International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 30 Issue 2 – March 2023.

HEALTHCURE DISEASE DETECTION

S.Prasanth[1], S.Siva[2], G.Rajaguru[3], Prof. S.Sujitha[4]

[1,2,3]B.E student, Dept. of CSE, P.S.R Engineering College, Sivakasi, TamilNadu, India

[4]Assistant professor, Dept. of CSE, P.S.R Engineering College, Sivakasi, TamilNadu, India

Abstract: People face various diseases due to the state of the environment and their living habits. The prediction of the disease in the early phase thus becomes an important task. However, accurate detection based on symptoms is too difficult for doctors. The correct detection of the disease is the most challenging task. To overcome this problem, data mining plays an important role in disease prediction. However, supervised machine learning (ML) algorithms have shown significant potential in outperforming standard disease diagnosis systems and medical professionals in the early detection of high-risk diseases. The goal is to recognize trends across different types of supervised ML models in disease detection by examining performance metrics.
Keywords: *CNN,ML, XGBoost.*

I. INTRODUCTION

A disease inference system is machine learning used to predict human diseases by providing relevant symptoms. Our system uses powerful machine learning algorithms to predict diseases based on symptoms provided by users. The healthcare industry uses and produces quite a large amount of data that can be used to obtain information about a particular disease for a patient. This healthcare information will be further used for the effective and best possible treatment of the patient's health. This area also needs some improvement using informative data in health sciences. But the main challenge is to extract the information from the data because the data is present in huge amount so some data mining and machine learning techniques are used. The expected result and this project is to predict the disease in advance, so that it can prevent the threat of life in time and save people's lives and reduce the cost of treatment to some extent. Detective diseases are 1. Breast cancer detection 2. Diabetes detection 3. Heart disease detection Recent work on deep learning has been in the disparate areas of machine learning, leading to a shift towards machine learning models that can learn and understand hierarchical representations of raw data with some preprocessing. With the development of this concept called big data technology, more attention is paid to disease prediction.

BREAST CANCER DETECTION

Breast cancer is considered a multifactorial disease worldwide and the most common cancer in women, accounting for approximately 30% of all cancers in women (i.e. 1.5 million women are diagnosed with breast cancer and 500,000 women die from the disease worldwide each year). Over the past 30 years, the disease has increased while the death rate has decreased. However, the reduction in mortality from mammography screening is estimated to be 20% and the improvement in cancer treatment is estimated to be 60%.

DIABETES DETECTION

Diabetes is a harmful disease in the world. Diabetes caused by obesity or high blood glucose and so on. It affects the hormone insulin, which results in abnormal crab metabolism and improves blood sugar levels. Diabetes occurs when the body does not produce enough insulin. Diabetes is the leading cause of death in the world. Early prediction of diseases like diabetes can be controlled and save human life. To achieve this, this work investigates the prediction of diabetes using various diabetes-related attributes.

HEART DISEASE DETECTION

Heart disease is often used interchangeably with cardiovascular disease. These types of diseases mainly relate to conditions of blocked or narrowed blood vessels that result in stroke, chest pain or angina pectoris, and heart attack. Other types of heart disease, such as those that affect the rhythm, valve, or muscle of the heart, are other types of heart disease.

II. LITERATURE REVIEW

Megha Rathi et al. proposed a model based on a hybrid approach using machine learning.

The approach using MRMR feature selection with four classifiers to find the best results. The author used four classifiers SVM, Naïve Bays, Function tree and End Meta and made a comparison between all of them. SVM is found to be a good classifier.

To find out better results. The classifier used by the author was Extreme Learning Machine, SVM, KNN and ANN. A small change was made in the classifier to find better results. Accordingly, Extreme Learning Machine gave better results. Support Vector Classifier, Random Forest, Gradient Boosting, Naive Bayes, Cart Model, Neural Network and Linear Regression algorithm on WisconsinBreast Cancer datasets (original) and compared the effectiveness and efficiency of these algorithms in terms of accuracy, precision, sensitivity and specificity. best classification accuracy. The support vector has demonstrated its effectiveness in the prediction and diagnosis of breast cancer; achieves the best performance in terms of accuracy and low error rate.

A survey on breast cancer detection using image processing techniques.

To examines various techniques that are part of medical image processing and are prominently used in the detection of brain tumors from MRI images. Based on this research, this document was written which lists the various techniques used. A brief description of each technique is also provided. Also, of all the different steps involved in the tumor detection process, segmentation is the most important. One of the most important tasks in breast cancer detection Interestingly, the domain of brain tumor analysis effectively uses concepts to detect different types of brain tumors. system is the isolation of abnormal tissues from normal breast tissues. processing of medical images, especially MR images, for the automation of basic steps, i.e. extraction, segmentation, classification for proximal tumor detection. Research is more in favor of MR for its non-invasive imaging properties. Computer-aided diagnostic or detection systems are becoming challenging and still an open problem due to the variability of tumor shapes, regions, and sizes. Past works by many researchers in medical image processing and soft computing have provided a remarkable review of automatic brain tumor detection techniques focusing on both segmentation and classification and their combinations. The manuscript reviews different brain tumor detection techniques for MR imaging along with strengths and for detecting different types of brain tumors.

Sun and Zhang discussed several deep learning methods and classification methods such as artificial neural network, decision trees, random forest, and support vector machine.

Health care systems offer customized services in a wide range of areas to help patients integrate into their normal lives. Diabetes mellitus is among the most serious serious problems in the medical profession. Classification is one of the most important decision-making methods in today's practical conditions. The primary goal is to categorize the data as diabetic or non-diabetic and increase classification accuracy. Machine learning in diabetes diagnosis is mostly about understanding patterns from a diabetes data set that would be provided. Machine learning has always been an evolving, reliable and supportive technology in the medical sector in recent times. This study is aimed at identifying patients' diabetes types based on personal and clinical information using machine learning classifiers. This section contains a summary of the works proposed by various researchers during the last decade. It is beneficial to identify the shortcomings of the proposed works in the field of machine learning classifiers of treatment regimens of diabetic patients. Diagnosing diabetes is an ever-growing field of study. implemented a logistic regression classification technique to classify diabetes data. The training data includes 459 patients and the test data includes 128 patients. The classification accuracy achieved by the authors was 92% using logistic regression. The main drawback of the model was that it was not compared with other diabetes prediction models and therefore could not be validated on a 50% training set and 50% test set. The model was designed using a combination of Naive Bayesian and Support Vector Machine algorithms for diabetes prediction. A dataset was collected from three different locations and the proposed model was validated on this dataset. Eight attributes were present in the dataset and consisted of 402 patients, 80 of whom were type 2 diabetics. Ensemble of Naive Bayes and Support Vector Machine achieved an accuracy of 97.6%, which is much better than the algorithms when run separately on the dataset, that is, Naive Bayes achieves an accuracy of 94.52 and Support Vector Machine achieves an accuracy of 95.52%. The authors did not mention any preprocessing technique to filter out unwanted values from the dataset.

Santana Krishnan. J,et,al proposed an article Prediction of Heart Disease Using Machine Learning Algorithms" using decision tree and Naive Bayes algorithm for prediction of heart disease. In a decision tree algorithm, a tree is built using certain conditions that give a True or False decision. Algorithms like SVM, KNN are results based on conditions of vertical or horizontal distribution depends on dependent variables. But a decision tree for a tree structure that has a root node, leaves and branches based on the decision made in each tree Decision tree also helps in underestimating the importance of attributes in a dataset. They also used the Cleveland dataset. The dataset is divided into 70% training and 30% testing using some methods. This algorithm provides 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, non-linear, dependent data, so it is suitable for heart disease dataset because this dataset is also complicated, dependent and non-linear in nature. This algorithm provides 87% accuracy. Some analysis was conducted to consider performing a data mining strategy on the same data set and the result decided that the decision tree has the highest accuracy than the Bayesian classifier. In this algorithm, there will be one input layer and one output layer and one or multiple layers are the hidden layers between these two input and output layers. Through hidden layers, each input node is connected to the output layer. Some random weights are assigned to this connection. The second input is called the bias, which is assigned a weight based on the requirement whether the connection between the nodes can be forward or backward.

III. METHODOLOGY

EXISTING METHODOLOGY

The system predicts chronic diseases that are for a specific region and for a specific community. Disease prediction is done only for specific diseases. In this system, Big Data & CNN algorithm is used for disease risk prediction. For type S data, the system uses machine learning algorithm i.e. K-nearest Neighbors, Decision Tree, Naive Bayesian. The accuracy of the system is up to 94.8%. The current paper streamlines machine learning algorithms for effective prediction of chronic disease outbreaks in high-disease communities. We experiment with modified prediction models on real hospital data collected from central China. We propose a novel convolutional neural network-based multimodal k-disease prediction algorithm (CNN-MDRP) using structured and unstructured hospital data. The current system predicts chronic diseases that are region-specific and community-specific. This system only predicts certain diseases. In this system, Big Data & CNN algorithm is used for disease risk prediction. For type S data, the system uses machine learning algorithm i.e. K-nearest Neighbors, Decision Tree, Naive Bayesian. The accuracy of the current system is up to 94.8%. In the current paper, they streamline machine learning algorithms for effective prediction of chronic disease outbreaks in high-disease communities. They experiment with modified prediction models on real hospital data collected from central China. They propose a convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured hospital data.

PROPOSED METHODOLOGY

The system is used to predict most chronic diseases. It accepts structured and text type data as input to a machine learning model. This system is used by end users. The system will predict disease based on symptoms. This system uses machine learning technology. To predict the power of the random forest and the XGboost algorithm used for prediction, the final output will be in the form of 0 or 1, for which the Logistic tree is used. Diseases are predicted by our system. It accepts a structured data type as input to a machine learning model. This system is used by end users - patients/any user. In this system, the user enters all the symptoms they suffer from. These symptoms will be fed to a machine learning model to predict the disease. Algorithms are then applied that provide the best

accuracy. The system will predict disease based on symptoms. The final output of this system will be the disease predicted by the model.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Collection of dataset

Initially collecting a dataset for our heart disease prediction system. After collecting the dataset, we split the dataset into training data and test data. The training data set is used to learn the prediction model and the test data is used to evaluate the prediction model. For this project, 70% training data is used and 30% data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; of which 14 attributes are used for the system.

Selection of attributes

Attribute or feature selection involves selecting appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various patient attributes are selected for prediction, such as gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc. Correlation matrix is used to select attributes for this model.

Data preprocessing

Data preprocessing is an important step for building a machine learning model. Initially, the data may not be clean or in the desired format for the model, which can cause misleading results. During data preprocessing, we transform the data into the format we require. It is used to deal with noise, duplicates and missing values of the data set. Data preprocessing includes activities such as dataset import, dataset partitioning, attribute scaling, etc. Data preprocessing is required to improve model accuracy.

Data balancing

Unbalanced datasets can be balanced in two ways. They are under-sampling and over-sampling (a) Under-sampling: In under-sampling, the balance of the data set is achieved by reducing the size of the ample class. This process is considered when the amount of data is sufficient. (b) Over Sampling: In Over Sampling, the balancing of the data set is done by increasing the size of rare samples. This process is considered when the amount of data is insufficient.

Disease prediction

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification.

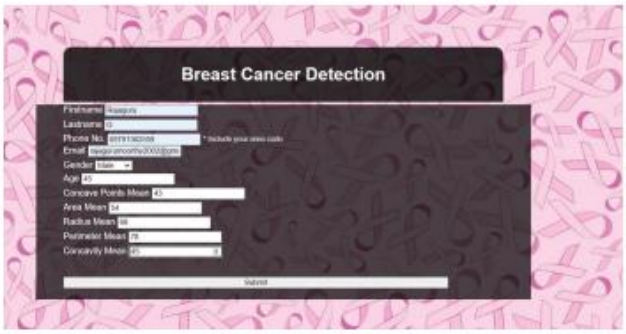

FIGURE 1: BREAST CANCER LOGIN PAGE

FIGURE 2: DIABETES LOGIN PAGE

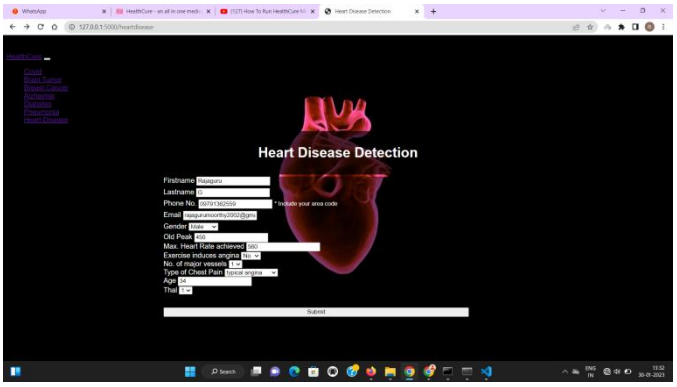

FIGURE 3: HEART DISEASE LOGIN PAGE

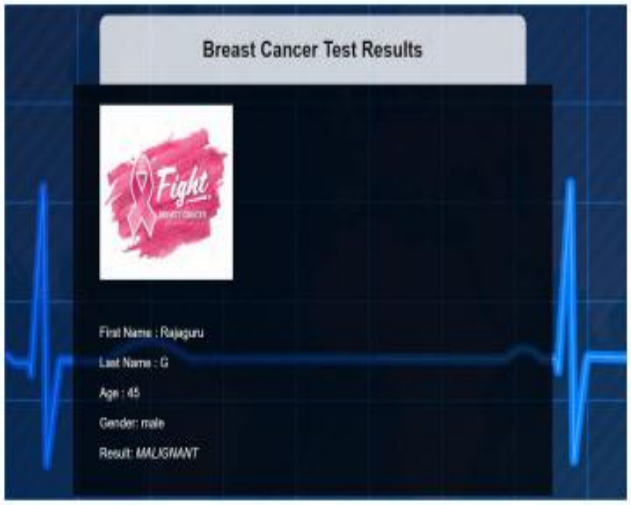

FIGURE 4: BREAST CANCER RESULT

FIGURE 5: DIABETES RESULT

FIGURE 6: HEART DISEASE RESULT

V. CONCLUSION

With Relatively good and higher accuracy is achieved with our proposed system. This is then used by researchers, doctors or physicians to provide the best treatment and medical care to patients. Hence, machine learning when used in healthcare can lead to effective treatment and the patient is also well taken care of. Here we are trying to implement some healthcare machine learning features into our system. Instead of direct diagnosis, when a patient is predicted to have a disease, machine learning is implemented using certain machine learning algorithms and then healthcare can be smarter and better. When we compare the different algorithms used for disease prediction from our dataset and the output we expect, we get the best accuracy with the logistic regression algorithm and the KNN algorithm, while the LDA algorithm had the lowest performance compared to the other algorithms. Machine learning (ML) provides methods, processes and certain techniques that can help solve the problem of diagnostic problems in a simpler and modernized range of medical fields. ML is currently used for prediction and analysis of clinical trials. ML is currently also used for the data analysis process such as error detection in the dataset and to deal with incorrect data present in our system. It is a debatable topic that the perfect use and implementation of ML algorithms can be a great source of help in the integration of computer systems in the field of healthcare to facilitate and improve the work of doctors and ultimately lead to an improvement in the efficiency and quality of our medical care for the respective patients. This study is based on a comparison and evaluation of recent work on the prognosis and prediction of Alzheimer's disease using machine learning methods. Recent trends with respect to machine learning are explicitly revealed, including the types of data used and the performance of machine learning methods in predicting the early stages of Alzheimer's disease. Clearly, machine learning tends to improve prediction accuracy, especially compared to standard statistical tools. However, based on the review, the clinical diagnosis was not 100% accurate because pathological verification was not provided, which subsequently introduces uncertainty into the predicted results.

VI. FUTURE ENHANCEMENT

As today we can clearly observe the increase in the use of computers and technology to consider huge amounts of data, computers are being used to perform various complex tasks with a commendable degree of accuracy. Machine learning (ML) is a collection of different techniques and algorithms that enable computers to perform such complex tasks in a simplified manner. It is also used in academia, which is for students or students, and also in industry to make accurate predictions and use these different sources of data sets and information. So far we can say that we have grown in the field of big data, machine learning and data science etc. and we have been part of one of those industries that have been able to collect such data and employees to transform their goods and services in the desired way. The learning methods developed for these industries and researches offer excellent potential for further improvisation of medical research and clinical patient care in the best possible way. Machine learning uses mathematical algorithms and procedures that are used to describe the relationship between the variables used in the model and others. Our article will explain the process of training a model and learning an appropriate algorithm to predict the presence of a specific disease from a tissue sample based on its properties. Although these algorithms work in different and unique ways depending on how they are developed and used by researchers. One way is to consider their highest goals. The goal of our paper and statistical methods is to reach a conclusion about data that is collected from a wide range of samples of our population. Although many techniques such as linear and logistic regression are able to predict diseases. Consider, for example, the case where, if we can create a model that describes and understands the relationship between clinical variables and their transience, we can track organ transplant surgery, i.e. we need factors and features that differentiate low mortality from high mortality, if we can develop such outcomes and in in the near future to reduce the death rate to the desired level and also cannot be said to be better than such a situation.

REFERENCES

1. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

2. K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, ”Performance comparison of three data mining techniques for predicting kidney disease survivabil- ity”, International Journal of Advances in Engineering Technology, Mar. 2014.

3. Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Anal- ysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.

4. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

5. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

6. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

7. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.

8. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

9. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

10.Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.