THE APPLICATION OF MACHINE LEARNING TO THE PREDICTION OF DISEASE

[#1]Ms. J. SWATHI, *Associate Professor,*
[#2]Ms. G. LAKSHMI, *Assistant Professor,*
[#3]SD MEER SUBAN ALI, *Associate Professor,*
Department of Computer Science & Engineering
TRINITY COLLEGE OF ENGINEERING AND TECHNOLOGY, PEDDAPALLY

ABSTRACT: The increased usage of digital data collection systems has resulted from the growth of electronic medical equipment. The current deluge of data available to healthcare practitioners makes it more difficult to distinguish between symptoms and arrive at an accurate diagnosis of diseases as soon as feasible. Previous research has demonstrated that supervised machine learning (ML) algorithms outperform traditional diagnostic processes, making them valuable for the early detection of potentially fatal diseases. The goal of this study is to assess the efficacy of multiple supervised machine learning models for disease diagnosis by contrasting and comparing their degrees of performance. Three of the most well-known algorithms for supervised machine learning are Naive Bayes (NB), Decision Trees (DT), and K-Nearest Neighbors (KNN). Support Vector Machines, or SVMs, have been shown to be effective in the identification and diagnosis of Parkinson's disease and kidney diseases. Logistic Regression (LR) has been carefully explored and proven for its capacity to effectively predict cardiovascular disease. During the course of this investigation, the Random Forest (RF) and Convolution Neural Networks (CNN) models were used to anticipate breast cancer and other major diseases.

1. INTRODUCTION

Motivation

The advent of artificial intelligence (AI) has enabled the development of robotic systems with human-like perception, cognition, and behavior. The phrase "artificial intelligence" (AI) is an umbrella word that covers a wide range of specific subfields. Machine learning, computer vision, deep learning, and natural language processing are a few examples of these subfields. The purpose of the many diverse statistical, probabilistic, and optimization methodologies used by machine learning algorithms is to improve future decision-making based on acquired data. These algorithms have applications ranging from disease management and the detection of fraudulent credit card activity to the analysis of consumer spending trends and network intrusion detection. Some of these systems were created using the supervised learning method. To be successful, prediction models must be challenged to forecast for unlabeled scenarios using labels that they are already familiar with. According to the aforementioned theory, using supervised learning as a powerful diagnostic tool may help medical professionals make more accurate diagnoses.

According to data collected by the Medicaid and Medicare departments in the United States, more than half of the population has a chronic health condition. This proportion is determined by the number of people who qualify for Medicaid and Medicare. As a result, the nation's healthcare system is under unprecedented financial duress, with costs expected to reach an all-time high of $3.3 trillion in the United States alone in 2016. It is expected that the average cost for each individual will be around $10,348. According to reports from both the World Health Organization and the World Economic Forum, India incurred a 236.6 billion dollar loss in economic output in 2015. This tragedy was caused by the widespread spread of diseases that can lead to mortality as a result of poor nutrition and inactivity. The preceding figures highlight the importance of early disease detection in lowering the death rate associated with various diseases, the expenses of which have already been highlighted. Early disease prediction has the potential to result in a variety of favorable outcomes, including enhanced community health management and lower medical care expenditures.

According to Yuan (year), the poor performance of machine learning algorithms can be attributed to two major issues. To acquire results that are reliable and meaningful, it is necessary to perform an in-depth investigation of, and debate about, the databases and their contents. It is widely acknowledged that selecting suitable qualities from a dataset for use in machine learning algorithms can be difficult due to the problem's complexity, resource allocation considerations that must be made, and the amount of time that must be spent. The existence of these variables reduces the learning model's effectiveness and leads to serious errors that endanger patients' lives and health. Ismaeel argued that the combination of standard statistical methodologies, professional knowledge accumulated over years of experience, and clinicians' intuitive evaluations could distort the identification of risk variables for the condition. All three of these elements are based on years of clinical expertise. As more health records are maintained electronically, diagnosing diseases in their early stages becomes more challenging. Machine learning (ML) has seen the development of a variety of algorithms and cutting-edge computer technologies that aid in the discovery and extraction of significant patterns and hidden insights from data, with the ultimate goal of assisting intelligent decision-making. The reduction in the quantity of work required of medical professionals coincided with an increase in the percentage of patients who lived.

Aim

The goal of this project is to explore the hypothesis that using supervised machine learning algorithms results in better health outcomes by allowing for faster and more accurate disease diagnosis. We undertake a literature review on the various supervised machine learning models and how they could be applied to the problem of disease diagnosis in this work. It is impossible to prevent bias when analyzing how well a particular algorithm performs in a range of scenarios. This technique, on the other hand, produces more exhaustive and precise findings. The research on machine learning models will primarily focus on difficulties concerning the cardiovascular system, kidneys, breasts, and brain system. The software will test a variety of machine learning algorithms, including K-Nearest Neighbors, Naive Bayes, Decision Trees, Convolutional Neural Networks, Support Vector Machines, and Logistic Regression, to see how well they function as diagnostic tools. The essay concludes with a list of the most effective machine learning models that may be applied to a variety of use cases.

2. LITERATURE REVIEW

Common Diseases

Dahiwade and his colleagues proposed machine learning as a tool for predicting the spread of epidemic diseases. Because to the UCI Machine Learning repository, the symptoms of a wide range of common diseases may now be found in one place. The system employs both the Convolutional Neural Network (CNN) and the K-Nearest Neighbors (KNN) classification algorithms to forecast probable medical issues. Additional information about one's lifestyle was added into the planned course of therapy in order to assess the level of risk associated with the projected disease. Dahiwade et al. compared and contrasted the processing speeds and results of the K-Nearest Neighbors (KNN) and Convolutional Neural Network (CNN) techniques. CNN's processing speed was timed at 11.1 seconds, and the network's accuracy was determined to be 84.5 percent. A recent statistical analysis found that the K-nearest neighbors (KNN) method fared worse than the convolutional neural network (CNN) method. The findings of this study corroborate the premise proposed by Chen et al., namely that CNNs outperform more traditional supervised algorithms such as K-Nearest Neighbors, Naive

Bayes, and Decision Trees. According to the authors, the improved performance of the recommended model is attributable to its ability to uncover tiny nonlinear correlations between features. The model gets its name from this ability. Because CNN performs in-depth assessments of relevant parameters, a clearer picture of the current state can be formed, leading to an improvement in illness prognostication accuracy in scenarios with a high level of complexity. The prior assumption is supported by a large amount of evidence. Nonetheless, the models presented in the study did not sufficiently reflect the attributes of the neural networks employed in the study. These characteristics include a wide range of aspects such as network size, architecture, learning rate, and back propagation. Because precision is such a high emphasis in performance evaluation, the question of whether the data acquired is reliable arises. When discussing algorithms, the study's authors did not address the issue of bias. When an algorithm first performs poorly, it is common to detect a significant improvement in efficacy after adding extra feature variables. This improvement can be attributed to the algorithm's ability to handle the new variables more effectively.

Kidney Diseases

The Kidney Function Test (KFT) dataset was used in this work to assess the performance of different classifiers in the diagnosis of Chronic Kidney Disease (CKD). In this study, the F-measure, precision, and accuracy metrics of three distinct classifiers are compared and contrasted. K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF) are the names of these classifiers. NB had a higher level of precision, although RF scored higher on both the F-measure and accuracy measures, according to the data. Vijayarani believed that by combining the NB and SVM algorithms, she would be able to detect babies and adolescents at risk of developing renal disease. After doing classifier analysis on their data, patients with renal illness were categorized into four unique categories. To mention a few, this group of diseases includes acute nephritic syndrome, acute renal failure, chronic glomerulonephritis, and chronic kidney disease (CKD). The secondary purpose of the study was to identify the classification approach that was both helpful and effective. Based on the evidence, it appears that the Support Vector Machine (SVM) algorithm is superior to the less precise Naive Bayes (NB) technique. NB, on the other hand, was able to categorize data at hitherto unheard-of speeds. Both Charleonnan et al. and Kotturu et al. concluded that the support vector machine (SVM) classifier is the best option for kidney diseases. They attribute its outstanding performance to its ability to handle semi-structured and unstructured data well. Several empirical studies have been undertaken to determine how to diagnose chronic kidney disease (CKD). The Support Vector Machine (SVM) has been demonstrated to be better to other approaches in efficiently diagnosing complex renal illnesses due to its capacity to include larger data sets. This is due to the fact that other methods are available. Despite the fact that this finding is supported by data, there are claims that not all relevant hyper-parameters were included in the evaluation of the effectiveness of machine learning algorithms. These claims call the veracity of this finding into question. Experimenting with different hyper-parameter elements may result in outcomes with greater accuracy variability and better overall performance within ML systems, according to Uddin (year).

Heart Diseases

Marimuthu et al. (year) investigated the prospect of using supervised machine learning algorithms to predict cardiac disease. Based on a variety of criteria, the researchers decided not to consider some pieces of data. Decision Trees (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Naive Bayes (NB) are four of the most often used machine learning methods nowadays. According to the study's findings, the LR algorithm had an accuracy rating of 86.89%, which was much higher than the other algorithms studied. Dwivedi worked hard in 2018 to improve the accuracy of heart disease prognosis by including other parameters such as resting blood pressure, serum cholesterol in milligrams per deciliter, and peak heart rate. This was done in order to increase the prognosis' accuracy. The UCI Machine Learning group generously shared their dataset with the scholarly community. The collection included 270 samples, 120 of which were found to be positive for heart disease and the remaining 150 to be negative for the disease. Dwivedi investigated Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), and Classification Tree. After 10 rounds of cross validation, Logistic Regression (LR) was shown to have the highest classification accuracy and sensitivity. This study implies that LR can identify heart problems more precisely. Logistic Regression outperformed other methods such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Adaboost, according to the outcomes of a research project conducted by Polaraju and Vahid et al. The research was conducted properly, most notably in terms of how thoroughly it evaluated a range of machine learning models. An in-depth examination of the interactions between various hyperparameters was conducted in order to determine which technique to machine learning achieves the best possible balance of accuracy and precision. Despite this advantage, the restricted number of imported datasets limits the accuracy and precision of the learning models employed in disease identification.

Breast Diseases

Shubair used a range of machine learning algorithms to detect breast cancer in her patients. Among these were Support Vector Machines (SVM), Bayesian Networks, and Random Forests (RF). The major dataset of breast cancer patients in Wisconsin has been granted to researchers at the University of California, Irvine (UCI). Using the provided dataset, several different learning models were assessed, and the results were rated based on parameters such as accuracy, recall, precision, and ROC area under the curve. K = 10 folds were used in a K-fold cross-validation test to determine the accuracy of the classifiers. As a result of the simulation, Support Vector Machines (SVM) were found to be effective in terms of recall, precision, and accuracy. The findings demonstrate this. Despite the fact that both models had a fair likelihood of detecting malignancies, the ROC curve showed that the random forest (RF) model performed better than the other. Yao conducted an experiment using both of the well-known data mining techniques of Random Forest (RF) and Support Vector Machine (SVM) to assess the usefulness of these techniques in predicting breast cancer. With a classification rate of 96.27 percent, sensitivity of 96.78 percent, and specificity of 94.5 percent, the Random Forest technique proved highly efficient. The SVM technique outperformed the alternatives in terms of accuracy (95.85 percent), sensitivity (95.7 percent), and specificity (95.5 percent). According to Yao's research, the Random Forest (RF) strategy, as opposed to the Support Vector Machine (SVM) approach, can generate more accurate data predictions for each feature attribute. When compared to other popular methods to the problem, the Random Forest (RF) method exhibits lower sensitivity to volatility and data overfitting. Furthermore, when used to massive amounts of data, it has good scalability characteristics. As a result, RF has the potential to become the method of choice for classifying breast diseases. The study's inclusion of such a huge number of performance metrics added to the assertion's believability. Pre-processing was intended to prepare raw data for training, but it was revealed that it reduced the performance of machine learning models. The purpose of pre-processing was to prepare raw data for training. Yao claims that the machine learning system's performance is hampered since the image quality that results from disregarding some input is lower than expected.

Parkinson’s disease

Because support vector machines (SVMs) and fuzzy k-nearest neighbors (FKNNs) are both considered cornerstone methods, the goal of this study was to compare and contrast the two approaches. Chen et al. (year) developed a cutting-edge Parkinson's disease (PD) diagnosis system using the Fuzzy k-Nearest Neighbor (FKNN) algorithm. Principal Component Analysis (PCA) was used to incorporate the most discriminatory qualities, and as a result, an enhanced Fine K-Nearest Neighbors (FKNN) model was constructed. The UCI collection contained 31 participants for whom biological speech measurements were taken; 24 of these patients had Parkinson's disease. The findings of experiments support the conclusions, which show that the FKNN technique outperforms the SVM approach in terms of sensitivity, accuracy, and specificity. I'm going to start looking for this person as soon as I possibly can. When Behroozi's innovative framework for the Parkinson's disease diagnosis procedure was combined with a filter-based technique for feature selection, the classification accuracy increased significantly. This improvement could be as much as 15%. The problem of insufficient core data was overcome by using classifiers of various types to confirm the dataset's division into geographical regions. This provided a solution to the problem. Many classification algorithms, including as k-nearest neighbors, support vector machines, discriminant analysis, and naive bayes, are used in the work covered here. According to the study's conclusions, the Support Vector Machine (SVM) fared better than the many other statistical approaches considered. In addition to Support Vector Machines (SVM), Eskidere has spent a large amount of time and effort evaluating the performance of a range of alternative categorization techniques. Different approaches that can be employed include the Least Squares Support Vector, also known as LS-SVM, the General Regression Neural Network, also known as GRNN, and the Multi-layer Perceptron Neural Network, also known as MLPNN. Based on the results, it appears that the LS-SVM model should be used. This discovery is backed up by a considerable amount of research on decoders, which includes a comparison of their capabilities and an examination of how they perform in contrast to specified performance benchmarks. Machine learning algorithms, according to Lavesson, are tailored to maximize the usefulness of various performance measurements. Unlike KNN and SVM, which prioritize accuracy above the amount of false positives, neural networks are designed to minimize squared error. Furthermore, the authors have an exceptional ability to describe complex models. The kernel and regularization parameters of support vector machines (SVMs) have been extensively studied. The machine learning models, on the other hand, were not calibrated prior to data analysis. Calibration of a range of learning models, such as Naive Bayes, Support Vector Machines, and Random Forests, can be advantageous, according to Caruana (year).

3. CONCLUSION

Several machine learning algorithms have proven to be useful in the early detection of certain cardiovascular, renal, mammary, and neurological disorders. According to the published research, the prediction algorithms Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR) are the most commonly used. Precision is the single most important component in determining total performance, according to all of the study. When it came to recognizing common conditions, the CNN model outperformed other approaches by a wide margin. Because of its consistency in dealing with datasets with high dimensions, semi-structured formats, and unstructured data, the Support Vector Machine (SVM) model has consistently demonstrated outstanding accuracy in the detection of renal issues and Parkinson's disease (PD). This is due to the model's consistency in dealing with high-dimensional datasets, semi-structured formats, and unstructured data. In the field of breast cancer prediction, studies have demonstrated that the Random Forest (RF) algorithm has a better chance of accurately detecting disorders. According to the conclusions of these investigations, this is the case. This is owing to its capacity to process large datasets quickly while reducing the possibility of overfitting. The logistic regression (LR) approach outperformed the competition when it came to forecasting cardiovascular problems.

The development of complicated machine learning systems should be the key focus of future research in order to increase disease prediction accuracy. Following the completion of the training phase, it is recommended that the learning models be updated more often in order to increase their overall performance. This would be done in order to increase the learning models' performance. To limit the likelihood of overfitting and improve the quality of the models used, datasets should be supplemented with a wide range of demographic and visual information. The adoption of more relevant feature selection strategies is expected to improve the effectiveness of learning models.

REFERENCES

[1] Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[2] Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatolog- ical disorders through image-processing and machine learning," in 2017 International Conference on Intelligent Sustainable Systems (ICISS),2017, pp. 1047–1051.

[3] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

[4] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305.

[5] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in 2018 4th International Conference on Computing Com- munication and Automation (ICCCA), 2018, pp. 1–4.

[6] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra,"A proposed model for lifestyle disease prediction using support vector machine," in 2018 9th International Conference on Computing, Com- munication and Networking Technologies (ICCCNT), 2018, pp. 1–6.

[7] F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Confer- ence on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907.

[8] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning ma- chine (elm) technique for heart disease diagnosis," in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015,pp. 1–3.

[9] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," Proceedings of the 3rd Inter- national Conference on Computing Methodologies and Communication,ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.

[10] S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities," International Journal of Scientific Research in Science and Technology, pp. 29–35, 2019.