ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-7, Issue-3, June-2023.

Machine Learning Based Disease Classification: A State-of-the-Art Survey

Rahul Kumar[1], Chetan Agrawal[2], Prachi Tiwari[3]
Department of CSE, Radharaman Institute of Technology & Science, Bhopal, India[1, 2, 3]
rahulkumar110296@gmail.com[1], chetan.agrawal12@gmail.com[2], prachi.38@gmail.com[3]

Abstract: *Huge volumes of data are frequently handled in the medical industry. Results may be impacted if large amounts of data are handled using traditional ways. In specifically, for disease prediction, machine learning algorithms can be utilized to gather data for medical research. For the evaluation of patient medications and specialists, early disease detection is essential. The diagnosis of many diseases is done using machine learning algorithms including decision trees, support vector machines, multilayer perceptron, Bayes classifiers, K-Nearest Neighbors ensemble classifier techniques, etc. The use of machine learning algorithms can result in quick and accurate disease prediction. The application of machine learning approaches to forecast various diseases and their types is examined in this research article. This study looked at studies that largely dealt with the prediction of diabetes, heart disease, breast cancer, chronic kidney disease, and machine learning. The hybrid strategy that improves the performance of individual classifiers is also examined in the paper.*

Keywords: Machine Learning, Classification, Disease Classification, SVM, ANN.

Introduction

Machine Learning is an artificial intelligence branch that aims to provide computer methods to accumulate change and update intelligent systems knowledge. Artificial intelligence (AI) allows systems to observe from environments, execute certain features and increase the likelihood of success in fixing real world challenges. AI turns out to be an interesting field with technological improvements and scientific growth. It therefore leads to a growing attention on ML techniques.

Machine learning (ML) is a significant method for data analysis that iteratively learns from the available data with the aid of learning algorithms. In words of Ethem Alpaydın machine learning can be defined as "Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future or descriptive to gain knowledge from data, or both" Machine learning is an artificial intelligence (AI) application that enables systems to learn and improve automatically without being explicitly programmed. Machine learning focuses on developing programs that are able to use the data and to learn it for them.

The learning process begins with observations or information, such as examples, direct experience or instruction, in particular look for data patterns and make better decisions based on the examples we provide in the future. The key goal is to allow the machines learn automatically without human interference or assistance and adjust response accordingly [1] The intended contribution of AI in the field of medical science is to develop programs that can actually help a medical expert in practicing expert and more accurate diagnosis.

The forecast of diseases plays an important role in machine learning. Various types of diseases can be predicted using ML techniques Here we examine how machine learning techniques that are used to predict various disease types. This paper focused on the prediction of chronic kidney disease, heart disease, diabetes and breast cancer lymphatic system and lung disorders. Next we present a brief description of a few diseases.

Chronic kidney disease is a common term for diverse disorders affecting the kidney's structure and function [2]. The definition of chronic kidney disease is centered on kidney damage (i.e. albuminuria) or reduced kidney function for a period of 3 months or more [2]. Kidney failure is among the most serious outcomes of chronic renal disease, with complications of decreased kidney function being the primary reasons.

The Complications can occur at any stage, which often lead to death. From the analysis of World Health Organization, worldwide, an estimated 12 million deaths occur every year due to the Heart disease. Breast cancer has become a common form of cancer in women. Mammography is the conventional method for detecting breast cancer. Here we evaluate various machine learning techniques majorly used in the assessment of breast cancer. Insulin is among the primary hormones in the body if insulin is not produced or used properly by the body, the unneeded amount of sugar is driven out by urination. This disease is known as diabetes.

The objectives of disease classification using machine learning methods:

- To improve the accuracy and efficiency of disease diagnosis. Machine learning algorithms can be trained on large datasets of medical images, symptoms, and other data to learn to identify diseases with greater accuracy than humans. This can lead to faster and more accurate diagnosis, which can improve patient outcomes.
- To reduce the cost of disease diagnosis. Machine learning algorithms can be used to automate many of the tasks involved in disease diagnosis, such as image analysis and symptom interpretation. This can reduce the need for human experts, which can save money.
- To improve access to healthcare. Machine learning algorithms can be used to develop mobile apps and other tools that can be used to diagnose diseases in remote areas or in countries with limited healthcare resources. This can help to improve access to healthcare for people who would otherwise not be able to afford or obtain it.
- To develop new treatments and cures for diseases. Machine learning algorithms can be used to analyze large datasets of medical data to identify new patterns and correlations that can help to develop new treatments and cures for diseases.

In addition to these objectives, machine learning methods can also be used to:

- Identify risk factors for diseases. Machine learning algorithms can be used to analyze large datasets of medical data to identify factors that are associated with an increased risk of developing certain diseases. This information can be used to develop interventions to reduce the risk of disease.
- Personalize treatment. Machine learning algorithms can be used to analyze individual patient data to personalize treatment plans. This can help to ensure that patients receive the most effective treatment for their specific condition.
- Monitor disease progression. Machine learning algorithms can be used to monitor the progression of diseases over time. This information can be used to adjust treatment plans as needed and to identify patients who are at risk of developing complications.

Overall, machine learning methods have the potential to revolutionize the way diseases are diagnosed, treated, and monitored. By improving the accuracy, efficiency, and accessibility of healthcare, machine learning can help to improve patient outcomes and save lives.

The paper also reviews a set of hybrid approaches used in the field of medical science for disease prediction. The use of the hybrid method increases the overall performance by incorporating the classification capacity of individual classifiers and the chances of misclassification of a specific instance are reduced significantly. The different apprentices can be combined in various ways. They can work in parallel with all inputs and their output can somehow be combined. If an instance is incorrectly classified by a classifier, by an individual classifier, the error is corrected by the right classification done by other individual classifiers. Such hybrid Algorithm first trains an initial learner, and then trains subsequent learners on data that the first learner misclassifies and in this way, the weaknesses of each base-learner are covered up by the next learner.

II. Related Work

MS Uzer et al. [1] built a hybrid breast cancer detection system The Artificial Neural Networks was used as classification algorithm. Cancer is the state where the body cells lose their functions, Start multiplying and dividing uncontrollably Cancer is the state in which body cells lose their functionalities, begin to multiply uncontrollably Cancer based on malignant tumors is the main cause of death in the world. Authors opted for the method of hybrid function selection based on Sequential Backward Selection, Sequential Forward Search and PCA, after this a ten-fold cross validation approach was also applied to produce the results with 98.57% accuracy.

In the study author[2] introduced the use of a local linear neural wavelet network to detect breast cancer by training its parameters using the Recursive Least Square(RLS) approach to improve its individual performance The results were estimated and compared with other experiments and demonstrated that the approach proposed is very efficient and provides a good classification[21][22].
In research [3] the author modified the SVM classification based on RS for prediction of medical diseases. The approach used the benefits of RS to eliminate redundant information and the benefit of SVM in training and testing data. The technique was used on three sets of data, which shows an increase in accuracy as compared to other approaches.

To increase performance of classification a decision making support system which is based on Fuzzy Logic and Fuzzy Decision Trees was suggested by author [4] FDT which is simple to understand and apply. The split ratio of 70:30 was used to test the results on two breast cancer datasets and the error rates were 0.3661 and 0.1414 respectively Increasing the classification performance of the decision - making support system based on Fuzzy Logic and Fuzzy Decision Trees. FDT, which is both easy to understand and applicable was used. The 70:30 split ratio was used to test the results for two data sets of breast cancer where the error rates came out to be 0.3661 and 0.1414 respectively.
In order to enhance the accuracy of CAD-based techniques, the author [5] proposed the use of boosting to resolve the problem of breast cancer characterization, a hybrid boosting algorithm combining the benefit of various boosting techniques was used. The outcome of the application of the various Boosting techniques examined on real breast cancer illustrate that the hybrid boosting algorithm exceeds 48 percent of the other boosting technique.
In this paper [6] a model was presented to estimate the risk score assessment of heart disease in the state of Andhra Pradesh with a decreased number of attributes. In addition, characteristic selection measures such as IG, SU and genetic search have been used to find out the specific characteristics that add more to the prediction of heart disease and thus in some way reduce the number of diseases. The author generated class association rules using feature subset selection. These generated rules would assist doctors predict a

patient's heart disease, use this very approach on public datasets and have it compared it with other existing solutions.

Three algorithms were compared by the author [7] and prediction models were building to predict the risk of type II diabetes. Type II diabetes is a state in which the pancreas is not able to produce the needed amount of insulin leading to abnormal sugar level in the blood the three algorithms that were used were artificial neural networks (ANN), NaiveBayes and K-nearest neighbors (KNN). The results showed that the neural network with 96 % prediction accuracy outperforms prediction accuracy of Naïve Bayes which showed 95 % accuracy and KNN which showed 91 per cent accuracy.

To help doctors predict the heart disease T. Manju et al. [8] introduced a hybrid multi layer feed forward neural network (MLFFN) and genetic algorithm. The heart attack is the main cause of death; its main causes are diabetes, unhealthy diet, high blood pressure, obesity and smoking. The author introduces the use of the Multi Layer Feed Forward Neural Network, which integrates the Genetic Algorithm and Back Propagation Network (BPN) for the prediction of heart attack. The data set consisted of 13 factors, only 6 of which were used to train the ANN. The main aim for ANN then was to predict the possibility of heart attack in a patient.
Levenberg–Marquardt (LM) feed-forward MLP neural network was anticipated by Babak Sokouti et.al. [9] to classify cervical cell images obtain from 100 patients .The semi - automated diagnostic system for cervical cancer consists of two phases: image pre processing/processing and feed forward MLP neural network. The results showed that cervical cell images were successfully classified according to the proposed method with a good classification rate. The proposed semi-automated system will also help in the identification of precancerous and cancer cells.

Four machine learning methods, including the k- nearest neighbors, the support vector machine, logistic regression and decision tree classifiers were evaluated to forecast the chronic kidney disease by Charleonnan et al. [10] the performance of these models, to forecast chronic kidney disease was compared with each other to pick the best classifier . The experimental results show that the SVM classifier has achieve the maximum accuracy of 98.3 percent. After training and testing using the proposed method, SVM also has the highest sensitivity. It can therefore be inferred that the SVM classifier is appropriate for the forecast of chronic kidney disease.
A different approach has been introduced and presented in a research conducted by Asif Salekin and John Stankovic [11], to determine the CKD using machine learning techniques As classifiers they used; k-nearest neighbors, random forests and neural networks to find a appropriate solution. Authors performed a function reduction analysis using a wrapper method to find the attributes that detect CKD with high accuracy To identify cost - effective, highly accurate CKD detection classifier by considering only 5 attributes. Using this approach authors achieved a detection accuracy of 0.993 and a root mean square error of 0.1084 according to the F1 measure.

The classification model were built with various different classification algorithms here, Wrapper subset attribute evaluator and best first search method for predicting and classifying CKD and non- CKD patients. In the classification of CKD and non- CKD cases, the models showed better performance. Authors compared the results of different models [12]. It was observed from the comparison that the classifiers scored better on reduced data than the original data set.
Sahil Sharma, Vinod Sharma and Atul Sharma [13] assessed 12 classification methodology using a data set. The outcome of the candidate methods was evaluated with the actual medical results of the subject in order

to calculate the efficiency. As performance assessment metrics, authors used predictive accuracy, sensitivity, precision and specificity. According to the observed test results, the decision tree technique performed strongly with almost 98,6 percent accuracy, %, sensitivity of 0.9720, precision of 1 and specificity of 1.

The authors [14] suggested an extremely efficient two - stage hybrid ensemble method. The two stage hybrid ensemble classifier incorporates the capability of particular classification algorithms. The final results show that the two-stage hybrid ensemble is quite an effective way for chronic kidney disease classification. The outcome of this ensemble classifier on the efficiently selected decreased feature set and the complete feature set in terms of different performance metrics such as accuracy(2-class), sensitivity, precision, specificity. were obtained and the proposed method was found to be effective additionally GUI-based diagnostic tool was developed that could help doctors to validate their findings.

R Rao R Bharat et al. [15] built a computer aided diagnosis (CAD) system named lungCAD up a system called lungCAD for computer - aided diagnostics (CAD). From the CT thorax studies, the system used a classification algorithm to detect pulmonary nodules. The medical approach for diagnosis used chest x-ray and CT scan.
LungCAD helped the clinician significantly to strengthen the accuracy of their diagnosis. LungCAD was also accepted by FDA in 2006.
Shivajirao M.Jadhav et al.[16] suggested a system that would use ECG recordings to spot Athymias in the human heart by training an artificial neural network (ANN) multilayer perceptron (MLP) on an ECG data set. The system proposed classified the patterns into two kinds a normal and an abnormal class. The data set was also used for training a modular artificial neural (ANN) network. The MLP system showed 86.67 percent accuracy of classification and 93.75 percent sensitivity, while the Modular ANN showed 93.1 percent accuracy.

Ö. Akin [17] had selected a resampling strategy which is based upon Random Forest (RF) ensemble classifier to improve diagnosis of cardiac arrhythmia The approach was found to be 90.0 percent accurate and the experiments revealed the efficiency of random sampling strategy in training RF ensemble classification algorithm was successful an approach.
In Pinar Yildirim's research [18], the effect of class imbalance was investigated in training data by taking into account the building of a neural network classifier for medical decisions on chronic kidney disease. A comparative study was carried out in the research by means of sampling algorithms based on multilayer perceptron with separate learning rate values for the prediction of CKD by considering neural networks. This study shows that sampling algorithms can improve the performance of classification algorithms. It also shows that the learning rate is a important parameter that has a considerable impact on multilayer perceptron.

In this study [19] the author has presented a hybrid system that integrates the genetic algorithm with the random forest for lymphatic disease prognosis. The lymph system is part of the body's immune system and carries fluids all across the body. The lymph system is part of the body's immune system and carries fluids all through the body. The genetic algorithm used to reduce the size of the lymph disease dataset and the random forest technique was used as a classifier. The suggested system performance was compared to that of other feature selection algorithms paired with an RF classifier such as the principal component analysis

(PCA),RF etc. The results of this study showed that GA-RF attained a higher level of classification accuracy of about 92.2 percent.

The authors [20] presented a system that was produced to assess the health of sick patients who were suffering from heart failure remotely. It also included data mining functionalities for continuous patient monitoring. A "Heart Failure severity assessment" approach was proposed which used Data Mining via CART classifiers. The results were obtained testing the system. The system accomplished precision and accuracy percentage of 96.39 and 100.00 respectively in HF detection and percentage of 79.31 and 82.35 in differentiating between severe and mild HF respectively.

III. Discussion

The results of reviewing previous studies presented in this work have several important implications:
In this review, we found that Machine Learning techniques have contributed significantly in developing methods for disease classification. These methods have been evaluated to classify diseases based on a number of real-world datasets, however, most of the methods developed by supervised methods do not support the incremental learning and only about 1% of methods can support incremental version of data mining techniques.
Developing efficient frameworks for big data analytic in the healthcare is one of the most important tasks, which should be taken into consideration. Developing a well-designed method for disease diagnosis from big data in healthcare is important. Disease diagnosis system developers need to consider the data sharing mechanism, privacy of data, the security of data, and the growth of data size. Moreover, in Machine Learning it is important to develop methods for disease diagnosis to be modified for big data environment easily.

IV. Conclusion

Some of the research gaps in disease classification using machine learning methods:

- **Data availability and quality:** One of the biggest challenges in machine learning for disease classification is the availability of high-quality data. Many diseases are rare, so there may not be enough data to train a machine learning model with high accuracy. Additionally, the data that is available may be noisy or incomplete, which can also affect the accuracy of the model.
- **Algorithm selection and tuning:** There are many different machine learning algorithms that can be used for disease classification. The choice of algorithm depends on the specific characteristics of the data and the desired performance of the model. However, there is no one-size-fits-all approach, and it can be difficult to select the best algorithm for a particular problem. Additionally, the parameters of the algorithm must be tuned to achieve optimal performance. This can be a time-consuming and challenging process.
- **Model interpretability:** Once a machine learning model has been trained, it is important to be able to interpret the results of the model. This is especially important for medical applications, where it is important to understand why the model made a particular prediction. However, many machine learning models are black boxes, and it can be difficult to understand how they make their predictions.
- **Bias and fairness:** Machine learning models can be biased, which means that they may not perform equally well for different groups of people. This is a particular concern for medical applications, where it is important to ensure that all patients have equal access to accurate diagnosis and treatment.

These are just some of the research gaps in disease classification using machine learning methods. As research in this area continues, these gaps will hopefully be addressed, and machine learning will become an

even more powerful tool for improving healthcare. In addition to the above, here are some other research gaps that are worth mentioning:

- The need for more robust and generalizable models: Current machine learning models for disease classification are often trained on datasets that are specific to a particular disease or population. This means that they may not perform well when applied to new data. There is a need for more robust and generalizable models that can be trained on a wider variety of data and that can perform well on new data.
- The need for models that can handle uncertainty: In the real world, there is often uncertainty about the diagnosis of a disease. This is due to factors such as the variability of symptoms, the presence of co-morbidities, and the limitations of medical tests. There is a need for models that can handle uncertainty and that can provide probabilistic predictions.
- The need for models that can be used in real-world settings: Many machine learning models for disease classification are developed in research settings. However, these models often need to be adapted for use in real-world settings, where there are different constraints and requirements. There is a need for research on how to adapt machine learning models for real-world use.

As stated in this work, machine learning techniques are applied in the realm of medical prediction. The main emphasis is on the use and synthesis of several methods for machine learning-based disease prediction. Algorithms including Decision Trees (J48), Support Vector Machines, Multilayer Perceptron, Bayes classifiers, and K-Nearest Neighbors Ensemble classifier approaches are covered. to facilitate doctors' work by identifying prediction principles from medical data sets The importance of machine learning in the medical field is recognised, and actions are being taken to incorporate pertinent methods in disease prediction. We looked at several research studies that were conducted using certain useful methodologies by diverse people.

Other diseases could be included in this study, which might yield different findings. Additionally, only studies that employed open-access UCI datasets in their research on Parkinson's, heart disease, diabetes, and breast cancer were taken into consideration for this study. As a result, in order to fully cover the application of data mining techniques for different types of diseases, additional databases must also be taken into account. As a result, depending on the limits of this study, we will continue our review in the future to categories research publications on an ongoing basis. For researchers' upcoming research, the review in this paper offers suggestions. Based on the findings of our evaluation, we advise researchers to employ huge datasets from other machine learning repositories in addition to UCI datasets in their experiments to guarantee the efficacy of the created method. Concentrating on large datasets enables them to assess the precision of the chosen classification and can inspire them to create methods for incremental learning to address the problem of temporal complexity. It is our aim that this study will help researchers create medical decision support systems with helpful data on data mining techniques, their usage in illness diagnosis.

References

[1]. Zhang G., “A Modified SVM Classifier Based on RS in Medical Disease Prediction”, 2009.
[2]. Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong,Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques,” Proc. Management and Innovation Technology International Conference (MITiCON-2016) , IEEE, Oct. 2016.

[3]. Uzer S. Mustafa, Inan O., Yilmaz N.," A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS, and PCA. Springer Journal of Neural Computing and application, 2012.
[4]. Senapati MR, Mohanty AK, Dash S, Dash PK , "Local linear wavelet neural network for breast cancer recognition",Neural Computing and Applications,Volume 22, Issue 1, 2013, pp. 125-131.
[5]. Levashenko V. Zaitseva E.,"Fuzzy Decission Trees in Medical Decision Making Support System",Proceedings of the Federated Conference onComputer Science and Information Systems, 2012, pp. 213–219.
[6]. Hilal A,R., Basir O., "Combination of enhanced AdaBoosting techniques for the characterization of breast cancer tumors", International Conference on Future BioMedical Information Engineering, 2009.
[7]. Jabbar M.A., Chandra P. Deekshatulu B.L., "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection", 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012.
[8]. Sapon M.A., Ismail K., Zainudin S. and Ping C.S., "Diabetes Prediction with Supervised Leaming Algorithms of Artificial Neural Network," Intemational Conference on Software and Computer Applications, Kathmandu, Nepal, 2011.
[9]. B. Sokouti, S. Haghipour, and A. D. Tabrizi, A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features in Neural Comput. Appl.,Vol. 24, No. 1 (2014), pp. 221–232.
[10]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016.
[11]. Salekin, Asif & Stankovic, John. (2016). Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. 262-270. 10.1109/ICHI.2016.36.
[12]. Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," July18, 2016.
[13]. Sahil Sharma, Vinod Sharma, Atul Sharma, " A Two Stage Hybrid Ensemble Classifier Based Diagnostic Tool for Chronic Kidney Disease Diagnosis Using Optimally Selected Reduced Feature ", (2018)
[14]. R.Bharat Rao, Jinbo Bi, Nancy Obuchowski and David Naidich, LungCAD: A Clinically approved Machine Learning System for Lung Cancer Detection in International conference on knowledge discovery and data mining (San Jose, California, USA, 2007)
[15]. Shivajirao M. Jadhav , Sanjay L. Nalbalwar and Ashok A. Ghatol, Artificial Neural Network Models based Cardiac Arrhythmia Disease Diagnosis from ECG Signal Data in International Journal of Computer Applications, Vol.44 ,No 15 (2012), pp. 8-13.
[16]. Akin O., "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," Computers in Biology and Medicine, vol. 41, 2011, pp. 265-271
[17]. Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE International Conference on Computer Software and Applications (COMPSAC), IEEE, Jul. 2017,
[18]. Elshazly H., Azar A.T., El-korany A., Hassanien A.E., "Hybrid System for Lymphatic Diseases Diagnosis", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013.
[19]. Pecchia L., Meililo P., and Bracale M.,"Remote Health Monitoring of Heart Failure with Data Mining via CART Method on HRV Feature". IEEE Transaction on Biomedical Engineering, Vol.58, 2011.