A Survey of Disease Classification Methods Using Machine Learning Algorithms

Shrusti Malviya[1], Chetan Agrawal[2], Darshna Rai[3]
Department of CSE, Radharaman Institute of Technology & Science, Bhopal, India[1, 2, 3]
shristimalviya77@gmail.com[1], chetan.agrawal12@gmail.com[2], darshna.rai@gmail.com[3]

Abstract: On a consistent basis, those working in the medical field must deal with enormous amounts of data. When dealing with enormous amounts of data, adopting the conventional approaches can have an effect on the findings. In the field of medical research, for instance, machine learning algorithms can be utilized to forecast the onset of disease. The review of a patient's medications and specialists cannot begin unless the sickness has been diagnosed at an early stage. The diagnosis of disease at an early stage Several different types of machine learning algorithms, including Decision Trees, Support Vector Machines, Multilayer Perceptrons, Bayes classifiers, and K-Nearest Neighbors Ensemble classifiers, are applied in order to efficiently diagnose a wide range of health problems and identify a variety of health problems. Using algorithms that are based on machine learning, it is feasible to swiftly and reliably forecast diseases. In this study, several types of diseases and their manifestations were predicted with the help of machine learning algorithms. This work covers a variety of problems, some of which include the prediction of chronic renal disease, the detection of diabetes, and the detection of breast cancer. Additionally, hybrid strategies for the enhancement of classifiers are investigated in this paper.

Keywords: Machine Learning, Classification, Disease Classification, Chronic Disease, Artificial Intelligence.

Introduction

In the field of artificial intelligence, machine learning is concerned with developing computer algorithms for storing and updating intelligent systems' knowledge as it changes. There are a number of ways in which Artificial Intelligence (AI) can be used to improve the chances of solving real world problems. Technology and scientific advancements have made AI an exciting field. As a result, machine learning approaches are getting more and more attention. In the field of data analysis, machine learning (ML) is a key technique that uses learning algorithms to iteratively learn from the data at hand. For Ethem Alpaydin, machine learning involves "programming computers to maximize performance criteria using example data or past experience." Learning is the execution of a computer program to optimize the parameters of a model based on training data or previous experience. Models can be predictive or descriptive, or both, depending on the goal of the data analysis. Machine learning is an application of artificial intelligence (AI) that allows computers to learn and can improve on their own without the explicit programming. The goal of machine learning is to create programs that can take data and learn from it themselves.

There are several ways to learn, but the most important one is by observing patterns and making better judgments based on what you've learned from previous experiences. Allowing machines to learn on their own, without human intervention or support is a critical goal. [1] The goal of artificial intelligence (AI) in medical science is to create tools that can assist doctors in making more precise diagnoses. Machine learning is heavily reliant on the prediction of disease. Artificial intelligence (AI) approaches can be used to anticipate a variety of diseases. Various illness kinds are predicted using machine learning techniques, which we explore in this section. This study focuses on the prediction of chronic renal disease, heart disease, diabetes, and breast cancer lymphatic system and pulmonary problems. Next, we'll give a brief overview of a few illnesses.

Various illnesses affecting the kidney's structure and function are together referred to as "chronic renal disease" [2]. It takes kidney damage (i.e. albuminuria) or decreased kidney function for a period of three months or more to be considered chronic kidney disease [2]. There are several significant side effects of chronic renal illness, but one of the most common is kidney failure. Complications can occur at any point in the process, and when they do, it's generally fatal. Heart disease kills an estimated 12 million people around the world each year, according to the World Health Organization. Women are more likely than men to develop breast cancer. Breast cancer is typically detected through mammography. Our focus here is on a variety of machine learning methods that are commonly used to detect breast cancer. Blood sugar can be flushed away through pee if insulin levels are too low or not being utilized appropriately by our bodies. Diabetes is the medical term for this illness. Additionally, the study discusses a variety of hybrid methodologies utilized in the medical industry to forecast disease. Incorporating the classification capacity of each classifier into the hybrid method improves overall performance while reducing the likelihood of a given instance being incorrectly classified. Various combinations of apprentices can be made. They're able to process all inputs simultaneously and aggregate their output in some way. If a classifier, an individual classifier, mistakenly classifies an instance, the error is repaired by other classifiers, an individual classifier. For example, in the case of a hybrid Algorithm that teaches an initial learner and then teaches successive learners on data that the first learner misclassifies, the shortcomings of each base-learner are covered up by the following learner in this manner. Figurative approach of diseases diagnosed by Machine Learning Techniques is shown in Figure 1.

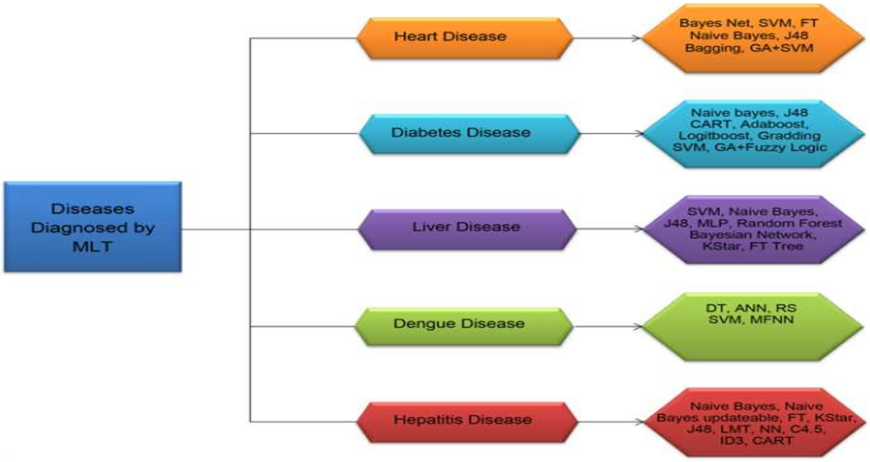
Diseases Diagnosed by MLT
Heart Disease
Bayes Net, SVM, FT Naive Bayes, J48 Bagging, GA+SVM
Diabetes Disease
Naive bayes, J48 CART, Adaboost, Logitboost, Gradding SVM, GA+Fuzzy Logic
Liver Disease
SVM, Naive Bayes, J48, MLP, Random Forest Bayesian Network, KStar, FT Tree
Dengue Disease
DT, ANN, RS SVM, MFNN
Hepatitis Disease
Naive Bayes, Naive Bayes updateable, FT, KStar, J48, LMT, NN, C4.5, ID3, CART

Figure 1: Diseases diagnosed by MLT.

The research objective of this survey is to provide a comprehensive overview of disease classification using machine learning algorithms. The survey aims to achieve the following objectives:

- **Review and summarize existing literature:** The first objective is to conduct an extensive review of the literature on disease classification using machine learning algorithms. This involves identifying and analyzing relevant research papers, articles, and studies published in the field. The goal is to provide a comprehensive overview of the existing approaches, methodologies, and techniques employed in disease classification.
- **Categorize disease classification methods**: This objective involves categorizing the different machine learning algorithms and techniques used for disease classification. It includes classifying the algorithms based on their underlying principles, such as supervised learning, unsupervised learning, semi-supervised learning, and transfer learning. The objective is to provide a structured classification framework that allows researchers and practitioners to understand the diversity of methods available for disease classification.
- **Identify challenges and limitations:** This objective involves identifying and discussing the challenges and limitations associated with disease classification using machine learning algorithms. It includes challenges related to data quality, class imbalance, feature selection, interpretability, generalization, and scalability. The objective is to provide a comprehensive understanding of the obstacles researchers may face and potential areas for improvement in future studies.
- **Emerging trends and advancements:** The objective is to highlight the emerging trends and advancements in disease classification using machine learning algorithms. This includes exploring the use of deep learning, ensemble methods, transfer learning, and other advanced techniques in the field. The goal is to provide insights into the latest developments and their potential impact on improving disease classification accuracy and efficiency.
- **Practical implications and future directions:** The final objective is to discuss the practical implications of disease classification using machine learning algorithms and provide insights into future research directions. This includes identifying potential applications of disease classification systems in clinical practice, healthcare decision-making, and personalized medicine. The objective is to stimulate further research and innovation in the field.

By accomplishing these research objectives, the survey aims to provide a comprehensive and up-to-date overview of disease classification using machine learning algorithms. It aims to serve as a valuable resource for researchers, practitioners, and healthcare professionals interested in understanding the current state of the field, identifying research gaps, and exploring opportunities for advancing disease classification methodologies.

II. Literature review

A hybrid breast cancer detection system was developed by MS Uzer et al. [1]. As a classification algorithm, Artificial Neural Networks were employed. In the case of cancer, the body's cells cease to function and begin to grow and divide inexplicably. When cells in the body lose their functions and begin to reproduce uncontrolled, they are in the state of cancer. The most common cause of death worldwide is cancer, namely cancer caused by malignant tumors. For the hybrid function selection, authors used Sequential Backward Selection, Sequential Forward Search, and PCA; this was followed by a 10-fold cross-validation strategy that yielded findings with 98.57% accuracy.

[2] Proposed a local linear neural wavelet network for breast cancer detection by training its parameters using Recursive Least Square (RLS) to increase its individual performance. [2] Based on comparisons with

other studies, it was determined that the suggested method is highly effective and yields an accurate classification.

In a study [3], the author used RS to alter the SVM classification for disease prediction. The strategy made use of RS's advantages in removing redundant data and of SVM's advantages in training and testing data. Three sets of data were utilized to test the method, and the results suggest that it is more accurate than other alternatives.

An easy-to-use decision support system based on Fuzzy Logic and Fuzzy Decision Trees (FDT) was recommended by author [4] FDT to improve categorization performance. Two breast cancer datasets were tested using a 70:30 split ratio, with error rates of 0.3661 and 0.1414, respectively. Decision support systems based on Fuzzy Logic and Fuzzy Decision Trees can be improved by enhancing their categorization performance Use of the FDT, which is both simple and relevant, was made. Both the breast cancer data sets with error rates of 0.3661 and 0.1414 were tested using a 70:30 split ratio.

Author [5] suggested the use of boosting in order to improve the accuracy of CAD-based procedures, and a hybrid boosting algorithm was developed to combine the benefits of several boosting techniques. The hybrid boosting algorithm outperforms the other boosting algorithms when applied to real breast cancer cases, according to the results of the tests.

We describe in this study [6] an approach to estimating Andhra Pradesh's cardiovascular risk score using fewer attributes. The use of other characteristic selection methods like IG, SU, and genetic search has also been used to identify the specific qualities that improve heart disease prediction and, as a result, help to minimize the number of diseases in the population.. The author used feature subset selection to build class association rules. Predicting a patient's risk of heart disease using these created criteria would be helpful to clinicians, and this approach could be tested on public datasets and compared to other solutions.

Prediction models for type II diabetes risk were built using three algorithms compared by the author [7]. Artificial neural networks (ANN), Naive Bayes, and K-nearest neighbours were the three algorithms used to model Type II diabetes, a condition in which the pancreas fails to generate enough insulin to keep blood sugar levels normal (KNN). The neural network beats the prediction accuracy of Nave Bayes (95%) and KNN (91%).

To aid in the diagnosis of cardiovascular illness Multi-layer feed-forward neural networks (MLFFNs) and genetic algorithms were introduced by T. Manju et al. [8]. Diabetes, a poor diet, high blood pressure, obesity, and smoking all contribute to heart attacks, which are the leading cause of mortality. Predicting heart attacks can now be done using the Multi Layer Feed Forward Neural Network, which combines the Genetic Algorithm with the Back Propagation Network (BPN). Only six of the 13 parameters in the data set were used to train the neural network. When ANN was first developed, its primary goal was to predict whether a patient was at risk of having a heart attack.

It was predicted by Babak Sokouti et.al. [9] That a Levenberg–Marquardt (LM) feed-forward MLP neural network would be used to classify cervical cell pictures obtained from 100 patients.

Pre-processing/processing of images and feed-forward MLP neural network make up the semi-automated cervical cancer diagnostic system. Cervical cell pictures were successfully classified using the proposed method and had a high classification rate, according to the findings. Cancer and precancerous cells can be identified with the help of the suggested semi-automated approach.

To predict chronic renal disease, Charleonnan et al. investigated four machine learning algorithms, including the k-nearest neighbors (KNN), SVM, logistic regression (LR), and decision tree classifiers. The performance of these models, to forecast the disease, was compared with each other. The SVM classifier achieved a maximum accuracy of 98.3% in the experiments. According to the presented method, SVM also demonstrates the best accuracy. Because of this, the SVM classifier can be used to predict chronic renal disease.

Research undertaken by Asif Salekin and John Stankovic [11] has offered an alternative method for determining the CKD using machine learning techniques. They used k-nearest neighbors, random forests, and neural networks as classifiers to come up with a suitable answer. Researchers used a wrapper method to run a function reduction analysis to identify the characteristics most accurately indicative of CKD. Only five parameters are needed to identify a cost-effective, highly accurate CKD detection classifier. A detection accuracy of 0.993 and an RMS error of 0.1084 were attained with this strategy, according to the F1 metric.

Wrapper subset attribute evaluator and best first search method were used to build the classification model for CKD and non-CKD patients, respectively. The models performed better in the classification of CKD and non-CKD cases. [12] Authors compared the outcomes of several models. During the comparison, it was found that the classifiers performed better when compared to the original data set when it was reduced in size.

Classification methods were evaluated using a data set provided by Sahil Sharma, Vinod Sharma, and Atul Sharma [13]. Efficacy was calculated by comparing the candidate methods' results to the actual medical outcomes of the subject. Predictive accuracy, sensitivity, precision, and specificity were utilised as performance indicators by the authors. Testing revealed that the decision tree technique functioned well, reaching an accuracy of nearly 98%, a specificity of 1, an accuracy precision ratio of 0.9720, and an accuracy sensitivity ratio of 1.

A two-stage hybrid ensemble approach was proposed by the authors [14]. The two-stage hybrid ensemble classifier includes the capabilities of various classification techniques. When it comes to classifying chronic renal illness, a two-stage hybrid ensemble is an excellent option. Accuracy(2-class), sensitivity, precision, specificity, and other performance measures for the reduced feature set and the whole feature set were evaluated using this ensemble classifier and found to be successful. Additionally There has been work done on a diagnostic GUI tool that could assist doctors in validating their findings.

Computer-aided diagnosis (CAD) system dubbed lungCAD was designed by R Rao R Bharat et al. [15] (CAD). A classification technique was utilized to identify pulmonary nodules from the CT thorax investigations. A chest x-ray and a CT scan were used to make a medical diagnosis.

The use of LungCAD considerably improved the clinician's ability to make an accurate diagnosis. In 2006, the Food and Drug Administration (FDA) approved LungCAD.

An artificial neural network (ANN) multilayer perceptron (MLP) was trained on an ECG data set by Shivajirao M.Jadhav et al.[16] to detect Athymias in the human heart. According to the proposed method, patterns might be divided into two categories: normal and abnormal. A modular artificial neural (ANN) network was also trained using the data set. The MLP system had an accuracy rate of 86.67 percent and a sensitivity rate of 93.75 percent, while the Modular ANN had an accuracy rate of 93.1 percent.

Random Forest (RF) ensemble classifier Re-sampling was chosen by. Akin [17] is to improve cardiac arrhythmia diagnosis. The random sampling technique was found to be effective in training the RF ensemble classification algorithm and the trials showed that it was a successful approach with an accuracy of 90%.

Research by Pinar Yildirim [18] examined the impact of class imbalance in training data on the development of a neural network classifier for medical decisions related to chronic kidney disease. In order to compare sampling methods based on multilayer perceptron with distinct learning rates for the prediction of CKD, neural networks were taken into consideration in the research. Classification algorithms can benefit from sample algorithms, according to a new study. Multilayer perceptron performance can be significantly impacted by the learning rate, as this study demonstrates.

In this study [19], the author presented a hybrid method for lymphatic illness diagnosis that combines the genetic algorithm with the random forest. The lymphatic system transports fluids throughout the body and is an integral aspect of the immune system. The lymphatic system transports fluids throughout the body and is an integral aspect of the immune system. To decrease the lymph disease dataset, genetic algorithms were utilized in conjunction with random forest classification techniques. There was different feature selection algorithms linked with RF classifiers such as the PCA, RF etc. that were tested to see how well the proposed system performed in comparison. As a consequence of this study, the classification accuracy of GA-RF was found to be 92.2 percent.

III. Problem Statement

Using machine learning methods, the problem at hand is to create a system for disease classification that is both accurate and efficient. Because there is now more medical data available than ever before, there is an ever-increasing demand to automate the process of disease detection and classification. Traditional methods of disease classification frequently rely on manual interpretation by healthcare experts. This method can be laborious, fraught with subjectivity, and rife with the possibility of making mistakes. By utilizing extensive medical datasets and utilizing automated pattern recognition, machine learning algorithms have the potential to greatly increase the accuracy as well as the efficiency of disease classification.

Nevertheless, the creation of a reliable disease classification system based on machine learning algorithms presents a number of obstacles. Because of the variety of clinical measurements, imaging data, patient demographics, and medical records that are included, it can be challenging to isolate important characteristics and patterns from medical data. Additionally, the presence of class imbalance, which occurs when specific diseases may have limited cases or infrequent occurrences, adds an extra barrier to the process of achieving balanced and correct classification. Additionally, the success of machine learning algorithms is strongly dependent on the quality as well as the representativeness of the training data. It is possible that there will be a shortage of labeled medical data to use for educational purposes, and the acquisition of such data may be time-consuming and expensive. In addition, concerns about privacy and rules designed to protect personal data may impose limitations on how medical records can be used and shared.

Therefore, the main objective of this project is to develop a disease classification system that addresses these challenges and achieves high accuracy, robustness, and generalizability. The system should be able to successfully utilize numerous medical data sources, be able to deal with class imbalance, and be able to accommodate restricted labeled data availability. It is possible that the created disease classification system will be able to assist medical professionals in correct diagnosis, the planning of therapy, and the management of patients, which will ultimately lead to improvements in the quality of healthcare results.

IV. Discussion

This study's findings have a number of critical consequences, including:
These methodologies have had a considerable impact on the development of disease classification methods. Classifying diseases using a variety of real-world datasets, these methods have been tested for their ability to enable incremental learning, although only around 1% of the systems produced via supervised techniques do so. Big data analytic framework development in the healthcare industry should be prioritized because it is critical. It is critical in the healthcare industry to build a mechanism for accurately diagnosing disease using large amounts of data. The data sharing mechanism, privacy of data, security of data, and the expansion of data size must be taken into consideration by disease diagnosis system developers.

Furthermore, in Machine Learning, methods for disease diagnosis should be developed that can be easily adapted to a big data environment.

V. Conclusion

This study discusses the use of machine learning algorithms in the medical prediction field. Using machine learning to make predictions about various diseases, the key focus is on combining various algorithms. We talk about algorithms like decision trees (J48), support vector machines, multilayer perceptron, Bayes classifiers, and K-Nearest Neighbors Ensemble classifiers. To identify patterns in medical data that can be used by clinicians to speed up their workflow. Steps are taken to employ applicable machine learning techniques in disease prediction, acknowledging their importance. We looked at a variety of research projects, and some of the most effective strategies used by others.

Other diseases could be included in this study, resulting in different findings. In addition, we only looked at studies that used UCI public datasets in their research on Parkinson, Heart, Diabetes, and Breast Cancer when compiling this review. Data mining strategies for various sorts of diseases will require new datasets, therefore those should be explored as well. Because of this, we plan to conduct a follow-up evaluation in order to classify research papers, based on the constraints of this review, on an ongoing basis. For future research, the findings of this study can be used as a guide. As a result of our review, we recommend that researchers employ huge datasets from other machine learning repositories as well as UCI datasets in their experiments to ensure the efficacy of the method created. Using huge datasets not only allows them to assess the accuracy of the accepted classification, but it can also drive them to develop incremental learning approaches for the temporal complexity issue. Hopeful outcomes include a better understanding of the methodologies used in illness detection and the development of medical decision support systems based on data mining.

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., Meilillo 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.