Prediction of Stroke Disease Using Deep CNN Based Approach

Md. Ashrafuzzaman[1], Suman Saha[2], and Kamruddin Nur[3]
[1] Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[2] Department of Information and Communication Technology, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh, Gazipur, Bangladesh
[3] Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh
Email: assrafuzzamans@gmail.com, suman@ict.bdu.ac.bd, kamruddin@aiub.edu

***Abstract*—Stroke is a medical condition that occurs when there is any blockage or bleeding of the blood vessels either interrupts or reduces the supply of blood to the brain resulting in brain cells starting to die. It causes the disability of multiple organs or unexpected death. The time of cure in stroke patients relies on symptoms and injury of organs. The stroke is avoided in up to 80 percent of cases if the patients identify and relieve the dangers in due time. With the advancement of machine learning in medical imaging, the early recognition of stroke is very much possible that plays a vital role in diagnosis and getting read of this life-taking disease. Considering the above case, in this paper, we have proposed a Convolutional Neural Network (CNN) model as a solution that predicts the probability of stroke of a patient in an early stage to achieve the highest efficiency and accuracy. The model is an improvised variant of a multi-layer perceptron and it comprises info, a yield layer, and many secret layers. The data set used in the prediction model is the health care data set which has eleven features and only one target class as the outcome. Therefore, we have also applied some feature selection methods for extracting the most contributed features in the classification. The model accuracy is compared with other machine learning models and found the model is better than others with an accuracy of 95.5 percent.**

***Index Terms*—stroke prediction, machine learning approach, data mining, neural network, CNN**

I. Introduction

Stroke is one of the most life-threatening diseases for people in worldwide [1]. It injures the brain like a "heart attack" and has been ranked the second leading cause of death in the United States and developing countries like Bangladesh. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11 percent of total deaths [2]. The percentage of increasing stroke disease is increasing every day due to laziness, consumption of medicines, and bad dietary practices [3]. A stroke happens when there is a loss of blood flow to part of the brain. Man's brain cells cannot get the oxygen and nutrients they need from blood, and they start to die within a few minutes. This can cause lasting brain damage, long-term disability, or even death. Stroke can be separated into three types: ischemic stroke, hemorrhagic stroke, and transient ischemic attack. According to the World Health Organization (WHO), 87 percent of the stroke patient died on ischemic stroke and the rest of the patients died on hemorrhagic stroke and transient ischemic attack [4]. Therefore, early prediction is required for detecting and curing the stroke [5].

Artificial Intelligence is now widely used in the healthcare and medical industry as well. With the rapid development of deep learning-based machine learning algorithms in recent years, the application of AI in diagnosis, risk stratification, and therapeutic decision-making has become ever- more widespread. In the medical industry, the occurrence of a stroke can be easily predicted using Machine Learning algorithms [6] [7]. Although ML models have still not been widely adopted in clinical practice, the utility of risk scores for the prediction of stroke in the contemporary world's population is not up to mark. On the other hand, deep learning-based techniques have achieved popularity for extracting features automatically from raw data with little or no preprocessing in the field of medical applications [8]. Since the volume of data is increasing in healthcare systems continuously, such an approach has been used successfully to predict various other health applications, such as heart failure and osteoporosis, etc., and automatic recognition of diabetic retinopathy. Therefore, this technique outperforms other conventional ML algorithms, such as Logistic Regression (LR) and Support Vector Machine (SVM), etc. on prediction [9].

Convolution Neural Networks (CNNs) are very promising for medical applications and have achieved satisfactory results in disease prediction. The model employs a variation of multilayer perceptron's and includes one or more convolutional layers that can be either fully interconnected or pooled. It provides higher accuracy than other machine learning algorithms by automatically identifying the valued features without any human supervision. Therefore, in this article, we have built an intelligent CNN model on HealthCare Problem:

Manuscript received January 21, 2022; revised June 30, 2022; accepted July 18, 2022.

Prediction Stroke Patients dataset collected from Kaggle for early prediction [10].

The key contributions of this work are summarized below.

- Building an intelligent 1D-CNN model which can predict stroke on benchmark dataset.
- Identifying the best features for the model by Performingdifferent feature selection algorithms.
- Evaluating the performance of the model with different Machine Learning models through a rigorous simulation and presenting the comparative result based on different performance parameters.

The rest of this paper has the following sections. Section II gives a brief description of some previous research works related to Stroke prediction. Section III briefly explains the proposed methodology including the information of the data set, features of attributes, the training and validation sample, and classical machine learning approaches. In Section IV, we have presented our experimental results. Section V highlights the important observation of analysis of stroke with respect to different attributes. Finally, Section VI concludes the paper.

II. Related Work

The use of Machine learning and Deep Learning in the health care setting is becoming a powerful tool for the diagnosis or prediction of intricacies and patient outcomes in several diseases. Stroke is the leading reason for the long-term disabilities of elderly people which creates multiple social or monetary difficulties. However, machine learning techniques have been employed to aid in the diagnosis and findings of treatment for stroke patients from early diagnosis to patient outcome after treatment. The most recent research analyzing the applications of machine learning in stroke is summarized below.

Shoily *et al.* [11] investigated on person's physical state and medical report data to identify the possibility of a future stroke or not using four classical machine learning algorithms over the Weka toolkit. From the performance analysis, they have found that Naïve Bayes performs better than other algorithms. But the problem is the dataset they have used is not perfectly symmetrical. Therefore, it has no impact on the predicted accuracy of the other techniques. On the other hand, in [12], stroke prediction was carried out by applying data mining techniques including k-nearest neighbor and C4.5 decision tree algorithms in WEKA tools. The experimental result showed that the C4.5 decision tree algorithm and K-nearest neighbor predict stroke nearly 95.42% and 94.18% respectively.

In [13], the authors have performed the stroke prediction by using various machine learning algorithms including Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier with a range of physiological parameters for reliable prediction. The experiment is conducted over an open-access stroke prediction dataset. In their analysis, the Random forest technique topped with 96% percent classification accuracy. In another research work [14], Bandi *et al.* employed some machine learning techniques to recognize, categorize, and predict stroke from medical history and proposed an improvised Random Forest ensemble technique to build a prediction model for analyzing different types of risk obtained within the strokes.

R. S. Jeena and S. Kumar in [15] investigated various physiological parameters considering the risk factors for the prediction of stroke disease and applied Support Vector Machine (SVM) with various kernel functions over the dataset International Stroke Trial database and got an accuracy of 90 % by the linear kernel. However, considering more input features to this approach can improve the system performance. Hakim *et al.* in [16] proposed an ensemble-based Modified Bootstrap Aggregating (Bagging) technique for the prediction of brain stroke and compared the performance of their proposed method with the traditional bagging technique based on different evaluation metrics. They have found the modified proposed technique is more accurate than the traditional bagging technique in predicting brain stroke with more than 96% accuracy.

In [17], stroke prediction was made using different Artificial Intelligence methods over the Cardiovascular Health Study (CHS) dataset. They have used a decision tree algorithm for the feature selection process, a PCA algorithm for reducing the dimension, and adopted a backpropagation neural network classification algorithm to construct a classification model. However, their proposed approach needs further improvement for getting up to mark accuracy rate. In [18], Kansadub *et al.* developed a model for predicting stroke over demographic data. The model is trained separately using three classification algorithms: Decision Tree, Naive Bayes, and Neural Network, and then construct a sound model for identification. From their experiment, although the decision tree algorithm provides better accuracy than the other two algorithms, it is found that Neural Network was the most suitable approach by accuracy, FP, and FN. Peng et.al introduced an artificial neural networks (ANNs) model for predicting stroke over patients' physiological data in [19]. Although the proposed method has reached about 98 percent classification accuracy but test and validation accuracy for the minority class is poor. The dataset used in the experiment was imbalanced which requires further preprocessing to address the issue of data imbalance. Cheon *et al.* [20] have used a deep neural network approach over medical service usage and health behavior data for predicting stroke. To extract appropriate features from medical records, they have applied Principle Component Analysis (PCA), and later those features are used in predicting stroke. The proposed model can be used by both patients and doctors to determine the possibility of stroke with accuracy up to 83.48%.

However, the proposed approach needs some auto-fine-tuning methods to improve training time and better performance.

From the above study, we have seen that most of the research works are conducted by classical machine learning algorithms for the prediction of stroke based on medical history data. The deep neural network is yet a new concept in the health industry which gives outstanding performance in the classification task. Therefore, this research differs from others in the sense that we have introduced a noble deep neural network CNN model to determine the possibility of future stroke. The performance of the proposed model is compared with other existing classical machine learning models. Furthermore, we also have applied feature selection algorithms to extract the most contributing features to be most valuable to a model to predict the target class.

III. Methodology

In this section, we discuss our proposed methodology. The working procedure of the system is shown in Fig. 5. However, the main steps of the procedure are explained below.

A. Dataset

To conduct this research, we have collected the health care dataset from Kaggle. Each row in the data set provides relevant information about the patient. The goal of the dataset is to analytically foresee whether a patient gets a stroke or not, in light of certain indicative estimations remembered. Fig. 1 shows the sample dataset used in the experiment. There was no identifying information in the dataset such as the patient's name, address or SSN, etc. Therefore, the dataset used in the experiment has no risk of patient confidentiality information. However, the major points of the dataset are listed below.

- The dataset contains 5110 samples.
- Each sample has 11 features and one target value.
- Target feature indicates whether a person is likely to get a stroke or not. '0' refers to a person who did not get a stroke and '1' refers to a person who got a stroke.
- In between the class, 249 observations are likely to get a stroke and the rest of the 4861 observations did not get a stroke.
- The dataset is divided into two partitions - training and testing. The model is trained with 80% data and tested with the rest of 20% data.

B. Features of the Dataset

There is a total of 11 features of the data set we have used. A short description of those features is defined in Table I below:

TABLE I. Feature of the Data Set

Feature	Description
ID	Unique identifier for a person
Gender	A patient's gender
Age	Age of the patient
Hypertension	O (zero) means the patient doesn't have hypertension and 1 (one) means the patient has hypertension
Heart disease	O (zero) means the patient doesn't have any heart disease and 1 (one) means the patient has a heart disease
Ever married	The patient is married or not
Work type	A patient's work type
Residence type	About a patient's residential area
Avg. glucose level	The average glucose level in blood
BMI	Patient's body mass index
Smoking status	The patient is a smoker or not or which type of smoker

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1

Figure 1. Snapshot of dataset.

C. Data Pre-processing

Data Pre-processing is needed before model structure to eliminate the undesirable noise and outliers from the dataset, bringing about a deviation from legitimate training. Anything that intrudes on the model from performing with less effectiveness is dealt with in this stage. In this article, we have performed the following data preprocessing steps.

1) *Handling Null Value:* Missing Data can happen when no data is accommodated for at least one thing or an entire unit. It is an exceptionally large issue in a genuine situation. Firstly, we have found whether the data set has any missing value or not.

If any missing value is encountered, the attribute value is filled or replaced with the mean value of that particular column.

2) *Label Encoding:* Since the dataset contains some string data, the string values are required to convert to a float value to fit in the model. To do so, we have applied the label encodingthe dataset.

D. Features Selection

Feature selection algorithms are used to lower the number of input variables to those that are considered to be the most advantageous to a model to predict the target variable. In this article, we have applied the following 3 feature selection algorithms to find the best possible features for our model.

1) Univariate selection

The first feature selection method we have used is the univariate feature selection method. The univariate feature selection process studies each feature individually to identify the strength of the association of the feature with the result variable. It works by choosing the best features based on univariate statistical tests. There are different options for univariate selection. Among them, we have used SelectKBest which removes all but K's highest-scoring features. However, the Scikit learn library gives the SelectKBest class that can be utilized with a set-up of various statistical tests to choose a particular number of features. For feature scoring the univariate feature selection with F-test is used. We have used the default selection function and the score of every feature is given below in Table II.

TABLE II. FEATURE UNIVARIATE SCORE

Sl	Specs	Score
1	age	3635.226911
7	avg_glucose_level	1718.285446
3	heart_disease	87.987436
2	hypertension	75.449498
4	ever_married	20.622787
8	bmi	15.894122
9	smoking_sratus	3.369423
5	work_type	2.925901
6	residence_type	0.600717
0	gender	0.239001

2) Feature importance

Feature importance refers to procedures that compute a score (or value) for the input features of the predictive model. The score denotes the "significance" of each feature during prediction. The high score value of a specific feature indicates it has a larger impact on the model to predict a particular variable. This technique is used to better understand the data and model with decreasing the number of input features. By utilizing the feature significance property of the predictive model, this technique determines the feature score of each feature of the dataset and deletes the low score features to get the highest score for simplifying the model and enhancing its performance. Also, feature importance is an inbuilt class that accompanies Tree-Based Classifiers. Here, we utilized an extra tree classifier for extricating the main 10 elements from the data set as shown in Fig. 2.

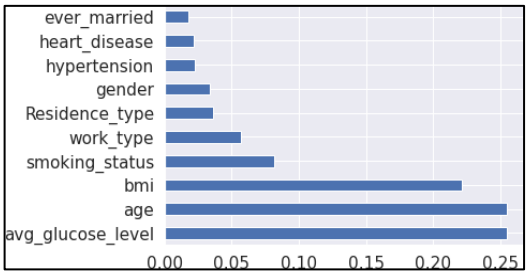


Figure 2. Feature importance score.

3) Correlation matrix with heatmap

A correlation matrix with heatmap is a 2D matrix demonstrating the correlation coefficients between the features. It understands how the features are related to each another and the strength of the relationship to identify the most related feature in the data set with the target variable. A correlation plot generally includes a number of numerical variables, with individually variable shown by a column. The rows depict the association between each pair of variables. The cells value exhibits the strength of the association, with positive values showing a positive association and negative values giving a negative association. Fig. 3 shows the correlation matrix with heat map over the dataset used in our experiment. Here the correlation value can accept any value from -1 to 1. From the color-coding of the cells, the associations between variables can be easily identified.

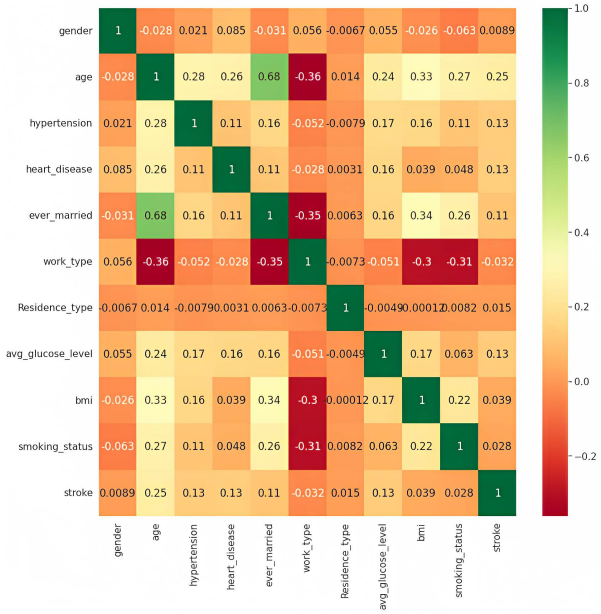


Figure 3. Correlation matrix with heatmap.

E. Training and Validation Sample

The dataset we have collected is partitioned into two subsets, one for training and another for testing. The first one is used to fit the learning model while the other is utilized to assess the fit model. However, in our dataset, we have used 80 percent data (i.e. 4088 samples) for training and 20 percent data (i.e. 1022 samples) for testing. Before splitting the dataset, we shuffle the dataset for more generic validation and training.

F. Classical Machine Learning Algorithms

1) Logistic regression

Logistic Regression [21] is a supervised learning algorithm utilized for foreseeing the likelihood of the result variable. This algorithm is the best fit when the result variable has output as (0 or 1). When the dataset has just two potential output values then, Logistic Regression is selected.

2) Decision tree classification

The Decision Tree classification algorithm [22] is used to take care of Regression and classification problems. This algorithm is additionally a supervised learning strategy where the info factors as of now have they're comparing yield variables. It is designed as a tree. In this algorithm, the information constantly parts as indicated by a specific boundary. A decision tree algorithm has two sections: Decision Node and the Leaf node. The information is split at the previous hub, and the last option is the node that gives the result.

3) Random forest classification

Random Forests [22] are made from numerous autonomous choice trees prepared freely on an arbitrary subset of the data set. These trees are produced at the time of training, and the results are generated from every decision tree. For the last forecast from this classification algorithm, a strategy called voting happens. This technique implies that every decision tree votes in favor of a result class. The random forest picks the class with the most extreme number of votes as the last expectation.

4) Support vector machine

The Support Vector Machine (SVM) [23] is one of the most famous Supervised Learning algorithms, which is utilized for Classification just as Regression problems. Nonetheless, essentially, it is utilized for Classification issues in Machine Learning. The objective of the SVM calculation is to make the best line or choice limit that can isolate n-layered space into classes so we can without much of a stretch put the new informative item in the right classification later on. This best choice limit is known as a hyper-plane.

5) Naive Bayes classifier

Naive Bayes algorithm [24] is a supervised learning algorithm, which depends on the Bayes hypothesis and is utilized for taking care of classification problems. It is primarily utilized in text characterization that incorporates a high-layered training data set. The naive Bayes algorithm is one of the basic and best classification algorithms which helps in building quick AI models that can make speedy expectations. It is a probabilistic classifier, which implies it predicts based on the likelihood of an object.

G. Proposed CNN Approach

In the modern advancement of AI, modern tool Convolution Neural Networks (CNNs) contribute significantly in the area of medical applications. CNN started achieving popularity day by day and defeating all other models with good accuracy and low error rate. This motivates to apply the CNN model (Fig. 4) in stroke disease prediction. A CNN is made from stacking a few structure blocks: convolution layers, pooling layers, and complete associated layers. In our proposed CNN approach, we used a 1D-CNN to use structured 1D features from the dataset and the model produces 1D result also. In our dataset, there are 10 features that are used to train the model.

A model's exhibition under specific portions and loads is determined with a misfortune work through sending engendering on a preparation dataset. And then learnable boundaries, pieces, and loads are refreshed according to the disaster regard through backpropagation with a tendency to improve estimation. CNN doesn't need hand-created highlight extraction. Its architectures do not necessarily require the segmentation of tumors or organs by human experts. It is undeniably more information-hungry on account of its great many learnable boundaries to gauge, and, subsequently, is even more computationally costly, bringing about requiring graphics processing units (GPUs) for model preparation [25].

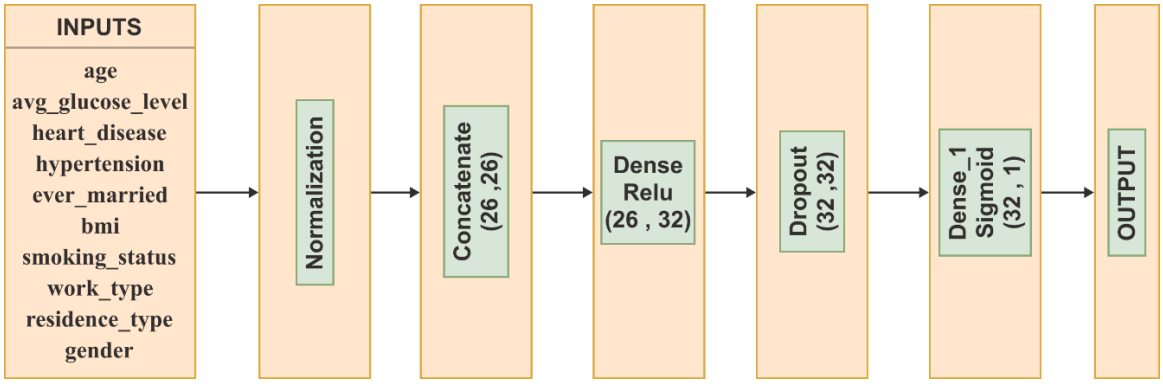


Figure 4. Proposed CNN model.

H. Workflow of the Proposed Approach

At first, for constructing the project we collected a data set which is called "Health care dataset for stork prediction" from Kaggle which has 10 features and 1 target column. Then we perform some data prepossessing like removing the null values and deleting the not necessary data from the dataset. For removing the null values, we calculate the mean value of that column and replace the null value with that to get a more generic

performance. Then we use 2 normalization functions to normalize the categorical and String data. After that, we use all the 10 features as our parameters to train the model. Then we build our model. In our model, we use the Dense layer as input layers and use the RELU activation function. Then we set the dropout to 0.5 for the input layer. For output, we also use the Dense layer but this time we use the sigmoid activation function, and we use binary-cross entropy as a loss function. In our model all the features get concatenated after the Normalization and string lookup process and the input shape when it goes to the first dense layer is (none, 26), but when it goes through the final dense layer after the dropout layer the input shape changed to (none, 32) and coming out from the final layer with an output shape (none, 1) what we expected to have. Then we split the data for training and testing. We used 80 percent of the data for training and the rest of the 20 percent is for testing. After the splitting, we fit the data in our model and set the batch size to 32 and the learning rate to 0.5. And we fit the data for 200 epochs. Finally, we calculate the accuracy of the test data for the model. However, in Fig. 5, the flowchart of the proposed approach has been shown.

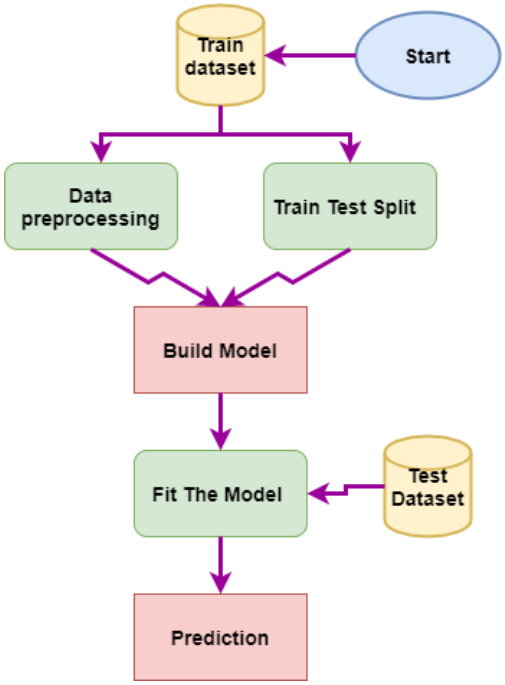


Figure 5. Flow chart of the proposed system.

IV. Experimental Result

This section presents the experimental results for the performances of our proposed CNN model including other classical machine learning models. All experiments are run on GPU- empowered TensorFlow with Keras structure. We extricate highlights in a profound learning organization, involved CNN engineering. The batch size is set to 32 to feed our model and set the learning rate to 0.5. We have utilized Relu activation function work for each layer except the sigmoid initiation work for the yield layer. We used the binary cross-entropy loss function to model our dataset. In our experiment, the stroke parameter is taken as a needy variable. The remainder of the boundaries is taken as autonomous factors. The stroke parameter takes only binary values, where 0 represents no stroke and 1 represents stroke. The result is shown as a percentage which gives the probability of getting a stroke. The dataset is divided into two partitions: training data (80 percent) and validation data (20 percent). The model is trained for 200 epochs. After training the model, we have seen that a there is 15.1% validation loss and 95.5% validation accuracy. we have set the dropout rate to 0.5. The dropout layer arbitrarily sets input units to 0 with a recurrence of the rate at each progression during preparing time, which forestalls over-fitting. Information sources, not set to 0 are scaled up by 1/ (1 - rate) with the end goal that the whole overall information source is unaltered.

A. *Training Accuracy vs Validation Accuracy*

We set the epoch to 200 and the accuracy of training and validation is shown in Fig. 6.

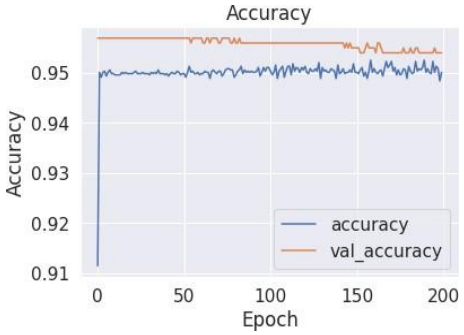


Figure 6. Training accuracy vs validation accuracy.

B. *Training Loss vs Validation Loss*

Fig. 7 shows the training and validation loss after 200 epochs.

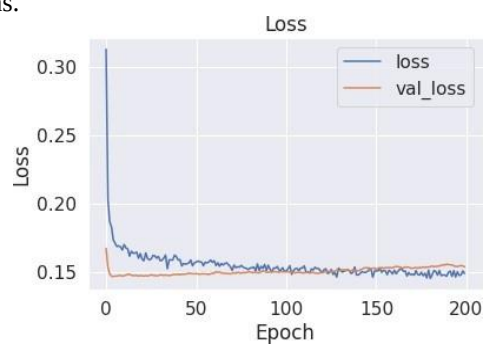


Figure 7. Training loss vs validation loss.

C. *Confusion Matrix of the Proposed Model*

The performance of machine learning techniques is measured with respect to a few performance measure parameters. A confusion matrix is an outline of forecast results on a classification problem. It summarized the correct and incorrect predictions and count values for each class. It shows our model the way it is confused when it makes predictions. Also, it gives insight into errors and which type of errors it makes A confusion matrix including A, B, C, and D for actual data and predicts data is formed to evaluate the parameters. Here,

A =True Positive, B = True Negative, C = False Positive, and D = False Negative. Fig. 8 and Fig. 9 show the confusion matrix of the proposed model and other classical machine learning models respectively.

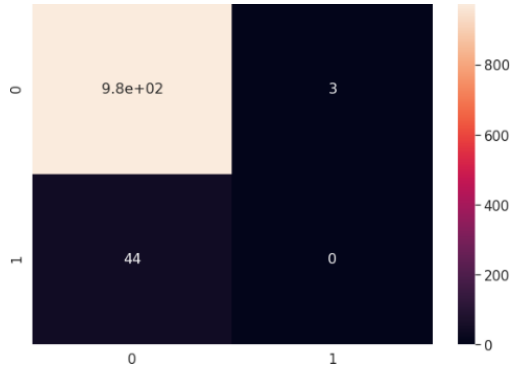
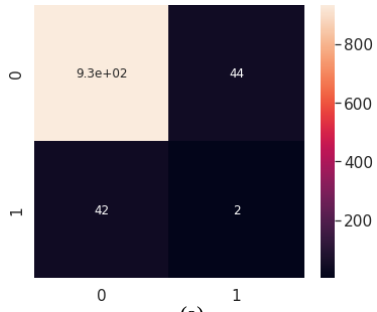
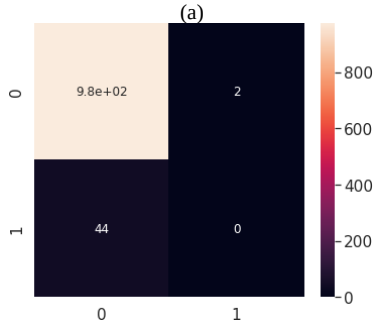


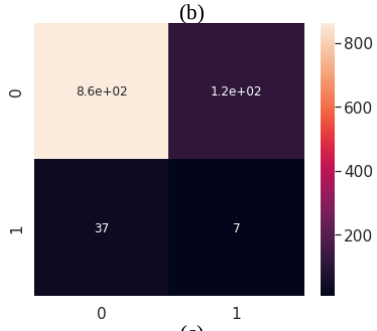
Figure 8. Confusion matrix of the proposed model.



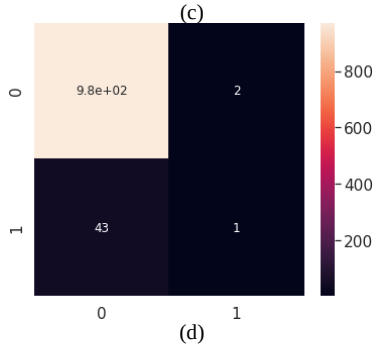
(a)



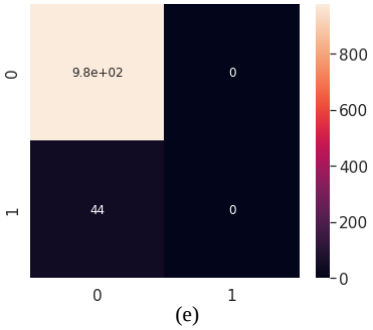
(b)



(c)



(d)



(e)

Figure 9. Confusion matrix of (a) decision tree (b) logistic regression (c) naive bayes (d) random forest (e) support vector machines.

D. *Performance of the Proposed Model*

In this section, we have presented the performance of the proposed CNN model. The performance matrices are discussed below.

1) *Accuracy*

The accuracy of a model is equivalent to the extent of expectations that the model grouped effectively.

$$Accuracy = \frac{A + B}{A + B + C + D}$$

2) *Sensitivity*

Sensitivity, otherwise called the recall, hit rate, or the Genuine Positive Rate (TPR), is the extent of the aggregate sum of significant examples that were really recovered.

$$Sensitivity = \frac{A}{A + D}$$

3) *Precision*

Precision is also known as positive predictive value and is the extent of important examples among the recovered occurrences.

$$Precision = \frac{A}{A + C}$$

4) *F1 score*

The F1 score is a proportion of a test's exactness — It is the symphonies mean of accuracy and review. It can have a most extreme score of 1 and at least 0.

$$F1\ Score = \frac{2A}{2A + C + D}$$

The Accuracy score, Sensitivity, Precision, and the F1 Score of the model are presented in Fig. 10.

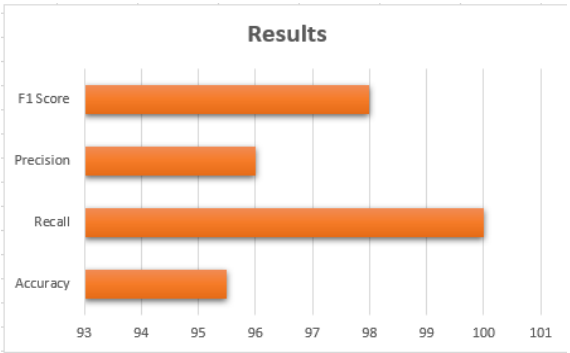


Figure 10. Result of the proposed model.

E. Performance Comparison

We have compared the result of the model with other classical machine learning models. The comparison result is shown in Table II to better understand the performance. From Fig. 11, it is clearly visible that the proposed CNN model outperformed other classical machine learning models. However, the performance of Logistic Regression is nearby the proposed model. On the other hand, GNB gives the worst performance among the mentioned algorithms.

TABLE III. RESULT ANALYSIS

Algorithm	Accuracy	Recall	Precision	F1 Score
Logistic Regression	95%	100%	95%	98%
Random Forest	94.7%	100%	96%	98%
Decision Tree	92.6%	96%	96%	96%
Naïve Bayes	87.5%	88%	96%	92%
Support Vector Machine	95%	100%	96%	98%
Proposed Model	95.5%	100%	96%	98%

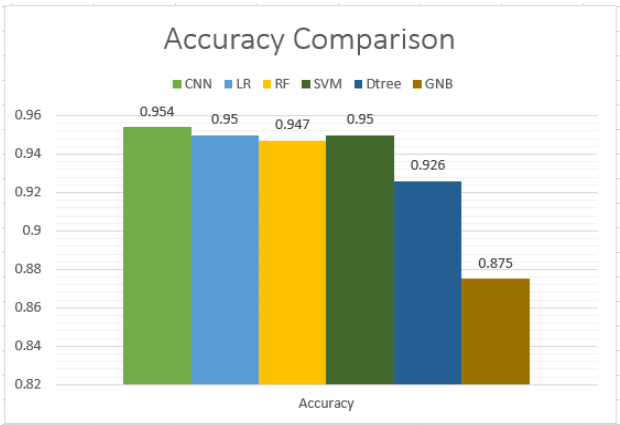


Figure 11. Comparison of accuracies of classical machine learning algorithms.

V. OBSERVATIONS

We have examined the correlation of the risk of stroke disease with other aspects. From the investigation, we have noticed that the risk of having a stroke is highly associated with average glucose level, age, and BMI as shown in Fig. 12. From the figure, we can see that most stroke patient has a glucose level of around 75 to 175. This means the chance of stroke increases for high glucose levels. Also, between 58 and 78 age, the patients are greatly suffered from stroke disease with some outliers. This is also the same as the average glucose level. The risk of stroke rises with the age of the person to up to a particular range. On the other hand, the occurrence of strokes increases with BMIs of 27 or greater. We have also analyzed the association of stroke disease with other factors such as gender, work type, and hypertension as shown in Fig. 13. From the figure, it can be concluded that 57% of the males and 43% of the females got suffered by stroke. Therefore, males have a higher chance of being affected by stroke diseases than females. In the case of jobholders, the employees of the private job are at greater risk of getting a stroke than other jobs. This is because a private job demands time pressure, mental stress, and coordination burdens. This high-stress job is connected to an increased chance of getting a stroke. Generally, the risk of getting a stroke among children is very low. From the correlation of hypertension, we also have seen that people having hypertension have a higher risk of getting a stroke disease. This is due to people's unhealthy behavior, smoking, poor eating habits, lack of exercise, being overweight, stress, genetics, etc. are linked to increased risk of stroke disease. Therefore, the probability of occurring stroke disease is found to be higher in private jobs than in other work types.

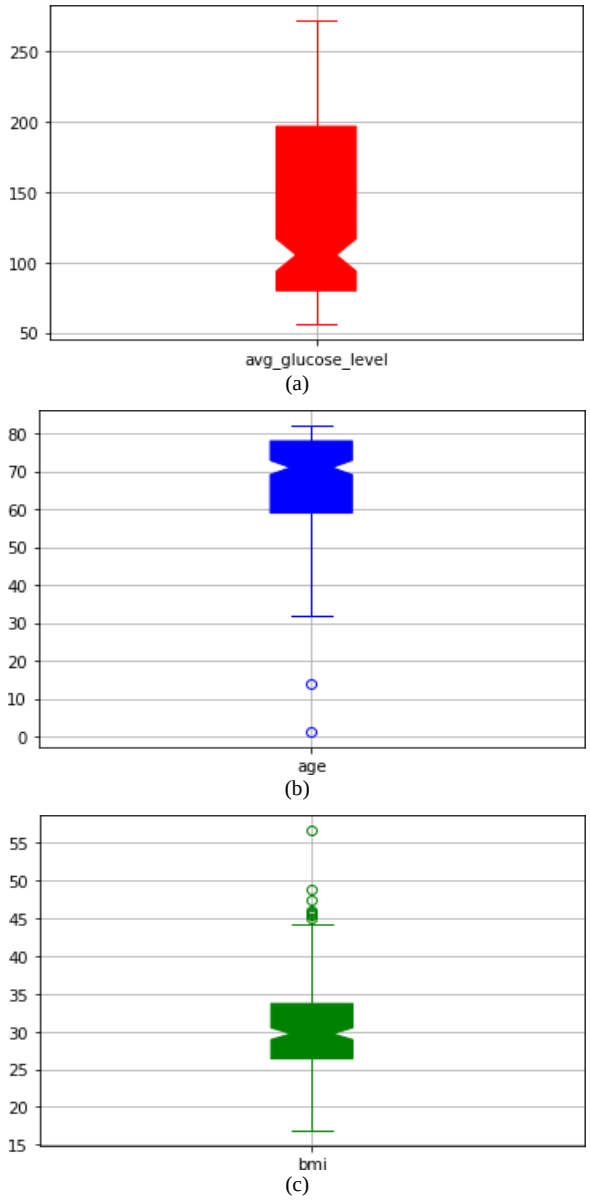


Figure 12. The risk of stroke with respect to (a) average glucose level (b) age and (c) BMI.

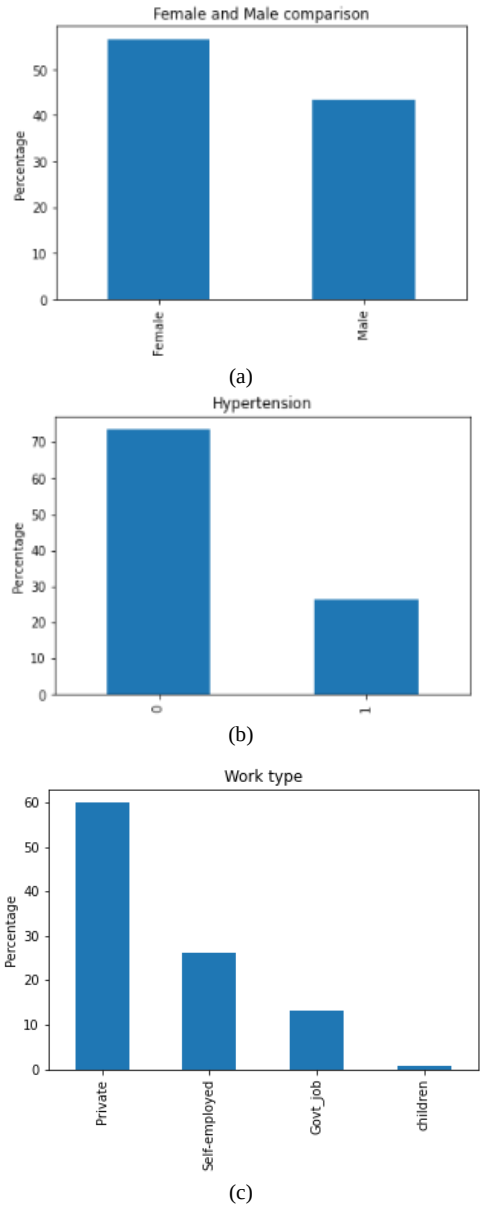


Figure 13. Stroke analysis based on (a) gender (b) hypertension and (c) work type.

VI. Conclusion

In this paper, we have analyzed the different attributes of people to determine the chance of stroke disease. The experiment is performed over health care data set collected from Kaggle. To investigate the dataset, we have used different classification models. Besides we have developed a CNN model to predict the possibility of whether a person will be getting a stroke or not followed by evaluating the model's performance. The experimental result shows that the proposed model is more effective than some other existing models with a promising 95.5 percent accuracy. We can use this model to diagnose a patient to determine the possibility of getting a stroke at an early stage. Finally, the analysis of stroke concerning different attributes discovered a general pattern among the attributes which have a greater risk of developing stroke disease.

Conflict of Interest

The author declares that there is no conflict of interest in this research.

Author Contributions

Md. Ashrafuzzaman, Suman Saha, and Kamruddin Nur participated in the research, analysis, and writing of this paper. All the authors had approved the final version of the paper.

References

[1] D. Pastore, F. Pacifici, B. Capuani, *et al.*, "Sex-Genetic interaction in the risk for cerebrovascular disease," *International Journal of Environmental Research and Public Health*, vol. 24, no. 24, pp. 2687-2699, 2017.

[2] The top 10 causes of death. (2020). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] U. R. Acharya, S. L. Oh, Y. Hagiwara, *et al.*, "Automated EEG-based screening of depression using deep convolutional neural network," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 103-113, 2018.

[4] Stroke facts. (2021). [Online]. Available: https://www.cdc.gov/stroke/facts.htm

[5] R. Ramyea, S. Preethi, K. Keerthana, *et al.*, "An intellectual supervised machine learning algorithm for the early prediction of hyperglycemia," in *Proc. Innovations in Power and Advanced Computing Technologies*, 2021, pp. 1-7.

[6] N. S. Adi, R. Farhany, R. Ghina, *et al.*, "Stroke risk prediction model using machine learning," in *Proc. International Conference on Artificial Intelligence and Big Data Analytics*, 2021, pp. 56-60.

[7] R. Kavitha, W. Jaisingh, and S. Sujithra, "Applying machine learning techniques for stroke prediction in patients," in *Proc. International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation*, 2021, pp. 1-4.

[8] S. Chauhan, L. Vig, M. D. F. D. Grazia, *et al.*, "A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from mri lesion images," *Frontiers in Neuroinformatics*, vol. 13, no. 53, pp. 1-12, 2019.

[9] C. Y. Hung, W. C. Chen, P. T. Lai, *et al.*, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *Proc. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2017, pp. 3110-3113.

[10] Stroke prediction dataset. (2022). [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[11] C. C. Peng, S. H. Wang, S. J. Liu, *et al.*, "Artificial neural network application to the stroke prediction," in *Proc. IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability*, 2020, pp. 130-133.

[12] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *Proc. 8th Annual Industrial Automation and Electromechanical Engineering Conference*, 2017, pp. 158-161.

[13] M. A. Hakim, M. Z. Hasan, M. M. Alam, *et al.*, "An efficient modified bagging method for early prediction of brain stroke," in *Proc. International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering*, 2019, pp. 1-4.

[14] T. Kansadub, S. Thammaboosadee, S. Kiattisin, *et al.*, "Stroke risk prediction model based on demographic data," in *Proc. 8th Biomedical Engineering International Conference*, 2015, pp. 1-3.

[15] R. S. Jeena and S. Kumar, "Stroke prediction using SVM," in *Proc. International Conference on Control, Instrumentation, Communication and Computational Technologies*, 2016, pp. 600-602.

[16] L. Amini, R. Azarpazhouh, M. T. Farzadfar, *et al.*, "Prediction and control of stroke by data mining," *International Journal of Preventive Medicine*, vol. 4, suppl. 2, pp. S245-249, 2013.

[17] S. Cheon, J. Kim, and J. Lim, "The use of deep learning to predict stroke patient mortality," *International Journal of Environmental Research and Public Health*, vol. 16, no. 11, pp. 1-12, 2019, 1876.
[18] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," *Revue d'Intelligence Artificielle*, vol. 34, no. 6, pp. 753-761, 2020.
[19] T. Tazin, M. N. Alam, N. N. Dola, *et al.*, "Stroke disease detection and prediction using robust learning approaches," *Journal of Healthcare Engineering*, vol. 2021, p. 12, 2021.
[20] T. I. Shoily, T. Islam, S. Jannat, *et al.*, "Detection of stroke disease using machine learning algorithms," in *Proc. 10th International Conference on Computing, Communication and Networking Technologies*, 2019, pp. 1-6.
[21] Logistic regression. (2021). [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-Learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[23] Support vector machine algorithm. (2021). [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
[24] Naïve Bayes classifier algorithm. (2021). [Online]. Available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier
[25] R. Yamashita, M. Nishio, R. Do, *et al.*, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611-629, 2018.

Md. Ashrafuzzaman was born in Pabna, Rajshahi, Bangladesh. He received a B.Sc. degree in Computer Science and Engineering from the Bangladesh University of Business and Technology (BUBT). He is currently working as an Executive Administrator at a well-known English Medium School in Bangladesh. He is continuously ready to learn new things with full excitement and enthusiasm. His research focuses on deep learning algorithms, computer vision and IOT. He has experience working in Python, Keras, TensorFlow, Sklearn, NumPy, etc.

Suman Saha is currently serving as a Lecturer in the Department of ICT at Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh. Before that, he served as a faculty member in the Department of Computer Science and Engineering at Bangladesh University of Business and Technology (BUBT) from June 1, 2014 to May 30, 2021 and Institute of Information and Communication Technology (IICT), Dhaka University of Engineering and Technology (DUET) from June 01, 2021 to October 04, 2021. Mr. Saha completed his B.Sc Engg. in CSE from the University of Chittagong and M.Sc in CSE from Bangladesh University of Engineering and Technology (BUET). His research area includes AI, Machine Learning, Deep Learning, NLP, Data Mining, Wireless Sensor Networks, Blockchain, etc.

Kamruddin Nur (Senior Member, IEEE) is currently serving as an associate professor in the Department of Computer Science at American International University-Bangladesh (AIUB). He also served as the Chairman in the Department of Computer Science and Engineering at Stamford University Bangladesh (SUB) and Bangladesh University of Business and Technology (BUBT). Dr. Nur completed his PhD from UPF, Barcelona, Spain, Masters from UIU, and Bachelor from Victoria University of Wellington (VUW), New Zealand. Dr. Nur authored many prestigious journals and conferences in IEEE and ACM, served as TPC members, and reviewed articles in IEEE, ACM, Springer journals, and conferences. His research area includes Ubiquitous Computing, Computer Vision, Machine Learning, and Robotic Automation.