WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multiple disease Detection Using Machine Learning Algorithms

Gokulakrishnanv[1],Dhinakaran S[2] ,Asothrajan T[2], Chandru A[2],Dinesh M[2]

[1]Department of CSE, Assistant Professor , Dhanalakshmi Srinivasan Engineering College , Perambalur

[2]Department of CSE , UG Student, Dhanalakshmi Srinivasan Engineering College , Perambalur

ABSTRACT

Using Machine learning, our article proposes sickness prediction system. for tiny problems, the users got to go into person to the hospital for check-up that's longer intense. conjointly handling the medium entails appointments is reasonably agitated. Such a haul is also solved by sickness prediction application by giving correct steering with reference to healthy living. Over the past decade, the use of the actual sickness prediction tools aboard the relating to health has been exaggerated as a result of a variety of diseases and fewer doctor-patient quantitative relation. Thus, throughout this method, we've got a bent to unit of measurement concentrating on providing immediate and proper sickness prediction to the users regarding the symptoms they enter aboard the severity of sickness expected. Best applicable rule and doctor consultation area unit given throughout this project. For prediction of diseases, wholly completely different machine learning algorithms unit of measurement used to guarantee quick and proper predictions. In one channel, the symptoms entered area unit crosschecked with the data. Further, will be preserved among the data if the symptom is new that its primary work is and thus the various channel can provide severity of sickness expected. A web/android application is deployed for user for easy moveableness, configuring and having the flexibility to access remotely where doctors cannot reach merely. sometimes users do not appear to be aware about all the treatment with reference to the particular sickness, this project in addition look forward to providing medication and drug consultation of sickness expected. Therefore, this arrangement helps in easier health management. Machine Learning Approach for distinctive sickness Prediction victimization Machine Learning relies on prediction modelling that predicts sickness of the patients in keeping with the symptoms provided by the users as associate degree i/p to the system. This paper offers an inspiration of predicting multiple diseases victimization Machine Learning algorithms. Here we are going to use the thought of supervised Machine Learning within which implementation are done by applying call Tree, Random Forest, Naïve Bayes and KNN algorithms which can facilitate in early prediction of diseases accurately and higher patients care. The results ensured that the system would be purposeful and user destined for patients for timely diagnoses of diseases in a very patient. Despite the actual fact that varied data processing classification algorithms exist for predicting cardiovascular disease, there's inadequate knowledge for predicting cardiovascular disease in a very United States intelligence agency betic individual. as a result of the choice tree model systematically beat the naive Bayes and support vector machine models, we tend to fine-tuned it for best performance in statement the chance of cardiovascular disease in polygenic disorder people.

Keywords: —Machine Learning, KNN algorithm, SVM, Decision Tree Algorithm, Naive Bayes Algorithm

INTRODUCTION

The Earth is passing through a violet patch of technology, where there is increasing demand of intelligence and accuracy behind it. Today's people unit of measurement a lot of possible regarding|keen about|addicted to|dependent on|obsessed on|smitten by net but they are not concerned about their personal health. throughout this twenty 1st Century humans unit of measurement encircled with technology as they are the constituent of our day to day life cycle. With this we tend to tend to unit of measurement invariably specializing within the health for ourselves and our earned valuables severally. people avoid to travel in hospital for small draw back which may become a major malady in future. Establishing question answer forums is popping into a simple due to answer those queries rather than browsing through the list of most likely relevant document from cyber web. Our basic set up is to develop a system which may predict and provides the most points of the illness expected beside its severity that as symptoms unit of measurement given as input by the user. The system will compare the symptoms with the datasets provided at intervals the data. If the symptom matches the datasets then it got to raise various relevant symptoms specifying the name of the symptom. If not, the symptom entered got to be notified as wrong symptom. once this a prompt will come up asking whether or not or not you'd prefer to still save the symptom at intervals the data. If you click on affirmative, it will be saved at intervals the data, if not it will attend the recycle bin. Machine Learning could be a set of AI that's chiefly affect the study of algorithms that improve with the employment of knowledge and skill. Machine Learning has 2 phases i.e. coaching and Testing [7]. Machine Learning provides AN economical platform in medical field to unravel varied aid problems at a way quicker rate. There area unit 2 styles of Machine Learning -Supervised Learning and unattended Learning. In supervised learning we tend to frame a model with the assistance of knowledge that's well labeled . On the opposite hand, unattended learning model learn from unlabeled knowledge. there's a requirement to form such a system that may facilitate finish users to predict diseases on the idea of symptoms given in it while not visiting hospitals. By doing therefore, it'll decrease the frenzy at OPD's of hospitals and convey down the employment on medical employees. Not solely this, this technique can cut back the expensive treatment and panic moment at the top stages in order that correct medication are often provided at the correct time and that we will lower down the death rate additionally. this technique additionally consists of a feature of info that stores the info entered by the top users and also the name of the illness the patient is plagued by which will be used as a past record and can facilitate in any treatment in future. The analysis accuracy is enhanced by victimisation Machine Learning algorithms. Altogether this technique can facilitate in easier health management. method of determinative a condition supported a person's symptoms and indicators is thought as diagnosis. within the diagnostic method, one or a lot of diagnostic procedures, like diagnostic tests, area unit performed. identification of chronic diseases could be a very important issue within the medical business since it's supported several symptoms. it's a posh procedure that often results in incorrect assumptions. once designation diseases, the clinical judgment is predicated totally on the patient's symptoms additionally because the physicians' data and skill. As a result, it's tough for a doctor to form the correct judgment on an identical basis if he's not supported by clinical tests and patient history data. Even experienced physicians will get pleasure from a computer-aided diagnostic system in creating sound medical judgments [8]. Thus, medical professionals area unit terribly curious about automating the identification method by desegregation machine learning techniques with doctor experience [9]. data processing and machine learning approaches area unit creating vital efforts to showing intelligence translate accessible knowledge into valuable data so as to boost the diagnostic process's potency. many studies are conducted to explore the employment of machine learning in terms of diagnostic skills. it had been discovered that, when put next to the foremost experienced doctor, UN agency will diagnose with seventy nine.77% accuracy, machine learning algorithms might establish with ninety two.1% correctness [10]. Machine learning techniques area unit expressly wont to sickness datasets to extract options for optimum sickness identification, prediction, prevention, and medical care.

RELATED WORK

A structural model and a group of conditional possibilities area unit utilized by Bayesian classifiers. they create the idea that the contributions of all factors area unit freelance. It 1st calculates the previous chance for every category, so applies the incidence of every variable price to an unknown situation. A Thomas Bayes network classifier is made on a Bayesian network, that reflects a probability distribution over a group of class characteristics. The SVM methodology and also the area Thomas Bayes technique were accustomed predict uropathy [11]. The authors tried to categorise numerous stages of uropathy mistreatment the advised ANFIS algorithmic rule. The study's purpose was to style an efficient categorization algorithmic rule mistreatment many assessment metrics like accuracy and execution time. whereas the SVM algorithmic rule provided higher classification accuracy, the area Thomas Bayes fared higher since it made leads to less time. The results show that SVM outperforms the area Thomas Bayes Approach in predicting excretory organ sickness. The fuzzy technique with a membership operate was accustomed forecast viscus sickness [12]. mistreatment the Fuzzy KNN Classifier, the authors tried to eliminate ambiguity and uncertainty from information. The 560-record dataset was separated into twenty six categories, with every category having twenty four things. The dataset was separated into 2 equal parts: coaching and testing. The fuzzy KNN methodology was enforced once pre-processing techniques were used. this system was examined mistreatment many assessment metrics like accuracy, precision, and recall, among others. supported the info, it had been discovered that the fuzzy KNN classifier outperformed the KNN classifier in terms of accuracies or the prediction of viscus sickness, a unique technique supported the ANN algorithmic rule was devised [3]. The researchers created AN interactive prediction methodology supported categorization mistreatment a man-made neural network algorithmic rule and taking into consideration the 13 most significant clinical parameters. The advised methodology tested effective for predicting cardiopathy with AN accuracy of eighty one and may be terribly helpful for tending practitioners. within the existing system the data set is commonly very little, for patients and diseases with specific conditions. These systems area unit in the main designed for the extra prodigious diseases like upset, Cancer etc. The pre-selected characteristics may typically not satisfy the changes among the malady and its influencing factors that will cause quality in results. As we have a tendency to tend to sleep in endlessly evolving world, the symptoms of diseases together evolve over a course of it slow. together most of the current systems build the users expect long periods by making them answer prolonged questionnaires.

PROPOSED METHOD

We proposing such a system that may flaunt an easy, value effective , elegant computer program and even be time economical . Our planned system bridges the gap between doctors and patients which can facilitate each categories of users to attain their goal. this method is employed to predict diseases in line with symptoms. during this planned system we have a tendency to square measure about to take down 5 symptoms from the users and measure them by applying algorithms like call Tree, Random Forest , Naïve Thomas Bayes and KNN which can facilitate in obtaining correct prediction .Our system can explore and merge additional datasets which incorporates massive diversity of population to urge more practical results and therefore our system can improve and enhances the accuracy of the results. together with the redoubled accuracy rate, we'll proliferate the responsibility of our system for this job and might gain the trust of patient during this system. with the exception of of these, our system can comprise of a information for storing the information entered by the users and therefore the name of the unwellness the patient is plagued by which might be used as a reference in future for additional treatment. thence this method can contribute in easier health management with higher satisfaction to the users. n this framework, machine learning algorithms- support vector machine, naïve Thomas Bayes, call tree square measure used. The Naive catchword The {bayes|Bayes|Thomas Thomas Bayes|mathematician} classification [4] refers to a basic probabilistic classification supported sturdy freelance assumptions within the application of the Bayes theorem. The existence or absence of a specific category feature doesn't depend upon the presence or absence of the other feature. It operates on the idea of conditions.

It uses Bayes' theorem that determines the likelihood that an occurrence happens once another event happens. If B represents the dependent event and A represents the last event the theory Thomas Bayes is also phrased as follows: Sample (B equipped in A) = Sample (A and B)/Sample (A and B) (A) . The approach divides the amount of events during which A and B occur along by the amount of circumstances during which A happens to urge the chance of B given A alone. so as to estimate the parameters (variable media and variances), the Naive Thomas Bayes Classifier advantages from solely some coaching knowledge. thanks to the idea of freelance variables, all the variances should be computed for every category. it's relevant to binary yet as multi-class issues. SVM [5] may be a technique typically used for kernel learning to handle problems with massive prediction. The SVM classifier has shown larger generalization and a well-scaling of each linear and nonlinear knowledge as compared to alternative classifiers. additionally, the SVM classificator delivers terribly sturdy pattern recognition performance in conjunction with numerous oftentimes used approaches in applied math learning and improvement theory. characteristic an summary that separates positive examples from negative knowledge with the best error margin is that the main aim of the SVM arrangement. once the information is linearly divisible, it's straightforward to settle on the optimum hyper-plane cacophonous 2 categories of information. For non-inlinear mapping to massive dimension area for non-separable issues, SVM applies 'Kernel Functions' on the opposite facet. There square measure variety of kernels functions as well as Linear Kernel perform (LKF), Polynomial Kernel perform (PKF) and Sigmoid Kernel perform (SKF), Exponential Radial Basis Kernel perform (ERBKF) (GRBKF). The Radial Basic perform (RBF) has been known because the finest kernel perform among the many kernel functions. call trees square measure used [6] extensively for categorizing immense datasets. call trees reason knowledge between the basis node.

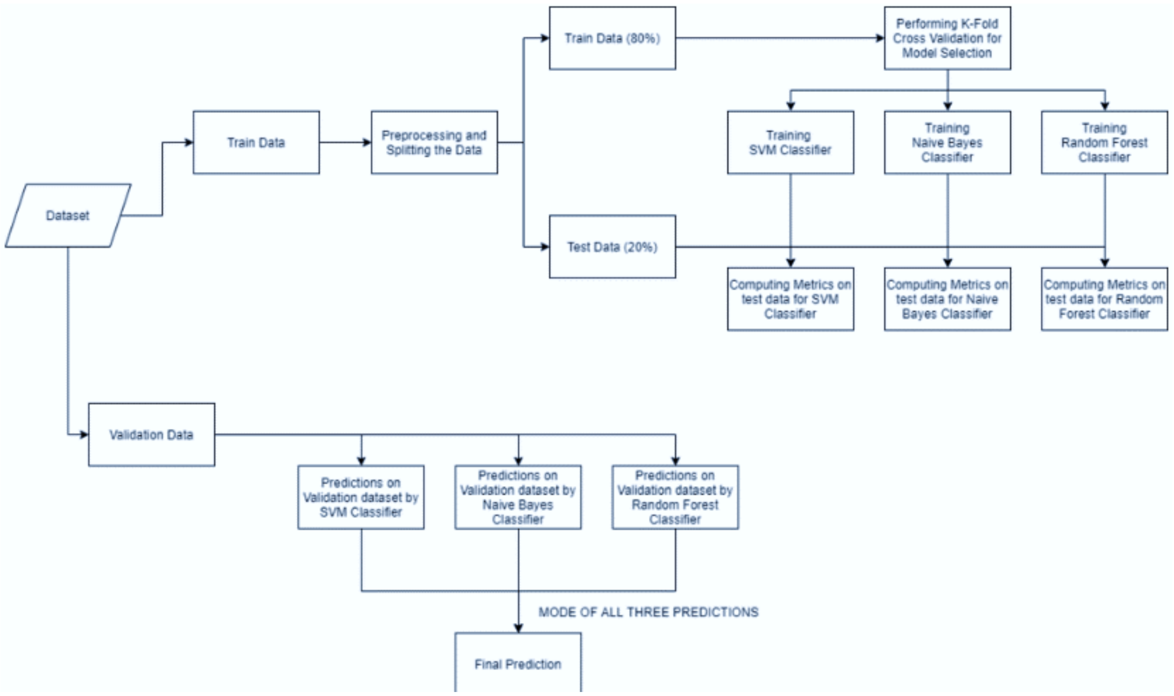
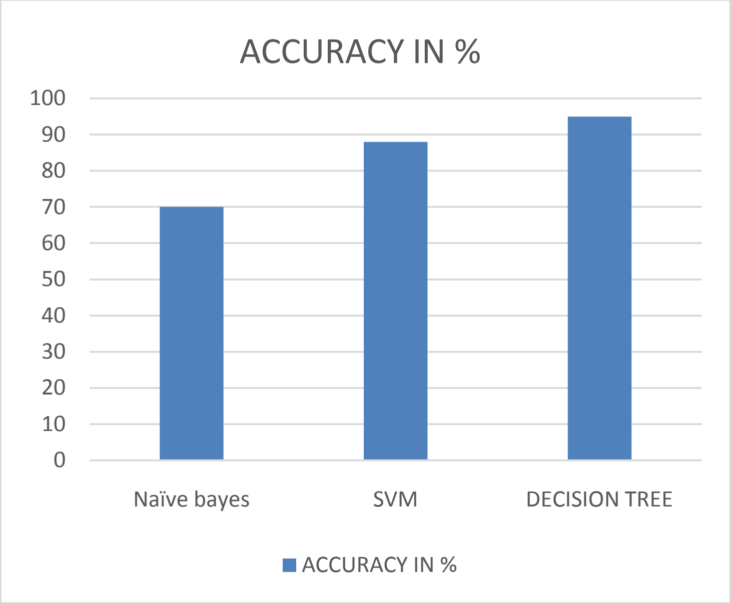

Dataset
Train Data
Preprocessing and Splitting the Data
Train Data (80%)
Test Data (20%)
Performing K-Fold Cross Validation for Model Selection
Training SVM Classifier
Training Naive Bayes Classifier
Training Random Forest Classifier
Computing Metrics on test data for SVM Classifier
Computing Metrics on test data for Naive Bayes Classifier
Computing Metrics on test data for Random Forest Classifier
Validation Data
Predictions on Validation dataset by SVM Classifier
Predictions on Validation dataset by Naive Bayes Classifier
Predictions on Validation dataset by Random Forest Classifier
MODE OF ALL THREE PREDICTIONS
Final Prediction

FIG.1.Proposed Workflow Architecture

DISEASE PREDICTION USING MACHINE LEARNING

Gathering the Data: Data preparation is that the primary step for any machine learning downside. we are going to be employing a dataset from Kaggle for this downside. This dataset consists of 2 CSV files one for coaching and one for testing. there's a complete of 132 columns within the dataset out of that one hundred thirty five columns represent the symptoms and also the last column is that the prognosis.Cleaning the Data: cleanup is that the most significant step in an exceedingly machine learning project. the standard of our knowledge determines the standard of our machine learning model. therefore it's continually necessary to wash the info before feeding it to the model for coaching. In our dataset all the columns area unit numerical, the target column i.e. prognosis could be a string kind and is encoded to numerical type employing a label encoder.Model Building: when gathering and cleanup the info, the info is prepared and may be wont to train a machine learning model. we are going to be mistreatment this cleansed knowledge to coach the Support Vector Classifier, Naive mathematician Classifier, and Random Forest Classifier. we are going to be employing a confusion matrix to work out the standard of the models. Inference: when coaching the 3 models we are going to be predicting the unwellness for the input symptoms by combining the predictions of all 3 models. This makes our overall prediction additional strong and correct. foremost we are going to be loading the dataset from the folders mistreatment the pandas library. whereas reading the dataset we are going to be dropping the null column. This dataset could be a clean dataset with no null values and every one the options accommodates 0's and 1's. Whenever we have a tendency to area unit determination a classification task it's necessary to envision whether or not our target column is balanced or not. we are going to be employing a bar plot, to envision whether or not the dataset is balanced or not. currently that we've got cleansed our knowledge by removing the Null values and changing the labels to numerical format, It's time to separate the info to coach and check the model. we are going to be cacophonic the info into 80:20 format i.e. eightieth of {the knowledge|the info|the information}set are used for coaching the model and 2 hundredth of the data are wont to appraise the performance of the models. From the higher than plot, we are able to observe that the dataset could be a balanced dataset i.e. there area unit precisely one hundred twenty samples for every unwellness, and no more leveling is needed. we are able to notice that our target column i.e. prognosis column is of object datatype, this format isn't appropriate to coach a machine learning model. So, we are going to be employing a label encoder to convert the prognosis column to the numerical datatype. Label Encoder converts the labels into numerical type by distribution a singular index to the labels. If the whole variety of labels is n, then the numbers appointed to every label are between zero to n-1. when cacophonic the info, we are going to be currently functioning on the modeling half. we are going to be mistreatment K-Fold cross-validation to judge the machine learning models. we are going to be mistreatment Support Vector Classifier, Gaussian Naive mathematician Classifier, and Random Forest Classifier for cross-validation. Before stepping into the implementation half allow us to get acquainted with k-fold cross-validation and also the machine learning models. K-Fold Cross-Validation: K-Fold cross-validation is one among the cross-validation techniques within which the complete dataset is split into k variety of subsets, conjointly called folds, then coaching of the model is performed on the k-1 sets and also the remaining one subset is employed to judge the model performance.Support Vector Classifier: Support Vector Classifier could be a discriminative classifier i.e. once given a labelled coaching knowledge, the rule tries to seek out Associate in Nursing optimum hyperplane that accurately separates the samples into completely different classes in hyperspace.Gaussian Naive {bayes|Bayes|Thomas mathematician|mathematician} Classifier: it's a probabilistic machine learning rule that internally uses Bayes Theorem to classify the info points.Random Forest Classifier: Random Forest is Associate in Nursing ensemble learning-based supervised machine learning classification rule that internally uses multiple call trees to create the classification. in an exceedingly random forest classifier, all the interior call trees area unit weak learners, the outputs of those weak call trees area unit combined i.e. mode of all the predictions is because the final prediction. From the higher than output, {we can|we will|we area unit able to} notice that every one our machine learning algorithms are playacting alright and also the mean scores when k fold cross-validation also are terribly high. to make a sturdy model we are able to mix i.e. take the mode of the predictions of all 3 models in order that even one among the models builds wrong predictions and also the different 2 make correct predictions then the ultimate output would be the proper one. This approach can facilitate North American country to stay the predictions way more correct on utterly unseen knowledge. within the below code we are going to be coaching all the 3 models on the train knowledge, checking the standard of our models employing a confusion matrix, then mix the predictions of all the 3 models.



CONCLUSION

The main aim of this article is to detect the illness in accordance with symptoms place down by the patients with correct implementation of Machine Learning algorithmic program. during this paper we've used four Machine Learning algorithmic program for prediction and achieved the mean accuracy of over 96% that shows outstanding rectification and high accuracy than previous work and conjointly makes this technique additional reliable than the prevailing one for this job and therefore provides higher satisfaction to the user compared with the opposite one. It conjointly stores the info entered by the user and therefore the name of the illness the patient is affected by within the information which might be used as past record and can facilitate in future for future treatment and therefore tributary in easier health management .We have conjointly created a interface for higher interaction with the system by users that is incredibly simple to work .This paper shows that Machine Learning algorithmic program may be wont to predict the illness simply with totally different parameters and models. within the finish will|we will|we are able to} say that our system has no threshold of the users as a result of everybody can use this technique. There square measure several attainable enhancements that would be explore to diversify the analysis by discovering and considering further options. because of time boundation , the subsequent work needed to be performed in future. there's conceive to use additional classification techniques or ways, totally different discretization techniques, multiple classifier pick ways. would really like to use totally different rules like association rule and varied algorithms like logistical regression and agglomeration algorithms. In future, willing to form use of filter primarily based feature choice ways so as to attain additional applicable likewise as purposeful result. Despite the existence of the many data processing classification ways for pre dicting heart condition, there's poor knowledge to predict heart Dis ease during a diabetic individual. we have a tendency to fine-tuned the choice tree model for optimum performance in statement the possibility of heart condition in diabetic patients since it systematically outperformed the naive Bayes and support vector machine models.

REFERENCES

[1] R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https:// doi.org/10.9734/cjast/2021/v40i631320.

[2] M. Sivakami, P. Prabhu. Classification of algorithms supported factual knowledge recovery from cardiac data set, Int. J. Curr. Res. Rev. 13(6) 161-166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).

[3] M. Sivakami, P. Prabhu. A Comparative Review of Recent Data Mining Techniques for Prediction of Cardiovascular Disease from Electronic Health Records. In: Hemanth D., Shakya S., Baig Z. (eds) Intelligent Data Communication Technologies and Internet of Things. ICICI 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 38. Springer, Cham 477-484. ISSN 2367-4512 ISSN 2367-4520 (electronic), ISBN 978-3-030-34079-7 ISBN 978-3-030-34080-3 (eBook) 2020.

[4] P. Prabhu, S. Selvabharathi. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019 (pp. 138–142) Institute of Electrical and Electronics Engineers Inc.

ISBN:9781728136639. 2019.

[5] N. Jothi, N.A. Rashid, W. Husain, Data mining in healthcare – A review, Procedia Comput. Sci. 72 (2015) 306–313.

[6] H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) 2017 55.

[7] K.B. Wagholikar, V. Sundararajan, A.W. Deshpande, Modeling paradigms for medical diagnostic decision support: a survey and future directions, J. Med. Syst. 36 (5) (2012) 3029–3049.

[8] E. Gürbüz, E. Kılıç, A new adaptive support vector machine for diagnosis of diseases, Expert Syst. 31 (5) (2014) 389–397.

[9] M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.

[10] Y. Kazemi, S.A. Mirroshandel, A novel method for predicting kidney stone type using ensemble learning, Artif. Intell. Med. 84 (2018) 117–126.

[11] H. Barakat, P. Andrew, Bradley, H. Mohammed Nabil Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, IEEE Trans. Inf. Technol. Bio Med. J. 14 (4) (2009) 1–7. [12] R. Tina Patil, S.S. Sherekar, Performance analysis of Naive bayes and J48 classification algorithm for data classification, Int. J. Comput. Sci. Appl. 6 (2) (2013) 256–261.

[13] Shruti Ratnakar, K. Rajeswari, Rose Jacob, Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes, Int. J. Adv. Comput. Eng. Netw. 1 (2) (2013) 51–55.

[14] S. Grampurohit, C. Sagarnal, Disease prediction using machine learning algorithms, 2020 Int. Conf. Emerg. Technol. (INCET) (2020) 1–7, https://doi. org/10.1109/INCET49848.2020.9154130.

[15] R.J.P. Princy, S. Parthasarathy, P.S. Hency Jose, A. Raj Lakshminarayanan, S. Jeganathan, Prediction of Cardiac Disease using Supervised Machine Learning Algorithms, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570–575, https://doi.org/10.1109/ ICICCS48265.2020.9121169.