Disease Prediction Using Machine Learning

Prof. Neelam Malyadri[1], Sanchit Raj[2], Sitesh Choudhary[3], Harshvardhan Singhal[4]
[1]*Assistant Professor, School of Computer Science Engineering, Reva University,* Bangalore, India
[2,3,4]*School of Computer Science Engineering, Reva University,* Bangalore, India

***Abstract*— For the purpose of preventing and treating disease, an accurate and speedy assessment of any health-related problem is vital. If the issue is serious, the conventional diagnostic procedure might not be enough. The creation of a machine learning (ML)-based medical diagnosis system for illness prediction can lead to a diagnosis that is more accurate than the traditional approach. We have created a system for predicting diseases using several ML algorithms. Around 230 diseases may be processed using the data. The diagnosis algorithm produces the ailment that the patient may be experiencing based on their symptoms, age, and gender. The user's disease is predicted by a "Disease Prediction" system based on predictive modelling using the symptoms they provide as input. The technology calculates the probability that the disease will manifest by analyzing the user's symptoms as input.**

Keywords— Predictive Modeling, Naïve Bayes Classifier, Disease prediction, machine learning, symptoms

I. INTRODUCTION

Typically, when a patient has a specific illness, they must make a costly and time-consuming doctor appointment. Also, because it cannot identify the user's sickness, it may be difficult for them if they are located far from medical facilities and physicians. So, it can be easier for the patient and simpler if the aforementioned procedure can be finished by an automated programme, saving time and money. Systems that deal with heart-related diseases are among those that analyse the patient's risk level using data mining techniques. Based on users' reported symptoms, a web tool called Disease Predictor predicts their health. Several health-related websites provided data sets for the Disease Prediction System. The disease predictor will enable the user to estimate the likelihood of the disease based on the stated symptoms.

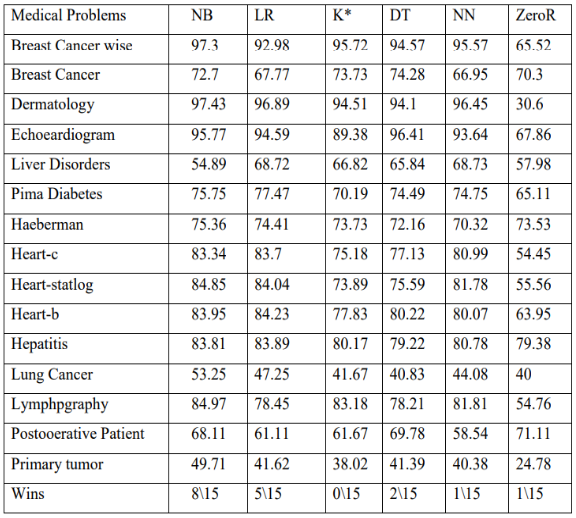
People are constantly interested in learning new things, especially given the daily increase in internet usage. When an issue develops, people typically try to consult the internet for help. Unlike medical facilities and practitioners, everyone has access to the internet. Someone does not instantly have an option when they have a certain disease. As a result, since the general public has continuous access to the internet, this method can be advantageous. One of the most important aspects of both the economy and human living is medicine and healthcare. There have been significant changes between the worlds we live in today and those from a few weeks ago. Everything has morphed into something odd and terrifying.

Board-certified medical experts referred to as virtual physicians prefer to perform phone and video consultations over in-person visits, while this is not an option in an emergency. Since they can accomplish tasks more quickly, consistently, accurately, and without the risk of human mistake, machines are viewed as being superior to mankind. Without the assistance of a person, a disease predictor, sometimes referred to as a virtual doctor, may precisely forecast any patient's illness. There are many virtual doctor models, but they don't have the precision that is needed since not all of the required characteristics are taken into consideration. The major goal was to develop a number of models and identify which one produces the most precise projections. The structure for ML efforts is the same, despite the fact that their scale and complexity differ. Many rule-based machine learning approaches were employed to recall the development and implementation of the prediction model. Several machine learning (ML) techniques were used to launch a number of models, which collected raw data and sorted it into groups according to symptoms, gender, and age.

Knowing how to accurately diagnose patients through clinical examination and evaluation is crucial. Making judgements that are convincing may need the use of computer-based decision support tools. The medical field produces a lot of data on a variety of subjects, including clinical evaluation, patient reports, treatments, follow-up meetings, and drugs. To execute properly, precise orchestration is required. Ineffective information management has had an influence on the quality of the data association. As data volumes increase, a legal method must be developed to concentrate and process information in a practical and efficient manner. A classifier that can separate the data into categories based on qualities is created using one of the many machine learning programmes. There are at least two classes in the data set. These classifiers are used to analyses medical data and predict diseases.

II. LITERATURE SURVEY

Researchers K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman have investigated the best technique for mining medical diagnosis. In this study, Nave Bayes was contrasted against five classifiers: a straightforward rule-based method, Logistic Regression (LR), KStar (K*), Decision Tree (DT), and Neural Network (NN) (ZeroR). 15 real-world medical cases were chosen from the UCI machine learning collection to assess the effectiveness of each strategy (Asuncion and Newman, 2007). In 8 of the 15 data sets included in the trial, Nave Bayes was shown to perform better than the other algorithms, which led to the conclusion that Nave Bayes provides higher prediction accuracy outcomes than other methods.



Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echoeardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	84.97	78.45	83.18	78.21	81.81	54.76
Postooerative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	8\15	5\15	0\15	2\15	1\15	1\15

Table 1-Predictive Accuracy of Bayes and other Technique

Treating chronic diseases internationally is neither time- nor money-efficient, according to study by Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi. The scientists conducted this research in an effort to predict the likelihood of upcoming diseases. To predict potential disease risks, CARE, which solely examines a patient's medical history and ICD-9-CM codes, was utilized. Based on each patient's individual medical history and the experiences of other patients like them, CARE estimates each patient's biggest disease risks using collaborative filtering algorithms and clustering. The ICARE iterative version, which uses ensemble principles for superior performance, is also described in depth by the authors. For thousands of ailments, some of which may emerge years in advance, ICARE's exceptional future disease coverage offers more accurate early warnings. The CARE framework may be used to study a larger range of illness histories, bring up overlooked concerns, and promote discussions about early identification and prevention when it is utilized to its full potential.

To present an overview of current knowledge discovery strategies in databases employing data mining techniques used in contemporary medical research, particularly in the prediction of heart disease, Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni did this work. The results of several tests comparing the effectiveness of predictive data mining algorithms on the same dataset show that Decision Trees perform the best, with Bayesian classification occasionally matching Decision Tree accuracy. Other prediction techniques, such KNN, Neural Networks, and classification based on clustering, don't function as well, though.

By comparing user-provided data to a trained set of values, researchers Shadab Adam Pattekari and Asma Parveen utilized the Naive Bayes Algorithm to forecast cardiac diseases. For this study, patients were allowed to provide some basic information, and when the information is compared to the results, heart disease is predicted.

M.A. Nishara Banu and B. Gomathy used medical data mining techniques such association rule mining, classification, and clustering I to study the various heart-related illnesses. A decision tree is made to display all possible consequences of a decision. The best outcomes are achieved by developing a number of rules. The parameters used in this study to make judgements were age, sex, smoking, being overweight, consuming alcohol, blood sugar, heart rate, and blood pressure. Using a variety of ids, the risk level for various parameters is kept (1-8). An average level of prediction is included in a weighted ID of less than 1, but a weighted ID of more than 1 carries a greater risk. Finding patterns in the dataset is done using the K-means clustering technique. The technique separates the data into k groups. Every point in the dataset is assigned to a closed cluster. The average of the points in each cluster is recomputed as the cluster center.

The author has come to the conclusion that the presented data set is particularly beneficial for illness identification using machine learning techniques like Naive Bayes and Apriori. In this instance, predictions are made using small-volume data, such as symptoms or prior knowledge gained through the physical diagnosis. This work's limitations include its inability to consider very large datasets. Additionally, it is challenging to categorise medical data due to its current growth.

A CNN-MDRP approach was created by Shraddha Subhash Shirsath [15] to forecast diseases utilising a substantial amount of structured and unstructured healthcare data. CNN-MDRP uses a machine learning algorithm (Neavi- Bayes) to focus on both structured and unstructured data, which improves illness prediction accuracy and speed compared to CNNUDRP, which only analyses structured data. Here, big data is considered.

The optimum clinical decision-making approach, according to Ajinkya Kunjir, Harshal Sawant, and Nuzhat F. Shaikh [6], is to use patient history to anticipate disease. This predicts a wide range of illnesses as well as an unexpected pattern of patient status. A best clinical decision-making system that accurately predicts disease using previous data. It also established the notion of several illnesses and an unidentified pattern. Pie charts and 2D/3D graphs were used in this illustration. Pie charts and 2D/3D graphs are also utilized to show data.

III. PROPOSED WORK

We plan to create a tool that predicts diseases, helps with online appointment scheduling, and enables direct communication with doctors through websites in order to address all issues related to disease in people. We offer both patients and physicians a number of features, including login. Who would utilise it through a web application, in line with the requirements of our application. Developing an online platform for disease prediction based on a variety of symptoms is the aim of this research. The user may pick from a number of symptoms and check for illnesses using their probability values. the development of a web-based disease prediction platform.

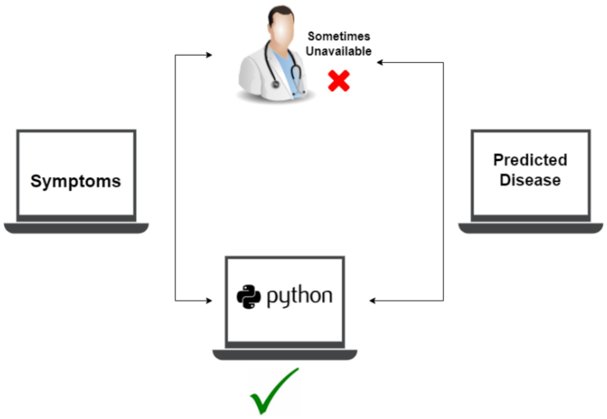


Figure 1. The proposed approach of disease prediction. The doctor might not always be available on call. But, in the modern world, this prediction mechanism is constantly accessible whenever it is required. The age, gender, and symptoms of a person may be given to the ML model for further processing. In order to train and test the algorithm that would lead to the anticipated disease after basic data processing, the ML model uses the most recent input.

The proposed task of disease prediction using machine learning for all diseases comprises the development of a predictive model that can accurately identify the risk of developing various diseases based on patient data. The model would be trained using enormous databases of medical records, which would contain patient demographics, medical histories, genetic data, lifestyle variables, and other essential information.

The proposed effort would require a number of processes, including feature selection, model building, model validation, and data collecting and preparation. Large volumes of medical data would need to be gathered from numerous sources, such as patient surveys, medical databases, and electronic health records.

Cleaning and transforming the data to make it appropriate for analysis would be considered data preparation. This would entail filling in any missing values, dealing with outliers, and scaling the data to make sure that all variables are scaled equally.

The process of feature selection entails determining the key elements that predict the target disease. It would be essential to use a number of statistical and machine learning approaches to discover the factors that are most strongly related with the condition in order to do this.

In order to create a predictive model that can precisely determine the risk of getting specific diseases based on patient data, powerful machine learning techniques would be used. To increase the model's accuracy, it would be trained on a sizable dataset of medical records using a variety of methods, including deep learning, ensemble learning, and feature engineering.

Last but not least, model validation would entail testing the predictive model's precision on fresh data to make sure it is dependable and robust. This would entail assessing the model's effectiveness and pinpointing areas for development using a variety of statistical and machine learning techniques.

Overall, by enabling early diagnosis and treatment of many diseases, the suggested study of disease prediction using machine learning has the potential to revolutionise healthcare. Healthcare professionals can take proactive steps to prevent or treat the condition before it worsens by precisely identifying patients at risk of developing particular diseases.

The dataset was divided into input and output, with the input factors being age, gender, and symptoms, respectively. We divided the given data into train and test sets at random. Next, using various methods, these sets were further trained and encoded. The accuracy of various ML algorithms is then determined by testing the training set and making predictions about the values. The output representing the disease was then obtained by decoding the expected values.

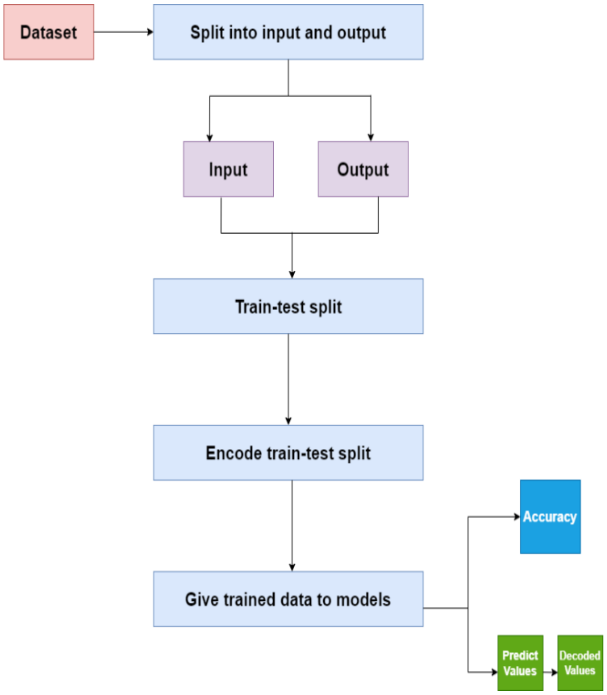


Figure 2: Functioning of the ML models

IV. METHODOLOGY

All diseases might be predicted using a multi-step method that involves data collection, feature selection, model building, and model validation.

A. *Data collection and pre-processing*

Large amounts of medical data must initially be gathered from numerous sources, such as electronic health records, medical databases, and patient surveys. To make the data appropriate for analysis, it must be cleaned and processed. The data must be scaled so that all variables have the same scale, missing values must be filled in, and outliers must be dealt with.

B. *Feature selection*

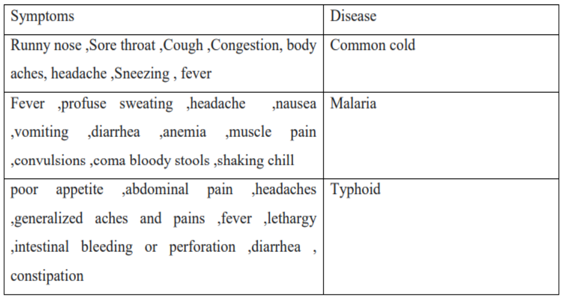
The following stage is to determine which factors are most crucial for predicting the target disease. To discover the factors that have the strongest correlation with the condition, various statistical and machine learning techniques are utilized. The process of feature selection is crucial since it lowers the number of dimensions in the data and increases the model's precision.

C. *Model development*

The following stage is using cutting-edge machine learning algorithms to create a predictive model that can precisely assess the probability of contracting particular diseases based on patient data. To increase the model's accuracy, it would be trained on a sizable dataset of medical records using a variety of methods, including deep learning, ensemble learning, and feature engineering. The right machine learning algorithm must be selected, and its parameters must be modified, to maximise the performance of the model.

D. *Model validation*

The predictive model must be tested for accuracy using fresh data as the last stage to assure its durability and dependability. This entails assessing the model's effectiveness and pinpointing areas for development using a variety of statistical and machine learning techniques. Model validation is crucial since it determines how generalizable the model is and guarantees that it can correctly forecast the risk of disease in new patients.



Symptoms	Disease
Runny nose ,Sore throat ,Cough ,Congestion, body aches, headache ,Sneezing , fever	Common cold
Fever ,profuse sweating ,headache ,nausea ,vomiting ,diarrhea ,anemia ,muscle pain ,convulsions ,coma bloody stools ,shaking chill	Malaria
poor appetite ,abdominal pain ,headaches ,generalized aches and pains ,fever ,lethargy ,intestinal bleeding or perforation ,diarrhea , constipation	Typhoid

Table 2: Sample datasets

V. CONCLUSION AND FUTURE WORKS

A promising area of research in the healthcare industry is the prediction of diseases using machine learning algorithms. Predictive models for several diseases, including diabetes, cancer, heart diseases, and others, have been created using a variety of machine learning methods. These models can assist healthcare workers in identifying people who are very susceptible to contracting a disease and in taking preventive action in accordance.

The choice of machine learning algorithm, the process of feature selection, and the quality and quantity of data used all affect how accurate illness prediction models are. In order to create a prediction model that is accurate, it is crucial to carefully choose these variables. Disease prediction using machine learning algorithms has the potential to revolutionize healthcare by providing early diagnosis and personalized treatment options.

This research aims to identify a disease from its symptoms. The project is set up so that the system either predicts disease or gets user symptoms as input and output. The average forecast's accuracy is found to be 55%. Disease Predictor was successfully built using the Django framework.

Some potential future research in the area of disease prediction using machine learning algorithms includes the following:

Use of advanced machine learning algorithms: Currently, we are using classic machine learning techniques like logistic regression, decision trees, naive bayes, and others in this disease prediction model. In the future, more advanced machine learning algorithms such as deep learning, reinforcement learning, and others can be used to develop more accurate predictive models.

Integration of proteomics and genomes data: Data from proteomics and genomics can be incredibly illuminating in identifying the underlying molecular diseases. Integrating these data with clinical data can help to develop more accurate disease prediction models.

Development of personalized predictive models: Every individual is unique, and their risk of developing a disease may vary based on various factors such as age, gender, lifestyle, and others. Developing personalized predictive models that take into account these factors can help to improve the accuracy of disease prediction.

Validation of predictive models: Validating the predictive models on large and diverse datasets is important to ensure their accuracy and reliability. Therefore, future works should focus on validating the predictive models on larger datasets from different regions and populations.

Integration of electronic health records: Electronic health records are a goldmine of knowledge regarding the health state, medical background, and other pertinent details of patients. Using these data in conjunction with machine learning algorithms can aid in creating disease prediction models that are more precise.

This project does not take user recommendations for drugs into account. Hence, pharmaceutical suggestions can be provided in the project. A log can be used to keep track of a user's medical history and to implement pharmaceutical suggestions.

ACKNOWLEDGMENT

I would want to thank the CSE department of REVA UNIVERSITY BANGALORE -560064 INDIA from the bottom of my heart for your support. We also like to express our thanks to PROF. NEELAM MALYADRI, our guide, for her invaluable help, advice, and guidance.

REFERENCES

[1] D. Magheshkumar, M. Pavithra “Forming Assistant Web Service” www.ijraset.com,IC Value: 45.98, Volume 5 IssueIV, ISSN: 2321-9653, April 2020.

[2] Y. Khourdifi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm

optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12(1), 242 (2019)

[3] S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and na¨ıve bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2022)

[4] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7, 81542 (2019)

[5] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar,T.P.R. Vital, Intelligent parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3),1568 (2019)

[6] Different thresholds in the prediction of chronic obstructive pulmonary disease using neural network and Logistic model, Published in: 2021 International Conference on Public Health and Data Science (ICPHDS)

[7] Research of Heart Disease Prediction Based on Machine Learning, Published in: 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)

[8] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V.Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal 13, 8 (2022)

[9] T. Karayılan, O. Kılı¸c, in ¨ 2017 International Conference on Computer Science and Engineering (UBMK) (IEEE, 2017), pp. 719–723

[10] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities, Ieee Access 5, 8869 (2017)

[11] S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, International journal of environmental research and public health 15(8), 1596 (2018)

[12] A.U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, Mobile Information Systems 2018 (2018)

[13] M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M.Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7 (2020).

[14] B. Nithya , Dr. V. Ilango Professor, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques," International Conference on Intelligent Computing and Control Systems,2017.

[15] S.Leoni Sharmila, C.Dharuman and P.Venkatesan "Disease Classification Using Machine Learning Algorithms - A Comparative Study", International Journal of Pure and Applied Mathematics Volume 114 No. 6 2017, 1-10

[16] Allen Daniel Sunny1, Sajal Kulshreshtha, Satyam Singh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H " Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology (IJIET) Volume 10 Issue 2 May 2018.

[17] Shraddha Subhash Shirsath "Disease Prediction Using Machine Learning Over Big Data" International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018