Determining Chronic Diseases using Logistic Regression Technique

[1]Shivani Deshmukh
[1]Department of Computer Science and Engineering
[1]Deogiri Institue of Engineering and Management Studies

Introduction

Due to their dreadful clinical manifestations, such as a protracted onset cycle, sneaky symptoms, and numerous sequelae, chronic illnesses rank among the world's most serious health problems. Machine learning has recently emerged as a viable method to help with the early detection of chronic illnesses. Unfortunately, current research ignores the issues with feature masking and unbalanced class distribution in datasets for chronic diseases. In this research, we provide a general and effective diagnostic paradigm to address the aforementioned two issues for prompt and accurate diagnosis of chronic illnesses. To be more precise, we first suggest the Kernel Based Classifier (KBC) method, which is data enrichment in term of its subspace and can also prevent over fitting, to effectively capture high-level information buried in chronic illness datasets. Secondly, in order to address the issue of class imbalance, we further provide an attention-enhanced KBC algorithm to enhance the sample space of the data as well as the diagnostic precision for ill patients. Using nine available datasets and two actual chronic illness datasets, we assess the suggested framework (partly with class imbalance).

The global burden of chronic illnesses has been a serious health concern. The World Health Organization noted that the top 10 global causes of mortality in 2019 include seven chronic illnesses [1,2]. More than 63% of all fatalities worldwide are brought on by chronic illnesses. Heart disease, diabetes, hypertension, and other common chronic illnesses are mostly brought on by people's bad lifestyles [3]. The damage that chronic illnesses produce to a person's key organs (such as the eye, brain, heart, kidney, etc.) makes it simple to develop a number of major consequences that have an impact on both job and personal life [4]. Individuals who suffer from chronic conditions are more susceptible to infectious illnesses such the corona virus illness 2019 (COVID-19)[5]. A history of chronic illnesses affects and over 48% of COVID-19 patients, and these individuals are more prone to experience severe symptoms [6, 7]. Chronic illnesses will also result in high medical costs.

According to the Center for Disease Control and Prevention, the nation's 3.8 trillion in annual health care expenses are mostly driven by chronic illnesses. Chronic illnesses have certain unpleasant clinical manifestations, such as a protracted onset cycle, sneaky symptoms, permanent development, and numerous consequences, which is the major cause of the high mortality rate and high medical costs10. The information presented above serves as a reminder that we urgently need to improve chronic illness prevention, diagnosis, and treatment. Thus, it is necessary and crucial to diagnose chronic diseases early. By doing so, high-risk patients may be inspired to modify their unhealthy lifestyles, which would lower the likelihood of problems and further enhance their well-being and standard of living. Since the onset of chronic diseases is imperceptible and there are no obvious clinical symptoms in the early stage,

it is difficult for doctors to determine the risk of patients with chronic diseases. Nowadays, machine learning has become the hottest promising technology for the assisted diagnosis of diseases with its advantages of autonomous learning and low error rate11–13. It is challenging for doctors to assess the risk of people who have long-term illnesses since their beginning is imperceptible and they don't exhibit clear clinical signs in the early stages. With its benefits of learner autonomy and low mistake rates, machine learning has now emerged as the most exciting and promising technique for the aided diagnosis of illnesses. Modern machine learning algorithms like support vector machines (SVM), logistic regression (LR), k-nearest neighbour (KNN), decision trees (DT), and the ensemble of certain algorithms[18] have all been extensively used in the earlier detection of numerous chronic diseases, such as chronic kidney disease and diabetes. Nevertheless, current research focuses mostly on data preparation (such as data regularisation and extraction of features) to enhance the performance of early diagnosis of a single chronic condition. Moreover, they disregard the issues with feature masking and unbalanced class distribution in datasets for chronic diseases. So, these techniques do not support enhancing the diagnostic model's performance and are not appropriate for a rapid and accurate diagnosis of chronic illnesses.

The characteristic in the data may not be directly connected to decision-making, according to the feature hiding problem. To extract the features directly associated with decision-making23, it has to be thoroughly examined in combination with other factors. For instance, it is not feasible to determine if a patient has cardiac disease based just on their heart rate and body mass index. However neither doctor nor the deep learning may be capable of making a sound choice if the observable original features of the data are employed directly. To capture the data's potential features linked to the diagnosis of chronic illnesses, we must thus broaden overall subspace of the data. Also, the class imbalance problem, also known as the large skew between the sample sizes for the various classes, is referred to as the unbalanced classification process of the dataset.

The other classes are referred to as the minority class, while the dominating class is referred to as the majority class. The learnt model, which is more concerned with properly recognising the majority class and disregarding the minority class, will become unreliable after learning from the dataset with the class imbalance problem. The number of sick instances (minority class) is typically less than the amount of healthy cases, particularly in the chronic illness dataset (majority class). Therefore, it is much more expensive to misdiagnose a sick person as a healthy person than it is to misdiagnose a healthy person as ill. The patient could miss the most beneficial therapy time if the former occurs.

Consequently, it is crucially important and also a very difficult task to reliably identify ill patients from the class unbalanced chronic disease dataset with hurting the overall diagnostic performance. Deep neural networks, which can extract high-level characteristics from data to provide improved classification performance, offer a lot of potential for tackling different technical challenges in a variety of sectors. The majority of deep neural network methods, however, are not tolerant of tiny datasets and are susceptible to data overfitting. Also, current present deep neural network techniques cannot train a properly diagnostic models for chronic diseases due to the fact that the data gathered on chronic diseases is typically sparse (i.e., small-scale datasets). Lately, several academics have been interested in the deep polynomials neural network (PNN). When compared to certain other deep neural network

methods, PNN is far more accommodating to classification jobs on tiny datasets, according to our investigation of its advantages. Remarkably, the perfect PNN has no parameters and can repeatedly achieve a training error of zero.

The input is a linear transformation at each node of the PNN network. Every polynomial number so over input data may therefore be represented by PNN. Layer by layer, PNN's network architecture is built, much like other deep neural network methods, to reflect more and more higher level (hidden) aspects of the input data. In other words, PNN can efficiently capture characteristics linked to the diagnosis of chronic diseases by expanding the subspace of its inputs in a hierarchical manner. The output layer of the PNN may then be created by resolving a straightforward convex optimization problem.

Objective:

- Using small-scale datasets, we investigate a general and effective diagnostic paradigm for making an early, accurate diagnosis of chronic illnesses.
- To minimise the issue of over fitting, we suggest using the KBC method, which can effectively capture high-level characteristics concealed in datasets on chronic diseases and achieve excellent accuracy in classification.
- In addition, we suggest an sparse and density matrix technique to address the issue of class imbalance, which significantly boosts the diagnostic model's recall—that is, its ability to correctly identify unwell patients.
- Using nine chronic diseases datasets (partially with minority class) and extensive experimental results, we evaluate and compare the proposed methodology against other state-of-the-art methods. The findings indicate that the suggested two diagnostic model performs better state-of-the-art machine learning algorithms and can obtain greater accuracy and recall.

Related Work

Using machine learning, a research study suggested a prediction model to identify three major chronic diseases: diabetes, kidney disease, and heart disease. The most useful characteristics are chosen using the adaptive probabilistic divergence-based feature selection approach. The study found that by maximising the most crucial variables for illness diagnosis, the suggested technique provided the maximum accuracy (2). A machine learning approach based on feature selection is suggested to forecast diabetes, heart attacks, and cancer, three chronic illnesses. Convolutional Neural Network (CNN) incremental feature selection is used to detect the presence of diseases. The suggested approach demonstrated 93% classification accuracy with faster calculation (8).

Further investigation is made on the key characteristics of often chronic illnesses. Applications of Information Gain, Gain Ratio, and correlation-based feature selection methods are made. After that, the Random Forest prediction model is constructed using a number of subsets of the top-ranked features. It

demonstrated how important it is for the process of making a medical diagnosis to consider the most important elements (19). The Stacked Generalization strategy is employed to enhance the efficacy of classification algorithms in another research work for the prediction of chronic diseases (20). To outperform five chronic illness prediction models, five classification algorithms—Decision Tree (DT), k-NN, SVM, Logistic Regression (LR), and Naive Bayes—are tested. The model performance is shown to be improved, and the Stacked Ensemble technique reaches the maximum accuracy of 90%. Using categorization methods, illnesses like as diabetes, breast cancer, and kidney disease are predicted in (21). Uncertain items in datasets are investigated and eliminated using rough K-means clustering. It was shown that using a classification model using extracted characteristics produced better results than using a traditional model.

An emerging field of artificial intelligence study is early breast cancer prediction. Using online methods and prediction models, several investigations are carried out for early-stage breast cancer prediction. Fuzzy temporal rules are employed to identify highly influential aspects for online breast cancer prediction, whereas fuzzy rule-based classification is used for classification purposes. The findings demonstrate that feature selection and fuzzy rule-based categorization increase classifier accuracy. A different research concentrated on feature filtering methods for early breast cancer detection. To choose the most important characteristics in patient datasets, frequent item-set mining is utilized. SVM performs better than other models when compared to the decision tree, Naïve Bayes (NB), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM). In a study article, the reduction of false positive and false negative prediction findings was the main focus. To rank highly relevant characteristics in the dataset, the Genetic Algorithm technique is used to obtain information. SVM is used to categorise results into positive and negative categories.

Five classification models are used to predict early breast cancer in a different study (25), including SVM, K-NN, ANN, random forest, and logistic regression. In this study, the most influential features are extracted using the Pearson correlation, and when the most significant features are chosen, the ANN model achieves the greatest accuracy of 98%. In order to distinguish between cancer patients and healthy persons, two breast cancer datasets are employed in (26). In order to choose the most important characteristics in datasets, an evolutionary algorithm is utilised. Breast cancer is classified using three algorithms: multilayer perceptron (MLP), probabilistic neural network (PNN), and radial based function (RBF) (27). Compared to RBF and PNN, MLP takes more processing time for the training model and weighting of its neurons. MLP, RBF, and PNN yield accuracy values of 97%, 98%, and 100%, respectively. To predict breast cancer, a different study (28) uses SVM with several kernels. The key characteristics are extracted using an energy-based shape histogram. The proposed model had a 99% accuracy rate, which was the highest. Early diabetes diagnosis has the potential to save patients' lives.

In a study report, the clustering method k-means clustering is compared to classification techniques like Artificial Neural Network (ANN) and Random Forest. Significant components are chosen using Principal Component Analysis (PCA). The study shown that while body mass index (BMI) and blood glucose levels closely correlate with diabetes, PCA improves the accuracy of diabetes prediction (28). Diabetes retinopathy, or long-term diabetes, can lead to visual loss. The consequences of diabetes on vision are predicted through a scientific study. The prediction of feature selection is performed using KNN,

Decision Tree, Multilayer Perceptron, and SVM. It is shown that feature selection approaches greater accuracy and sensitivity by comparing the amount of accurate predictions made prior to and following feature selection (9). There are certain methods for treating diabetic retinopathy that make use of PCA and deep neural networks.

Here (10) the computer vision and machine learning technology are used to estimate BMI from the photographs. In large data, federated learning has several applications. For censorious operations, the probabilistic method with sensors is employed. A study is done to predict type II diabetes and hypertensive in both people since some people have diabetes together with hypertension. Problems with data distribution are resolved by synthetic minority oversampling, and type II diabetes and hypertension are predicted using the ensemble approach. This study shown that pre-processing data before developing a model improves prediction accuracy [11]. Other medical conditions identification, text mining, and network security all make use of machine learning techniques (12–14). To forecast the hazards connected to diabetes, further study is performed. Classification techniques include Logistic Regression, Decision Tree, ID3, C4.5, k-NN, and Naive Bayes. PCA and PSO algorithms are used to identify and remove irrelevant characteristics. We compare the processing speed and enhanced accuracy of the two feature selection methods. It demonstrates that associated with visual is a potent method for improving the accuracy of prediction models[15].

Using decision trees & SVM, physiological and iris-based characteristics are employed to predict type II diabetes. Convolutional Neural Networks for Sickness [16], fuzzy for object recognition and classification techniques, among other techniques [17]. Principal Component Analysis (PCA) is used with three ensemble approaches, AdaBoost, Bagging, and K-NN, to improve classification performance. The most important characteristics yield the maximum accuracy of 95.81%. [18]. A feed-forward and features extractor neural net is a convolutionary neural network (CNN). In a research (54), CNN is used to make an early diagnostic of Diabetes Type II. CNN is implemented using a database of steel mill workers that include information on their demographics, physical activity, and hypertension. CNN is implemented using a dataset of steel plant workers that include information on their demographics, physical activity, and hypertension. Among all of the workers' datasets, the diagnosis of diabetic patients had the highest accuracy, at 94.5%. Patients with diabetes frequently have additional chronic illnesses. A research project outlined the characteristics that heart and diabetic patients have in common. By utilising classification and regression algorithms, it is shown that heart disease may be predicted early in diabetic patients.

Proposed System

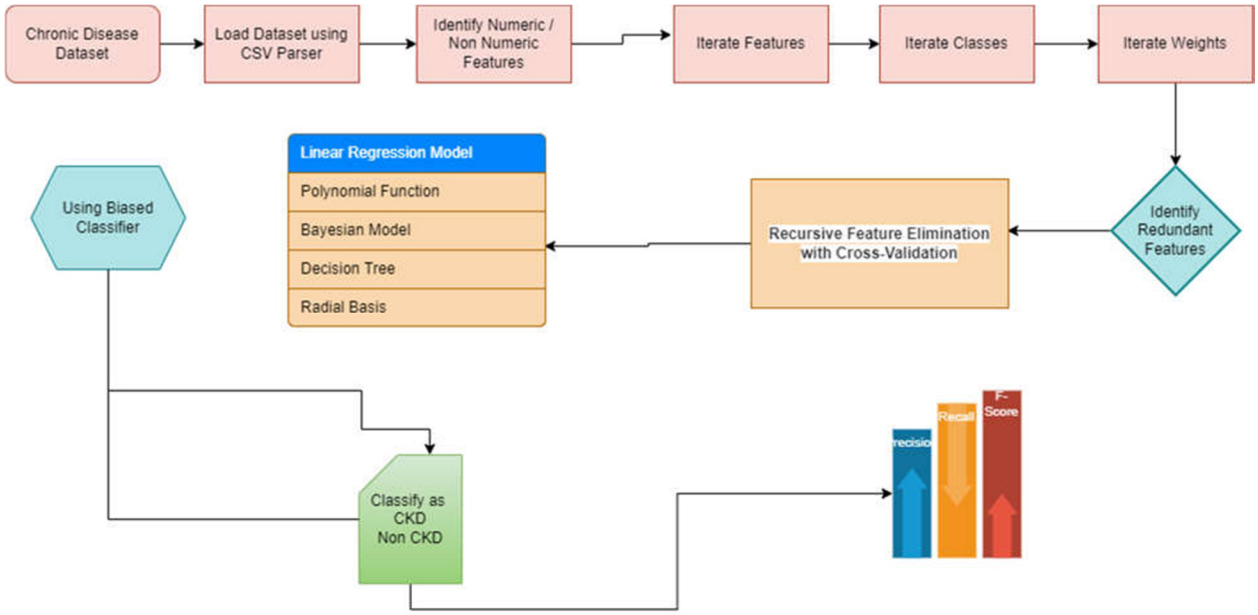

Figure 1 Proposed System Architecture

Dataset Selection

Breast cancer, diabetes, heart disease, hepatitis, and renal disease are among the five chronic diseases for which data sets have been gathered from a variety of internet sites, such as Kaggle, Dataworld, Github, and the UCI machine learning repository. To forecast various illnesses, four datasets were previously employed in a research project (19), while other datasets were used in several research studies. Each dataset contains data on several individuals who have different conditions; as a result, dataset instances and attributes vary.

Based on patient demographics and the outcomes of medical tests, all datasets utilised in this study include both numerical and categorical data attributes. The features taken from digital pictures of fine-needle fine - needle aspiration of various breast tissues are included in the breast cancer data attributes. Datasets on diabetes are derived from patient demographics and the outcomes of medical tests, including gender, age, blood pressure, body mass index, and glucose levels.

To determine whether a person has heart disease or not, two datasets comprising heart disease signs and patient data are employed. Demographics, chest pain, cholesterol, hyperglycemia, ECG readings, maximal heart rate, depression, and exercise-induced angina are among the heart disease dataset features utilised for identification. Demographic information and the results of medical lab tests are used to predict hepatitis early. Categorical dataset characteristics, such as age, gender, antivirals, liver size, tiredness, pain, bilirubin, albumin, etc., are employed for diagnosis.

Using variables such as blood urea, blood glucose, red and white blood cell count, blood sugar level, diabetes presence, and appetite, the kidney failure dataset is used to identify the existence of kidney disease. All of the aforementioned datasets contain a variety of properties; however, not all aspects are equally important for diagnosing diseases. This study discovers that by taking into account just useful characteristics, feature selection methodologies aid to maximise classification outcomes. Thus, the classification findings are validated using two publicly accessible datasets for each condition. In several web platforms, only one dataset is available for the categorization of renal disease, though.

Overall Steps in Proposed System

Data collecting comes first in the process. Data from both organised and unstructured sources is gathered by our suggested method. Data sets are divided into cleaning and test data sets when preprocessing is applied to the collected data. Finally, in order to increase the precision of the prediction findings, the training data set is trained using machine learning methods such as Sparse Multinominal Logistic Regression and neural network across a number of epochs. After the intended objective has been reached after several epochs, the generated model is prepared for testing. At this stage, the model is put to the test using a new set of data that was not used for training in order to assess how well it performs. The suggested model is suitable for deployment if it achieves the requisite accuracy in test data.

Data Collection

Real-world data consists of both structured and unstructured information, such as demographics, a patient's place of residence, and the results of lab tests. Structured information contains the patient's fundamental health information. In order to protect the patient's privacy, the data set does not include any of their identifying information, including name, ID, and location.

Preprocessing

Most structured data is preprocessed to account for the possibility of missing values. So, it is crucial to add the missing data, eliminate it, or change it in order to improve the quality of the data collection. The commas, punctuation, and white spaces are also removed during the preprocessing stage. When the data has undergone preprocessing, feature extraction and illness prediction are applied to it.

Model Description

The data set includes both organized and semistructured, as was already mentioned. The structured data includes tabularized information about the patient's living environment, laboratory test results, and the disease that they are suffering from, as well as demographic information about the patient's age, gender, height, weight, and other characteristics that are related to the disease's cause. The patient's medical symptoms and text-based information about the doctor interview make up the unstructured data. The prediction task benefits from the addition of unstructured input by obtaining more accurate results. 80% of the data set is used for training, while 20% is used for testing.

Disease Prediction Using Sparse Multinominal Logistic Regression

In order to forecast chronic illness, the suggested system employs the Sparse Multinominal Logistic Regression technique. The data set is first transformed into vector form, then word embedding is used to adopt zero values for the data's fill. After that, the convolution layer receives it.

The convolution layer provides the input to the pooling layer, which then performs the max pooling process. The fully connected layer receives the max pooling output before providing the classification outcomes to the output layer.

The hospital's computer systems and internet archives have a large number of datasets available for all ailments. These datasets include various features for a variety of applications; not all of the traits are useful for detecting a particular ailment. Before using a system for the categorization of data, data which was before is a crucial step. The model may produce false results if it was built on a data with incorrect entries. Similar to how not every attribute in a dataset equally contributes to detection. Including unnecessary features lengthens the model's processing time and reduces model performance. The performance of the classifier is severely impacted if all characteristics are employed in the prediction analysis. It depends on the doctor's experience to make a diagnosis of a disease from various symptom input data in hospitals and medical labs. The primary purpose of feature selection approaches is to examine the role that each characteristic plays in the output prediction process. Prior to building a model, selection of features is a crucial strategy for reducing data complexity by removing pointless and unnecessary elements.

Approaches for feature selection decrease the amount of data so that the model's training and testing take less time. Feature selection is advantageous since it decreases the amount of data required for processing, freeing up more room and power (15). The categorization model is easier to understand and produces more useful results with fewer features. With feature selection, features to zero or very little contribution are removed. There are different ways to choose features. One of the best techniques for feature selection is Information Gain (IG). Entropy, which refers to the uncertainty of selection, measures information gain. The likelihood of getting chosen as the final class label increases with a low entropy number (16). Each attribute's weight is determined by Information Gain in order to determine

how much it contributed to the final class decision. The feature with the highest weight is the one that offers the most insight into the final class selection and is thus seen as having the most influence.

Conclusion

A significant scientific problem is early detection of chronic illnesses. With the use of our suggested model, we hope to anticipate additional chronic illnesses. Particularly, ensemble feature selection techniques may lead to improved chronic illness prediction. There hasn't been much study effort documented in the field of developing systems that can detect different ailments in people. For the categorization of medical data and the prognosis of diseases, many artificial intelligence algorithms are employed in literature. Such methods are frequently used to identify particular diseases using a small number of variables for specific datasets. In this study, a method based on enhanced artificial neural networks is used to forecast chronic illnesses.

References

[1] May HT, Anderson JL, Muhlestein JB, Knowlton KU, Horne BD. Intermountain chronic disease risk score (ICHRON) validation for prediction of incident chronic disease diagnoses in an australian primary prevention population. Euro J Intern Med. (2020) 79:81–87. doi: 10.1016/j.ejim.2020.06.009

[2] Hegde S, Mundada MR. Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. Int J Pervasive Comput Commun. (2020) 20:145. doi: 10.1108/IJPCC-04-2020-0018

[3] Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked. (2019) 16:1–9. doi: 10.1016/j.imu.2019.100203

[4] Howard N, Chouikhi N, Adeel A, Dial K, Howard A, Hussain A. BrainOS: a novel artificial brain-alike automatic machine learning framework. Front. Comput. Neurosci. (2020) 14:1–15. doi: 10.3389/fncom.2020. 00016

[5] Bi X, Zhao X, Huang H, Chen D, Ma Y. Functional brain network classification for Alzheimer's disease detection with deep features and extreme learning machine. Cognit Comput. (2020) 12:513–27. doi: 10.1007/s12559-019-09688-2

[6] Guo L. Under The background of healthy china: regulating the analysis of hybrid machine learning in sports activities to control chronic diseases. Measurement. (2020) 164:1–10. doi: 10.1016/j.measurement.2020.107847

[7] W.H.O. NonCommunicable Diseases. (2018). Available online at: https://www. who.int/newsroom/fact-sheets/detail/noncommunicable-diseases (accessed December 12, 2021).

[8] Hemanth Reddy K, Saranya G. "Prediction of cardiovascular diseases in diabetic patients using machine learning techniques," in Artificial Intelligence Techniques for Advanced Computing Applications, (New York, NY: Springer), p. 299–305 (2020).

[9] W.H.O. Cardiovascular diseases (CVDs). (2016). Available online at: https:// www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds) (accessed December 12, 2021).

[10] Diabetes - A Major Risk Factor for Kidney Disease. National Kidney Foundation. (2020). Available online at: https://www.kidney.org/atoz/content/ diabetes (accessed December 12, 2021).

[11] Parisi L, RaviChandran N. Evolutionary feature transformation to improve prognostic prediction of hepatitis. Knowl Based Syst. (2020) 200:1– 10. doi: 10.1016/j.knosys.2020.106012

[12] Abd El-Salam SM, Ezz MM, Hashem S, Elakel W, Salama R, ElMakhzangy H, et al. Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. Inform. Med. Unlocked. (2019) 17:1–7. doi: 10.1016/j.imu.2019.100267

[13] Chugh G, Kumar S, Singh N. Survey on machine learning and deep learning applications in breast cancer diagnosis. Cogn Computat. (2021) 13:1451– 70. doi: 10.1007/s12559-020-09813-6

[14] Raj RS, Sanjay DS, Kusuma M, Sampath S. "Comparison of support vector machine and naïve bayes classifiers for predicting diabetes," In 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). (Mew York, NY: IEEE) (2019).

[15] Aada A, Tiwari S. Predicting diabetes in medical datasets using machine learning techniques. Int J Scientific Eng Res Vol. (2017) 8:257–67. Available online at: https://ijsret.com/wp-content/uploads/2019/03/IJSRET_ V5_issue2_154.pdf

[16] Yuvaraj N, SriPreethaa KR. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Comput. (2019) 22:1– 9. doi: 10.1007/s10586-017-1532-x

[17] Elhoseny M, Shankar K, Uthayakumar J. Intelligent diagnostic prediction and classification system for chronic kidney disease. Sci Rep. (2019) 9:1– 14. doi: 10.1038/s41598-019-46074-2

[18] Sandhiya S, Palani U. An effective disease prediction system using incremental feature selection and temporal convolutional neural network. J Amb Intell Hum Comput. (2020) 11:5547–60. doi: 10.1007/s12652-020-01910-6

[19] Alam MZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Inform Med Unlocked. (2019) 15:1–12. doi: 10.1016/j.imu.2019.100180

[20] Maini E, Venkateswarlu B, Marwaha D, Maini B. Upgrading the performance of machine learning based chronic disease prediction systems using stacked generalization technique. Int J Comput Digit Syst. (2020) 10:1– 9. doi: 10.12785/ijcds/100192