IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING-BASED CLINICAL DISEASES PREDICTION USING XGBOOST AND CNN

[1]Chilukala Mahender Reddy, [2]Yash Nawani

[1]Assistant Professor, [2] B. Tech Computer Engineering Student

[1,2]Computer Engineering Department,

[1,2]Marwadi University, Rajkot, India

Abstract: The great influence of science and technology in the medical industry resulted in the collection of large quantities of data. As a result of such a huge accumulation of data, physicians are finding it a lot more complex to identify or predict the presence of a disease in a patient at an early stage. Luckily, advancements in supervised machine learning algorithms have showcased a huge impact in the processing of the collected data and assist medical practitioners in fast and accurately anticipating the presence of high-risk diseases early on. This will not only help in the prevention of the spreading of the disease but can also save huge medical costs that patients might incur. This paper aims to evaluate multiple supervised machine learning models in disease detection and analyze them through performance benchmarks. The predominantly discussed supervised learning algorithms were and K-Nearest Neighbor (KNN), Decision Trees (DT), Naïve Bayes (NB), Random Forest, XGBoost, CNN, DNN. The XGBoost performed highly at the prediction of heart diseases and diabetes. XGBoost predicted the precision of diabetes and heart disease and Convolutional Neural Networks (CNN) predicted the precision of brain tumors, respectively.

Index Terms **– ML, Diabetes, Heart disease, Brain Tumor, XGBoost, CNN.**

I. INTRODUCTION

The quality of medical services available is at its peak today. With the huge augmentation of technology in medical sciences, doctors are capable of finding cures for almost any ailment. However, doctors have always needed to accurately identify, quantify and interpret the diseases in order to improve the patient's health. Diabetes, one of the leading causes of death in the world today is still not completely curable. Nearly 643 million people are predicted to be diagnosed with diabetes by the year 2030, which might right to around 783 million by 2045. Similarly, heart disease and brain tumors are the other two major ailments that are growing to be very common among people. All such diseases can be easily treated and cured at a comparatively cheap price if they are predicted accurately at an early stage. Here, Machine learning comes into the picture. It has brought together a wide range of methods to algorithmically identify the presence of such diseases in a patient. Machine learning is a subset of AI that deals with the study of algorithms and models that improve with data and experience. On a broader view, Machine Learning provides an efficient platform in the medical field to predict and solve various healthcare issues at a much faster rate.

The proposed system is helpful in predicting the possible risk of diabetes, heart disease, and a brain tumor in a patient. This is implemented using XGBoost and CNN models which provide highly accurate and reliable results. XGBoost and CNN have supervised learning algorithms that are mainly used for classification and regression problems. XGBoost works by creating multiple decision trees and finally deciding the majority of the trees and the final decision. CNN, on the other hand, works on the input images by assigning importance to various aspects of the image.

II. RELATED WORK

In the study by Shivani et al. (2020), using Random Forest Classifier they have divided the data into the set of decision trees taking the inputs of subsets. The majority of votes from each decision tree is used further for the prediction of Diabetes. They used a dataset that is formed by initially filling a form by the user/patient. Finally, they gained an accuracy of 80%.

In the study by Jackins et al. (2020), Random Forest Algorithm and Naïve Bayes Algorithm are used for two diseases mainly Diabetes and Heart Disease. Their diabetes dataset is from NIDDK containing data of 769 patients of which all are females of at least 21 years old and from Pima Indian Heritage. Moreover, their heart disease dataset is from the Framingham heart study which

has multiple factors in it. They achieved 74.46%, and 82.35% for diabetes and heart disease using Bayes Classification. Similarly, Random Forest achieved 74.03%, and 83.85% for diabetes and heart disease respectively.

In study by Nai-arun & Moungmai, (2015), they collected initial data from 26 Primary Care Units (PSU) located in Sawanpracharak Regional Hospital in the year 2012-2013. The dataset contains data of 30,122 people of which 19,145 are normal people and the rest 10,977 people are diagnosed with diabetes. Decision Tree (DT) provided them with an accuracy of 85.090%, Artificial Neural Network (ANN) reached 84.532% accuracy, Logistic Regression (LR) achieved 82.308% accuracy, and Naïve Bayes (NB) marked an accuracy of 81.010%. Later they used Bagging (BG) with the above four models (BG+DT), (BG+ANN), (BG+LR), (BG+NB), and achieved accuracies of 85.333%, 85.324%, 82.318%, 80.960% respectively. Later they again used the initial four models with Boosting (BT). Hence forming (BT+DT), (BT+ANN), (BT+LR), (BT+NB). Finally reaching accuracies of 84.098%, 84.815%, 82.312%, 81.019% respectively. The Random Forest model achieved an accuracy of 85.588% which was the highest of them all.

In the study by J. Seetha & S. Selvakumar Raja (2018), they have performed brain tumor classification using three different models that are SVM, DNN, and CNN. The dataset used here is Benchmark (BRATS) 2015 testing dataset. It contains a total of 274 brain scan images of which 220 are High-Grade Gliomas (HGG) and the rest 54 are Low-Grade Gliomas (LGG). Out of the three models SVM achieved the lowest accuracy at approximately 83%. DNN and the proposed CNN models performed well with approximately the same accuracies. DNN marked an accuracy of approx. 97% and CNN achieved 97.5% accuracy.

In the study by Hossain et al. (2019), Convolutional Neural Network is used for the prediction of a brain tumor of a patient. They have segmented image processing techniques in order to process a brain scan MRI image to predict the brain tumor in patients. The dataset used by them is the BRATS dataset which consists of 217 images in total out of which 187 are tumor and 30 are non-tumor brain images. Finally, for the 70:30 splitting ratio, they achieved an accuracy of 92.98%. Whereas for the 80:20 ratio 97.87% accuracy was achieved.

In the study by Febrianto et al. (2020), two different CNN models were implemented one with more additional layers than the previous one. The dataset they used is provided by Kaggle the dataset to a total of 2065 images out of which 1085 had a tumor and the rest 980 were tumor-less brain images. The first CNN model consisted of 6 layers that are one Conv2D layer, three MaxPooling2D layers, flatten layer and two dense layers. On the other hand, the second CNN model contains 9 different layers. Upon dividing the dataset in 70:30 splitting the ratio in terms of training to testing purpose. In the testing, they used half for testing, and the rest is used for data validation. The first CNN model reached an accuracy of 85%. Whereas the second CNN model showed an accuracy of 93%, which was considerably higher than the first CNN model.

In the study by Kaur et al., (2019), they have used multiple models that are KNN, Linear-SVM, Decision Tree, Random Forest, and MLP in order to predict multiple diseases two of them are diabetes and heart disease. The dataset they used for diabetes and heart disease consists, of class 2 of 768 samples and class 5 of 303 samples respectively. They have used the mobile device as an IoT agent in order to collect and improve performance. Finally, for diabetes they reached K-NN: 74.67%, Linear-SVM: 79.87%, DT: 75.97%, MLP: 78.57%, RF: 81.16% for heart disease they achieved accuracies of K-NN: 55.73%, SVM: 57.37%, DT: 52.45%, MLP: 47.54%, RF: 55.73% respectively.

In the study by Pal & Parija (2021), heart disease is predicted using a Random Forest algorithm. The dataset used by them is from Kaggle which contains 303 different samples and 14 different attributes. Later various parameters were used in order to evaluate the performance of the model which are accuracy, sensitivity, and specificity in percentages. They achieved an accuracy of 86.9%, a sensitivity value of 90.6%, and a specificity value of 82.7%.

In the study by C et al. (2019), they have heart disease prediction using two different models that are Random Forest and Logistic Regression. The dataset they have used is from the UCI repository. It contains 5 target values (0,1,2,3,4), 0 specifies no heart disease. Upon training the models they reached an accuracy of 98% and 80% for Random Forest and Logistic Regression respectively.

In the study by Chari et al. (2019), they have performed classification of diabetes using various different models that are Decision Tree, Bagging with Decision Tree, Radom Forest, and Random Forest with Feature Selection. Decision Tree achieved an accuracy of 75.2%, Bagging with Decision Tree showed an accuracy of 81.3%, Random Forest displayed 85.6% accuracy and finally, Random Forest with Feature Selection showed the highest accuracy of 92.02%.

III. PROPOSED METHOD

We have proposed a system that helps in the prediction of the risk of multiple diseases at a single platform. There are various existing researches where the trained models are used to predict only of the disease. In the proposed system three different models are collectively bought together at a single place in order to predict multiple diseases. Flask module of python is then used to construct a website-based platform where users can input the required variables and get the prediction results of the simultaneous disease.

Prediction of diabetes, and heart disease is performed by XGBoost which works on a gradient boosting framework, and brain tumor prediction is implemented using the CNN model. The collected dataset is initially pre-processed: through which all the null values and errors are detected and removed so that the dataset is ready to use.

Right after the pre-processing step pair-wise correlation of columns in the dataset is computed. This is done through the implementation of a heat map which provides a broader and clearer view of the dataset as well as in finding the correlation.

Next, the dataset is split into training and testing groups in the ratio of 80:20. This greatly increases the accuracy of the models and augments the overall performance and reliability. Then, XGBoost and CNN algorithms are used to train the final models. Finally, these models are stored, as a result, repeatedly training of dataset is minimized. Hence, it saves a lot of time.

3.1 Dataset Collection

Dataset for diabetes is collected from Mendeley data (Rashid, 2020), which consists of over 1000 instances with three different classes (Diabetic, Non-Diabetic, and Predicted- Diabetic). Various patients' files were collected and data was input into the diabetes dataset.

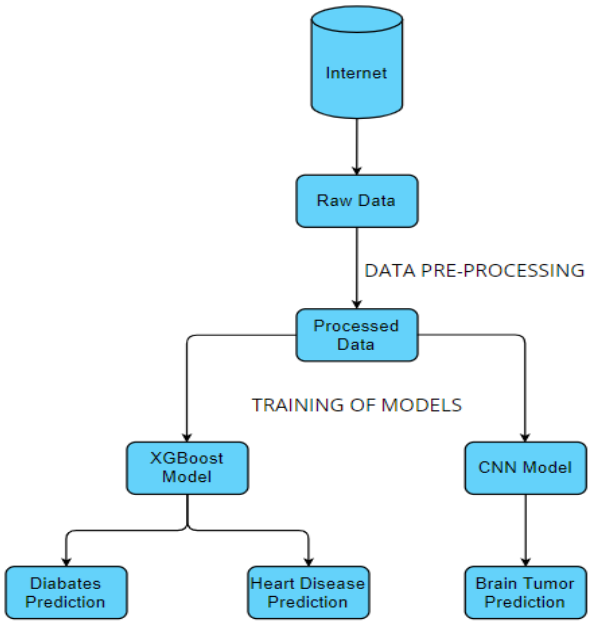

Figure III.1: General Methodology

Multiple parameters are used in the diabetes dataset the No. of Patient, Sugar Level Blood, Age, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides (TG) and HDL Cholesterol, HBA1C, Class.

Dataset for heart disease is collected from Kaggle. It contains over 304 instances and contains 14 different attributes. The target attribute shows the possibility of heart disease where 0 refers to no disease and 1 refers to the presence of heart disease.

Dataset for brain tumor is four different groups that are glioma tumor, meningioma tumor, pituitary tumor, and no tumor. Each sub-dataset contains approximately 800 images and hence an overall collection of 3000 images approximately. This dataset was collectively formed from the dataset available on Kaggle and Google images.

3.2 Training of models

Step 1: Initially input variables (Training Dataset is preloaded in the system). Then target variable and objective are set.

Objective function = training loss + regularization.

Step 2: Then, the number of iterations is set i.e., the number of decision trees that are needed to be added is set.

Step 3: Finally Early fitting of the model is instantiated to avoid the problem of over-fitting the model.

XGBoost is a supervised learning algorithm that works on the gradient boosting framework. It is a highly optimized model that runs through parallel processing, tree-pruning, and regularization in order to avoid situations such as over-fitting of the model. XGBoost is used in creating diabetes prediction and heart disease models. These models have shown much higher accuracies when compared to any other model trained using different machine learning algorithms.

Convolutional Neural Network is a feed-forward artificial network that utilizes spatial correlations which present in the existing input data. A five-layered CNN is used for brain tumor detection. The proposed model contains 7 stages including some hidden layers which highly improve the accuracy and performance of the trained model. Hence, providing us with the most prominent results.

IV. EXPERIMENTAL WORK

4.1 Software and Setup Tools

Google Collab is used for code implementation in python language. Models are trained using TensorFlow 2.8.0. Other packages that are needed for proper implementation of the system are Pandas, NumPy, seaborn, cv2, and Sklearn.

4.2 Implementation

Diabetes and heart prediction:

Pre-processing of data is done where the missing values are detected and changed to zero and other errors from the data are removed. Hence the data is ready to use further.

Then the seaborn package is used to plot a heatmap of the dataset which computes the pairwise correlation between the columns of the dataset excluding the null values.

Then XGBoost algorithm is used to train the final model. The estimators are set in order to decide the number of decision trees needed. Here estimators are set to 100.

After finding the accuracy it shows, that the final model is stored in. pkl format in the device.

Brain tumor prediction:

For brain tumor prediction, a Convolutional Neural Network algorithm is used.

The image augmentation is implemented that artificially expands the size of an image in the training dataset by creating a new modified version of the images. ImageDataGenerator function present in Keras is used to implement this method.

A deep convolutional neural network model takes a huge amount of time something in days or weeks to train epochs on large datasets. Hence, we have used the pre-trained model 'EfficientNetB0'. It will use weights from the ImageNet dataset.

The top layer is set to false so that the outer layer of the pre-trained model is not included. This way we can include our own top/outer layer in the model.

Few other parameters are set that are necessary like dropout rate = 0.5, dense activation = 'softmax', and using some call-back function like tensorboard, ModelCheckPoint, ReduceLROnPlateau.

Then, we compile our model using the 'Adam' optimizer and in a batch size of 32.
Finally, we save our model in .h5 extension format.

V. EXPERIMENTAL RESULTS

A multiple Disease prediction system is developed using XGBoost and Convolutional Neural Network. Collectively, Both the models have provided high accuracies. The figure shows the accuracies showed by both the models:

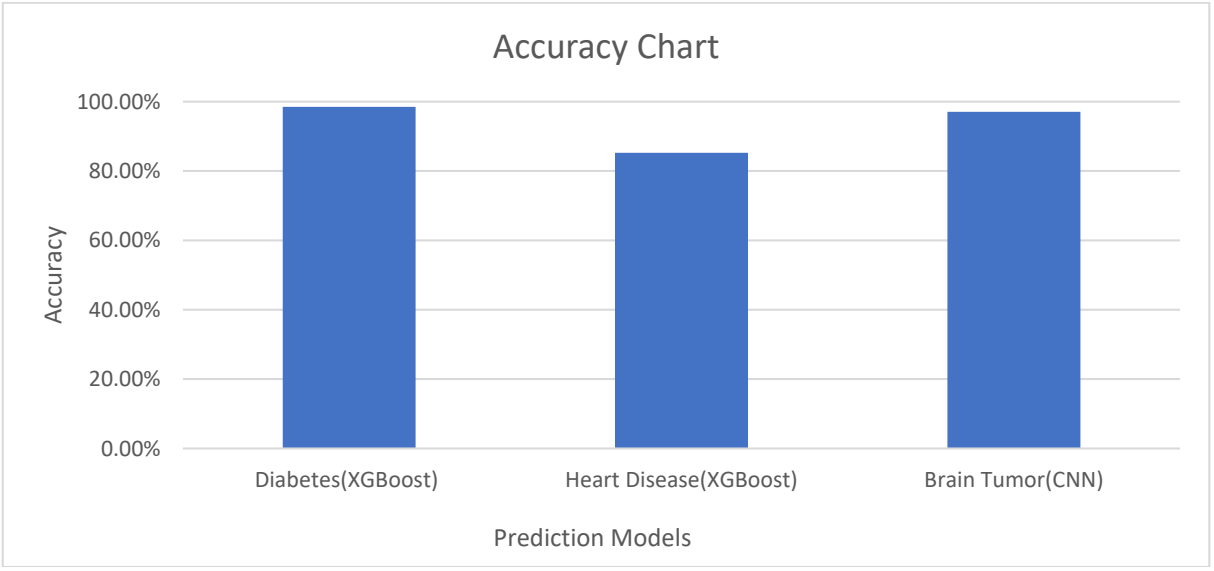

Figure V.1: Accuracy Chart

The diabetes prediction model (XGBoost) achieved an accuracy of 98.5%. The Heart Disease prediction model (XGBoost) achieved an accuracy of 85.25% and the Brain tumor prediction model (CNN) achieved an accuracy of 97.06%. XGBoost provided higher accuracies when compared with other models like Random Forest, Linear Regression, K-NN, SVM, and Linear-SVM.

VI. CONCLUSION

In this paper, a multiple disease prediction system is proposed which predicts the possibility of risk of multiple diseases like diabetes, heart disease, and a brain tumor in a patient. Three different datasets are collected for each disease out of which two are available on Kaggle and one is available on Mendeley Data.

Two different machine learning algorithms are used to train models for accurate prediction. XGBoost algorithm is used for diabetes prediction and heart disease prediction and the CNN algorithm is used for brain tumor prediction.

Various existing researches have implemented the prediction of such diseases using different models like K-NN, SVM, Random Forest, Linear-SVM, and Naïve Bayer for diabetes and heart disease prediction. Many of them achieved an accuracy ranging between 75-82%. Whereas the proposed diabetes prediction model has an accuracy of 98.5%. Similarly, on studying the existing research, the proposed XGBoost model for the heart disease prediction model has 85.25% accuracy. The existing research where DNN models and BRATS datasets have been used for brain tumor predictions achieved an accuracy of approximately 90-92%. However, the proposed CNN Brain Tumor prediction model has 97.6% accuracy.

REFERENCES

1) C, A. G., S, A. M., Deepthi N, Dhanushree V, & Rummana Firdaus. (2019). Heart Disease Diagnosis Using Machine Learning. *International Journal of Engineering Research & Technology*, *7*(10). https://www.ijert.org/heart-disease-diagnosis-using-machine-learning

2) Chari, K., babu, M., & Kodati, S. (2019). Classification of Diabetes using Random Forest with Feature Selection Algorithm. International Journal Of Innovative Technology And Exploring Engineering, 9(1), 1295-1300. https://doi.org/10.35940/ijitee.l3595.119119

3) Febrianto, D. C., Soesanti, I., & Nugroho, H. A. (2020). Convolutional Neural Network for Brain Tumor Detection. *IOP Conference Series: Materials Science and Engineering*, *771*(1), 012031. https://doi.org/10.1088/1757-899x/771/1/012031

4) Hossain, T., Shishir, F. S., Ashraf, M., Al Nasim, M. A., & Muhammad Shah, F. (2019). Brain Tumor Detection Using Convolutional Neural Network. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. https://doi.org/10.1109/icasert.2019.8934561

5) J. Seetha, & S. Selvakumar Raja. (2018). Brain Tumor Classification Using Convolutional Neural Networks. *Biomedical and Pharmacology Journal*, *11*(3), 1457–1461. https://biomedpharmajournal.org/vol11no3/brain-tumor-classification-using-convolutional-neural-networks/

6) Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2020). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x

7) Kaur, P., Kumar, R., & Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, *78*(14), 19905–19916. https://doi.org/10.1007/s11042-019-7327-8

8) Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science*, *69*, 132–142. https://doi.org/10.1016/j.procs.2015.10.014

9) Pal, M., & Parija, S. (2021). Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*, *1817*(1), 012009. https://doi.org/10.1088/1742-6596/1817/1/012009

10) Rashid, A. (2022). Diabetes Dataset. Mendeley Data. Retrieved 11 May 2022, from https://data.mendeley.com/datasets/wj9rwkp9c2/1

11) Singh, S., Bait, S., Rathod, J., & Pathak, P. (2022). DIABETES PREDICTION USING RANDOM FOREST CLASSIFIER AND INTELLIGENT DIETICIAN. International Research Journal Of Engineering And Technology (IRJET), 07(01), 2155-2157.