International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 2, March-April-2021, ISSN (Online): 2395-566X

Literature Survey on IOT and Machine Learning Based Disease Predictor

Charmi Zala
PG Software Engineering,
Birla Vishvakrama Mahavidyalaya
charmimzala@gmail.com

***Abstract*- The world is moving with a fast speed and in Order to keep up ourselves with the whole world we tend to ignore the symptoms of disease which can affect our health at a large extent. Healthcare is one of the important parts for each human being in the world. Health care is given extreme importance with conspicuous novel corona virus. Spreading of disease such as Covid19 has become a global pandemic due to fast spreading of virus in all the countries. Knowing the current situation Internet of Things (IoT) with machine learning will help a lot in serving the best to healthcare and saving many lives around us. Predicting disease according to the symptoms can reduce unnecessary rush to the hospital. This could help to treat the patient early and save many lives getting affected by different disease and virus. As it is rightly said "prevention is better than cure" so predicting disease can help to prevent the occurrence of disease. The Internet of Things (IoT) is very helpful as it can work with real time data as well as the past data that was recorded earlier. This IoT sends data through Wireless Sensor Network (WSN) to the computational devices so that the result can be generated. Machine learning with help of different prediction algorithm such as Decision tree, Naive Bayes etc can help to predict the disease fast and accurate. IOT and Machine Learning are on trend in medical field as it helps both patient and the doctor. Early prediction leads to save time, cost and prevent humans getting affected by diseases.**

***Keywords*- IoT, healthcare, decision tree, naive bayes, random forest.**

I. INTRODUCTION

Due to environmental changes new diseases are getting active which gradually affects many human being. Healthcare is the one of the major factor that everyone takes care of and it requires continuous monitoring of health factors such as pulse rate, heat rate, oxygen level of blood, blood pressure, glucose level etc. since the 2019 outbreak, coronavirus disease has spread worldwide causing many death in every country. This disease started spreading fast that every country was affected by this and gradually the economy was also affected. This disease was deadly and in such time early prediction of disease with symptoms could lead to save many lives. Prediction of diseases at time could break many chains of the disease.

Predicting the accurate disease is also one of the challenging factors. IOT and machine learning both could help to detect various health care problems. Machine learning has made easy to identify the disease in human body. Using prediction algorithm in machine learning could lead to get more accurate prediction of disease and could help human to get treated accordingly. Iot refers to a relationship of practical and functional between different objects in internet of things and on other hand machine learning has great capacity to work with bulk data and make accurate prediction with help of different algorithm. In addition Iot devices and sensor reading that are related to medical parameter can be used effectively to detect the disease in human body that could help the patient condition not getting severe. With Iot large amount of real data can be generated accurately. Looking to the current situation predicting disease in early stage is one of the important tasks.

Machine learning is of many types that is supervised, unsupervised learning, reinforcement, evolutionary and deep learning. Such types of learning help in classifying huge data in short period of time. Majority author have used deep learning and supervised learning that uses training data as a label so to train the model. An Iot system can be built using different sensors. The system consists of sensor like temperature, pulse rate, oxygen, pressure etc.

The system can be integrated to the sensors and the board which will help to send the data each time the temperature, pulse rate or oxygen reaches the threshold value. Temperature sensor used to detect the temperature level of human body. The oxygen sensor checks the Spo2 level of human body and the pulse rate checks the rate of pulse. The Arduino module will help to connect the internet and keep track of data on a daily basis. If a person oxygen level gets lower than threshold value the immediately user will get alert through the system. Similarly temperature and the pulse rate. Even with the help of Arduino the data

© 2021 IJSRET

can be stored in cloud and whenever the temperature, pulse rate and oxygen value reaches the threshold value a mail message can be sent to doctor so that immediate treatment can be given to the patient.. It will serve humanity in the reduction of various disease infections and save healthcare members around the world. I t also saves hospital places for emergency cases. The model uses IoT device called Arduino. Many authors has used this because it does not require much knowledge to get started, Fairly low cost, depending on shields you need, lots of sketches and shields available, no external programmer or power supply needed.

Below is the figure 1 provides a brief work flow of how does disease prediction model works using different machine learning algorithm.

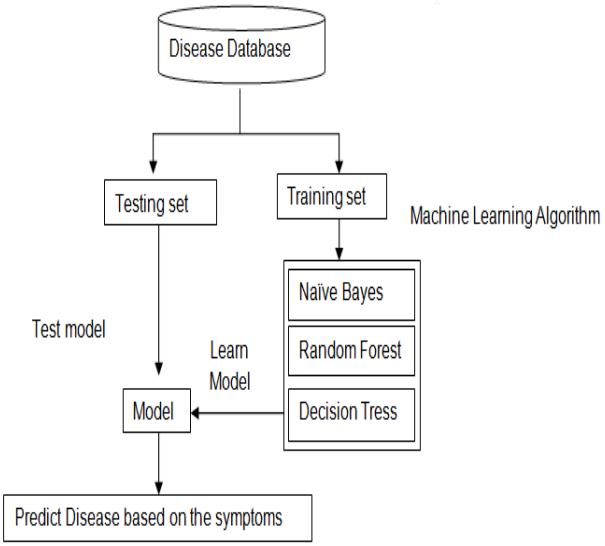

Fig 1. Flow chart of how disease prediction works using different Machine learning algorithm.

II. LITERATURE SURVEY

The author has proposed Article called "prediction of liver disease using a classification algorithm". The classification algorithm such as logistic regression, SVM and KNN used to predict liver disease. Every algorithm is compared based on their accuracy through a confusion matrix. Here the dataset of liver disease is downloaded from UCI with an instance of 567, and the data is collected from the ILPD (Indian Liver Patient dataset).

After running several algorithm the collected result and after analyzing the result, I was observed that the tt two algorithms, i.e. KNN and Logistic regression was having the best accuracy result compared with other remaining algorithms and out of these two algorithms, logistic regression was the best and highly responsive in term of true positive rate or recall. From the confusion matrix KNN accuracy comes out to be 73.97% and the SVM accuracy comes out to be 71.97%. So, the best model for liver disease prediction is logistic regression as it gives the best accuracy among all other algorithms. [1]

The author has proposed an Article called "Iot Based Health Monitoring System". They have proposed a system which includes the use of sensor i.e. Temperature sensor, pulse rate sensor, temperature sensor of outer environment and sensor of outer humidity. The system uses Arduino controller. Here sensors are embedded on the patient body and other sensor is placed at home to check the humidity and the temperature of the room. The sensor value is also displayed on the LCD display.

The system collects the data, calculates the value and then through an IoT cloud it is transmitted to the base station and from base station values are accessed by the doctor. If the framework finds any sudden change in heart rate or the temperature then it would immediately inform the patient about it using alarm and also the system shows live data of the from sensor and can be seen from the web. The heart rate sensor is based on principle of photo phlethysmography. This whole system helps in reducing the cost of physical visit to hospital, testing etc. [2]

In this manuscript that is "Prediction of Heart Disease Using Machine Learning Algorithms" works with two algorithms, i.e. decision tree and Naive Bayes. Both of the algorithms are used for prediction and classification. Here comparison is done with these two algorithms based on their accuracy. In this paper author has used the python programming language which helped them to find the accuracy with the appropriate result and helped them to figure out which algorithm works best for their proposed model. Dataset used here was downloaded from the UCI machine learning repository. This dataset contain 300 instances.

With help of data mining classification technique the entire dataset was classified into two categories i.e. yes or no. The ultimate goal of this system is to predict the possibilities of occurring heart disease of the patients in terms of percentage .The goal of the result is to check whether a person having heart disease or not and to know this it has used two classifiers. Both of the algorithms had produced accuracy i.e. decision tree with 91% and Naive Bayes with 87% [4].

In this manuscript that is "Machine learning-based prediction of COVID-19 diagnosis based on symptoms" includes a machine-learning model which predicts a positive SARS-CoV-2 infection in a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the first months of the COVID-19 pandemic. The data used was released by Israeli Ministry of Health has limitations and biases. The dataset contains initial records, on a daily basis, of all the residents who were tested for COVID-19 nationwide. Based on these data a model was

© 2021 IJSRET

developed that predicts COVID-19 test results using eight binary features: sex, age 60 years or above, known contact with an infected individual, and five initial clinical symptoms. Predictions were generated using a gradient-boosting machine model which was trained with lightGBM Python package and built with decision-tree base-learners and missing value were inherently handled by the gradient-boosting predictor.

To identify the principal features driving model prediction, Shapley Additive explanations (SHAP) values were calculated. The models scored on the test set using the auROC (area under the receiver operating characteristic curve). It has also plots graph of PPV against the sensitivity (precision–recall curve) were drawn across different thresholds. Metrics were calculated for all the thresholds from all the ROC curves, including sensitivity, specificity, PPV and negative predictive value, false-positive rate, false-negative rate, false discovery rate and overall accuracy. Confidence intervals (CI) for the various performance measures were derived through resamppling, using the bootstrap percentile method with 1000 repetitions. [5]

In this manuscript that is "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules" the first step was to preprocess the data for elimination of missing value is applied. Further carry out generation of weight fuzzy and develop a fuzzy rule based decision support system. The method selected suitable attribute to generate fuzzy rules which are weighted based on the frequency in the learning and understanding database. Then fuzzy rules were used to built the decision support system. The proposed system was tested on three types of dataset those were Cleveland dataset collected from V.A medical center. The dataset contains total of 303 instances in which 202 are training data and 101 are used as test data.

Hungarian dataset were collected from Hungarian Institute of Cardiology, Budapest includes 196 of training data, 98 of test data which has total instance of 294 records. And total instance of 123 records were taken in which 41 were test data and 82 were training data from Switzerland dataset collected from University Hospital, Zurich, Switzerland. This method when compared to neural-network based system has achieved highest accuracy. The method reported accuracy of 57.851% for Cleveland data. Hungarian dataset gave the accuracy of 50.583%. Switzerland dataset reported the accuracy which is 20%, higher then neural network based system. [6]

In this manuscript that is "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities" a new CNN based multimodal disease risk prediction algorithm is proposed by using structured and unstructured data of hospital. The disease prediction system was invented for the numerous regions. Here prediction is performed on three diseases like diabetics, cerebral infraction and heart disease. The disease prediction was carried out on structured data. Prediction of heart disease, diabetes and cerebral infraction was carried out by using different machine learning algorithm like naïve bayes, Decision tree and KNN algorithm.

The result of Decision tree algorithm was better than Naïve bayes and KNN algorithm. Also, predicted that whether a patient has experiences from the high risk of cerebral infarction or low risk of cerebral infarction. For the risk prediction of cerebral infraction, CNN based multimodel disease risk prediction was utilized on text data. The accuracy comparison is done between CNN based unimodel disease risk predictions against CNN based multimodal disease risk prediction algorithm. The accuracy of disease prediction achieved here was up to the 94.8% with faster than CNN based unimodal disease risk prediction algorithm.

The CNN based multimodal disease risk prediction algorithm steps was similar as of the CNN-UDRP algorithm only the testing steps consisted of two additional steps. This paper worked on both type of dataset like structured and unstructured data. While previous work only based on structured data, none of the author worked on unstructured and semi- structured data. But this paper depends on structured as well as unstructured data.[7]

In this manuscript that is "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," a clinical decision-making system which can predict the disease on the basis of historical data of patients. In this it predicts multiple diseases and unseen pattern of patient condition. It has best clinical decision- making system used for the accurate disease predictor on the historical data.

The platform used here is WEKA. Here KDD that is knowledge discovery from data process is used. Data mining used which is a part of KDD process so to analyze technique in which previously undiscovered patterns relation and interesting concepts are mined or extracted from the set of data. Here the dataset contains 20 medical related attributes for each diseases and maximum up to 500 algorithms training data examples.

The medical dataset selected for project were heart disease, breast cancer, arthritis and diabetes. Here diagrammatic representation was done for medical dataset. Here few dataset were downloaded from UCI repository and some were built by taking advice of the doctor. In that also determined multiple diseases concept and unseen pattern. For the visualization purpose in this used 2D/3D graph and pie Charts. And 2D/3D graph and pie charts designed for visualization purpose. Here algorithm used were Naïve Bayes so that prediction can be done accurately and Decision tree's J48. Comparing the

© 2021 IJSRET

accuracy of both the algorithm Naïve bayse performed well with latency of 0sec to 0.01sec as Decision tree min latency was 0 to 4sec.

In this manuscript that is "Predicting Disease by Using Data Mining Based on Healthcare Information System" applies the information mining process to predict high blood pressure of the patient from patient medical records with eight alternative diseases.

The data was extracted from true world health care system information containing all the medical records in which each patient's medical profile is represented by 284 statistical features or attributes such as total number of outpatient visits, number of Emergency visits, has diabetes, and so on derived from the historical records. Among this dataset nine Boolean feature denoting the presence or the absence of nine diseases.

The majority patients were healthy so to balanced the data under sampling technique was used to generate five balanced training set and testing set Under- sampling technique has been applied to come up with coaching knowledge sets, and data processing tool WEKA has been used to generate the Naive Bayesian and J48 classifiers so that to improve the prediction performance, and rough set tools were used to scale back the ensemble supported the concept of second- order approximation.

Experimental results showed a bit improvement of the ensemble approach over pure Naive Bayesian and J-48 in accuracy, sensitivity and F-measure.

Initially they'd an classification and so ensemble the classifiers and so the reduction of Ensemble Classifiers is employed. But the choice trees generated by J-48 is typically lacking within the leveling therefore the overall improvement of victimization ensemble approach is a smaller amount. [9]

III. METHODOLOGY

1. Decision Tree:
It is sort supervised learning algorithm program that is widely used for classification issues. Surprisingly, it works for every categorical and continuous dependent variable and it can automatically handle the missing value. It is used for classification and prediction.

A tree has several analogies in real world, and seems that it's influenced a large area of machine learning, covering each classification and regression. In call analysis, a choice tree is used to visualize and expressly represent selections and higher cognitive process. As the name itself suggest, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it's also widely used in machine learning. It is not suitable for small database.

2. Random Forest:
Random Forest is a great algorithm to train early in the model development process, to see how it performs and it's hard to build a "bad" Random Forest, because of its simplicity.

This rule is additionally an excellent alternative, if you would like to develop a model during a short amount of your time. On prime of that, it provides a fairly sensible indicator of the importance it assigns to your options. Random Forests are terribly onerous to ram down terms of performance. And on prime of that, they'll handle tons of various feature varieties, like binary, categorical and numerical.

Overall, Random Forest may be a quick, easy and versatile tool, though it's its limitations. Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the categories (classification) or mean prediction (regression) of the individual trees Random call forests correct for call trees' habit of over fitting to their training set.

3. Naive Bayes:
Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. For instance, if you are trying to identify a fruit based on its color, shape and taste, then an orange colored, spherical, and tangy fruit would most likely be an orange.

All these properties individually contribute to the probability that this fruit is an orange and that is why it is known as "naive". As for the "Bayes" part ,it refers to statistician and philosopher, Thomas Bayes and the theorem named after him , Bayes' theorem , which is the base for Naïve Bayes Algorithm. More formally, Bayes 'Theorem is stated as the following equation;

$$P(A/B) = (P(B/A)*P(A)) / P(B)$$

4. Logistic Regression:
The logistic regression is a predictive analysis. Logistic regression is one of the supervised learning classification algorithm used to predict the binary outcomes of all the given set of independent variable. Sometimes logistic regressions is difficult to interpret.

Types of question binary logistic regression can answer is, do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)? In logistic regression the main focus is the probability of outcomes dependent variable. The odds of success is by odd = P/1-P.
In logistic regression, the dependent variable is a logit that is natural log of odd given by;

© 2021 IJSRET

Log (odds) = logit (P) = ln (P/1-P).

5. K-nearest Neighbors:
K-nearest neighbor can be said as classification, non-parametric algorithm which stores all available cases and its works is to classify new cases based on a similarity measure. It is non parametric as it is does not make any assumption on underlying distribution.

KNN uses Euclidean distance to predict the class as;

$$\mathbf{d(x,y) = \sum_{i=0}^{k} (xi - yi)\text{^}2}$$

It works like this, a case is classified by majority vote of its neighbour, then the case is assigned to the class which is most common amongst its K nearest neighbour which is measured by measured by Euclidean distance.

6. Support Vector Machines:
Support Vector Machines is a classification algorithm which is based on the idea of finding out the hyper plane which divides the dataset into 2 classes. Support vectors are the data points that are nearest to the hyper plane. Support Vector Machine is non-probabilistic, so they assign a data point to a class with 100% certainty.

IV. ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my guide Prof. KirtiKumar J. Sharma and Prof. P. B. Swadas who gave me the golden opportunity to do this wonderful project on the topic "IOT and Machine Learning Based Disease Predictor". Which also helped me in doing a lot of Research and I also came to know about so many new things for which I am really thankful to them.

I would like to express my gratitude to my respected H.O.D of Software Engineering Dr. D G Thakore for giving me this opportunity and for motivating me to do innovative things that will be beneficial for my future.

I would also like to thank our respected principal Dr. Indrajit N. Patel for giving me this golden opportunity to study in this great college and also helping us in various things. This would not have been possible without the opportunity. I am thankful to all who provided me an opportunity to complete this project.

REFERENCES

[1] Thirunavukkarasu K, Singh AS, Irfan M, Chowdhury A (2018) Prediction of liver disease using classification algorithms. In: 2018 4th international conference on computing communication and automation (ICCCA)

[2] Prajoona Valsalan, Tariq Ahmed Barham Baomar, Ali Hussain Omar Baabood. "Iot Based Health Monitoring System". http://www.jcreview.com/ ?mno=96095 [Access: April 10, 2021]. doi:10.31838/jcr.07.04.137.

[3] Xu, H., Huang, S., Qiu, C., Liu, S., Deng, J., Jiao, B., Tan, X., Ai, L., Xiao, Y., Belliato, M., & Yan, L. (2020). Monitoring and Management of Home-Quarantined Patients With COVID-19 Using a WeChat-Based Telemedicine System: Retrospective Cohort Study. Journal of medical Internet research, 22(7), e19514. https://doi.org/10.2196/19514.

[4] SanthanaKrishnan J, Geetha S (2019) Prediction of heart disease using machine learning algorithms. In: 2019 1st international conference on innovations in information and communication technology (ICIICT).

[5] Zoabi, Yazeed & Deri-Rozov, Shira & Shomron, Noam. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine. 4. 10.1038/s41746-020-00372-6.

[6] P. K. Anooj, "Clinical Decision Support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University- Computer and Information Sciences (2012) 24, 27-40.

[7] P K, Anooj. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University - Computer and Information Sciences. 24. 27-40. 10.1016/j.jksuci.2011.09.002.

[8] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[9] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," in IEEE big data analytics and computational intelligence, Oct 2017 pp.2325.

[10] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System" , in IEEE 2012.

[11] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical History.

© 2021 IJSRET