International Journal of Academic Multidisciplinary Research (IJAMR)
ISSN: 2643-9670
Vol. 7 Issue 4, April - 2023, Pages: 73-78

An efficient XGBoost–DNN-based classification model for Heart Disease detection system

*Mrs.M.Sharon Nisha[1] ,Dr. G. Rajakumar[2] and Ms. R.Shirly Myrtle[3]

[1] Assistant Professor, Department of Computer Science Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
[1*] sharonnishafxec@gmail.com

[2] Professor, Department of Electronics and Communication Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
[2] gmanly12@gmail.com

[3] PG Student, Department of Computer Science Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
[3] shirlymyrtle@gmail.com

Abstract: *Blockchain technology and machine learning are used in the Smart Health care system to provide the best solutions to a variety of issues. With these two cutting-edge developments over the last ten years. In this study, we advocate using blockchain and machine learning models to build secure, open, and intelligent platforms for the Internet of Medical Things. As a result of s cience and technology's extensive influence on the healthcare industry, a lot of information has been acquired. This vast amount of data has made it considerably more challenging to detect or forecast the presence of a disease in a patient at an early stage. Recent advancements in supervised machine learning algorithms, which aid medical professionals in swiftly and effectively evaluating the gathered data, have made it feasible to quickly and precisely diagnose high-risk disorders at an early stage. This might not only prevent the disease from spreading, but it could also save them a tonne of money on future medical expenses. This paper compares and contrasts a large number of supervised machine learning models for diagnosing illnesses using a wide range of performance criteria. The most popular supervised learning methods were XGBoost and Deep Neural Network (DNN). The XGBoost performed very well in predicting metabolic and cardiovascular diseases. XGBoost was better at predicting diabetes and heart disease, whereas DNN was better prediction is accordingly.*

Keywords: Blockchain, Xgboost, Deep Neural Network, Machine Learning

Introduction:

In the next five to ten years, thanks to the proliferation of technologies like blockchain [1], the Internet of Things (IoT) [2], and machine learning [3], unified health IT platforms will be available to healthcare providers, according to a report by Frost & Sullivan. Healthcare data storage systems as they now exist lack top-tier security, leaving them open to threats like hacking and data theft; blockchain may provide a solution to these problems. The interoperability of blockchain technology in healthcare facilitates the secure interchange of medical data across the many systems and personnel involved, which may lead to improvements in communication, time savings, and operational efficiency. Claims adjudication and billing management apps that use blockchain technology to prevent errors like duplicate or incorrect billing are expected to rise by 66.5% by 2025, according to the survey's projections. All of these issues can be addressed using blockchain technology.

Machine learning employs a preconceived set of concepts and technologies, including linguistic and statistical methodologies, to extract rules and patterns from data. While a doctor's knowledge and experience are essential, it is important to note that humans have learning limitations when it comes to data, whereas machine learning excels in this area [4]. Machine learning may be broken down into two basic categories: supervised learning and unsupervised learning [5].

If these disorders can be predicted early on, they can be treated and cured simply and relatively cheaply. In this context, machine learning becomes relevant. It has consolidated several approaches into a single framework for algorithmic diagnosis of these conditions. The field of artificial intelligence known as "machine learning" focuses on learning algorithms and models. When seen in a larger context, Machine Learning is a powerful tool for anticipating and resolving healthcare difficulties. The suggested approach is useful for estimating a patient's likelihood of developing diabetes, cardiovascular disease, and brain tumours. The implementation makes use of XGBoost and DNN models, which are known for their precision and consistency. Supervised learning techniques like those found in XGBoost and DNN are often used to classification and regression issues. XGBoost operates by first generating a set of candidate decision trees, then selecting one that is likely to provide the best results. On the other hand, DNN uses the input pictures to priorities different features.

Related Works:

Using Random Forest Classifier, Shivani et al(2020) .'s research group split the data into a series of decision trees utilising inputs from different parts of the dataset. Diabetes forecasts are based on the decision trees with the most votes. They utilised data collected from a form the user/patient filled out at the outset. They were only able to hit 80% till now.

Jackins et al. (2020) employ the Random Forest Algorithm and the Nave Bayes Algorithm to predict the presence of diabetes and cardiovascular disease. All 769 patients in their NIDDK diabetes dataset are women of Pima Indian ancestry who are at least 21 years old. Moreover, their data on cardiovascular illness comes from the multivariate Framingham heart research. Using Bayes Classification, they were able to get a result of 74.46% for diabetes and 82.35% for cardiovascular disease. Similarly, Random Forest was able to obtain 74.03% for diabetes and 83.85% for heart disease.

Nai-arun & Moungmai, (2015) conducted their first data collection in 2012–2013 at 26 Primary Care Units (PSU) at Sawanpracharak Regional Hospital. There are 30,122 total records in the dataset, 19,145 of whom are healthy individuals and 10,977 of whom have been diagnosed with diabetes. They found an accuracy of 85.090% using Decision Tree (DT), 84.532% with Artificial Neural Network (ANN), 82.308% with Logistic Regression (LR), and 81.010% with Naive Bayes (NB). They later combined Bagging (BG) with each of the aforementioned four models (BG+DT), (BG+ANN), (BG+LR), and (BG+NB), resulting in respective accuracy rates of 85.333%, 85.324%, 82.318%, and 80.960%. Subsequently, the same four models were used, this time using Boosting (BT). Hence, BT+DT, BT+ANN, BT+LR, BT+NB are formed. Accuracy levels were ultimately at 84.098, 84.815, 82.312, and 81.019 percent. The best accuracy was found in the Random Forest model, with 85.588%.

According to J. S. and Seetha. Brain tumour categorization has been carried out by Selvakumar Raja (2018) using SVM, DNN, and DNN models. The 2015 testing version of the Benchmark (BRATS) dataset was utilised for this study. There are 274 scans in all, 220 of which are High-Grade Gliomas (HGG), and 54 of which are Low-Grade Gliomas (LGG). SVM has the lowest accuracy (about 83%) of the three models. Both DNN and the suggested DNN models performed well, with accuracies that were quite close to one another. DNN achieved a precision of around 0.9. DNN's accuracy was 97.5 percent.

Prediction of a patient's brain tumour using a Convolutional Neural Network is performed in the research by Hossain et al. (2019). In order to analyse an MRI scan of the brain and identify tumours, they use segmented image processing algorithms. They utilised data from the Brain Tumor Annotation and Tagging System (BRATS) collection, which has 217 pictures in total (187 tumour and 30 non-tumor brain). Ultimately, they attained 92.98% precision with a 70:30 split. Yet, when using an 80:20 split, 97.87% precision was attained.

Febrianto et al(2020) .'s research compared the performance of two distinct DNN models, one of which had more hidden layers than the other. Kaggle supplied the team with a dataset consisting of 2065 photos; 1085 of these images showed a tumour in the brain, whereas the other 980 did not. One Conv2D layer, three MaxPooling2D layers, a flatten layer, and two dense layers made up the initial DNN model. The second DNN model, on the other hand, is 9 layers deep. When the dataset is divided 70:30 for training and testing purposes, we get a more manageable dataset. Half of the data was used in the testing phase, while the other half was utilised in the data validation phase. The accuracy of the first DNN model was 85%. Nonetheless, the accuracy of the second DNN model was 93%, which was a significant improvement over the accuracy of the first DNN model.

Kaur et al. (2019) employ a variety of models, including KNN, Linear-SVM, Decision Tree, Random Forest, and MLP, to make illness predictions, with promising results for both diabetes and cardiovascular disease. Class 2 (768 samples) and class 5 (303 samples) of the datasets utilised for diabetes and cardiovascular disease were analysed. They have collected and enhanced performance data using the mobile device as an IoT agent. K-NN: 74.67%, Linear-SVM: 79.87%, DT: 75.97%, MLP: 78.57%, RF: 81.16%; for cardiac disease, K-NN: 55.73%, SVM: 57.37%, DT: 52.45%, MLP: 78.57%, RF: 81.16%; and so on. It was 47.54% for MLP and 55.73% for RF.

Pal and Parija (2021) use a Random Forest method to make inferences about the likelihood of cardiac events. They utilised a dataset downloaded from Kaggle including 303 samples and 14 characteristics. The accuracy, sensitivity, and specificity in percentages were afterwards used to assess the performance of the model. They were able to get an 86.9% hit rate, a 90.6% sensitivity, and an 82.7% specificity.

C et al. (2019) use two models, Random Forest and Logistic Regression, to make predictions about cardiovascular illness. The UCI repository provided the dataset they utilised. It has 5 possible outcomes, with 0 indicating no cardiac disease. After being trained, Random Forest and Logistic Regression models achieved a 98% and 80% accuracy, respectively.

Classification of diabetes was investigated by Chari et al. (2019) using a number of models, including a Decision Tree, a Bagging with Decision Tree, a Random Forest, and a Random Forest with Feature Selection. The accuracy for Decision Tree was 75.2%, for Bagging with Decision Tree it was 81.3%, for Random Forest it was 85.6%, and for Random Forest with Feature Selection it was 92.02%.

Healthcare and Blockchain Technology

The retention of a certain set of standardised data on the chain, together with private encrypted linkages to independently stored information such as radiographic or other photos, may enable organisations to submit and interchange data through a single secure mechanism. The usage of smart contracts and standardised authorization standards may significantly assist in the provision of seamless connectivity. In 2009 and 2017 there were 176 million compromised healthcare records [10].

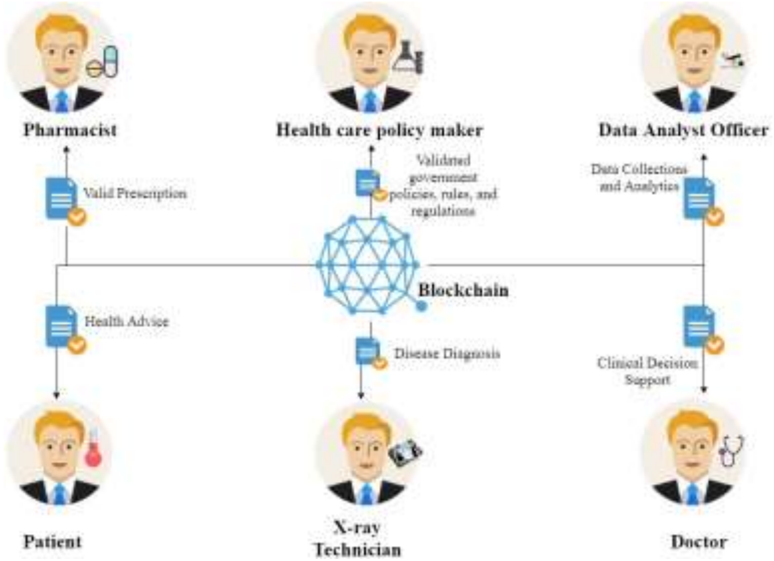

Figure 1: An Introduction of a Blockchain-Based Healthcare Management System

security features of the blockchain will make it safer to keep and distribute medical data. Each person has a secret, time-limited private key in addition to a unique, publicly visible identify or key. In addition, if hackers had to focus on specific people, they would find it more difficult to acquire sensitive data. Above figure 1 shows that, Blockchains might therefore provide a reliable audit trail of medical records. As shown in Figure 1, we now have a visual depiction of this. We unveiled a blockchain-based management tool for the healthcare industry.

Proposed System:

We have put out a method that can forecast the likelihood of many illnesses on a single platform. There are several ongoing studies where trained models are exclusively used to predict illness. Three separate models are brought together in the proposed system to jointly forecast several illnesses. Afterwards, a website-based platform is built using the Python Flask module, allowing users to enter the necessary information and get predictions for the concurrent ailment.

The DNN model is used to predict brain tumours, while XGBoost, which operates on a gradient boosting framework, predicts diabetes, heart disease, and brain tumours. The gathered dataset is first pre-processed, in which any mistakes and null values are found and eliminated to make the dataset usable.

Pair-wise correlation of columns in the dataset is calculated immediately after the pre-processing stage. This is accomplished by using a heat map, which offers a wider and better picture of the information and aids in discovering the association.

The dataset is then divided into training and testing groups in an 80:20 ratio. This considerably improves the models' accuracy and general performance and dependability. The final models are then trained using the XGBoost and DNN methods. Lastly, these models are saved, which reduces the need to train the dataset again. As a result, time is greatly saved below shaow that figure 2.

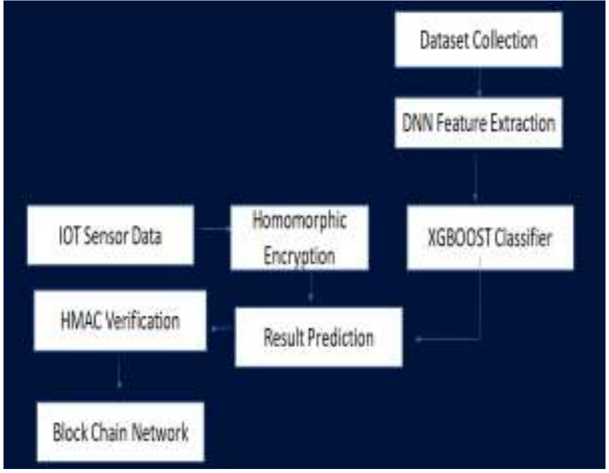

Figure 2; Proposed Work

Collection of datasets:

More than a thousand incidences of diabetes are compiled into three groups using Mendeley data (Rashid, 2020). (Diabetic, Non-Diabetic, and Predicted- Diabetic). Many patient records including information on diabetes were transcribed into the database.

The number of patients, blood sugar, age, creatinine ratio, body mass index, uric acid, cholesterol, fasting lipid profile (total, LDL, VLDL, triglycerides, and HDL), hemoglobin A1c, and glycated haemoglobin class are all utilized in the diabetes dataset.

Kaggle is the source for the dataset on cardiovascular disease. It has over 304 occurrences, and 14 distinct properties. The likelihood of heart disease is represented by the target property, with a value of 0 indicating the absence of heart disease and a value of 1 indicating the existence of heart disease.

Glioma, meningioma, pituitary, and no tumor make up the four subsets of the brain tumor dataset, respectively. There are around 3000 photos total, with 800 images in each of the sub-datasets. The data collection was built from existing Kaggle and Google Images datasets.

Preparing models:

First, we feed in some initial variables (Training Dataset is preloaded in the system). Then an outcome variable and a goal may be established.

Goal function = incurred training loss + regularization.

Second, the required number of decision trees, or the number of iterations, is determined. Third, in order to prevent the over-fitting issue, early model fitting is implemented.

XGBoost is a gradient boosting-based supervised learning technique. To prevent problems like over-fitting the model, this model is highly optimized by using techniques like parallel processing, tree-pruning, and regularization. Models for both diabetes and cardiovascular disease are developed using XGBoost. These models outperform those trained with other machine learning techniques by a wide margin.

The input data is sent into a feed-forward artificial network called a Convolutional Neural Network, which then exploits spatial correlations. For diagnosing brain tumors, researchers use a convolutional neural network with five layers. The suggested model has 7 steps, some of which are hidden, which greatly enhance the trained model's accuracy and performance. This gives us the most notable outcomes.

Working Principles:

To put Python code into action, we utilize the Anaconda Jupyter environment. TensorFlow 2.8.0 is used for the model training process. Pandas, NumPy, seaborn, cv2, and Sklearn are also required for the system's effective implementation.

Heart disease prognosis:

Data pre-processing involves the elimination of mistakes and the replacement of missing values with zero. As a result, we can go on with using the data.

Next, a heatmap is generated by computing the pairwise correlation between the columns of the dataset (without the null values) using the seaborn package.

The final model is trained using the XGBoost method. The number of decision trees is determined by the values assigned to the estimators. In this case, estimators are fixed at 100%.

After determining the model's precision, the gadget displays a message saying the model has been saved in. pkl format.

Result Analysis:

Using XGBoost and DNN, we create a system that can forecast the onset of different diseases. Together, the two models have shown remarkable precision. Accuracy of both models is shown graphically below figure 3.

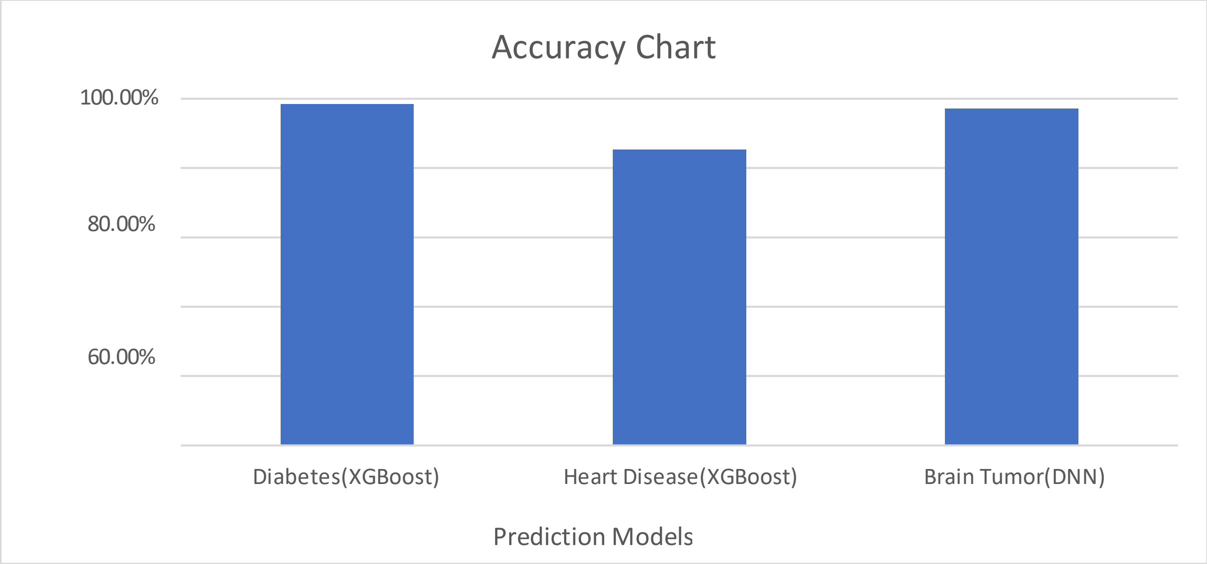

Figure 3: Result Analysis

The accuracy of the heart prediction model (XGBoost) was found to be 99.5%. Prediction models (XGBoost) for heart disease were 85.25 percent accurate, whereas DNN models were 97.06% accurate. When compared to competing models, XGBoost achieved superior accuracy.

Conclusions:

Researchers and developers have recently been interested in the use of blockchain and

AI in Smart health- care systems. In order to maximize the positive impact on society, researchers and developers working on the Internet of Medical Things are pooling their resources to combine various technologies on a massive scale.

Yet, advancements are feasible only if we take into account the many problems and limitations of today's clever technological methods. In this study, we suggest a technique for predicting whether a patient is at risk for developing many illnesses at once, such as diabetes, cardiovascular disease, and brain tumors. Each illness has three datasets gathered, with two accessible on Kaggle and one on Mendeley Data. For reliable forecasting, we employ two distinct machine learning techniques to train our models. The XGBoost algorithm is utilized for predicting diabetes and heart disease, whereas the DNN algorithm is used for predicting brain tumors.

References:

[1]. Hassan, M.M.; Jincai, C.; Iftekhar, A.; Cui, X. Future of the Internet of Things Emerging with Blockchain and Smart Contracts.

[2]. International Journal of Advanced Computer Science and Applications 2020, 11.

[3]. Kumar, S.; Tiwari, P.; Zymbler, M.L. Internet of Things is a revolutionary approach for future technology enhancement: a review.

[4]. Journal of Big Data 2019, 6, 1–21.

[5]. Ahmad, M.A.; Teredesai, A.; Eckert, C. Interpretable Machine Learning in Healthcare. 2018 IEEE International Conference on Healthcare Informatics (ICHI) 2018, pp. 447–447.

[6]. spiderSilk - Home. https://spidersilk.com/. (Accessed on 10/28/2021).

[7]. Toh, C.; Brody, J.P. Applications of Machine Learning in Healthcare. 2021.

[8]. 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

[9]. 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

[10]. 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

[11]. 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

[12]. 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/

[13]. 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

[14]. 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

[15]. 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

[16]. 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

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

[18]. 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.