Science in Information Technology Letters
Vol. 1., No. 1, May 2020, pp. 24-39
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Healthcare analytics by engaging machine learning

Pragathi Penikalapati [a,1,*], A Nagaraja Rao [b,2]

[a] Research Scholar, School of Computer Science and Engineering, VIT University, Vellore, India
[b] Associate Professor, School of Computer Science and Engineering, VIT University, Vellore, India
[1] pragathipradeep.ch@gmail.com; [2] anagarajaraoa@vit.ac.in;
* corresponding author

ARTICLE INFO

Article history
Received June 1, 2019
Revised June 12, 2019
Accepted November 29, 2019

Keywords
Electronic Health Records
Feature Selection
Machine Learning
Classification Models
Clustering Models

ABSTRACT

Precise prediction of chronic diseases is the very basis of all healthcare informatics. Early diagnosis of the disease is crucial in delivering any healthcare service. The modern times witness our general vulnerability to several health disorders due to a stressful lifestyle causing anxiety and depression, or susceptibility to hypertension and diabetics or major diseases such as cancer or cardiovascular ailments. Hence, we should undergo periodic screening and diagnostic tests for such possible disorders to lead healthy lives. In this context, Machine Learning technology can play a pivotal role in developing Electronic Health Records (EHR) for implementing quick and comprehensively automated procedures in disease detection among the at-risk individuals at an early stage, so that accelerated processes of referral, counseling, and treatment can be initiated. The scope of the current paper is to survey the utilization of feature selection and techniques of Machine Learning, such as Classification and Clustering in the specific context of disease diagnosis and early prediction. This paper purposes of identifying the best models of Machine Learning duly supported by their performance indices, utility aspects, constraints, and critical issues in the specific context of their effective application in healthcare analytics for the benefit of practitioners and researchers.

1. Introduction

Chronic diseases are those that persist for a very long time, and their early diagnosis is certainly a dire necessity in the medical field. Some of the very common chronic diseases are diabetics [1], depression [2], cancer [3], strokes, hepatitis C, arthritis, and cardiovascular diseases. The primary reason for the growing adverse health issues can be attributed mostly to urban living. As per the available data from the United Nations, 54% of earth's population dwell in urban locations, which in all possibility may rise to a whopping 66% by the year 2050, leading to escalating levels of chronic diseases [4]. Government reports point to the fact that more than 50% of the population suffers from one or more chronic diseases, and about 80% of the people spend on the treatment of chronic diseases amounting to billions. This underlines the impact of

chronic diseases on the world population. The Chinese reports on nutrition and chronic diseases in 2015 reveal that the main cause of death there is the presence of chronic diseases to a stunning 86.6% [5]. Hence, it is highly required to risk-assess chronic diseases. Preventive measures and effective treatment processes can be initiated if these chronic diseases are identified at an early stage, reducing the possibility of mortality among the patients [6].

As most of the patient's history and health status are registered through the Electronic Health Records (EHR), analytics solutions emerge by way of Machine Learning, using the extensive health data available. Analytics is a method of developing insights by effectively interpreting data and employing qualitative as well as quantitative analysis [7]. As per the modest US estimations, healthcare services can save up to $450 billion a year by applying Machine Learning [8] to the clinical data. In the context of the huge generated healthcare data, the major challenges would be the collection and effective use of this data through analysis and increase efficiency in prediction and treatment. Early diagnosis and prevention of the disease is the current approach to healthcare issues than being compelled to go for treatment after a delayed diagnosis. In the conventional approach, the doctors or the practitioners employ risk calculators for the estimation of disease development possibility. These calculators make use of the basic information regarding routine of life, medical condition, demographics, and so on for a computational assessment of the possibility of getting a particular disease. Mathematical tools and methods based on equations are used for such computations.

The transfer approach of the EHR model based on ML facilitates the application of predictive models over diverse EHR systems. These models can be made adaptable for using datasets belonging to one EHR for the prediction of the results of specific other systems [9]. Such systems are basically heterogeneous in nature, containing both unstructured and structured data sources, including text, images, medical imaging, and the like [10]. Though the storage of such data does not cause any concern, deployment of this data for any sort of analysis or prediction becomes difficult owing to the inconsistent formats. Even though the data emerge from diverse sources and in multiple systems, the technologies of Machine Learning including optical character recognition, image processing, and natural language processing can assist in the transformation of this heterogeneous data into a uniform format

Application of Machine Learning can be very beneficial or the formulation of a useful model capable of processing diverse predictor variables and come out with accurate disease prediction followed by a designing of a fast, dependable and automated mechanism of screening for the identification of at-risk individual and duly refer him to the concerned clinicians. This survey mainly aims at summarization of all the peer-reviewed and published literature about the theories of clustering model classification. This study also attempts to classify the types of literature by categories of diseases and the types of Machine Learning applications for disease prediction. The organization of the paper is as follows: Section II describes Machine Learning in all its details, including its features. Section III is earmarked for dealing with diverse Machine Learning models in the specific context of healthcare data, and Section IV concludes the paper with current challenges and future possibilities of extended applications.

2. Method

2.1. Machine Learning Algorithms for Health Care

The ease in the applications of Machine Learning is in their flexibility to facilitate the development of such models capable of quick analysis of data to deliver accurate and fast results considering the real-time data as well as the previous history. The providers of healthcare services are now better equipped with the induction of Machine Learning for making better and faster diagnostic decisions and opinions and processes of treatment for enhanced delivery of general healthcare services to the patients [7], [10]. Earlier, in the absence of proper tools or technologies, healthcare professionals had to face several challenges in handling the extensive

volume of data in the stages of gathering, analyzing, and making effective predictions for appropriate treatment procedures. Prediction of risk to human health can be understood as the method of supervised learning in the context of Machine Learning, where the input value is depicted as the patient's attribute value, that includes all the personal information of the patient such as gender, age, the prevailing symptoms, lifestyle, smoking and drinking habits and other available both structured and unstructured data. The output value represented as C indicates the possibility of the patient belonging to the high-risk group of a particular disease.

$R = \{R_0, R_1\}$, where, R0, which represents that the patient belongs to the high-risk category with reference to cerebral infarction while R1 denotes that the patient belongs to the low-risk category of the same disease.

A prediction model f(x, φ): y is a function that maps a set of input variables x into a response variable y with a set of parameters φ. The perception of training Machine Learning model is to identify appropriate values of parameters that facilitate optimization of spresific criteria such as to maximize the accuracy in prediction, to minimize the size of the model, to minimize the complexity of time in the forecast, and to finalize the comprehensibility of the model. Predefinition of data in numerical formats such as feature vectors is a primary requisite for the training of a Machine Learning model. The pre-processing stage [2] can accomplish this redefinition of data.

As per the indications in the literature pertaining to medical data mining, several researchers use diverse classifier models for the prediction of chronic diseases to obtain accuracy in forecasting and valid diagnostic results. They have been employing several approaches from among a wide range of available classifier models, such as decision trees, naive Bayes, SVMs, and neural networks for chronic disease [6] prediction as well as diagnosis. Fig. 1 illustrates the application of the classification process on the processed data for obtaining predictive results.

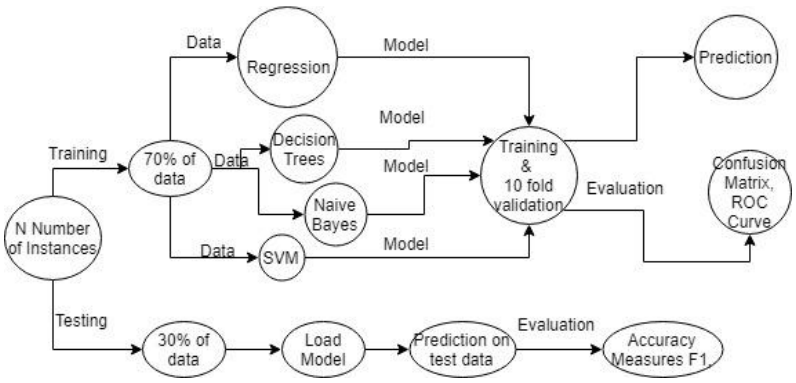

Fig. 1. Workflow of Machine Learning models employed for predicting chronic disease

2.2. Supervised Models with Cronic Disease Usecases

The algorithms in the supervised learning are represented along with a cluster of training examples, commonly referred to as instances. An instance is a combinational pair of an input and an output value. Normally, the input is depicted as an array that defines the instance in terms of the numerical format. This array is typically termed as the feature vector. The output is the desired value of prediction, such as a real number or a category. Normally, the output is termed as 'class' for the general purpose of classification. The algorithms strive to map duly considering the input and output values during the training period. After the accomplishment of training, the algorithm becomes capable of predicting unknown output values from the given set of new and unknown input values. The performance efficacy of the algorithm can be established by 'hiding' the output values of the instances and test the algorithm to predict them. Then a comparison between the real output, also known as 'ground truth' and the prediction of

the algorithm, can be made. Classification is when the to-be-predicted variable is categorical, and regression is when the output prediction is quantitative [11]. Though it is not feasible to exhibit every supervised and unsupervised model, this paper attempts to convey popular models with Chronic Disease Use cases.

2.2.1 Bayesian Probabilistic model

In general, the representation of a Bayesian network could be through the following three indispensable components:

i) Every individual node is associated with a random variable that could either be discrete or continuous.
ii) The pairs of nodes are connected through either arrows or a set of directed links indicating a direct dependency existing among the linked variables, with the directions of the arcs signifying the directions of the influence;
iii) Every individual node Xi contains a conditional probability distribution P(Xi | Parents(Xi)) for quantification of the strengths regarding these influences.

Let us consider one ordinary case for medical diagnosis with the target variable as the presence and risk of cardiovascular disease. Here, the influential aspects are the standard clinical metrics such as the levels of cholesterol, blood pressure, blood glucose and index of body mass, which can be impacted by certain socio-demographic aspects such as gender, age, the behavior of health including patterns of diet, levels of exercise and the medical history of the family. Fig. 2 depicts the illustration of this example by a Bayesian network.

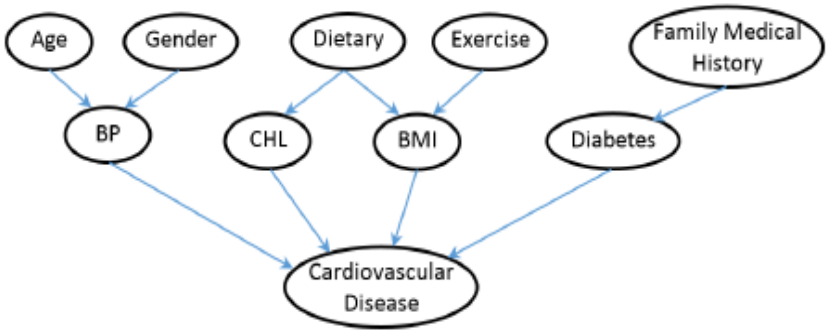

Fig. 2. Example BN of Medical Diagnosis Domain

Now, consider a formal introduction of some basic notations, which are generally employed in BNs. In a Bayesian network with n discrete-valued nodes $X = \{X_1, X_2, \ldots \mathrm{Xn}\}$, the structure is denoted as g, the parameters by $\theta_{ijk}(i \in \{1,2,\ldots,n\}, j\{1,2,\ldots,\mathrm{v}_i\}, k \in \{1,2,\ldots,r_i\})$, which means the conditional probability of Xi to take its kth value given the jth configuration of its parents i.e. $P(\mathrm{X}_i^k \mid pa_i^j)$. ri denotes the number of states of the discrete variable Xi; Pa(Xi) represents the parent set of node Xi; the number of configurations of Pa(Xi) is $v_i = \Pi x_t \in pa(X_i)^{r_t}$. The basic foundation for most of the mathematical properties and methods of perceptions and inferences in all the Bayesian networks is the principal conditional independence, which is implied by the Bayesian networks. The following is the further simplified version of the general chain rule as expressed in the BNs: $P(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i \mid pa(x_i))$

Thus, the BNs can precisely represent any comprehensive combined probability distribution based on the fusion of its structure (topology) and the other metrics such as conditional distributions. The employment of independence of casual influence (ICI) can further simplify all such factored representations. This ICI further reduces the feature space dimension and thus

relaxes the constraints of requirement on the number of training samples and enhances the accuracy in the prediction of the learned model.

If examples of medical diagnoses are considered with an added dimension {time}, it can be found that nearly all the attributes of a person transform over a period of time, such as the levels of cholesterol, blood pressure, index of body mass, and patterns of diet. Any model capable of representing the distribution of states over a period of time can make the optimum utilization of the data, thus generated by the considered dynamic processes. Dynamic Bayesian network (DBN) could be one very successful instance of this. The dynamic process is modeled by DBNs in such a way that the time is divided into diverse slots of time at a fixed rate and then, represents the probabilistic state transitions between diverse slots of time with a fragment of a Bayesian network, which permits intra-time-slice arcs (indicating the conditional dependencies existing among the values of the variables at the same instant) and inter-time-slice arcs (enabling the jumping through single or multiple slices of time pointing to the conditional dependencies existing among the values of the variables over time). DBNs are capable of bypassing the learning problem of NP-hard structure by just permitting provisions for edges between time slices. But, the true efficacy of the DBNs is in their potential to handle the temporal processes. CVD has been an established and primary reason for the growing mortality globally, with more than 17.3 million annual deaths, which is certainly more than the number of cumulative factors contributing to death. According to the available reports for the year 2013, more than 31 percent of deaths around the world can be attributed to CVDs. It is also an established factor that the development and impact of CVDs can drastically be prevented if some lifestyle modifications and interventions of healthy practices are made. Consequently, keeping this in view, Lung and Blood Institute has designed the Coronary Artery Risk Development in Young Adults (CARDIA) Study. This study is longitudinal in nature, where the data of 25 years has been considered (from 1985) by obtaining the data of the participants from their early adulthood and monitoring their physical condition all through the period of experimentation for over three decades.

The authors in [1] have developed and implemented a sequential set of experiments to model the advancement of the levels of Coronary Artery Calcification (CAC), which established itself as the most dependable predictive approach for the identification of subclinical Coronary Artery Disease (CAD) by using the data in CARDIA. A transitory model capable of easy interpretation of the longitudinal training data would become necessary to identify the risk factors on the advancement of CVD much later in adult life. As the participants' revisits could be very regular, and as a result of the most uncertain characteristics of the medical data, there is a dire need for a discrete-time temporal probabilistic model. Hence is the authors' choice of Dynamic Bayesian Networks (DBNs).

The clinical parameters, such as the levels of blood pressure, blood glucose, and cholesterol, have a direct relationship with the level of CAC. Consequently, only the fundamental socio-demographic and health behavior data has been considered for a more precise demonstration of the correlation that exists between diverse factors of lifestyle and the development of CVD. It is to eliminate the possibility of the interference of the additional irrelevant features of information with the training of the model. It means that these DBNs are trained only with the provided non-clinical data in the prediction of the occurrence of abnormality in the CAC level corresponding to the aging factor of the individual from an early age to middle age. The results of the experiments demonstrate that these features of behavior can be moderately predictive about the occurrence of enhanced levels of CAC. They provide insights to the doctors and permit them to identify the changes in lifestyle and behavior, which could drastically minimize the risks in cardiovascular ailments. Moreover, this probabilistic model, with its interpretative nature facilitates all likely interactions among the experts in the domain, such as the physicians on one side and the model on the other, and further furnishes them a simple method of modification/refinement of the model based on their expertise/experience.

2.2.2 Random Forests

A vast number of algorithms of supervised learning have been evolving to furnish adequate flexibility for the minimization of training error, simultaneously allowing generalization regarding the new data sets in a computationally efficient manner. The authors in [2] intend to highlight one such approach, namely, random forests as a highly effective instance of the innovative algorithm. This random forest algorithm has been in existence for over 15 years. It has been acclaimed as one of the most effective "off-the-shelf" available algorithms for classification. As is self-evident from the name itself, this algorithm is developed from trees, to be precise from the decision trees. Let it be assumed that the objective is the bifurcation and classification of individuals as responders or non-responders to statin. Here, the authors consider a cluster of training examples with established responders and non-responders to statin each with distinguishable by a set of attributes as sex, age status of diabetics or smoking, etc. In several instances, we encounter with thousands of available features. The authors have constructed a series or ensemble of decision trees which make use of these predictive attributes for the discrimination between two distinct groups. One individual attribute capable of efficiently accomplishing this split is selected at every individual node of each tree. A single variable may not be adequate for achieving the task of perfect separation, and there may be an additional requirement of subsequent nodes. One significant difference among the trees is that each tree can access only one subset of considered features, which is generally understood as 'bagging'. Besides, only one subset of attributes is taken into consideration at each node. The resultant stochasticity permits every individual tree to cast an independent vote favoring a final classification and thereby performs the role of regularization. Every single tree may not always be very accurate. But the collective final majority vote when hundreds of trees are considered is amazingly accurate.

Machine learning competitions have been extensively using the approaches of random forests with considerable success in a wide range of learning disciplines. Ishwaran, Lauer, and colleagues have adapted these approaches of random forests for the computational analysis of the data of survival and termed their methods aptly as "random survival forests (RSF)" [12]. They have adopted a binary variable for the computation of death and applied their approach to a wide range of problems, which include the prediction of survival in conditions of systolic heart failure and instances of women in postmenopausal conditions. Subsequently, they have considered 33,144 women in the Women's Health Initiative Trials and tried the conventional demographic and clinical variables along with 477 ECG biomarkers. They have made use of RSF to develop a survival model and have identified 20 predictive variables for determining the long-term mortality, which further includes 14 ECG biomarkers. Such models developed through the reduction of features in the subset have exhibited enhanced performance efficacy on raining data as well as on a held-out test set. It is interesting to note here that after the selection of the subset with 20 variables, an ordinary additive model (a regularized and improvised variation of the Cox proportional hazards model) has performed as good as RSF as far as the classification of the patient is concerned. It suggests that the chief merits of the RSF include the selection of features. In reality, most of these variables never have been involved in the mortality predictions previously.

2.2.3 Support Vector Machine

A support vector machine (SVM) can be taken as a category of the model generally adopted for data analysis and pattern-identification in the analysis of classification and regression. SVM becomes extremely useful when the data contains precisely two categories. It classifies the data by identifying the best hyperplane, which segregates all data points of one category from the data points of the other category. The efficacy of the model enhances corresponding to the increased margin between these two categories. A margin can not contain any points located in its interior region. The support vectors are here can be taken as the data points located on the

threshold of the margin. The SVMs can be implemented based on mathematical functions and are employed for modeling real-world complex problems.

For instance, take a task of binary classification task containing a set of linearly separable training samples $E = \{(x_1, y_1)...(x_m, y_m)\}$.

where $x \in R^d$, i.e., x lies in a d-dimensional input space, and yi is the class label such that yi ∈ {-1, 1}. This specific label depicts the category of the data to which it belongs. Then, an appropriate function of discrimination could be defined as: f(x) = sgn((w.x) + b).

Where vector w determines the orientation of a discriminant plane (or hyperplane), (w.x) is the inner product of the vectors, w and x, and b is the bias or offset. There are n finite possible planes that could accurately classify the training data.

SVMs map the data of training into specific kernel space. There are several kernel spaces such as Linear, Quadratic, Polynomial, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. used for distinct purposes. Besides, SVMs could be implemented through multiple methods such as minimal sequential optimization, quadratic programming, and least squares. The challenges present in the context of SVMs are mainly the selection of kernel and selection of methods in such a way that the proposed models can neither be over-optimistic or unduly pessimistic.

At present, CVD is the one most significant single cause accountable for the high rate of mortality, even in developed countries. Consequently, it causes a heavy financial burden on the healthcare industry and health services. Besides, it has been predicted that the prevalence of CVD would be continuously on the rise in the immediate future decades. Extensive research studies and findings in the domain of statistical and clinical studies have detected many contributory factors which enhance the CVD risk. It is always essential to identify patients of high-risk categories to prevent the future occurrence of CVDs. Factors such as gender, advanced age, and history of the family regarding CVD have been identified as contributory risk factors, about which science can do almost nothing. The other significant risk factors include high levels of cholesterol and blood pressure, smoking, obesity, diabetics, and left ventricular hypertrophy [13].

The authors have investigated 134 subjects in [13] recruited from a hypertension clinic based in South-east London with comprehensive waveform data of Digital Volume Pulse (DVP). A straightforward approach of measurement at the fingertip of the patient through DVP, placing a detector of infra-red light absorption on the index finger of the patient suffices the requirements for the prediction of his CVD risks. Appropriate attributes are extracted from the waveform, and an SVM classifier is established to have predicted with over 85% of accuracy.

2.2.4 Neural Networks

Artificial Intelligence (AI) is a broad spectrum containing several subfields of which Neural Networks (ANNs) is one prominent area. ANNs have become indispensable in diverse applications as a robust tool due to their inherent capability to correlate input and respective output data based on vector mapping. They have been established to be quite effective and useful techniques in their applications in diverse medical fields in real-time clinical practices, including oncology, cardiology, pathology, urology, radiology, endocrinology, pneumology, pediatrics, and pediatric surgery. The domain of medicine is a fertile area for testing, trying, and establishing ANN as a robust application tool for enhancing the existing medical practices and techniques.

This study has been made by [14] on the evolution and evaluation of the ANN models for pattern recognition based on both Multi-Layer Perceptrons (MLPs) and Probabilistic Neural Networks (PNNs) and further on the actual application of these approaches to the issues of prediction of osteoporosis risk factor in the Greek population. To be more precise, a support

tool for making decisions has been created to assist the clinicians for the identification of probable people at enhanced risk levels in terms of Osteoporosis and to advise such people towards further investigations with bone densitometry. This area of application has enormous importance as the early identification of the problem of Osteoporosis can be vital in preventing osteoporotic fractures, which are generally associated with an enhanced rate of mortality and the involved increased socio-economic costs.

Let us consider the very nature of Osteoporosis, which is a sort of bone-disease leading to a decreased mass of bone density and the alteration of its micro-architecture structure in such a way the tolerance of bone is drastically reduced enormously increasing the risk of fracture. In addition to the involved direct physical implications such as fractures, unendurable pain, and physical inconvenience, osteoporotic fractures in the regions of the hip or the spine have been known factors of morbidity and mortality in several instances. As per the findings of several interesting studies in the area, every two out of five people over the age of 75 years, who develop a hip fracture, have a strong possibility of suffering and eventual death in about a year as a direct consequence. There are several intricate health and practical issues related to both immediate and the long term effects caused by such fractures such as hospitalization, continuous dependence on support both at home and the institutions, etc. Studies indicate that in the European Union, on average, at least one person breaks bones due to Osteoporosis in fifteen seconds. In most of the cases, the broken bones are the apparent initial symptoms of Osteoporosis, as a result of which the condition is mostly known as "the silent crippler". It is because most of the affected people do not realize until it is very late or beyond any possible remedial measures that they have Osteoporosis. However, it is quite possible to minimize the fracture risk in an individual enormously, if Osteoporosis has been identified early and adequate treatment is initiated.

The data of Osteoporosis utilized at the designing of the models of ANN have been procured from the Orthopaedic Clinical Information System of Alexandroupolis' University Hospital, Greece. Four specific parameters, such as age, sex, height, and weight, have been considered for every individual case. The basis for the computation of risk factors of an osteoporosis risk factor is the T-score value, which is a comparative bone density value of the patient against a normal healthy adult of a similar age group. The current study considers the data sets obtained from 3426 cases, which is further divided into sets of 2426 and 1000 records. The first set has been taken up for the training of the MLPs besides the PNN construction. In contrast, the other set is considered for performance testing and evaluation of the neural networks, the outcome of which is depicted in [14].

The standard performance parameters for issues related to classification are:

a. Accuracy: Quantification of instances, precisely classified.
b. Sensitivity: A true positive rate, also known as recall, depicting the quantity and ratio of correctly classified positives.
c. Specificity: A true negative rate reflecting the quantity and ratio of the correctly classified negatives.
d. Precision: A positive predictive value depicting the fraction of true positives from among the possible classified positives

$$precision(p) = \frac{truepositives(tp)}{\text{truepositives(tp)} + \text{falsepositives(fp)}}$$

F1 score is A weighted average measure of precision and recall.

2.3 Unsupervised Learning with Usecase

As different from the supervised learning, the value of output variable y is unknown in the unsupervised learning during the period of training. Consequently, the task to be accomplished would be to discover diverse categories naturally arising out of the real similarities in the input

data, such as identifying groups of similar users. Clustering algorithms in unsupervised learning are employed mainly for the identification of either groups or hierarchies present within the data. At times clustering can also be used for pre-processing stages in supervised learning. K-means, DBSCAN k-medoids, and hierarchical clustering are some prevalent clustering algorithms [15]. More details on both supervised and unsupervised learning can be obtained from [16]. It is possible to employ unsupervised learning also in the pre-processing stages before the employment of supervised approaches.

There are no tested and tried therapies in existence for the extremely different condition of the ailment of heart failure in combination with preserved ejection fraction (HFpEF). One contributing factor to this absence of adequate and required clinical procedures in HFpEF could be the fact that the patients who have been enrolled for this ailment also reflect several interrelated overriding pathophysiologic processes, all of which may reflect diverse reactions to the same agent. It is required to identify such processes. The use of genetics has been advocated to redefine the disease accurately. But, complex conditions such as HEpEF cannot be appropriately classified through the use of genetic variations, as in all possibility, multiple feeble genetic factors unpredictably interact among themselves, making the eliciting of disease phenotype inconsistent. The authors have focussed in [2] on employing the approaches of unsupervised learning for the classification of HFpEF patients. As has been discussed earlier, the unsupervised learning intends to identify the internal structure of the data. Initially, it commences with the framework almost identical to that of supervised learning, considering appropriate illustrations, for instance, patients here, each of which is characterized by a feature vector. Here, the values are provided with specific features such as age, sex, and height, which can be well reflected through a matrix (Fig. 3). But, we have used this matrix to identify a group of similar patients instead of using it to comprehend a model where features correspond to the outcomes. It can be accomplished by using several algorithms of which agglomerative hierarchical clustering could probably be the simplest. This algorithm performs the first groups the most similar individuals together, and then combine them into the same pairs. Yet another category of unsupervised learning algorithm contains the aspects of chief component analysis and factorization of non–negative matrix, which can decompose a matrix to convert the matrix of patients' features into a result of two matrices, one binds together the similar attributes into super-attributes, known as dimensionality reduction and the other defining every patient in terms of a vector of weight applicable to these super attributes. Subsequently, based on the similarity of weight vectors, the patients are grouped.

One more set of approaches of unsupervised learning is in existence, such as k-medoids clustering and the algorithm of attractor metagenes, which attempt to identify different examples of training or a composite as the nucleus around which the other instances of data are clustered. Intra-cluster illustrations should be more similar than the inter-cluster ones.

Fig. 3. ML Overview

Based on the HFpEF analysis as executed in [2], the authors intend to group the patients considering the quantitative clinical and echocardiographic variables. The authors have started with 67 distinctly different features in [2] and eliminated all possibly correlated features to end up with just 46 predictors with minimal redundancy (A in Fig. 4). Then, they have employed a model-based clustering in its regularised form, further using the multivariate Gaussian distributions to define every individual patient cluster taking into consideration the means and standard deviation which have been allotted to each individual feature. In order to accomplish parsimony, regularization has been employed for the selection of the optimal number of clusters of patients besides the number of free parameters appropriate for defining each individual cluster (B in Fig. 4). The allotment of patients to different clusters is based on the computation of a combined probability existing across all the attributes, and the selection of such clusters containing the maximum membership probability of every patient. A comparative analysis of the resultant groups established considerable diversity across the broad range of phenotypic variables. In the same way as the BCC prize winner, the authors have employed the clusters of phenotype as attributes in a model of supervised learning to predict the survival of patients with HFpEF. They have detected a vast improvement in them on the application of standard clinical models for the assessment of risk in our training set as well as an independent test set.

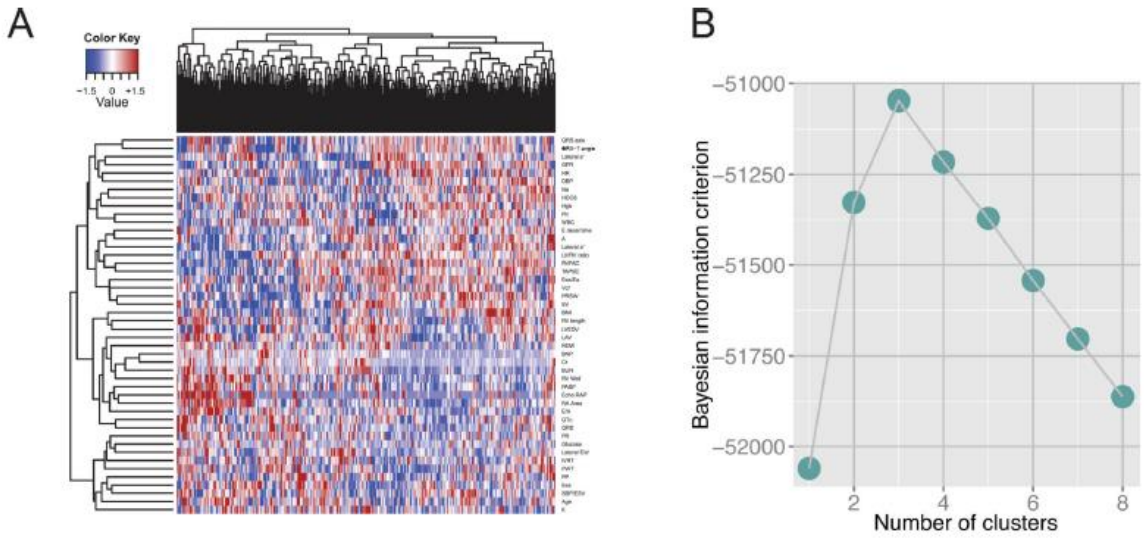

Fig. 4. Application of unsupervised learning to HFpEF

Based on Fig. 4, the part of A is the Phenotype heat map of HFpEF. The columns depict individual study participants and the rows of the individual features. The part of B is a criterion analysis of Bayesian information to identify the optimal number of phenotypic clusters (pheno-groups).

2.3 Learning Features

Data mining extensively makes use of the data processing technique of Feature selection or Variable Selection [17], primarily for data reduction through the process of eliminating redundant, superfluous and insignificant attributes from identified datasets [18]. Besides, this technique contributes to enhanced data comprehensibility and prediction performance and facilitates data visualization while decreasing the learning algorithm's training time. Eliminating inconsistent and noisy data before the application of any model to the data is desirable for obtaining fast and accurate results. A data set's dimensionality reduction is a prominent principle in implementing true-to-life applications. Besides, an exclusive selection of important features can result in complexity reduction exponentially. In recent years, the application of several approaches for feature identification on healthcare datasets can be perceived to obtain

more valuable and valid information. Numerous chronic diseases such as hypertension, diabetics [19], cardiovascular disorders, and thalassemia, can be effectively predicted through employing this method of feature selection on clinical databases.

The functional efficacy of diverse learning algorithms and the accuracy in disease prediction enormously enhance, if the data contains relevant significant attributes, systematically eliminating irrelevant and redundant information. In the context of the presence of a large quantity of redundant and often irrelevant attributes in the data sets, there is a dire need for an efficient approach for data selection to extract exclusively relevant information of a particular disease [6], [20], [21]. The authors have reviewed the basic taxonomy of feature selection and diverse methods of gene selection in [22]. These approaches are further classified as supervised, semi-supervised, and unsupervised feature selection. In the process, they also have addressed several existing challenges and obstacles involved in the knowledge extraction from the gene expression data. Fig. 5 presents a step-by-step illustration of the instances and process of feature selection.

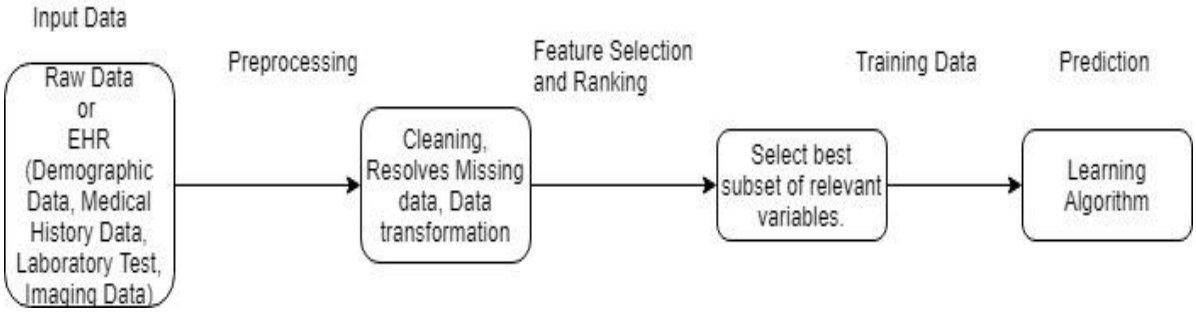

Fig. 5. Consecutive steps involved in preparing data to plug into Machine Learning model

Current researches concentrate on unsupervised learning over and above supervised learning as representation and extraction of features solely contribute to the success and accuracy of the predictive Machine Learning algorithms [23]. It is consequent to the limitations of supervised learning, and its feature selection approaches in identifying the sparse, high dimensional, noisy, and redundant data. Consequently, this method is not ideally suited for the modeling of complex and hierarchical data. Unsupervised learning, also known as representation learning, surmounts these constraints through automatic discovery of dependencies or complexities present in the data, to understand high-level and compact representation, which in turn furnishes enhanced features for useful data extraction during the application of classifiers and predictive models [21].

3. Results and Discussion

3.1. Comparision and Performance of Machine Learning Algorithms

Though extensive Machine Learning literature is present in the healthcare domain, this paper limits the scope to the study of a few disease categories related to cardiovascular and nervous systems, depression, and cancer. Since the supervised learning provides clinically more relevant and accurate results than the unsupervised learning, applications of Machine Learning in the healthcare domain, mostly prefer the former. However, unsupervised learning can be employed during the preprocessing stage for dimensionality reduction or sub-group identification, which assists in the subsequent stage of supervised learning to be more efficient. Linear regression, naïve Bayes, logistic regression, nearest neighbor, decision tree, random forest, support vector machine (SVM), discriminant analysis, and neural network are some of the relevant approaches.

The authors have surveyed and studied in [10] certain models of Machine Learning in the general backdrop of certain prevalent diseases. They have demonstrated the utilization of certain standard models of Machine Learning while handling chronic diseases. Certain popular models of Machine Learning and their respective contributions in the current literature on disease prediction are depicted in Fig. 6.

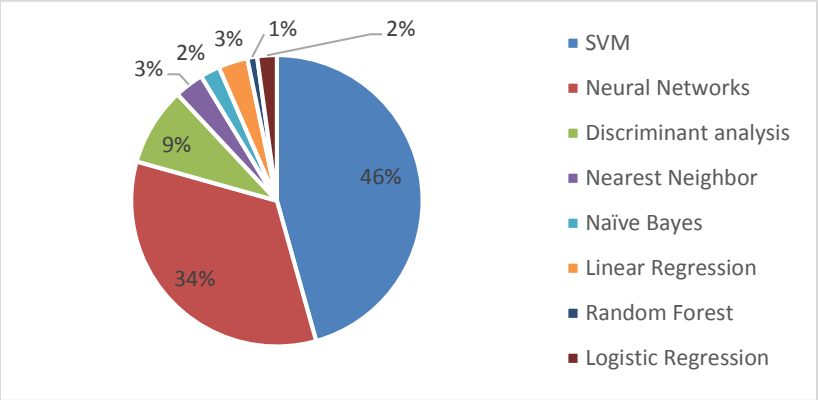

Fig. 6. The models of Machine Learning employed in predicting chronic disease

Fig. 6 takes its inputs from [10], which establishes that the focus of the current researches is on SVM as well as neural networks. The organizational constraints of the paper permit only the study of select models of Machine Learning, which are largely employed for disease prediction.

In [24], the authors classified individuals liable to be diabetic along with non-diabetic, employing the models of naïve Bayes, decision trees, and SVM. In 586 out of the used 786 instances, the classification was accurate. The results of the experimentation established that naïve Bayes had returned the most effective results.

The authors have considered [4] prediction of diabetics as a problem of binary classification and employed the methods of Machine Learning, which include kernelized and sparse support vector machines (SVMs), random forests and sparse logistic regression, obtaining excellent results. They have resorted to the use of the methods based on similarity ratio test and joint clustering and classification (JCC), which have the potential to find out the hidden patient clusters and readjusts the classifiers to suit the individual clusters.

The authors have considered 8756 instances for classification employing 50 various features for a successful prediction of diabetics [1]. A novel approach integrating SVM and genetic algorithm for the development of a method of risk prediction has proved to be superior to the literature results recording an accuracy rate of 81.02%.

Considering the popularity enjoyed by neural networks [5], the authors have employed convolutional networks for disease prediction with 94.8% accuracy, outperforming the literature results. In modern times depression and anxiety have become so common that one in every twenty suffers from this, which could lead to several medical side-effects and diseases and hence, need continuous monitoring and assessment. Researchers have analyzed as many as 470 instances having 14 features and were able to predict precisely with an accuracy rate of about 82.6% employing the cat boost algorithm.

In [23], the researchers have sought the analysis of the unsupervised methods for the prediction of the patient's future from the available information from EHR. The authors have taken into account as many as 76,214 instances for the prediction of about 78 diseases. The outcome of this proposed method demonstrated accurate results to the tune of 92.9 %.

Table 1. State of the art Disease prediction models using Machine Learning

Disease	Algorithm with reference	Accuracy with reference
Cardiac related	SVM ([22], [25], [26]), Neural Networks ([26]–[28]), Random Forest ([29], [30]), Naive Bayes ([31]).	92.07% [25], 76.54% [22], 77.63% [32], 95% [26], 82.51% [27], 98.57% [28], 82.7% [33], 98% [30].
Diabetic related	SVM ([21], [34]–[36]), Probabilistic models ([35], [37], [38]), Principle Component Analysis and Random Forest ([39]),Decision Trees ([37]).	91% [34], 95.38% [21], 62.61% [35], 85% [36],76.30% [37], 84% [38], 60.87% [35], 80.84% [39], 73.82% [37].
Cancer related	Probabilistic models ([40], [41]), Neural Networks([41]–[44]), SVM ([45], [46]), Bayesian ([47])	82.83% [48], 67.9% [41], 65.5% [41], 83.5% [42], 86% [43], 95.5% [44], 68% [45], 98.28% [46], 89% [47].
Depression	Neural Networks ([49], [50]), SVM ([51], [52]), Naive Bayes ([53])	82.35% [49], 97.2% [50], 90% [51], 82.10% [52], 74.6% [53].

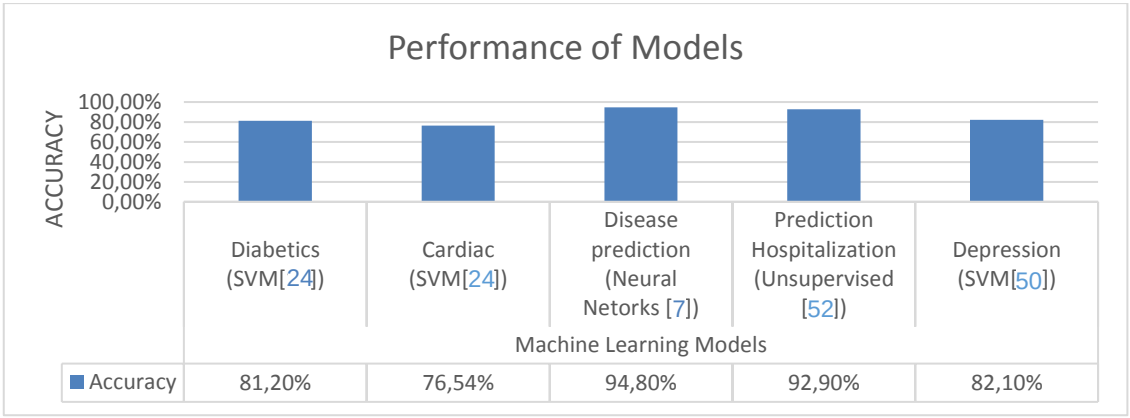

Fig. 7. Performance analysis of Machine Learning models in predicting different diseases

The accuracies represented in Fig. 7 establish the outperformance of the models of SVM and neural networks over the literature results in chronic disease prediction. In concerning the availability of literature in Healthcare Analytics employing Machine Learning, some potential state of the art models to predicting disease in earlier are projected with their references and accuracy scores in Table 1.

Since the vast percentage of researchers are leveraging Healthcare analytics by employing Machine Learning, Table 1 depicts the coverage of models with their accuracies in predicting disease. On observing different elements mentioned in this paper, it can be deduced that unsupervised models are more useful for feature enhancement. In contrast, neural network models could be preferred for leveraging predictive accuracies.

4. Conclusion

The ever-increasing EHR in the healthcare domain requires sophisticated and seamless models for efficiently employing Machine Learning analytics through the processes of supervised and unsupervised learning. The feature selection through the approaches of classification and clustering in the specific context of early prediction and diagnosis of chronic diseases deliver enhanced quality of healthcare services. Consequently, the most prevalent chronic diseases such as diabetics, cancer, depression, and cardiovascular disorders can very well be addressed much to the welfare of the patient saving time and money. The future challenges for the researchers and the practitioners would be to identify, develop or integrate robust and efficient models capable of accommodating Machine Learning processes for effective delivery of healthcare services in the prediction, diagnosis, and treatment of chronic diseases.

References

[1] S. Yang, *Effective Learning of Probabilistic Models for Clinical Predictions from Longitudinal Data*. ProQuest Dissertations Publishing, 2017, available at: Google Scholar.

[2] R. C. Deo, "Machine Learning in Medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[3] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Comput. Methods Programs Biomed.*, vol. 166, pp. 123–135, Nov. 2018, doi: 10.1016/j.cmpb.2018.10.012.

[4] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive Mob. Comput.*, vol. 51, pp. 1–26, Dec. 2018, doi: 10.1016/j.pmcj.2018.09.003.

[5] C.-C. Wu *et al.*, "Prediction of fatty liver disease using machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, Mar. 2019, doi: 10.1016/j.cmpb.2018.12.032.

[6] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams, and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," *Proc. IEEE*, vol. 106, no. 4, pp. 690–707, Apr. 2018, doi: 10.1109/JPROC.2017.2789319.

[7] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017, available at: Google Scholar.

[8] D. Jain and V. Singh, "Feature selection and classification systems for chronic iction: A review," *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, Nov. 2018, doi: 10.1016/j.eij.2018.03.002.

[9] A. F. Simpao, L. M. Ahumada, J. A. Gálvez, and M. A. Rehman, "A Review of Analytics and Clinical Informatics in Health Care," *J. Med. Syst.*, vol. 38, no. 4, p. 45, Apr. 2014, doi: 10.1007/s10916-014-0045-x.

[10] M. Islam, M. Hasan, X. Wang, H. Germack, and M. Noor-E-Alam, "A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining," *Healthcare*, vol. 6, no. 2, p. 54, May 2018, doi: 10.3390/healthcare6020054.

[11] J. Liu, Z. Zhang, and N. Razavian, "Deep EHR: Chronic Disease Prediction Using Medical Notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, 2018, vol. 85, pp. 440–464, [Online]. Available: http://proceedings.mlr.press/v85/liu18b.html.

[12] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008, doi: 10.1214/08-AOAS169.

[13] S. R. Alty, S. C. Millasseau, P. J. Chowienczyk, and A. Jakobsson, "Cardiovascular disease prediction using support vector machines," in *2003 46th Midwest Symposium on Circuits and Systems*, 2003, vol. 1, pp. 376–379, doi: 10.1109/MWSCAS.2003.1562297.

[14] D. H. Mantzaris, G. C. Anastassopoulos, and D. K. Lymberopoulos, "Medical disease prediction using Artificial Neural Networks," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, Oct. 2008, pp. 1–6, doi: 10.1109/BIBE.2008.4696782.

[15] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017, doi: 10.1136/svn-2017-000101.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009, doi: 10.1007/978-0-387-84858-7.

[17] Rui Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005, available at: Google Scholar.

[18] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011, available at: Google Scholar.

[19] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 1157–1182, 2003, available at: dl.acm.org/doi/10.5555/944919.944968.

[20] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*. Chapman and Hall/CRC, 2014, doi: 10.1201/b17320.

[21] K. Chandana, Y. Prasanth, and J. Prabhu Das, "A decision support system for predicting diabetic retinopathy using neural networks," *J. Theor. Appl. Inf. Technol.*, vol. 88, no. 3, pp. 598–606, 2016, doi: 10.1109/ERECT.2015.7499020.

[22] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record," *IEEE Access*, vol. 6, pp. 65333–65346, 2018, doi: 10.1109/ACCESS.2018.2875677.

[23] N. Sadati, M. Z. Nezhad, R. B. Chinnam, and D. Zhu, "Representation Learning with Autoencoders for Electronic Health Records: A Comparative Study," *arXiv Prepr. arXiv 1801.02961v2*, Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.02961.

[24] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 971–989, Sep. 2016, doi: 10.1109/TCBB.2015.2478454.

[25] A. Mustaqeem, S. M. Anwar, and M. Majid, "Multiclass Classification of Cardiac Arrhythmia Using Improved Feature Selection and SVM Invariants," *Comput. Math. Methods Med.*, vol. 2018, pp. 1–10, 2018, doi: 10.1155/2018/7310496.

[26] Q. K. Al-Shayea and MIS, "Artificial Neural Networks in Medical Diagnosis," *IJCSI Int. J. Comput. Sci.*, vol. 8, no. 2, pp. 150–154, 2011, doi: 10.1007/978-3-7908-1788-1_8.

[27] J. K. Kim and S. Kang, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis," *J. Healthc. Eng.*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/2780501.

[28] R. Narain, S. Saxena, and A. Goyal, "Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach," *Patient Prefer. Adherence*, vol. 10, pp. 1259–1270, Jul. 2016, doi: 10.2147/PPA.S108203.

[29] R. Mahajan, R. Kamaleswaran, J. A. Howe, and O. Akbilgic, "Cardiac Rhythm Classification from a Short Single Lead ECG Recording via Random Forest," *2017 Comput. Cardiol. Conf.*, vol. 44, pp. 2–5, 2018, doi: 10.22489/cinc.2017.179-403.

[30] C. Vimal and B. Sathish, "Random Forest Classifier Based ECG Arrhythmia Classification," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 1–10, Apr. 2010, doi: 10.4018/jhisi.2010040101.

[31] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, Oct. 2016, pp. 1–5, doi: 10.1109/CIMCA.2016.8053261.

[32] Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, "Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients," *Healthc. Inform. Res.*, vol. 16, no. 4, p. 253, 2010, doi: 10.4258/hir.2010.16.4.253.

[33] R. Mahajan, R. Kamaleswaran, J. A. Howe, and O. Akbilgic, "Cardiac Rhythm Classification from a Short Single Lead ECG Recording via Random Forest," in *2017 Computing in Cardiology (CinC)*, Sep. 2017, pp. 1–4, doi: 10.22489/CinC.2017.179-403.

[34] P. Janardhanan, L. Heena, and F. Sabika, "Effectiveness of support vector machines in medical data mining," *J. Commun. Softw. Syst.*, vol. 11, no. 1, pp. 25–30, 2015, doi: 10.24138/jcomss.v11i1.114.

[35] G.-M. Huang, K.-Y. Huang, T.-Y. Lee, and J. Weng, "An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients," *BMC Bioinformatics*, vol. 16, no. Suppl 1, p. S5, 2015, doi: 10.1186/1471-2105-16-S1-S5.

[36] S. Malik, R. Khadgawat, S. Anand, and S. Gupta, "Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva," *Springerplus*, vol. 5, no. 1, p. 701, Dec. 2016, doi: 10.1186/s40064-016-2339-6.

[37] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia*

Comput. Sci., vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[38] R. K. Leung *et al.*, "Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case–control cohort analysis," *BMC Nephrol.*, vol. 14, no. 1, p. 162, Dec. 2013, doi: 10.1186/1471-2369-14-162.

[39] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Genet.*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[40] K.-J. Wang, B. Makond, and K.-M. Wang, "Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan," *Comput. Biol. Med.*, vol. 47, pp. 147–160, Apr. 2014, doi: 10.1016/j.compbiomed.2014.02.002.

[41] O. Regnier-Coudert, J. McCall, R. Lothian, T. Lam, S. McClinton, and J. N'Dow, "Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers," *Artif. Intell. Med.*, vol. 55, no. 1, pp. 25–35, May 2012, doi: 10.1016/j.artmed.2011.11.003.

[42] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," *Comput. Biol. Med.*, vol. 48, pp. 1–7, May 2014, doi: 10.1016/j.compbiomed.2014.02.006.

[43] G. R. Hart, D. A. Roffman, R. Decker, and J. Deng, "A multi-parameterized artificial neural network for lung cancer risk prediction," *PLoS One*, vol. 13, no. 10, p. e0205264, Oct. 2018, doi: 10.1371/journal.pone.0205264.

[44] M. M. Khan, A. Mendes, and S. K. Chalup, "Evolutionary Wavelet Neural Network ensembles for breast cancer and Parkinson's disease prediction," *PLoS One*, vol. 13, no. 2, p. e0192192, Feb. 2018, doi: 10.1371/journal.pone.0192192.

[45] C.-J. Tseng, C.-J. Lu, C.-C. Chang, and G.-D. Chen, "Application of machine learning to predict the recurrence-proneness for cervical cancer," *Neural Comput. Appl.*, vol. 24, no. 6, pp. 1311–1316, May 2014, doi: 10.1007/s00521-013-1359-1.

[46] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction," *PLoS One*, vol. 12, no. 1, p. e0161501, Jan. 2017, doi: 10.1371/journal.pone.0161501.

[47] R. Agrahari *et al.*, "Applications of Bayesian network models in predicting types of hematological malignancies," *Sci. Rep.*, vol. 8, no. 1, p. 6951, 2018, doi: 10.1038/s41598-018-24758-5.

[48] K. J. Wang, B. Makond, and K. M. Wang, "Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan," *Comput. Biol. Med.*, vol. 47, no. 1, pp. 147–160, 2014, doi: 10.1016/j.compbiomed.2014.02.002.

[49] Filip Dabek and Jesus J. Caban, "A Neural Network Based Model for Predicting Psychological Conditions," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9250, pp. 252–253, 2015, doi: 10.1007/978-3-319-23344-4.

[50] A. Sau and I. Bhakta, "Artificial Neural Network (ANN) Model to Predict Depression among Geriatric Population at a Slum in Kolkata, India," *J. Clin. Diagn. Res.*, vol. 11, no. 5, pp. VC01–VC04, May 2017, doi: 10.7860/JCDR/2017/23656.9762.

[51] B. Mwangi, K. P. Ebmeier, K. Matthews, and J. Douglas Steele, "Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder," *Brain*, vol. 135, no. 5, pp. 1508–1521, May 2012, doi: 10.1093/brain/aws084.

[52] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, no. 1, p. 26094, May 2016, doi: 10.1038/srep26094.

[53] M. Tommasi, G. Ferrara, and A. Saggino, "Application of Bayes' Theorem in Valuating Depression Tests Performance," *Front. Psychol.*, vol. 9, p. 1240, Jul. 2018, doi: 10.3389/fpsyg.2018.01240.