IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 5 - V8I5-1245)
Available online at: https://www.ijariit.com

Medical diagnosis using Machine Learning

Shubham Chorage
shubham.s.chorage@gmail.com
Pune Institute of Computer Technology, Pune, Maharashtra

Manish Khodaskar
mrkhodaskar@pict.edu
Pune Institute of Computer Technology, Pune, Maharashtra

ABSTRACT

Health is certainly the most important asset of humanbeings which directly reflects in their progress or development.In this paper, we propose a machine-based system that can be used for medical diagnosis. The user would be able to upload hismedical data through the web application which would be further processed by a machine learning model for health diseasedetection. Here we focus on major diseases such as Diabetes,Heart and Kidney. The sole aim is to highlight and prove the benefits of Machine learning in the prediction of diseases and their diagnosis.

Keywords— *Machine learning, disease prediction, Data mining, Smart Health Prediction*

1. INTRODUCTION

With the progress of society, people's lifestyles and environmental conditions are gradually changing, which increases the possibility of hidden dangers of various diseases. Several diseases like Vision Impairment, Diabetes, Cancer, Kidney diseases, Heart Diseases have a serious impact all over the world. Diabetes in itself is deadly and has affected several hundred million people and death from heart disease comprises more than 30% of total deaths worldwide. These diseases are a serious concern to human beings and are drastically affecting human health all around the world. Therefore, disease prediction is highly necessary to avoid further complications, and hence in our system, we are predicting diseases like diabetes, heart disease and kidneydisease.

With the help of machine learning algorithms we can train our model to predict for different diseases like Heart diseases, Diabetes, Kidney Diseases.

The main purpose of such a system is to reduce efforts and make disease prediction effective and feasible for people using the system. Also, prediction of disease at an earlier stage is becoming an important factor with the fast changes in lifestyle happening all around the world. The predicted model also provides machine accuracy.

2. LITERATURE SURVEY

[1] The study uses healthcare data generated by various sources like electronic health records (EHRs), text and unstructured data, clinical records and sources belonging to government sectors, lab reports from medical centres and pharmacies. The study focuses on gathering data, storing and analyzing it for further feedback and prediction. Machine learning can help early detection of diseasesand deviation from a healthy state. Machine learning algorithms such as the Naive Bayes category and the random Forest category are used to provide accurate predictions. The Naive Bayesian classifier is basedon the Bayes' assumption that the forecasts are independent.

[2] The paper examined the performance of other class dividers most closely related to modern machine learning and related areas. The results of the paper provided doctors and health researchers with a valuable tool that would support them in making a decision about which classifier or groups of classifiers to use for the prognosis of specific diseases. The simple conclusion was that the SVM model is one of the best-performing algorithms for medical datasets. The authors propose to include more existing datasets in medical disease repositories and also classification algorithms such as convolutional neural networks, associative classifiers, and deep learning algorithms.

[3] The paper examined the historical data available from electronic health records (EHR) systems and to make an educated prediction about the risk of the patients' suicidal behavior. EHR data was taken fromlarge health care databases spanning across several years to determine future documented behavior. The model was very accurate to achieve some early, perceptive and definitive prediction. Bayesian models were developed using a retrospective cohort approach which helps serve as an early warning system to identify patients with high risk and then study them for further screening.

[4] The paper classifies machine learning algorithms based on disease specifications. A genetic classification technique was developed using historical medical databases to make intelligent clinical decisions pertaining to heart diseases. A study method of differentiation and integration was proposed to classify and define cancer cells. The study of various machine learning techniques made it possible to diagnose cancer with the information about the patient's attributes. An observation was made that machine learning models actually workbetter than neural network models in low resource setting situations.

[5] The study proposes a smart health prediction system that is intended to assist health professionals in their decision-making

process regarding medical situations. This system provides the guidance andinformation needed for doctors to diagnose patients on their medical illness and it eliminates the difficulties that the doctors need to encounter,particularly in their clinical decision-making process.The system would require to gather a whole lot of medical information that is valuable to be used in predicting a patient's health status, these patterns of information will be analyzed by using data mining techniques in order to find correlations and discover new pieces of information from unstructured data. Byusing machine learning tools, it will not only be able to produce reliable results with less time consumption and complexity but also with smart decision-making and useful information.

[6] The paper evaluates the performance of some of the most relevant classifiers in state of the art machine learning and related areas. The results of thepaper provided doctors and health researchers with a valuable tool which would support them in making a decision about which classifier or groups of classifiers to use for prognosis of specific diseases. The simple conclusion was that the SVM model is one of the best-performing algorithms for medical datasets. The authors propose to include moreexisting datasets in medical disease repositories and also classification algorithms such as convolutional neural networks, associative classifiers, and deep learning algorithms.

[7] The paper propose a system that allows users to get guidance on their health issues through an intelligent online health care functioning system. The purpose of this paper is to predict Chronic Kidney Disease (CKD), Heart Disease and Liver Disease using a combination of clustering technique, K- means algorithm.

[8] The paper was researched to generate a generic view of separate medical data analysis from the perspective of machine learning: a historical perspective, a modern perspective, and a vision of specific future trends in the sub-field of practical intelligence used. Comparison of other advanced systems, representatives of each branch of machine learning, as it is also used in various diagnostic tasks.Future styles are illustrated by two case studies. The first describes the newly developed approach to class distinction, which seems to be promising medicalexpertise. The second means using a study machine to diagnose specific unexplained events from concomitant medications, (currently) not yet approved in the general medical community but in the future may play a significant role in the overall diagnosis and treatment.

[9] Patients are in need of treatment and diagnosis that are accurate and precise in order for them to be able to recover back to their proper health and medical staff are required to be well-equipped in their clinical knowledge and communication skills to carefully assess their patients to ensure good health. Therefore, the application of data mining in health prediction is considered in this paper as the most righteous practice to facilitate better healthcare system than already present.The result of the system created will consist of the diseases and its respective accuracy level the patient is currently suffering from.The system will be implemented with data- mining algorithms that will help in deducing the disease that they bear by corresponding the information given by the patient with the health information the doctors and medical professional provide and that is stored in a database, the entire process would efficiently reduce the time consumption and challenging efforts that doctors put themselves into for making a clinical decision.

3. MATERIAL & METHOD

A. Dataset Collection

The database we use is a matrix with rows and columns where the lines represent patients and the columns represent the features or attributes that should be examined. Table 1 shows the data structure used for all 3 diseases consisting of 768,308, and 400 samples each for diabetes, heart, and kidney respectively.

Disease	Data Source	Features
Diabetes	National Institute of Diabetes and Digestive and Kidney Diseases.	Pregnancy, Glucose, Blood Pressure, Skin Stimulation, Insulin, BMI, Birth Control, Age, Effect
Heart	1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. 4. V.A. Medical	Age, Gender, Type of Breast Pain, Relaxation of Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Relaxation of Electrocardiographic Effects, Severe Heart Attack, Exercising Angina, Exercise Caused by

After processing the data when the data is ready we use different Machine Learning algorithms and different Machine Learning methods. The main reason behind using Machine Learning Techniques is to compare and also to evaluate the effectiveness of all methods and to determine their accuracy, and to get an estimate of the underlying factors in prediction.

The method proposed by our group takes the help of various classification methods and their combinations.

These methods are the most common mechanical learning methods used to obtain the best accuracy of data.

B. Data Preprocessing

Many healthcare-related data contain missing values and other impurities that may result in data failure. To improve thequality and performance of the system after the mining process, preliminary data analysis is performed. This processis very important as this helps to use Machine Learning Techniques on an effective website that helps to get accurate results and practical predictions. On our website, we integrate pre-processing in a series of two steps:.

1. Removal of Missing Values - In all cases where values are NaN, with no threshold. This instance is therefore deleted. So we remove unnecessary features and instances and this process is called subset selection features and this reduces the data size and also helps to work faster.
2. Data Splitting - When data is split up we train the algorithm on the training data set and keep the test data set aside. Basically, the general purpose is to bring all the attributes under the same scale. In our model, we have divided thetraining faction by 80% and the testing faction by 20%.

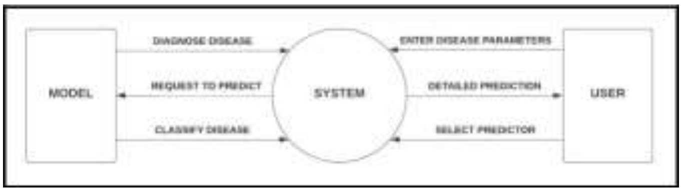

Figure 1: Data flow diagram

C. Modelling

1. Heart Disease : The widely used Logistic Classifier serves as the accurate predictor in cases of heartdisease. This category is based on classification and forecasting, considering real values based on theinput opportunities that are part of a particular class.In Logistic Classifier, the probability is determined using a sigmoid function which encompasses an interpreter function. Reversal ofobjects is used in machine learning due to its efficiency, as it does not require too manycomputational resources and is also one of the most common models of logistic regression as a result of classification has a binary value, i.e., values like trueor false, yes or no.

2. Diabetes : For Diabetes Prediction we have used a Vector Support Machine.The SVM model is a model of different classes in the hyperplane in a multidimensional environment. This type of hyperplane production is done in a repetitive order to minimize the error generatedIn SVM we divide the data sets into classes in order to detect hyperplane in the upper extremities (MMH). Primarily, SVM will produce hyperplanes that repeatedly divide classes in the best possible way. After that, after producing the hyperplane it will select the hyperplane that separatesthe classes accordingly.

3. Kidney Disease : A unique and useful Random Forest Classifier works as the correct basis for prediction of kidney disease. Unplanned forests, also called deciduous forests, are a way of isolating and retreating and other activities that work by building a lot of logging trees during training. This first builds decision trees on different samples and then takes their majority vote for classification and measurement in the case of recruits. In the random forest, the training algorithm uses a common method of integrating bootstrap, or bags, into tree readers. For example, if we begin with a training set X = x1,
..., xn with corresponding answers Y = y1, ..., yn, bags repeatedly then chooses a random sample that begins the process of training the data set and puts trees in it for samples.

4. After the training is done, the predictions of these invisible samples, x 'can be made by estimating the predictions in each of the retrospective trees in x'.

D. Results

The system uses a variety of methods and partitions. These methods are the most common mechanical learning methods used to obtain the best accuracy of data.
The accuracy after training, evaluation and implementation for the diabetes prediction model is
78.2 %.

The accuracy after training, evaluation and implementation for the heart disease prediction model is 85.1 %.

The accuracy after training, evaluation and implementation for the kidney disease prediction model is 100 %.

4. SYSTEM ARCHITECTURE

1. Input :The user is asked several questions based onthe disease he wants to check for that is heart diseases, diabetes or kidney diseases and this data is collected.

2. Processing: The data collected from user is validated and is checked if user has entered all details correctly and after validation, the data is processed and required data is extracted from it and is saved inthe database

3. Modeling: After the data is processed, the data is passed to a backend server and there we apply Machine Learning Technique on the data.We use avariety of machine learning algorithms and methods based on the disease we are testing and based on the model we predict whether the user suffers from that disease or not.

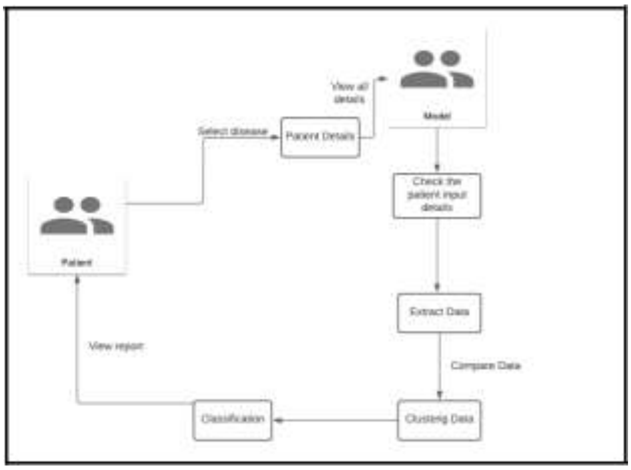

Figure 2

5. CONCLUSIONS

Each nation is taking more care and being considerate about human health issues in the present world. Many world organizations aid people in this regard by giving constant suggestions and updates on preventing and curing epidemics and disorders. The focus now has been shifted to identifying the root cause of the diseases and preventing them before they reach a critical stage by studying the symptoms underlying these diseases. In the era of Machine Learning and BigData, huge amounts of data are constantly being generated from various sources throughout the world.

Thus, Machine Learning plays an important role for structuring and analysing the data to generate results. The health-care industry is in a position to predict a disease based on certain given symptoms and provide valuable feedback with suggestions of curing the diagnosed disease. It helps to regulate the traffic of patients in hospitals and doctors can directly get various suggestions to treat the patients. The system is designed to accurately predict the disease infecting the patient so that an exact treatment method and medicine is available. The project benefits both the doctors and the patients, reducing the complications and paves a way for successful usage of machine learning in medical diagnosis.

6. ACKNOWLEDGEMENT

I would like to express my sincere thanks and gratitude to the Pune Institute of Computer Technology for providing me the opportunity to work on this project. I would like to extend my gratitude to the Department of Information Technologyfor their continuous support and a very special thanks to Prof. M. Khodaskar, my guide and mentor for his constant support and guidance.

7. REFERENCES

[1] B. Maram, K. Suresh Kumar, V. Gampala, Symptoms based disease prediction using big data analytics, in: 2021 Turkish Journal of Physiotherapy and Rehabilitation.

[2] Moreno-Ibarra, M.-A.; Villuendas-Rey, Y.; Lytras, M.D.;

Yáñez-Márquez, C.; Salgado-Ramírez, J.-C. Classification of Diseases Using Machine Learning Algorithms: A Comparative Study. Mathematics 2021, 9, 1817. https://doi.org/10.3390/math9151817.

[3] Machine learning for medical diagnosis: history, state of the art and perspective.18 December 2000, Revised 30 March 2001, Accepted 24 April 2001, Available online 16 July 2001.

[4] M. Gagliano, J. Pham, B. Tang, H. Kashif, Applications of Machine Learning in Medical Diagnosis, in November 2017.

[5] Smart Health Prediction System Using Data Mining July 2020 Conference: 4th Global Congress on Computing & Media Technology (GCMT - 2020)At: Kuala Lumpur, Malaysia.

[6] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, “An improved quantum-inspired differential evolution algorithm for deep belief network,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 10, 2020.

[7] R. Maskeliunas, R. Damaševicius, and S. Segal, “A review of internet of things technologies for ambient assisted living environments,” Future Internet, vol. 11, no. 12, 2019.

[8] S. Mohapatra, P. K. Patra, S. Mohanty and B. Pati, "Smart Health Care System using Data Mining," 2018 International Conference on Information Technology (ICIT), 2018, pp.

[9] 44-49, doi: 10.1109/ICIT.2018.00021.

[10] Smart Health Prediction System Using Data Mining July 2020 Conference: 4th Global Congress on Computing & Media Technology (GCMT - 2020)At: Kuala Lumpur, Malaysia.

[11] H. Zhao, S. Zuo, M. Hou et al., “A novel adaptive signal processing method based on enhanced empirical wavelet transform technology,” Sensors, vol. 18, no. 10, p. 3323, 2018.

[12] A. Keleş, “Expert doctor verdis: integrated medical expert system,” Turkish Journal of Electrical Engineering & Computer Sciences, vol. 22, no. 4, pp. 1032–1043, 2014.

[13] Machine learning for medical diagnosis: history, state of the art and perspective.18 December 2000, Revised 30 March 2001, Accepted 24 April 2001, Available online 16 July 2001.

[14] Machine learning for medical diagnosis: history, state of the art and perspective.18 December 2000, Revised 30 March 2001, Accepted 24 April 2001, Available online 16 July 2001.

[15] SVM-based anomaly detection in remote working: Intelligent software SmartRadar. M. Akpinar, M. Adak, Goker Guvenc

[16] Moreno-Sanchez, P.A. (2021). An Explainable Classification Model for Chronic Kidney Disease Patients. ArXiv, abs/2105.10368.