International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-13, Issue-3 (June 2023)
https://doi.org/10.31033/ijemr.13.3.40

Disease Prediction using Machine Learning

Mohd. Nadeem Khan[1] and Ankita Srivastava[2]
[1]Student, Department of Computer Science and Engineering, Integral University Lucknow, INDIA
[2]Professor, Department of Computer Science and Engineering, Integral University Lucknow, INDIA

[1]Corresponding Author: nadeemk0076@gmail.com

Received: 30-05-2023 Revised: 17-06-2023 Accepted: 30-06-2023

ABSTRACT

Based on predictive modelling, a disease prediction system determines the user's illness from the signs they provide as input to the system. The system evaluates the user's symptoms as input and outputs the likelihood that the illness will occur. Utilizing a decision tree classifier, disease prediction is accomplished. The likelihood of the illness is calculated by a decision tree classifier. More and more organisations in the biological and healthcare sectors are turning to big data to aid in early disease discovery and better serve their patients. The development of a system that will allow people to forecast chronic illnesses without having to see a doctor or medical professional for a diagnostic is necessary by watching patient signs and using a variety of machine learning modelling methods, different illnesses can be identified. Text and organised data processing do not follow any standard method. The suggested paradigm would examine both organised and random material. Prediction precision can be increased through machine learning. There is a need to research and develop a system that will allow an end-user to anticipate irreversible illnesses without having to consult a specialist or doctor for a diagnostic. To identify different diseases by analysing patient symptoms using various techniques of Machine Learning Algorithms. There is no proper technique for managing text and structured data. Both organised and unstructured material will be examined by the Suggested algorithm. Artificial Learning will improve prediction accuracy.

Keywords-- Disease Prediction, Machine Learning, Decision Tree Classifier, Healthcare, Chronic Illnesses, Predictive Model

I. INTRODUCTION

Programming machines to perform better using sample data or historical data is known as machine learning. Machine learning is the study of computer programs that gain knowledge from past experience and data. The training and testing stages of a machine learning system. disease diagnosis based on the signs and medical background of the patient Machine learning technology has improved over the years.

The medical field now has an incomparable platform thanks to machine learning technology, making it possible to fix healthcare problems quickly. Machine learning is being used to keep full hospital data. With the aid of machine learning technology, physicians can more accurately identify and treat patients, which improves patient healthcare services. Machine learning technology enables creating models to analyze data rapidly and deliver findings quicker.

The use of machine learning in the medical sector is best illustrated by the case of healthcare. To improve the accuracy from large amounts of data, work is currently being done on unstructured and written data. The current will use linear, KNN, and decision tree algorithms for illness prognosis. The collection of references at the conclusion of the paper should be cited in the same sequence as the main text.

II. LITERATURE REVIEW

There have been many research conducted on the topic of disease prediction utilizing various machine learning approaches and algorithms that medical institutions can apply. In this paper, some of those investigations are reviewed along with the methods and findings they employed.

In their study, **MIN CHEN et al. [1]** suggested a disease prediction system based on machine learning techniques. He employed approaches such as CNN-UDRP, CNN-MDRP, Naive Bayes, K-Nearest Neighbor, and Decision Tree to forecast disease. The proposed technique was 94.8% accurate.

Disease Risk Prediction was advised by **Sayali Ambekar et al., [2]** who carried out the task using a convolution neural network. Machine learning methods including the CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are employed in this research. The system employs structured data to be trained, and Naive Bayes is used to attain an accuracy of 82%.

A system that provides better outcomes for disease prediction was created by **Naganna Chetty et al. [3]** using a fuzzy approach. and employed methods such as fuzzy c-means clustering, fuzzy KNN classifier, and fuzzy KNN classifier. In this study, the accuracy of the predictions for the diseases of diabetes and liver disorders is 97.02% and 96.13, respectively.

A model for disease prediction was created by **Dhiraj Dahiwade et al. [4]** utilizing machine learning methods and KNN and CNN techniques. This research proposes disease prediction, based on the symptoms of

Peer Reviewed & Refereed Journal
Volume-13, Issue-3 (June 2023)
https://doi.org/10.31033/ijemr.13.3.40

the patient. CNN and KNN both have accuracy ratings of 98% and 95%, respectively.

Utilizing distributed machine learning classifiers, **Lambodar Jena et al. [5]** concentrated on risk prediction for chronic diseases using methods like Naive Bayes and Multilayer Perceptron.The accuracy of Naive Bayes and Multilayer Perceptron in this paper's attempt to forecast Chronic Kidney Disease is 95% and 99.7%, respectively.

Principal component analysis was used by **Dhomse Kanchan B. et al. [6]** to study the prediction of specific diseases using machine learning algorithms that included Naive Bayes classification, Decision Tree, and Support Vector Machine approaches. This approach has a 34.89% accuracy rate for diabetes and a 53% accuracy rate for heart disease.

Using machine learning applications and techniques, **Pahulpreet Singh Kohli et al. [7]** proposed disease prediction using methods like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and Adaptive Boosting. This essay focuses on the prognosis of diabetes, breast cancer, and heart disease. Logistic regression yields the highest accuracy rates, which are 95.71% for breast cancer, 84.42% for diabetes, and 87.12% for heart disease.

Using Naive Bayes and KNN algorithms, **Deeraj Shetty et al. [8]** investigated the applications of data mining for diabetic illness prediction. With this method, diabetes can be predicted with greater accuracy than with Naive Bayes thanks to KNN.

The Random Forest Algorithm was employed by **Rashmi G. Saboji et al. [9]** in an effort to find a scalable system that can predict heart disease via classification mining. The Nave-Bayes classifier is compared by this method, however Random Forest provides findings that are 98% more accurate.

Rati Shukla et al. [10] proposed using machine learning methods as Decision Tree, Support Vector Machine, Random Forest, Nave Bayes, Neural Network, and KNN to predict and detect breast cancer. The Support Vector Machine provides outcomes in this system that are more accurate than any other algorithm.
Following approaches over Disease Prediction :-

Common Diseases - A method built on machine learning that can foresee widespread illnesses was suggested. The UCI ML repository was mined for the symptoms dataset, which included information on the signs of numerous illnesses. The algorithm was able to forecast numerous illnesses using CNN and KNN categorization methods. Moreover, extra information about the assessed patient's lifestyle was included in the proposed response, which was helpful in understanding the gravity of the disease risk. The KNN and CNN algorithms were compared for their processing times and accuracy. The average working time for CNN was 11.1 seconds, and its accuracy was 84.5%.Statistical evidence shows that the CNN algorithm outperforms the KNN algorithm.

The results of this research corroborated previous ones showing that CNN is superior to the more common guided algorithms like KNN, NB, and DT. According to the authors' analysis, the suggested model's superior precision can be traced back to its superior ability to identify intricate nonlinear connections between features. In addition, CNN can correctly forecast illnesses of high intricacy because it can identify characteristics with high significance that give improved depiction of the disease.

Kidney Diseases - Using data from The Kidney Function Test, this research aimed to compare the efficacy of different algorithms for detecting CKD. The KNN, NB, and RF algorithms are used, and their precision, F-measure, and accuracy are compared to see which one provides the best results. When comparing F-measure and accuracy, RF performed better, while NB was more precise. **Vijayarani [11]** planned this research with the intent of determining if support vector machines (SVMs) and neural networks (NBs) were effective at identifying renal illnesses.

Further, the study aimed to identify the most efficient categorization method in terms of both precision and runtime. Based on the outcomes, SVM proved to be the superior performing algorithm by achieving significantly greater accuracy than NB. On the other hand, NB categorised material quickly. The inference is backed by the data, but it is undermined by the fact that it was previously suggested that the efficacy of ML algorithms should have been evaluated without experimenting with various hyper-parameters. Experimenting with various hyper-parameters allows for a wide range of precision outcomes and the possibility of improved ML algorithm performance.

Heart Diseases - The goal of this work was to use guided machine learning methods to make predictions about cardiac conditions. The characteristics of the data were organised by the writers according to gender, age, chest discomfort, gender, goal, and incline. There were four ML algorithms used: DT, KNN, LR, and NB. The study showed that the LR algorithm was the most efficient of the bunch, with a precision of 86.89%.

In 2018, **Dwivedi [12]** made an effort to better forecast cardiovascular illnesses by taking into consideration new factors like Average blood pressure, Blood Cholesterol in mg/dl, and Highest Heart Rate. The dataset used was downloaded from the UCI ML lab and included 120 samples with cardiac illness and 150 examples without the condition. The papers did an excellent job of analysing the ML models in depth. For instance, each ML algorithm's hyper-parameters were optimised through experimentation to achieve the highest achievable levels of accuracy and precision. Despite that benefit, the learning models' ability to accurately and precisely target illnesses is limited by the tiny size of the transferred datasets.

Breast Diseases - In an effort to diagnose breast cancer, Shubair turned to machine learning

Peer Reviewed & Refereed Journal
Volume-13, Issue-3 (June 2023)
https://doi.org/10.31033/ijemr.13.3.40

techniques such as Bayesian Networks, RF, and SVM. The Wisconsin original breast cancer dataset was accessed through the UCI Archive for the purpose of the research. This dataset was then used to evaluate the different learning models with regard to the essential measures. K-fold validation was used to evaluate the classifications, and the number of K that was selected was 10. The computer findings demonstrated that SVM performed exceptionally well in all three metrics tested. The ROC curve, however, suggested that RF had a greater chance of accurately classifying the growth.

Yao [13], on the other hand, tried out a number of different data mining techniques like RF and SVM to zero in on the most effective model for breast cancer prognosis. Compared to SVM's 95.85% accuracy value, sensitivity of 95.95%, and specificity of 95.53%, Random Forest algorithm's categorization rate, sensitivity, and specificity were 96.27%, 96.78%, and 94.57%, respectively.

Yao reasoned that the RF algorithm outperformed SVM because it gives more accurate approximations of the information obtained in each feature trait. Additionally, RF is the most effective method for classifying breast illnesses because it is scalable to big datasets and introduces less likelihood of variation and data overfitting. Multiple success measures were given, which served to bolster the underlying case in the studies.

Parkinson's Diseases - This research proved that a Parkinson's disease diagnosis strategy based on Fuzzy k-Nearest Neighbor (FKNN) was effective (PD). The study analysed the differences between the SVM-based and FKNN-based solutions. To compile the most distinguishing features for use in developing the best FKNN model possible. The UCI archive provided the sample with a large number of biological voice recordings from 31 people, including 24 with Parkinson's disease. According to the data, the FKNN technique achieves higher levels of sensitivity, accuracy, and specificity than the SVM approaches.

The goal of **Behroozi's [14]** study was to propose a new categorization paradigm for diagnosing Parkinson's disease; the use of filter-based feature selection increased classification accuracy by as much as 15%.

Separate models were developed for each subset of the dataset, which was used to define the framework and fill in the gaps caused by the lack of data. The indicators include SVM, Discriminant Analysis, KNN, and NB. Results indicated that SVM was superior to other methods in all cases.

Eskidere also compared the SVM's efficacy to the Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN), and Multi-layer Perceptron Neural Network (MLPNN) for tracking PD progression (MLPNN). LS-SVM performed best. A fair encoder optimal performance metric supports this.

III. DATASET AND MODEL DESCRIPTION

We use structured datasets in our proposed system, which may be constructed by gathering patient symptoms and diagnoses from local hospitals and opensource libraries available online. We use genuine datasets to get improved accuracy.

In the proposed approach, machine learning algorithms are used to forecast diseases based on patient symptoms. This algorithm predicts five diseases based on symptoms, but if we feed it datasets from other diseases, it can also predict more diseases.

IV. EXISTING SYSTEM

Although physicians still need technology in a variety of ways, such as surgery visualisation and x-ray imaging, it has perceptually lagged behind since the advent of sophisticated processing.

Due to additional variables including temperature, environment, blood pressure, and numerous other factors, the technique still needs the doctor's knowledge and expertise. Although there are a great deal of factors that are acknowledged as being necessary to comprehend the entire working process, no model has ever been able to assess them effectively. Physician decision support tools must be used to address this problem.

The physicians can use this method to help them choose wisely. Machine learning is being used to keep full medical statistics. With the aid of machine learning technology, physicians can make important decisions regarding patient evaluations and therapy options, improving patient healthcare services. Machine learning technology enables creating models to evaluate data rapidly and give findings quicker. The use of machine learning in the medical sector is best illustrated by the case of healthcare.

V. PROPOSED METHODOLOGY

Proposed System - Using signs, this method is used to forecast illness. This algorithm evaluates the model using a decision tree classification. End consumers make use of this technology. Based on signs, the algorithm will be able to anticipate illness. The technology used by this method is machine learning. The decision tree classification method is used to forecast illnesses. This technology is known as "AI Therapist" by us. This system is designed for those people who are constantly worried about their health, so we have included some features that recognise them and improve their happiness as well. As a result, the function "Disease Predictor" for health consciousness can identify diseases based on their signs.

Peer Reviewed & Refereed Journal
Volume-13, Issue-3 (June 2023)
https://doi.org/10.31033/ijemr.13.3.40

Our proposed methodology includes the following steps:

i. First, I will compile a database of symptoms and associated functioning issues in the body.
ii. Next, I will gather information that will link the symptoms to possible ailments, and so relevant disease information will be gathered.
iii. Then I will take the patient's symptoms as input and process them using Multilinear Regression.
iv. Following that, Multilinear Regression predicts the diseases that may be associated with the acquired symptoms.
v. The system will then display the diagnosis in the form of maximum possible disease and minimum possible disease.

The technique flow chart is shown below:

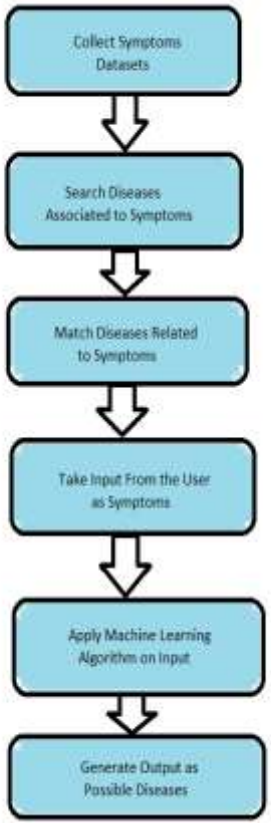
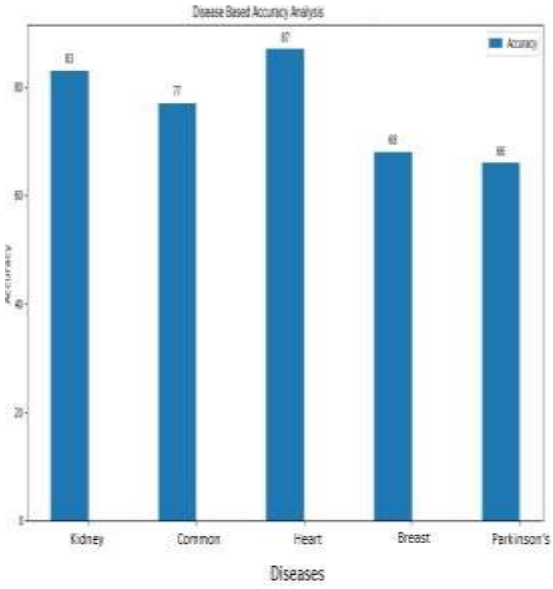


Figure 1: Flowchart of Proposed Model

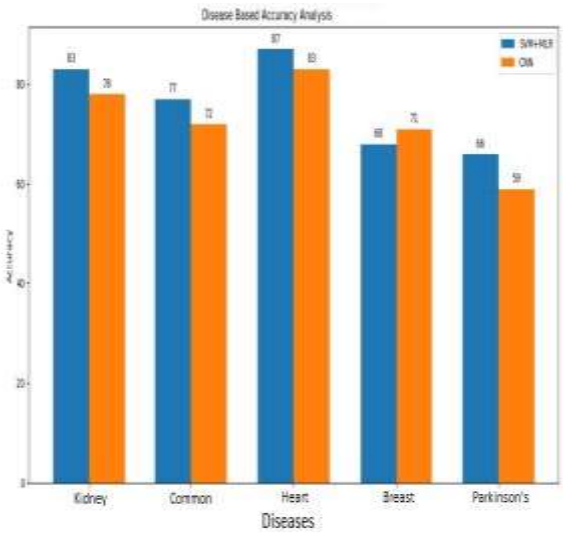
VI. RESULT ANALYSIS

This study paper's crucial component is the result analysis in our suggested system. We may compare how much better this proposed system is functioning by looking at the outcomes of the analysis. We will examine the accuracy of various diseases predicted using our suggested system in the outcome analysis. To analyse the results, we have 200 case datasets.

Accuracy analysis over 200 cases



Above diagram shows the accuracy of 5 diseases that are Common Diseases, Kidney Diseases, Heart Diseases, Parkinson's Disease, Breast Diseases.



Five diseases are included in the above graphic, along with their accuracies for each. For each consecutive bar, these five diseases are processed using two separate algorithms. Accuracy for the diseases processed using SVM is shown by the blue bar. The accuracy of diseases processed by CNN is displayed by the orange bar.

VII. CONCLUSION AND FUTURE WORK

Many illnesses, including those of the heart, renal, breast, and brain, were detected at an early stage with the help of various ML algorithms. It has been

Peer Reviewed & Refereed Journal
Volume-13, Issue-3 (June 2023)
https://doi.org/10.31033/ijemr.13.3.40

found that SVM, RF, and LR algorithms are the most popular for making predictions, with precision being the most common success measure. Among the models tested, CNN performed the best when asked to forecast the most prevalent illnesses.

As a result of its better precision and consistency when dealing with high-dimensional, semi-structured, and unorganised data, the SVM model has been widely adopted for the diagnosis of renal illnesses and PD. Since it scales well to large datasets and is susceptible to prevent overfitting, RF has demonstrated dominance in the likelihood of accurate categorization of illnesses for breast cancer prognosis.

Once again, the LR algorithm was found to be the most accurate in forecasting cardiac conditions. More advanced ML systems need to be developed in the future to improve illness forecast accuracy. Further, it is recommended that learning models be fine-tuned more frequently post-training to see if performance can be enhanced.

To prevent overfitting and improve the performance of distributed models, databases should be widened to include information from a wider range of populations. Lastly, the performance of the learning models can be improved through the application of more pertinent feature selection techniques.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang & L. Wang. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access, 5*(1), 8869–8879.

[2] **S**ayali Ambekar & Rashmi Phalnikar. (2018). Disease risk prediction by using convolutional neural network. *IEEE*. 978-1-5386-5257-2/18.

[3] Naganna Chetty, Kunwar Singh Vaisla & Nagamma Patil. (2015). An improved method for disease prediction using fuzzy approach. *IEEE*, 569-572. DOI: 10.1109/ICACCE.2015.67.

[4] Dhiraj Dahiwade, Gajanan Patle & Ektaa Meshram. (2019). Designing disease prediction model using machine learning approach. *IEEE Xplore Part Number: CFP19K25-ART*; ISBN: 978-1-5386-7808-4, pp. 1211-1215.

[5] Lambodar Jena & Ramakrushna Swain. (2017). Chronic disease risk prediction using distributed machine learning classifiers. *IEEE*, 978-1-5386-2924-6/17, pp. 170-173.

[6] Dhomse Kanchan B. & Mahale Kishor M. (2016). Study of machine learning algorithms for special disease prediction using principal of component analysis. *IEEE*, 978-1-5090-0467-6/16, pp. 5-10.

[7] Pahulpreet Singh Kohli & Shriya Arora. (2018). Application of machine learning in disease prediction. *IEEE*, 978-1-5386-6947-1/18, pp. 1-4.

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh & Nikita Patil. (2017). Diabetes disease prediction using data mining. *IEEE*, 978-1-5090-3294-5/17.

[9] Rashmi G Saboji & Prem Kumar Ramesh. (2017). A scalable solution for heart disease prediction using classification mining technique. *IEEE*, 978-1-5386-1887-5/17, pp. 1780-1785.

[10] Rati Shukla, Vikash Yadav, Parashu Ram Pal & Pankaj Pathak. (2019). *Machine learning techniques for detecting and predicting breast cancer*. *IJITEE, 8*, 2658-2662.

[11] Mohan, Vijayarani. (2015). *Kidney disease prediction using svm and ann algorithms*.

[12] Dwivedi, A.K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Comput & Applic., 29*, 685–693. DOI: 10.1007/s00521-016-2604-1.

[13] DePersia, Allison, Choi, Sarah, Yao, Katharine, Dunnenberger, Henry (Mark) & Hulick, Peter. (2023). Breast health assessment: a family health history tool using the electronic health record and clinical decision support to facilitate guidelines-driven hereditary breast cancer genetic testing at the time of screening mammogram. cancer research. *83*. P6-02. DOI: 10.1158/1538-7445.SABCS22-P6-02-03.

[14] Behroozi M. & Sami A. (2016). A multiple-classifier framework for Parkinson's disease detection based on various vocal tests *Int. J. Telemedicine Appl.*, 1–9.