IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLASSIFICATION OF PATIENTS USING DIFFERENT ML TECHNIQUES WITH RESULTS IN DISEASE PREDICTION

SHAIKH MOHD FAIZAN MOHD KALEEM[1], ASST.PROF. V. S. KARWANDE[2]

ME Student, Department of Computer Science & Engineering, EESGOI , India[1]
HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI,India. [2]

Abstract: The technology allows users to with the assistance of algorithms, users of the technology may predict whether or not they will develop diabetes mellitus and whether or not they will acquire cancer. This system uses a number of classification models, including association rules, logistical regression, artificial neural networks, decision trees, and Naive Bay. The accuracy of every model in the project is then determined using the Random Forest technique. The project is a smartphone application made to predict whether or not a person's class is at risk for diabetes and cancer. The dataset used is a Pima Indians Diabetes Data Set, which contains information on individuals, some of whom develop diabetes. We are investigating four popular classifiers for sickness risk prediction. These algorithms include Bay Naive, Regression Logistic, Artificial Neural Network, and Decision Tree. Subsequently, these algorithms are merged with bagging and boosting procedures to improve each model's solidity. Finally, the Random Forest algorithm is applied. The purpose of this research is to assess diabetes and cancer risk without requiring blood work or hospital stays for any individual. Encouraging and improving human health is another goal of the study.

Keywords: Extreme Learning Machine(ELM); Decision Tree (DT);Random Forest(RF);Convolution Neural Network(CNN); Confusion Matrix(CM);Artificial neural network(ANN);Super Vector Machine(SVM).

I INTRODUCTION

In Health care information systems like to gather data in databases for research and analysis in order to assist in medical decision-making. As a result, hospitals and other medical facilities are growing their medical information systems, which makes information retrieval more challenging. Computer-based analytical techniques are needed since traditional manual data analysis has become ineffective. Various methods have been devised and examined for the purpose of computerized data analysis. One significant advancement in the kind of analytical tools is data mining. Increases in diagnostic accuracy, cost savings, and human resource utilization have all been shown to be advantages of using data mining into medical analysis. Diabetes and cancer are two devastating illnesses in our civilization. Cancer claims a great number of lives each year. Taking the right medications might help the patient survive and in certain cases even cure them completely if the cancer is in a benign stage. Diabetes and cancer are two other diseases that kill people gradually. Both diabetes and cancer are becoming more and more commonplace on the planet. However, one Asian study claims that Asians account for 60% of the global diabetes population. Asian people are consequently more vulnerable to machine learning, which may identify diabetes and cancer in a patient. Therefore, it is anticipated that in this research, diabetes and cancer would have binary values of 1 or 0, which correspond to "YES" or "NO." We will test the algorithms to see whether they perform better in a different setting. A comparison of the performance is made across many classifications to assess how well they adhere to the same set of data and how long each classification model takes. Developing some techniques to apply curricular learning [6] to the data collecting is one of the study's challenges.

The two biggest health issues of the past few decades have been diabetes and cancer, both of which are extremely complicated and challenging to diagnose. It is brought on by the body's improper synthesis of insulin. The primary factor controlling glucose levels is insulin.

It also causes blindness, heart disease, kidney illness, and no injury, among other risks. A routine blood examination can be used to diagnose diabetes. Diabetes may be managed with the right food habits and exercise regimens to reduce the risks. Diabetes and cancer currently have no long-term treatments accessible. Diagnosing diabetes demands extra work for any physician who has studied symptoms in the past and thoroughly examined the patient's medical history. A number of machine learning algorithms have been developed to automatically identify cancer and diabetes in order to make diagnosis easier. Synthetic intelligence, sometimes referred to as artificial intelligence, is a branch of engineering that studies computer behavior. AI has become essential in recent years in mimicking human intelligence. The goal of the AI branch of machine learning is to impart information to these intelligent systems. A wide range of algorithms are used in machine learning to create and analyse data sets in any format. Machine learning can be done alone or under supervision. In supervised learning, data are taught and projected based on training. Tests on unidentified samples and exercise samples are used to build this function. Uncontrolled learning results in an untrained mechanism. In recent decades, there has been a growth in medical diagnostic decision support systems. Globally, experts in medicine are paying increasing attention to the design of medical systems. Machine learning techniques are used by medical systems to predict every illness based on its presence. When combined with these methods, pattern recognition and data mining allow for the useful collecting of medical data. The most popular data mining technique for generating decisions from real-world data is classification. The performance of the system may be directly impacted by the utilization of data. Performance and features or qualities are closely related. The predictive diagnostics system's accuracy will be more impacted by the best features being chosen.

II LITERATURE SURVEY

Diabetes, sometimes referred to as chronic illness, is a group of metabolic diseases brought on by an abnormally high blood sugar level over time. Achieving precise early prediction can significantly reduce the risk factor and severity of diabetes. Robust and accurate diabetes prediction is quite challenging since diabetes datasets include a low number of labelled data points and outliers (or missing values). This literature proposes a robust framework for diabetes prediction using outlier rejection, data standardisation, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP). where the matching Area Under ROC Curve (AUC) of the ML model is used to determine the weights. AUC is selected as the performance measure, and it is then optimised during hyperparameter tuning by applying the grid search approach. All of the investigations in this literature were conducted using the same experimental setups using the Pima Indian Diabetes Dataset. With sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively, the ensembling classifier outperforms the state-of-the-art results by 2.00 percent, making it the best classifier in the current system. For diabetes prediction, our proposed framework performs better than the other methods in the paper. Additionally, it can produce better results on the same dataset, which raises the accuracy of the diabetes prediction. Our diabetes prediction source code has been made public. [1].

One of the leading causes of mortality for women is breast cancer. The Moroccan Ministry of Health reports about forty thousand new cases annually. The death rate from disease is greatly reduced when diseases are diagnosed early enough for lifestyle modifications to be protective. Machine learning (ML) algorithms offer an alternative to the existing methods for predicting breast cancer, or at the very least, they can help radiologists with their reasoning process, which might prevent a breast cancer biopsy for a large number of women and some men. In the present study, different machine learning models are compared. In order to determine if a patient has a benign or malignant tumour, the study employs and evaluates four machine learning techniques (kNN, decision tree, Binary SVM, and Adaboost). The machine learning methods were taught and then evaluated on the Breast Cancer Wisconsin dataset. Neighbourhood Components Analysis (NCA) is used to include the features of the dataset into a feature selection model, hence reducing the number of features and, consequently, the complexity of the model. 99.12 percent was the greatest predictive accuracy for the kNN model, 98.86 percent was the best predictive specificity for the Binary SVM model, and one percent was the highest predictive sensitivity for both the kNN and Adaboost models.[2].

As part of healthcare practises, a range of patient data are gathered to help the physician diagnose the patient accurately. This information may include a patient's simple symptoms, a doctor's first diagnosis, or a thorough test result from a laboratory. Therefore, these data are only analysed by a medical professional, who then uses their own medical expertise to identify the condition. To identify whether or not a patient has an ailment, artificial intelligence has been used in conjunction with the Naive Bayes and random forest classification algorithms to categorise a range of sickness datasets, such as diabetes, heart disease, and cancer. For both methods, a performance analysis of the illness data is computed and compared. The simulation results show the nature and complexity of the dataset in addition as the effectiveness of classification algorithms on it. [3].

One of the most common and serious illnesses in Bangladesh and the rest of the globe is diabetes. In addition to harming the blood, it also causes a number of illnesses that lead to many deaths annually, including heart disease, renal sickness, blindness, and kidney issues. Designing a system that can reliably identify diabetes patients using medical data is therefore essential. This work suggests using a five-fold and ten-fold cross-validation method to train a deep neural network's characteristics in order to detect diabetes. The UCI machine learning repository database provided the Pima Indian Diabetes (PID) data set. The PID dataset findings demonstrate that deep learning can build a useful system for diabetes prediction, with an MCC of 97, an F1 score of 98, and a prediction accuracy of 98.35 percent. Furthermore, accuracy of 97.11 percent, sensitivity of 96.25 percent, and specificity of 98.80 percent were obtained using ten-fold cross-validation. The experimental findings demonstrate that the proposed method yields good results when applied with five-fold cross-validation. [4].

A group of metabolic diseases known as diabetes mellitus (DM) are characterised by consistently elevated blood glucose levels. It is brought on by either insufficient insulin synthesis or an improper reaction of the cells to the insulin that is generated. It is a serious global public health concern that impacts individuals everywhere. Diabetes arises from insufficient insulin production by the pancreas or from inadequate insulin utilisation by the organism. The diagnosis of diabetes (including its etiopathophysiology, treatment, and other aspects) requires the creation and analysis of a substantial quantity of data. The use of data mining techniques has demonstrated its efficacy and usefulness in assessing previously unidentified patterns or links present in vast datasets. This work describes five machine learning approaches: AdaBoost, LogicBoost, RobustBoost, Nave Bayes, and Bagging, for the analysis and prediction of diabetes patients. A data collection of Pima Indians with diabetes is used to test the suggested tactics. The calculated results show that the bagging and AdaBoost techniques achieve classification accuracy of 81.77 percent and 79.69 percent, respectively, which is highly accurate. Consequently, the methodologies for DM prediction that have been discussed are highly appealing, successful, and efficient. [5].

An overabundance of sugar that has accumulated in the blood is what causes diabetes. It is recognised as one of the worst illnesses in the world right now. Whether they realise it or not, individuals all across the world are impacted by this fatal illness. Heart attacks, paralysis, renal failure, blindness, and other problems can also result from diabetes. A variety of computer-based detection techniques have been developed and described for the purpose of predicting and evaluating diabetes. It costs more time and money to identify diabetics using the conventional technique. But now that machine learning has advanced, we may be able to provide a workable solution for this challenging issue. Consequently, we developed an architecture that is able to determine if a patient has diabetes. The main objective of this research is to develop a web application that leverages the increased prediction accuracy of a complex machine learning algorithm. We used the Pima Indian benchmark dataset, which uses diagnostics to forecast when diabetes may manifest. With an accuracy rate of 82.35 percent, the Artificial Neural Network (ANN) shows a significant improvement in accuracy, which encourages us to create an interactive web application for diabetes prediction. [9].

III. SYSTEMS ARCHITECTURE

The process starts with data modification. We will next look at four models for determining a prediction model. Next, the accuracy of each model is determined and contrasted with the optimal model. The ability to identify diseases like diabetes and cancer may be helpful for both patients and medical professionals. From a medical perspective, they can help patients decide what to do next by identifying factors like diabetes prevalence or a patient's susceptibility to cancer. In the end, the study produces a web application.

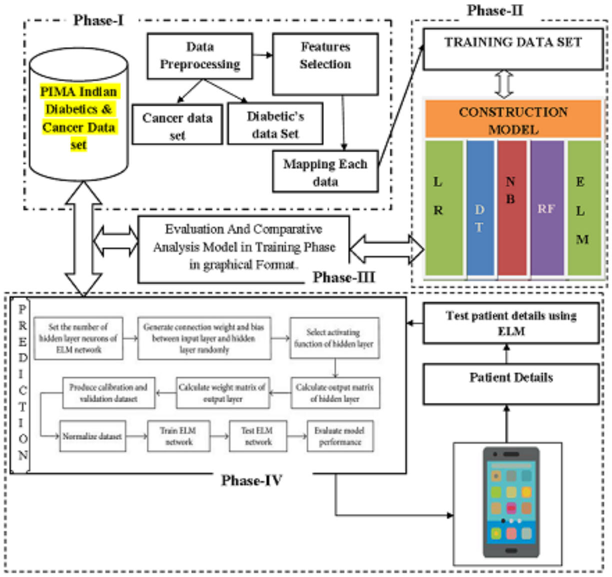
Phase-I
PIMA Indian Diabetics & Cancer Data set
Data Preprocessing
Features Selection
Cancer data set
Diabetic's data Set
Mapping Each data
Phase-II
TRAINING DATA SET
CONSTRUCTION MODEL
LR
DT
NB
RF
ELM
Evaluation And Comparative Analysis Model in Training Phase in graphical Format.
Phase-III
PREDICTION
Set the number of hidden layer neurons of ELM network
Generate connection weight and bias between input layer and hidden layer randomly
Select activating function of hidden layer
Calculate output matrix of hidden layer
Calculate weight matrix of output layer
Produce calibration and validation dataset
Normalize dataset
Train ELM network
Test ELM network
Evaluate model performance
Phase-IV
Test patient details using ELM
Patient Details

Figure No 3.1: System Architecture

There were four distinct sections to the architecture seen above. Here, we offered an Advanced modified user query ELM algorithm with a yes/no forecast for users with diabetes or cancer. There are two distinct modes throughout the full working module: Phase I, II, III, and IV of the testing module are included in the training module. To preprocess each attribute, we make reference to the validity or invalidity of the data set (data set specifics are supplied on the data structure). Verify the validity of each attribute by looking over its characteristics. Every field has been processed, and each field's value has been verified. Here, we have used a variety of algorithms, including LR, DT, NB, RF, and ELM and calculated the accuracy of each algorithm.

IV EXPERIMENTAL RESULTS

The methods were applied to improve the data set's validity, accuracy, receiver's curve, and logic. The prediction precision was increased by the proposed model that included the use of class and cluster approaches. One benefit is that algorithms may be included in the Pima Indian Diabetes Dataset and other datasets. However, the preprocessing stage is lengthier, which is a limitation. We have explained how certain models optimise the initialised cluster centre method with an emphasis on K-means. However, the predictions and matches of this improved model are based on diabetes and cancer. It guarantees optimal data retention and minimal time usage. Improving prediction model accuracy and modifying the model for a variety of datasets are the main challenges that have been overcome. Compared to other algorithms, the test and training mode's algorithm is incredibly accurate.

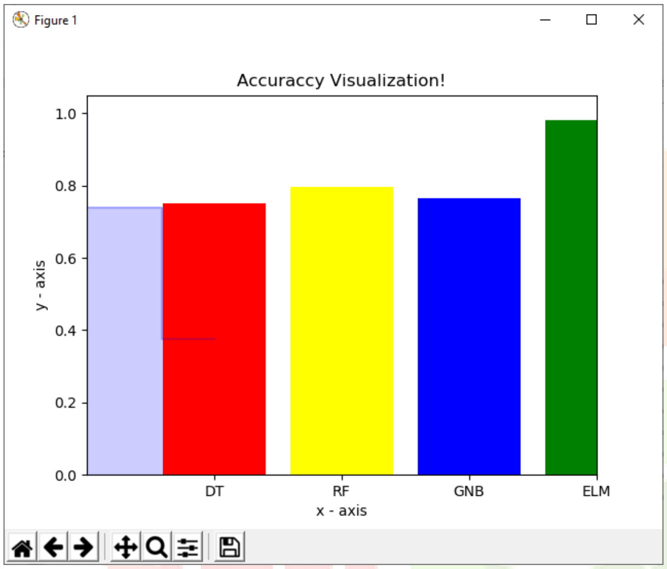

Figure No 4.1: Accuracy Visualization.

V CONCLUSION

The proposed system suggested framework in this proposed system, certain classification strategies are studied. Several optimisation attempts are being made to improve the algorithms' performance. The ability to identify diseases like diabetes and cancer may be helpful for both patients and medical professionals. From a medical perspective, they can help patients decide what to do next by identifying factors like diabetes prevalence or a patient's susceptibility to cancer. This allows doctors to evaluate the patient's condition and, in the case of a high-risk patient, provide medication and a better way of living.

REFERENCES

1. A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza- Albarrán, and K. L. Ramaiya, "Diabetes in developing countries," Journal of Diabetes, vol. 11, no. 7, pp. 522-539, Mar. 2019.

2. R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri,"Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in Proc. International Conference on Computing Networking and Informatics, Oct. 2017, pp. 1-5.

3. N. H. Choac, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, Apr. 2018.

4. P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karurangaa, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, R. Williams, and IDF Diabetes Atlas Committee, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation," Diabetes Research and Clinical Practice, vol. 157, pp. 107843, Nov. 2019.

5. M. Maniruzzaman, M. J. Rahman, M. A. M. Hasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," Journal of Medical Systems, vol. 42, no. 5, pp. 92, May 2018.

6. V. Jackins,S. Vimal,M. Kaliappan,Mi Young Lee,"AI-based smart prediction of clinical disease

using random forest classifer and Naive Bayes",Springer,2020.

7.Safial Islam Ayon, Md. Milon Islam,"Diabetes Prediction: A Deep Learning Approach",I.J. Information Engineering and Electronic Business, 2019.

8.Shahadat Uddin, Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni,"Comparing different supervised machine learning algorithms for disease prediction", BMC Medical Informatics and Decision Making,2019.

9.Muhammad Azeem Sarwar,Nasir Kamal,Wajeeha Hamid,Munam Ali Shah,"Prediction of Diabetes Using Machine Learning Algorithms in Healthcare",24th International Conference on Automation & Computing, Newcastle University,Newcastle upon Tyne, UK, 6-7 September 2018.

10.Samrat Kumar Dey,Ashraf Hossain,Md. Mahbubur Rahman,"Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm",21st International Conference of Computer and Information Technology (ICCIT),

21-23 December, 2018.

11.P. Suresh Kumar, S. Pranavi,"Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",IEEE,2017.

12.Deepika Verma,Dr. Nidhi Mishra,"Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques",International Conference on Intelligent Sustainable Systems(ICISS),IEEE Xplore,2017.

13.B. Nithya,Dr. V. Ilango,"Predictive Analytics in Health Care Using Machine Learning Tools and Techniques",International Conference on Intelligent Computing and Control Systems(ICICCS),IEEE Xplore,2017.

14.Ioannis Kavakiotis,Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras,Ioannis Vlahavas,Ioanna Chouvarda,"Machine Learning and Data Mining Methods in Diabetes Research",Computational and Structural Biotechnology Journal,2017.

15.Deepa Gupta,Sangita Khare,Ashish Aggarwal and Amrita Vishwa Vidyapeetham,"A Method to Predict Diagnostic Codes for Chronic Diseases using Machine Learning Techniques",International Conference on Computing, Communication and Automation(ICCCA),IEEE Xplore,2016.

16.Prof. Dhomse Kanchan B.,Mr. Mahale Kishor M.,"Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis",International Conference on Global Trends in Signal Processing, Information Computing and Communication,IEEE Xplore,2016.

17.Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat,"Comparative Studies on Breast Cancer Classifications with K-Fold Cross Validations Using Machine Learning Techniques",IEEE,2015.