), e1, 1–33, [Online]. Available: doi:10.1017/S026988891800036X

[18] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *INFORMATICA* **31** (2007), 249–268. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.9683

[19] Chen and Xu, "Least Squares Twin Support Vector Machine for Multi-Class Classification", *International Journal of Database Theory and Application* **8**(5) (2015), 65–76. [Online]. Available: http://dx.doi.org/10.14257/ijdta.2015.8.5.06

[20] D.E. Bloom and E.T. Cafiero-Fonseca, et al. Economics of Non-Communicable Diseases in India: The Costs and Returns on Investment of Interventions to Promote Healthy Living and Prevent, Treat, and Manage NCDs., World Economic Forum, Harvard School of Public Health, 2014. [Online]. Available: http://www3.weforum.org/docs/WEF_EconomicNonCommunicableDiseasesIndia_Report_2014.pdf

[21] Non communicable Diseases Progress Monitor, 2017. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.

[22] WG3 (2): "Non Communicable Diseases, Report of the Working Group on Disease Burden for 12th Five Year Plan (2012–2017)", Ministry of Health & Family Welfare, Government of India, May 2011.

[23] Indian Council of Medical Research, Public Health Foundation of India, Institute For Health Metrics And Evaluation, "India: Health Of The Nation's States – The India State-Level Disease Burden Initiative", New Delhi, India: ICMR, PHFI and IHME; 2017.

[24] R. Prakash Upadhyay, An Overview of the Burden of Non-Communicable Diseases in India, *IRANIAN JOURNAL OF PUBLIC HEALTH* **41**(3) (2012), 1–8. [Online]. Available: shttp://ijph.tums.ac.ir/index.php/ijph/article/view/2598

[25] Cardiovascular Diseases Prevention. & Treatment Program, Tamil Nadu Health Systems Project, Government of Tamil Nadu. [Online]. Available: http://www.tnhsp.org/cardiovascular-diseases-cvd-and-diabetes-prevention-and-treatment-programme

[26] Kaberi, et al.,"The 2014 Canadian Hypertesion Education Program Recommendations for Blood Pressure Measurement, Diagnosis, Assessment of Risk, Prevention, and Treatment of Hrpetension", *Canadian Jorurnal of Cardiology* **30** (2014), 485–501.

[27] F.S. Roque, P.B. Jensen and H. Schmock, et al., "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts", *PLoS Comput Biol* **7** (2011), e1002141.

[28] K. Hayrinen, K. Saranto and P. Nykanen, "Definition, Structure, Content, Use and Impacts of Electronic Health Records: A Review of The Research Literature". *Int J Med Inform.* (77) (2008), 291–304.

[29] B.W. Patterson and M.S. Pulia et.al, *"Scope and Influence of Electronic Health Record-Integrated Clinical Decision Support in the Emergency Department: A Systematic Review"*, Annals of Emergency Medicine 2019. [Online]. Available: https://doi.org/10.1016/j.annemergmed.2018.10.034

[30] M. Hajiaghaei-Keshteli and A. Mohammad Fathollahi-Fard, "A Set of Efficient Heuristics and Metaheuristics to Solve a Two-Stage Stochastic Bi-level Decision-Making Model for the Distribution Network Problem", Computers & Industrial Engineering (2018), [Online]. Available: https://doi.org/10.1016/j.cie.2018.07.009

[31] Amir Mohammad, Mostafa Hajiaghaei and Seyedali Mirjalili, "A set of efficient heuristics for a home healthcare problem", Neural Computing and Applications (2019), pp. 1–21. [Online]. Available: https://doi.org/10.1007/s00521-019-04126-8

[32] Bartosz Krawezyk, Michal Wozniak, Hypertension type classification using hierarchical ensemble of one-class classifiers for imbalanced data, *ADV INTELL SYST COMPUT* **311** (2015), 341–349. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09879-1_34

[33] Jordi Vilaplana, Francesc Solsona and Francesc Abella et al., H-PC: a cloud computing tool for supervising hypertensive patients, *J SUPERCOMPUT* **71**(2) (2015), 591–612. [Online]. Available: http://dx.doi.org/10.1007/s11227-014-1312-9

[34] S. Omboni and R. Ferrari, "The Role of Telemedicine in Hypertension Management: Focus on Blood Pressure Telemonitoring", *CURR HYPERTENS REP* **17**(4) (2015), 1–13. [Online]. Available: http://dx.doi.org/10.1007/s11906-015-0535-3

[35] Azian Azamimi, Zulkarnay Zakaria and Nur Farahiyah Mohamad, "Design and Development of Fuzzy Expert System for Diagnosis of Hypertension". 2011 Second International Conference on Intelligent Systems, Modelling and

Simulation, IEEE, pp. 113–117, 2011. [Online]. Available: http://dx.doi.org/10.1109/ISMS.2011.27

[36] A.M. Fathollahi Fard, M. Hajiaghaei-Keshteli and R. Tavakkoli-Moghaddam, "A Biobjective Green Home Health Care Routing Problem", *Journal of Cleaner Production* (2018), [Online]. Available: 10.1016/j.jclepro.2018.07.258.

[37] Paolo Melillo, Paolo Scala, Filippo Crispino and Leandro Pecchia, Cloud-Based Remote Processing and Data-Mining Platform for Automatic Risk Assessment in Hypertensive Patients, *LECT NOTES COMPUT SCI* **8868** (2014), 155–162. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13105-4_24

[38] D. Jiang, et al., "An Approach for Prediction of Acute Hypotensive Episodes via the Hilbert-Huang Transform and Multiple Genetic Programming Classifier", *INT J DISTRIB SENS N*, **501** (2015), 354807. [Online]. Available: http://dx.doi.org/10.1155/2015/354807

[39] A.M. Fathollahi-Fard, M. Hajiaghaei-Keshteli and R. Tavakkoli-Moghaddam, "A Lagrangian relaxation-based algorithm to solve a home health care routing", *International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics* **31**(10) (2018), 1734–1740.

[40] Bahadori-Chinibelagh, Fathollahi-Fard and Hajiaghaei-Keshteli, "Two Constructive Algorithms to Address a Multi-Depot Home Healthcare Routing Problem", *IETE Journal of Research* (2019), [Online]. Available: 10.1080/03772063.2019.1642802

[41] J. Wu, et al., A new approach to kinematic feature extraction from the human right ventricle for classification of hypertension: A feasibility study, *PHYS MED BIOL* **57** (2012), 7905. [Online]. Available: http://dx.doi.org/10.1088/0031-9155/57/23/7905

[42] Jing Xu, Jia Wu, Bahram Notghi, Marc Simon and John C. Brigham, A. Feasibility Study on Kinematic Feature Extraction from the Human Interventricular Septum toward Hypertension Classification, *LECT NOTES COMPUT SCI* **8641** (2014), 36–47. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09994-1_4

[43] Sarah Melville and James Brian Byrd, "Personalized Medicine and the Treatment of Hypertension", Current Hypertension Reports (2019) 21:13, [Online]. Available: https://doi.org/10.1007/s11906-019-0921-3

[44] Fernando Lopez-Martinez, Aron Schwarcz. MD Edward Rolando Nunez-Valdez and Vicente Garcia-Diaz, Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors, Expert Systems With Applications (2018), [Online]. Available: 10.1016/j.eswa.2018.06.006

[45] Muhammad Fazal Ijaz, Ganjar Alfian, Muhammad Syafrudin and Jongtae Rhee, "Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest", *Applied Sciences* **8** (2018), 1325; [Online]. Available: 10.3390/app8081325

[46] Goli Arji and Reza Safdari et al.," A systematic literature review and classification of knowledge discovery in traditional medicine", *Computer Methods and Programs in Biomedicine* **168** (2019), 39–57, [Online]. Available: https://doi.org/10.1016/j.cmpb.2018.10.017

[47] Chayakrit Krittanawong, Andrew S. Bomback, Usman Baber, Sripal Bangalore, Franz H. Messerli and W.H. Wilson Tang, "Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension", *Current Hypertension Reports* (2018) 20:75, [Online]. Available: https://doi.org/10.1007/s11906-018-0875-x

[48] LaFreniere, Daniel & Zulkernine, Farhana & Barber, David & Martin, Ken, Using machine learning to predict hypertension from a clinical dataset. *IEEE Symposium Series on Computational Intelligence (SSCI)* (2016), pp. 1–7. [Online]. Available: 10.1109/SSCI.2016.7849886

[49] E.I. Georga et al., *Artificial Intelligence and Data Mining Methods for Cardiovascular Risk Prediction. In: Golemati S., Nikita K. (eds) Cardiovascular Computing—Methodologies and Clinical Applications. Series in BioEngineering*, Springer, Singapore, 2019.

[50] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang and Guoyan Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review", *Journal of Healthcare Engineering*, Volume 2018, Article ID 4302425, 9 pages, [Online]. Available: https://doi.org/10.1155/2018/4302425

[51] Amir Mohammad Fathollahi-Fard, Mostafa Hajiaghaei-Keshteli and Seyedali Mirjalili, "Hybrid Optimizers to Solve a Tri-Level Programming Model for a Tireclosed-Loop Supply Chain Network Design Problem", *Applied Soft Computing* **70** (2018), 701–722, [Online]. Available: https://doi.org/10.1016/j.asoc.2018.06.021

[52] Anita Abdi, Andisheh Abdi and Amir Mohammad et al., "A Set of Calibrated Metaheuristics to Address a Closed-Loop Supply Chain Network Design Problem Under Uncertainty", *International Journal of Systems Science: Operations & Logistics* 2019, [Online]. Available: 10.1080/23302674.2019.1610197

[53] A.M. Fathollahi-Fard, M. Hajiaghaei-Keshteli and S. Mirjalili, "Multi-Objective Stochastic Closed-Loop Supply Chain Network Design with Social Considerations", *Applied Soft Computing Journal* (2018), [Online]. Available: https://doi.org/10.1016/j.asoc.2018.07.025

[54] Y. Fu, G. Tian and A.M. Fathollafi-Fard, et al., "Stochastic Multiobjective Modelling and Optimization of an Energy-Conscious Distributed Permutation Flow Shop Scheduling Problem with the Total Tardiness Constraint", Journal of Cleaner Production (2019), [Online]. Available: https://doi.org/10.1016/j.jclepro.2019.04.046

[55] Mojgan, Amir Mohammad et al. "A Multi-Objective Supplier Selection and Order Allocation Through Incremental Discount in a Fuzzy Environment", *Journal of Intelligent & Fuzzy Systems* **37**(1) (2019), 1435–1455. [Online]. Available: 10.3233/JIFS-182843.

[56] Bruha and F. Franek, Comparison Of Various Routines For Unknown Attribute Value Processing: Covering Paradigm. *Int J Pattern Recognition and Artificial Intelligence* **10**(8) (1996), 939–955. [Online]. Available: http://dx.doi.org/10.1142/S0218001496000530

[57] Jerzy W. Grzymala-Busse and Ming Hu, A Comparison of Several Approaches to Missing Attribute Values in Data Mining, *Lecture Notes in Artificial Intelligence* **2005** (2001), 378–385. [Online]. Available: http://dx.doi.org/10.1007/3-540-45554-x_46

[58] S.B. Kotsiantis, D. Kanellopoulos and P.E. Pintelas, "Data Preprocessing for Supervised Leaning", *International Journal of Computer Science* **1**(2) (2006), 111–117. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.5127&rep=rep1&type=pdf

[59] S. Ramaswamy, R. Rastogi and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets", Proceedings of the 2000 ACM SIGMOD international conference on Management of data – SIGMOD '00. pp. 427, 2000. [Online]. Available: http://dx.doi.org/10.1145/342009.335437

[60] Firas Ajil Jassim, "Image Denoising Using Interquartile Range Filter with Local Averaging", *International Journal of Soft Computing and Engineering* **2**(6) (2013). [Online]. Available: http://arxiv.org/ftp/arxiv/papers/1302/1302.1007.pdf

[61] Dragan Gamberger, Nada Lavrac and Saso Dzeroski, "Noise Detection and Elimination in Data Preprocessing: Experiments in Medical Domains", *Applied Artificial Intelligence: An International Journal* **14**(2) (2000), 205–223. [Online]. Available: http://dx.doi.org/10.1080/088395100117124

[62] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research* **16** (2002), 321–357. [Online]. Available: http://arxiv.org/abs/1106.1813

[63] Gao, Tianxiang, "Hybrid Classification Approach for Imbalanced Datasets" (2015). Graduate Theses and Dissertations. Paper 14331. [Online]. Available: http://lib.dr.iastate.edu/etd/14331/

[64] Mosley, Lawrence, "A Balanced Approach to the Multi-Class Imbalance Problem", Graduate Theses and Dissertations, Paper 13537, 2013. [Online]. Available: http://lib.dr.iastate.edu/etd/13537

[65] H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE T Knowl Data En* **17**(4) (2005), 491–502. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2005.66

[66] Mark A. Hall, Geoffrey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", *IEEE Transactions On Knowledge And Data Engineering* **15**(3) (2003), 1–16. [Online]. Available: http://dx.doi.org/10.1109/tkde.2003.1245283

[67] Chao-Ton Su, Chien-Hsin Yang, Feature selection for the SVM: An Application to Hypertension diagnosis, *Expert Systems with Applications* **34** (2008), 754–763. Online]. Available: http://dx.doi.org/10.1016/j.eswa.2006.10.010

[68] Ruxandra Stoean and Catalin Stoean, "Modeling Medical Decision Making by Support Vector Machines, Explaining By Rules Of Evolutionary Algorithms With Feature Selection", *EXPERT SYST APPL* **40** (2013), 2677–2686, [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.11.007

[69] Xianli Pan and Yitian Xu," Two Effective Sample Selection Methods for Support Vector Machine", *Journal of Intelligent & Fuzzy Systems* **30** (2016) 659–670, [Online]. Available: 10.3233/IFS-151785.

[70] Manjeevan Seera and Chee Peng Lim, A Hybrid Intelligent System for Medical Data Classification, Expert Systems with Applications, **41** (2014), 2239–2249. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.02.028

[71] Nazri Mohd Nawi, Walid Hasen Atomi and M.Z. Rehman, The effect of data pre-processing on optimized training of artificial neural networks, *PROCEDIA TECHNOLOGY* **11** (2013), 32–39. [Online]. Available: http://dx.doi.org/10.1016/j.protcy.2013.12.159

[72] M. Sokolova and G. Lapalme, A systematic analysis of performance measures for classification tasks, *Information Processing and Management* **45** (2009), 427–437.

[73] M. Hossin and M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process* **5**(2) (2015).