OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

CLASSIFICATION OF PATIENTS WITH DIABETES & CANCER PREDICTION RESULTS BY USING DIFFERENT ML TECHNIQUES

SHAIKH SUMAIRA ANJUM SHAIKH SALEEM[1], ASST.PROF. V. 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 utilize algorithms to forecast the likelihood of diabetes mellitus in their bodies and forecast whether or not they will get cancer. The several categorization models are employed in this system, such as Decision Tree, artificial neural networks, logistical regression, association rules and the Naive Bay. The Random Forest approach is then used to discover the exactness of each model in the project. The employed dataset is a Pima Indians Diabetes Data Set with the information of patients, some of whom acquire diabetes, which makes the project a mobile application designed to forecast whether or not the class of a person is at risk for diabetes and cancer. We are studying four common illness risk prediction classifiers. These algorithms consist of Decision Tree, Artificial Neural Network, Regression Logistic and Bay Naive. After that, strategies for bagging and boosting using these algorithms are integrated to increase the solidity of each model. At the end, the algorithm of Random Forest is used. The aim of this study is to estimate the risk of cancer and diabetes without blood testing or hospitalization for everybody. The study is also aimed at encouraging and promoting healthy human health.***

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

I INTRODUCTION

In order to help in medical choices, health care information systems prefer to collect data in databases for study and analysis. As a result, medical information systems in hospitals and medical institutes are expanding and making it more difficult to retrieve information. Traditional manual data analysis has become inefficient and computer-based analytical approaches are required. To this end, several techniques have been developed and studied for computerised data analysis. Data mining is a major improvement in the type of analytical tools. The benefits of bringing data mining into medical analysis have been demonstrated by increasing diagnosis accuracy, cost savings and human resources. In our culture, cancer and diabetes are two deadly diseases. Many individuals die of cancer every year. [1]. [2]. If the cancer is at a benign stage, taking suitable treatments can assist the patient to live and can in some circumstances even cure them entirely. Another illness that kills individuals slowly is cancer and diabetes. Almost everywhere in the globe, cancer and diabetes have grown prevalent [4]. However, according to an Asian research, 60 percent of all diabetes populations in the globe are from Asia [5]. Asian individuals are therefore at great danger Machine learning can predict the existence of cancer and diabetes in a patient. So cancer and diabetes is projected to have binary values such as 1 or 0 that means "YES' or "NO' in this research. The data collection utilised for cancer and diabetes contains patient features that might lead to cancer and diabetes. Machine learns the characteristics and then forecasts in "YES' or "NO'.' Some tests will be conducted to check whether the algorithms can work better in another scenario. The performance comparison is examined in several classifications to see how they comply with the same data set and how long each classification model takes. One of

the challenges of this study is to develop some methods to apply curricular learning [6] to the data collection.

Diabetes and cancer is one of the main health concerns over the last few decades, which have been very complex and difficult to detect. It is caused by inappropriate insulin production in the human body. Insulin is the main regulating parameter for glucose. It results in several other hazards such as renal disease, blindness, heart disease, and no harm. Diabetes diagnosis can be done by normal blood inspection. Diabetes can be treated by appropriate dietary habits and exercise programmes to lower the hazards involved. No permanent treatment is still available for diabetes and cancer. Diabetes diagnosis requires particular effort for any doctor with prior symptom knowledge and profound investigation of patient history. To facilitate and facilitate diagnosis, several machine learning algorithms are created to diagnose diabetes and cancer automatically. Artificial Intelligence, often known as Synthetic Intelligence, is an area of engineering connected with computer behaviour. In recent years, AI has played a fundamental role in simulating human intellect. Machine Learning is an AI field that tries to provide such intelligent systems with knowledge. Machine learning comprises of a large variety of algorithms for the construction and analysis of data sets of any form. Machine learning can be monitored or unattended. Data are taught and forecasted based on training in supervised learning. This function is developed on the basis of workout samples and tests on unknown samples. The mechanism stays untrained in uncontrolled learning. Medical diagnostic decision support systems have increased in recent decades. The design of expert medical systems has garnered more attention among scholars worldwide. Medical systems utilise machine learning approaches to anticipate every disease depending on its existence. Pattern recognition and data mining enable helpful medical data collection in conjunction with these approaches. Classification is the most frequent data mining methodology for decision making from real-world data. The use of data might have a direct impact on system performance. Features or characteristics have a lot to do with performance. Selection of the best characteristics will have more influence on the accuracy of the predictive diagnostics system.

II LITERATURE SURVEY

Diabetes, often known as chronic sickness, is a collection of metabolic illnesses caused by a persistently high blood sugar level. If exact early prediction is achievable, the risk factor and severity of diabetes can be considerably decreased. Due to the low number of labelled data and the existence of outliers (or missing values) in diabetes datasets, robust and reliable diabetes prediction is extremely difficult. 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) were used in this literature to propose a robust framework for diabetes prediction. In this research, weighted ensembling of different ML models is also proposed to improve diabetes prediction, where the weights are calculated using the ML model's corresponding Area Under ROC Curve (AUC). The performance statistic is chosen as AUC, which is then maximised using the grid search methodology during hyperparameter tweaking. Using the Pima Indian Diabetes Dataset, all of the experiments in this literature were carried out under the identical experimental settings. The ensembling classifier is the top performing classifier in this existing system, with sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively, outperforming the state-of-the-art findings by 2.00 percent in AUC. Our suggested framework outperforms the other approaches presented in the article for diabetes prediction. It can also deliver superior findings on the same dataset, resulting in higher diabetes prediction performance. Our diabetes prediction source code has been made public. [1].

Breast cancer is one of the most common causes of death among women. Every year, the Moroccan Ministry of Health reports approximately 40.000 new cases. Early diagnosis of diseases has a tremendous impact on disease mortality when lifestyle can be a protective pattern. Machine learning (ML) algorithms provide an alternative to current breast cancer prediction methodologies, or at the very least can aid radiologists in their reasoning flow, potentially saving many females and some males from breast cancer biopsy. The current research is a comparison of various machine learning models. The study uses and analyses four machine learning methods to identify if a patient has a malignant or benign tumour (kNN, decision tree, Binary SVM, and Adaboost). On the Breast Cancer Wisconsin dataset, the machine learning approaches were trained and subsequently tested. To reduce the amount of features and hence the model's complexity, the dataset's features are put into a feature selection model using Neighbourhood Components Analysis (NCA). The kNN model had the highest predictive accuracy of 99.12 percent, the Binary SVM model had the best predictive specificity of 98.86 percent, and both the kNN and Adaboost models had the highest predictive sensitivity of one.[2].

Healthcare practises include gathering a variety of patient data that will aid the doctor in appropriately diagnosing the patient's health status. Simple symptoms noted by the individual, an initial diagnosis by a physician, or a detailed test result from a laboratory could all be included in this information. As a result, these data are only used for analysis by a doctor, who subsequently diagnoses the ailment based on his or her own medical knowledge. Artificial intelligence has been used in conjunction with the Naive Bayes and random forest classification algorithms to classify a variety of illness datasets, including diabetes, heart disease, and cancer, in order to determine whether or not a patient is infected with the disease. A performance analysis of the

disease data is calculated and compared for both algorithms. The simulation results demonstrate the efficacy of classification strategies on a dataset, as well as the nature and complexity of the dataset. [3].

Diabetes is one of the most frequent and dangerous diseases in Bangladesh and throughout the world. It is not only damaging to the blood, but it also causes a variety of disorders such as blindness, renal illness, kidney problems, heart disease, and other conditions that result in a large number of deaths each year. As a result, it is critical to design a system that can accurately detect diabetes patients using medical information. This paper proposes an approach for employing a deep neural network to diagnose diabetes by training its properties in a five-fold and ten-fold crossvalidation method. The data set for Pima Indian Diabetes (PID) was obtained from the UCI machine learning repository database. With a prediction accuracy of 98.35 percent, an F1 score of 98, and an MCC of 97 for five-fold cross-validation, the results on the PID dataset show that deep learning can construct an advantageous system for diabetes prediction. Additionally, ten-fold cross-validation yielded accuracy of 97.11 percent, sensitivity of 96.25 percent, and specificity of 98.80 percent. The experimental results show that when using five-fold cross-validation, the suggested approach produces promising results. [4].

Diabetes mellitus (DM) is a category of metabolic disorders marked by persistently high blood glucose levels. It is caused by a deficiency in insulin production or by the cells' incorrect response to the insulin produced. It is a major public health issue that affects people all around the world. Diabetes develops when the pancreas fails to produce enough insulin or when the human body is unable to use the hormone adequately. Diabetes diagnosis (diagnosis, etiopathophysiology, medication, etc.) necessitates the generation and processing of a large amount of data. Data mining techniques have proven to be useful and successful in evaluating undiscovered relationships or patterns that may exist in large amounts of data. AdaBoost, LogicBoost, RobustBoost, Nave Bayes, and Bagging are five machine learning approaches described in this paper for the analysis and prediction of diabetic patients. The proposed strategies are tested on a data set of diabetic Pima Indians. The results computed are found to be quite accurate, with bagging and AdaBoost approaches achieving classification accuracy of 81.77 percent and 79.69 percent, respectively. As a result, the presented strategies for predicting the DM are very attractive, effective, and efficient. [5].

Diabetes is caused by an excess of sugar that has collected in the blood. It is currently regarded as one of the world's most deadly diseases. This deadly disease affects people all around the world, whether they are aware of it or not. Diabetes can also lead to heart attacks, paralysis, renal failure, blindness, and other complications. For predicting and assessing diabetes, a number of computer-based detection methods have been built and defined. The traditional method of detecting diabetes people takes more time and money. However, with the advancement of machine learning, we now have the potential to design a solution to this complex problem. As a result, we created an architecture that can predict whether or not a patient has diabetes. The major goal of this investigation is to create a web application based on a sophisticated machine learning algorithm's greater prediction accuracy. We employed the Pima Indian benchmark dataset, which is capable of predicting the onset of diabetes based on diagnostics. Artificial Neural Network (ANN) demonstrates a considerable improvement in accuracy with an accuracy rate of 82.35 percent, which motivates us to design an Interactive Web Application for Diabetes Prediction. [9].

Big Data and the Cloud are two examples of new technologies that are helping to solve healthcare issues. Healthcare data is growing at an exponential rate these days, necessitating an efficient, effective, and timely solution to cut mortality rates. Diabetes is one of the most serious chronic health issues. If wrong medicine is administered, this condition may result in damage to the eyes, heart, kidneys, and nerves of diabetic patients, as well as death. The purpose of this study is to examine and compare numerous machine learning algorithms in order to determine the best forecasting algorithm based on several metrics such as accuracy, kappa, precision, recall, sensitivity, and specificity. Random Forest (RF), SVM, k-NN, CART, and LDA algorithms are used in a comprehensive investigation on a diabetic dataset. When compared to other algorithms, the obtained findings suggest that RF provides more accurate predictions. [10].

Data mining is a method for extracting useful information from massive amounts of information. Nowadays, data mining has emerged as an essential topic in the healthcare business, delivering accurate disease prediction and deeper analysis of medical data.
Different data mining approaches are being used by the authors to identify various ailments such as stoke, diabetes, cancer, hypothyroidism, and heart disease, among others. In section two of this work, the literature review of various data mining strategies was presented. The breast cancer and diabetes datasets from the UCI machine learning repository were used in this study. The five classification algorithms were classified on the WEKA Explorer and WEKA Experimenter interfaces in this paper. The WEKA tool is a useful categorization tool that was used in this study. On the WEKA interface, the Nave Bayes, SMO, REP Tree, J48, and MLP algorithms are used to classify breast cancer and diabetes datasets. The performance of these five algorithms was evaluated using training data testing mode on a breast cancer and diabetes dataset. After examining the results of each method, it was discovered that nave bayes provides 72.70 percent accuracy on the breast cancer dataset while SMO provides 76.80 percent accuracy on the diabetic dataset.[11].

Machine learning is the way to go when we have a large data collection on which we want to perform predictive analysis or pattern identification. Machine Learning (ML) is the fastest-growing field in computer science, and health informatics is a particularly difficult problem to solve.

Machine Learning aims to create algorithms that can learn and improve over time and be used to make predictions. Machine learning techniques are widely employed in a variety of industries, and the health care business, in particular, has benefited greatly from machine learning prediction approaches.

It provides a number of alerting and risk management decision-making tools aimed at enhancing patient safety and healthcare quality. The healthcare business faces hurdles in critical areas such as electronic record management, data integration, and computer aided diagnostics and disease predictions as a result of the need to cut healthcare costs and the shift toward individualised treatment. To address these issues, machine learning provides a variety of tools, methodologies, and frameworks. This paper presents a study of several Machine Learning prediction methodologies and tools in practise. A look at Machine Learning's applications in many fields is also covered, with a focus on its importance in the health-care industry.[12].

III. SYSTEMS ARCHITECTURE

The procedure begins with alteration of data. Four models for finding a prediction model will next be examined. Then each model's accuracy is calculated and compared to the best model. It might be useful for individuals and clinicians to detect ailments such as cancer and diabetes. They can assist patients determine their next move from the doctor's viewpoint, by detecting a patient's vulnerability to cancer or prevalence of diabetes. The study finally creates a web application.

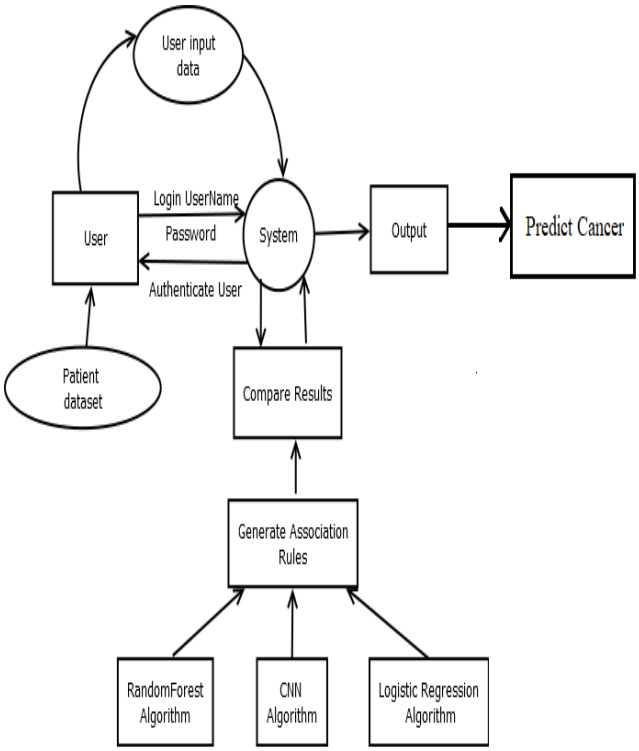

Figure No 3.1: System Architecture

The architecture above was separated into four segments. Here we provided an Advanced modified user query ELM algorithm and forecast that users have cancer/diabetics in the form of yes or no. The entire working module is separated into two modes: 1. Module of training & 2. Module testing Phase I, II, III & IV is included in the training module.We refer to the fact that the data set (data set details are specified on the data structure) is valid or invalid to preprocess each attribute. Check whether or not the characteristics of each attribute are legitimate. We have processed all fields and checked the value of each field in this processing. Here we have trained all data sets with various algorithms such as LR, DT, NB, RF, and ELM and calculated the accuracy of each algorithm.

IV EXPERIMENTAL RESULTS

The algorithms were used to increase the receiver's curve, validity, accuracy and logic of the data set. The suggested model consisting of both cluster and class methods improved prediction precision. The advantage is that the Pima Indian Diabetes Dataset and other datasets may include algorithms. Although the restriction is that it takes longer during the preprocessing stage. We have described that certain models focus on K-means by optimising the initialised cluster centre approach. But this enhanced model is based on diabetics and cancer predictions and matches. It ensures minimum time consumption and maximum data preservation. The key difficulties resolved are to improve prediction model accuracy and adapt the model to varied datasets. The algorithm of test and training mode is highly accurate in comparison with RF, LR, SVM, KNN & NB employing confusion matrix.

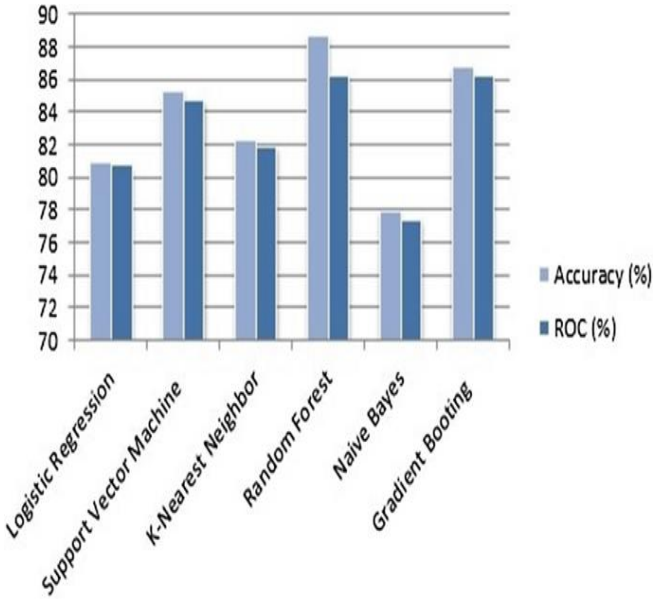

Figure No 4.1: Accuracy & ROC percent

V CONCLUSION

The proposed system some categorization techniques are explored in this suggested system. There are several optimization efforts to increase the performance of the algorithms. It might be useful for individuals and clinicians to detect ailments such as cancer and diabetes. They can assist patients determine their next move from the doctor's viewpoint, by detecting a patient's vulnerability to cancer or prevalence of diabetes. This is how physicians can assess the state of the patient and, if a person is highly at risk of cancer, physicians may decide on their medicine and a lifestyle for a healthier life.

REFERENCES

1. Md. Kamrul Hasan , Md. Ashraful Alam , Dola Das , Eklas Hossain,Mahmudul Hasan,"Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers",IEEE Access,2020.

2. Sara Laghmati,Bouchaib Cherradi,Amal Tmiri,Othmane Daanouni and Soufiane Hamida,"Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques ",IEEE Xplore,2020.

3. 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.

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

5.Vandana Rawat,Suryakant,"A Classification System for Diabetic Patients with Machine Learning Techniques",International Journal of Mathematical, Engineering and Management Sciences Vol. 4, No. 3, 729–744,2019.

6.Gopi Battineni, Getu Gamo Sagaro, Chintalapudi Nalini, Francesco Amenta and Seyed Khosrow Tayebati,"Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods",MDPI,December 2019.

7.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.

8.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.

9.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.

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

11.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.

12.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.

13.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.

14.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.

15.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.

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

17.Veena Vijayan V.,Anjali C. ,"Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach",IEEE Recent Advances in Intelligent Computational Systems (RAICS),10-12,December 2015.

18.C. Kalaiselvi, Dr. G .M. Nasira,"A New Approach for Diagnosis of Diabetes and Prediction of Cancer using ANFIS",World Congress on Computing and Communication Technologies,IEEE Xplore,2014.

19.Mr. Chintan Shah,Dr. Anjali G. Jivani,"Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction",4th ICCCNT,IEEE Xplore,July 4-6, 2013.