International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Virtual Expert -Disease Prediction Using Machine Learning

Prof Ashish.G.Nandre[1]
Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Jayashree Patil[3]
Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Sayama Pathan[5]
Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Prabodh Narkhede[2]
Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Yadnika Wagh[4]
Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Abstract - In addition to patient registration and data storage in the system, this project's Virtual Expert illness prediction helper for the Medicare system also features automated lab and pharmacy billing. The programme has the ability to assign a unique ID to each patient and records each patient's clinical information as well as results of automatically performed hospital tests. It has a search function so you can find out each patient's current status. Using the id, users may search a patient's information. With the use of a login and password, one may access the Virtual Expert prediction assistance for the Medicare system. The information is simple to get. The user experience is excellent. Data processing is quick since the data are adequately safeguarded for personal use.

Key Words: Predict, SVM, Naïve bayes, CBR, NGO etc.

1. INTRODUCTION

The Critical Patient Care or Monitoring System of today allows a clinician to continually monitor many patients for several parameters at once from a distant location while simultaneously controlling medication dose. These systems would make it much easier to develop and assess the ICU decision-support systems. Virtual Expert Prediction is an online application that offers a whole healthcare system and a dynamic user interface, allowing users to enter all the patient's vital signs using a variety of predefined alternatives. Additionally, this programme is created specifically to meet the user's demand to conduct health examinations efficiently and effectively. In the realm of medical science, this application may be utilised to minimise human mistake to the greatest extent feasible. The user doesn't require any special training to utilise this system. This alone demonstrates that it is user-friendly. As previously said, the Virtual Expert Prediction Web Application can anticipate the early stages of disease, which can result in safe, secure, dependable, and accurate systems that can save a specific individual from dying.

1.1 AIM

The aim is to provide as it is not intended for a particular organization. This project is going to develop a generic web application, which can be applied by any health organization or government in future. Moreover, it provides facilities to its citizens. Also, the web application is going to provide a huge amount of summary data.

1.2 OBJECTIVE

The main objective of this project is basically targeted to provide health related services to remote places and provide better health care and improve national health. Citizens can be instantly examined using this system which will be available to doctors, sevika, NGO, screeners, also it will provide proper diagnosis, maintain medical records and will be easily available to all.

2. LITERATURE SURVEY

A healthcare system that is renowned for offering complete digital medical services. India is a developing and expanding nation, but its medical infrastructure hasn't kept pace. resulting in inadequate access to medical institutions in rural areas, outmoded treatment methods, and sluggish feedback. When a patient relocates or sees a different doctor, it is crucial to keep track of their medical history and treatment plan. Careless behaviour toward the aforementioned points by a patient or a medical practitioner may be deadly. As a result, the Current System offers less flexibility.

[1] Support Vector Machine (SVM), KNN, CART, and Naive Bayes machine learning algorithms were utilised by Anusha Bharat et al. in their 2018 study for breast cancer

detection and prediction. Each algorithm worked differently, however SVM did better than the others with an accuracy rate of 99.01%.

[2] The prediction of Alzheimer's disease was carried out by J. Neelaveni et al. [2020] utilising machine learning algorithms including SVM, CNN, RNN, and Decision tree. With an accuracy of 85%, it is determined that SVM performed better than decision tree.

[3] In order to predict cardiac disease at an early stage, Rahul Katarya et al. [2020] reviewed a variety of models created utilising machine learning techniques such ANN, Random forest, Decision tree, and SVM. The best accuracy models, according to all the models, fall between 82 and 95%.

[4A model for heart disease prediction using machine learning was created by Ankita Duraphe et al. in 2020. They employed algorithms like Random forest, SVM, Logistic Model tree, and Hoffding decision tree in their investigation. They achieved 95.08% accuracy as a consequence, which is 3–4% more than what they achieved using SVM and the Logistic Model Tree.

[5] A machine learning-based model for the prediction of a covid-19 diagnosis based on symptoms was created by Yazeed Zoabi et al. in [2021]. They improved their accuracy by 88% utilising the confusion matrix.

[6] A machine learning-based methodology for identifying and forecasting chronic illnesses was suggested by Rayan Alanazi [2022]. In this model, the illness was predicted using CNN, and the distance was calculated using KNN. Additionally, they have applied a number of algorithms for better prediction. They used CNN and KNN to achieve accuracy of 95%, which is far higher than what they could have achieved using other methods.

[7] In 2021, Sumayh S. Aljameel et al. developed a model to forecast patients with COVID-19 illness severity and prognosis. Three classification algorithms—logistic regression (LR), SVM, random forest (RF), and extreme gradient boosting—were used to examine the data (XGB). Several preprocessing techniques are used during preprocessing.

Twenty clinical characteristics that were shown to be relevant for predicting patient survival vs patient death in the COVID-19 cohort were used in experiments. The findings indicated that SVM had superior accuracy than others, with a 95% accuracy rate.

[8] With the aid of CNN, Manav Sharma et al. [2021] created a model for brain tumour identification using machine learning. After the MRI noise was removed, the pictures were classified as tumour or non-tumor. The data was then divided into training, validation, and testing segments. Python is used for the whole implementation. The final model has a 97.79% accuracy rate.

[9] A model for brain tumour identification using ML and deep learning was put out by M Kavinraj et al. in 2020. In this model, the dataset has been divided, and MRI image characteristics have been retrieved. This shows that CNN has an accuracy rate of 89% while ANN has a rate of 65.21%. Thus, it can be said that CNN is the most effective method for studying the picture collection.

[10] A real-time cardiovascular health monitoring system for the identification of heart disease was created by Shadman Nashif et al in 2018. The system analysed several classification algorithms, and it was concluded that SVM provides accuracy of 97.53% by comparing Random Forest, SVM, and Naive Bayes. To improve user engagement, they integrated this ML model with an Android application.

[11] An innovative strategy for heart disease prediction utilising strength score was put out by Armin Yazdani et al. in [2021]. Using weighted related rule mining and the scores of relevant characteristics, the model predicted heart disease. The accuracy attained is 98%.

[12]Using a machine learning method, Swati Mukherjee et al. [2020] built a model for the detection of lung cancer. They used CNN with a variety of supervised learning approaches in that model. The model's accuracy according to CNN is 85.03%.

[13] Using machine learning methods, Nurul Azam Mohd Salim et al. [2021] created a forecast of dengue epidemic model. They have employed a number of approaches, including SVM, Decision Trees, and Random Forests, and it has been discovered that SVM delivers the best results with an accuracy of 70%.

[14] A illness prediction model was created by Dhiraj Dahiwade et al. in 2019 using a machine learning method. The dataset for this model was divided into training and testing sets, and the model was trained using CNN and KNN. In comparison to KNN, CNN was found to be more accurate and to take up less time. Accuracy as a result is 84.05%.

[15] RF, Bayesian Networks, and SVM are three machine learning (ML) techniques that Shubair [20] tried to use for the identification of breast cancer. The classifiers were evaluated using the K-fold validation technique, with K set to 10 [20].

SN	Name of the Author	algorithm	Description	Accuracy
1	Anusha Bharat, Pooja Natrajan , R Anishka Reddy	SVM, KNN,CART, navies bayes	Breast cancer prediction and diagnosis using machine learning	99.01%
2	J. Neelaveni, M.S. Geetha Devasana	SVM,CNN,RNN,Decision Tree	Alzheimer Disease using machine learning algorithm	85%
3	Rahul Katarya, Polip Ireddy Shreeniwas	ANN,SVM,Decision tree Random forest	Predicting heart disease at early stages using machine learning	82%-95%
4	Pranav Motarwar, Ankita Durahe,G. Suganya M. Premlatha	SVM, Random forest, logistic model tree,hoeffding decision tree	Cognitive approach for heart disease prediction using machine learning	95.08%
5	Rayan Alanazi	CNN,KNN	Identification and prediction of chronic diseases using machine learning	96%
6	Sumayh S. Aljameel,Irfan Ullah khan	SVM	Machine learning model to predict the disease severity and outcome in covid 19 patient	95%
7	Manav Sharma, Kamakshi Gupta	CNN	Brain tumour detection using machine learning	97.79%
8	Dhiraj Dahiwade, Ektaa Mishram	KNN, CNN	Designing disease prediction model using machine learning approach	84.05%
9	P Gokila Brindha, P Prasanth	ANN, CNN	Brain Tumour detection using ml and deep learning	65.21%
10	Caitlynn Reeves ,Rahmat Dapari	SVM	Prediction of Dengu using machine learning	70%
11	Swati Mukherjee,S.U.Bohra	CNN, Supervised learning	Lunge Cancer Disease Diagnosis using machine learning	85.03%
12	Armin Yazdani,Kasturi Varathan	ARM, Warm	A Noval Approach for heart disease prediction using strength scores	98%
13	Yazeed Zoabi , Shira Deri-rozod	Confusion matrix	Prediction of covid 19 diagnosis best on symptoms using machine learning	88%
14	Shadman nashif, Rakiv raihan	Naives Bayes, SVM, Random Forest	Heart disease detection using machine learning algorithm and health monitoring system	97.53%
15	Yusuf Aliyu Adamu,Jaspreet Singh	Random forest, AUC, Decision tree, logistic regression	Malaria prediction model using machine learning algorithms	97.72%

Chart1: Literature Survey

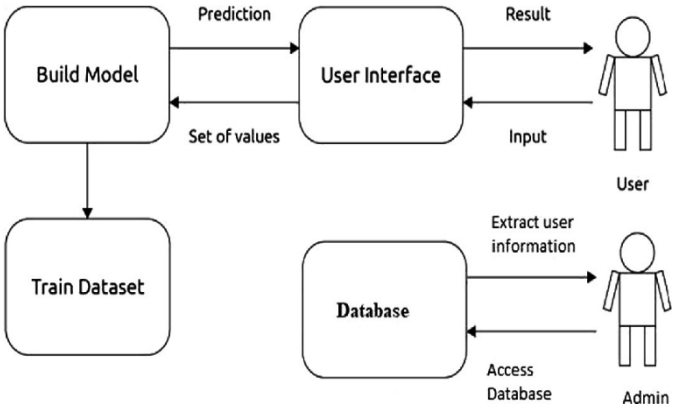

Fig -1: System Architecture

Our proposed system allows health professionals to reach any corner of the nation by having a remote clinic at their fingertips. Instant examination can be carried out from anywhere where a person registered first, his vital sign will be measured, symptoms will be gathered and a report will be generated based on the above inputs. Thus every registered member will be linked by a unique number that will be the AADHAR UID which in majority will reflect the national health scenario. Past medical records as well as the treatment procedure can be stored which will be further useful to provide better treatment based on past records. Computer generated prescription will further eliminate false prescription and irregularities in the pharmaceuticals.

3. METHODOLOGY

The methodology of the system is included in this chapter, as the title suggests. More specifically, methodology refers to the documentation of methods used to manage activities in a coherent, consistent, responsible, and repeatable manner with regard to system analysis and design. Methodology is a procedure that primarily entails intellectual activity; often, the output or outcome of the physical task is the sole way the methodology process manifests its final purpose. The term "methodology" in the context of software refers to a collection of actions or a process that regulates the activities of analysis and design guidelines, or to a structured, documented set of procedures and rules for one or more phases of the (software life cycle), such as analysis or design.

3.1 ALGORITHMS

3.1.1 Naïve Bayes Algorithm:

It is mainly used in text classification that includes a high-dimensional training dataset. one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

The essential naive bayes assumption is that each feature contributes equally and independently to the result. Due to the fact that it uses less processing resources, it has the benefit of operating quickly even on huge datasets.

1) Bayes theorem

Naïve Byes algorithm is based on Bayes theorem given by:

P(A|B)= P(A,B) / P(B)

Where,

P(A|B)= Posterior probability

P(B|A)=Likelihood

P(A)= Class prior probability

P(B)= Predictor Prior probability

In the previous formula, "s" stands for class and "h" for characteristics. The single term in P(hdenominator)'s is a function of the data (features); it is not a function of the class with which we are presently dealing. It will thus be the same for all classes. In order to create the prediction, we often disregard this denominator in naive Bayes Classification because it has no impact on the classification outcome:

$P(A|B) \propto P(B|A)P(A)$

Key terms:

- The percentage of the disease in the data set under consideration is the prior probability.
- The likelihood of classifying an illness in the presence of additional symptoms is known as likelihood.
- The percentage of symptoms in the dataset under consideration is known as marginal likelihood

Example:

We estimate the Nave Bayes findings for a collection of symptoms using the same medical data that was used for the decision tree

- Present (shown by vale'1') Indicates high fever
- Vomiting = Absent (value = "0")
- Present (represented with value "1"), shuddering
- Muscle wasting = Present (value '1' denotes this)
- 4 symptoms are taken into account: high temperature, vomiting, shivering, and muscle wasting.
- Dengue and malaria: classes

According to the dataset we used:

1. Probability = P (Feature: symptoms; Class: Dengue, Malaria)

Marginal Likelihood = P(Features=Symptoms)

2.Prior Likelihood = P

3. (Class)

Thus, after seeing the input symptoms, the prediction is generated by comparing the posterior probability for each class (i.e., for each disease). To do this, we'll employ expression (2). Consequently, let's think about the following notations to deflate our formula: 'F1' for 'High fever,' 'F2' for 'Vomiting,' 'F3' for 'Shivering,' 'F4' for 'Muscle Wasting,' and 'D' for 'Diseases(class)'.

First, the likelihood of having Dengue is calculated (i.e., the class is Dengue with the symptoms "high fever=present," "vomiting=absent," "shivering=present," and "muscle wasting=present" as input symptoms).

P (S=Dengue | F1=Present, F2=Absent, F3=Present, F4=Present) = = P (F1=Present, F2=Absent is the modified formula. F3 = Present, F4 = Present, and S = Dengue) * P (F1=Present | S=Dengue) = P (S=Dengue) P(S=Dengue) = P(F2=Absent | S=Dengue) * P(F3=Present | S=Dengue) * P(F4=Present | S=Dengue)

4/12 * 4/12 * 5/12 * 5/12 * 8/12 \s=0.01286

Second, the likelihood of malaria is calculated (using the class=Malaria and the same input symptoms as described in step one above).

P (F1=Present, F2=Absent, F3=Present, F4=Present | S=Malaria) = P (F1=Present, F2=Absent, F3=Present, F4=Present). Malaria = 3/12 * 0 * 2/12 * 2/12 * 4/12 = 0.002348 P(S=Malaria) = P(F1=Present | S=Malaria) * P(F2=Absent | S=Malaria) * P(F3=Present | S=Malaria) * P(F4=Present | S=Malaria)

According to the calculations above, 0.0023 > P(S = Dengue) > P(S = Malaria) = 0.0128.

The patient who exhibits the signs of a high fever, shivering, and muscle wasting is therefore more likely to have dengue than malaria, and we may therefore conclude that the data point under consideration belongs to the class "Dengue".

3.1.2 Support Vector Machine (SVM):

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in

the correct category in the future. This best decision boundary is called a hyper plane. SVM chooses the extreme points/vectors that help in creating the hyper plane.

Based on the actual outcome and expected outcome, ROC curves were created. To compare the discriminative abilities of the models, the AUCs for the test data sets were computed. In order to compare the AUCs based on the output of the SVM models and MLR models, we computed P-values using Delong's approach.

When the cutoff value in the SVM model was set to the default value ((0)), the following formulae were used to determine sensitivity, specificity, PPV, and NPV.

Sensitivity=TP /TP+FP

Specificity=TN/TN+FN

PPV=TP/TP+FP

NPV=TN/TN+FN

the number of true positives, false positives, true negatives, and false negatives are denoted, respectively, by TP, FP, TN, and FN.

3.1.3 Random Forest Algorithm:

The goal of the Random Forest algorithm is to deal with the overfitting and underfitting of data .It deals with this condition efficiently.

It works as follows:

1.Selects k symptoms from dataset (Medical record) with a complete of m symptoms indiscriminately.

2. Repeats n times, resulting in n decision trees constructed from various arbitrary combinations of k symptoms (or a different random sample of the data, called bootstrap sample)

3. Uses a random variable to forecast the disease while using each of the n decision trees that have been developed. so that a total of n diseases are predicted from n decision trees, stores the anticipated disease.

4. Computes the votes for each predicted disease and uses the mode (predicted Disease that occurs the most frequently) as the final prediction from the random forest method.

4. CONCLUSIONS

The main contributions of this research are the development of cooperative techniques for agents in a distributed medical environment, as well as the incorporation of agent technology and CBR in the healthcare system to aid medical actors in their processes to improve diagnostic capabilities, treatment protocols, prescriptions, and recommendations. The cost-effectiveness of this approach's implementation in terms of utilising and modifying the current healthcare services and information sources is another significant addition. Medical actors need to modify their business processes and alter how they behave while using new technologies in order to gain from adopting the suggested system. This strategy will raise the standard of healthcare generally and the productivity of medical staff.

REFERENCES

[1] "Genetic neural network based data mining in prediction of heart disease utilising risk variables," IEEE Conference on Information & Communication Technologies (ICT), 2013, vol., no.,pp.1227-31, 11-12 April 2013, by Amin, S.U., Agarwal, and Beg

[2] Intelligent cardiac disease prediction system utilising data mining approaches, IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2008, vol., no., pp. 108–115, March 31–April 4, 2008. Palaniappan S., Awang R.

[3] Bojana R.Andjelkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovic (2015). "Prediction models for calculation of survival rate and relapse for breast cancer patients." IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE).

[4] Telecom Italia Lab, Java Agent Development Framework, http://jade.tilab.com/.

[5] JACK Documentation by Agent Oriented Software Limited (AOS), available at http://www.aosgrp.com/products/jack/documentation and instruction/jack documentation.html.

[6] The article "Advancements and Trends in Medical Case-Based Reasoning: An Overview of Systems and System Development" by Salari Khajehpour and E. Raheleh (2013) can be found in Iran Journal of Medical Informatics, Vol. 2, No. 4, pp. 12–16.

[7] "Cardiac MR Images Segmentation For Identification Of Cardiac Diseases Using Fuzzy Based Approach," IEEE, ICSSIT 2020, pp. 1238-1246, August 2020; P. B. Chanda and S. K. Sarkar

[8] Implementing Wireless Body Area Networks for Healthcare Systems, Yuce, M.R. Physical Sensor and Actuators, 162, 116-129.

[9] Building a Cardiovascular Disease Predictive Model Using Structural Equation Model and Fuzzy Cognitive Map by Singh, M., Martins, L.M., Joanis, P., and Mago, V.K. (2016).

[10] Using big data, Ghadge, Girme, Kokane, and Deshmukh (2016) developed an intelligent system for heart attack prediction. Journal of Recent Mathematics, Computer Science, and Information Technology.

[11] Utilizing Data Mining Techniques in the Diagnosis and Treatment of Heart Disease, Shouman, Turner, and Stocker (2012). Alexandria, Electronics, Communications, and Computers, 173–177.

[12] Alemdar, H., and Ersoy, C. (2010). A Survey of Wireless Sensor Networks in Healthcare.

[13] Francis, S., and Antony, G. (2015) Utilizing a Raspberry PI, a patient monitoring system. International Journal of Science and Research.

[14] Comparison of Data Mining Classification Methods for Predicting Cardiovascular Disease, M. Kumari and S. Godara, 2011. Computer Science and Technology International Journal.