Proceedings of the Fourth International Conference on Smart Systems and Inventive Technology (ICSSIT-2022)
IEEE Xplore Part Number: CFP22P17-ART; ISBN: 978-1-6654-0118-0

A Machine Learning Approach to Detect the Brain Stroke Disease

Bonna Akter
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
bonna15-2585@diu.edu.bd

Aditya Rajbongshi
Dept. of CSE
Jahangirnagar University
Dhaka, Bangladesh
adityaraj.jucse@gmail.com

Sadia Sazzad
Dept. of CSE
National Institute of Textile Engineering and Research
Dhaka, Bangladesh
sadiakatha1993@gmail.com

Rashiduzzaman Shakil
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
rashiduzzaman15-2655@diu.edu.bd

Jahanur Biswas
Dept. of CSE
Jahangirnagar University
Dhaka, Bangladesh
jahanurbiswas.jucse@gmail.com

Umme Sara
Dept. of CSE
National Institute of Textile Engineering and Research
Dhaka, Bangladesh
ummesarapapri@gmail.com

2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) | 978-1-6654-0118-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSSIT53264.2022.9716345

***Abstract*—The brain, which comprises the cerebrum, cerebellum, and brainstem and is covered by the skull, is a very complex and intriguing organ in the human body. Stroke is the world's second-leading cause of mortality; as a result, it requires prompt treatment to avoid brain damage. Early detection of a brain stroke can help to prevent or lessen the severity of the stroke, which can lower death rates. Using machine learning algorithms to identify risk variables is a promising method. This paper proposed a model that included a methodology to achieve an accurate brain stroke forecast. The efficient data collection, data pre-processing, and data transformation methods have been applied to provide reliable information for our proposed model to be successful. A ”brain stroke dataset” was employed to build up the model. The standardization technique is used to standardize data. In the training and testing procedure, Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) classifiers are applied. The performance of each classifier has been estimated by adopting performance evaluation metrics such as accuracy, sensitivity (SEN), error rate, false-positive rate (FPR), false-negative rate (FNR), root mean square error, and log loss. Based on the outcome while using the RF classifier, we can determine that our proposed model provided the maximum accuracy, which was 95.30%.**

***Index Terms*—Brain Stroke, Disease Prediction, Machine Learning, Random Forest**

I. Introduction

The most severe and deadly disease in humans has long been thought to be brain stroke. The increased occurrence of brain stroke, which is associated with a high death rate, poses considerable risk and burden to healthcare systems worldwide. The brain is the most intricate element of the human body, as we all know. This three-pound organ is the brain's seat of intellect, as well as a sensation interpreter, movement creator, and behavior controller. It's a part of the brain that controls cognition, memory, emotion, touch, motor skills, vision, breathing, temperature, hunger, and other critical human activities. The brain, housed in a bone shell and kept clean by protective fluid, is the source of all the characteristics that define our humanity. Brain stroke occurs when blood flow to a part of the brain is restricted or reduced, depriving brain tissue of oxygen and nutrients. Brain cells begin to die in a minute under this circumstance. The number of people suffering from a stroke is increasing every day. Strokes in the brain are more common in males than women, especially in middle and older age. On the other hand, Stroke affects roughly 8% of children with sickle cell disease. A stroke affects 15 million individuals globally every year [1]. Five million of them die, and another five million are permanently crippled, putting a strain on families and communities.

To lower the morality of brain stroke patients, we built a prediction system that may help clinicians diagnose illnesses by feeding data from the patients into the processing system. The system analyzes the data, and the data is pre-processed by data pre-processing method. Finally, SVM, RF, and DT classifiers were used to identify brain stroke patients. This research aims to create a system that can predict brain stroke correctly. The stages that must be accomplished are as follows:

- Early diagnosis of diseases can aid in the reduction of risk.
- Brain stroke can destroy the global healthcare system.
- This method can be used to assist the physician.
- Early discovery of a brain stroke will also lead to advice on how to manage it in a more secure manner.
- Reduce the morality through brain stroke.
- The overall findings are used to assess the performance of the various models.

II. Related Work

In terms of early illness prediction, machine learning is a promising technology. Many studies have been conducted to predict diseases such as cancer, skin disease, stroke illness, and so on. There is virtually little study on the prediction of brain stroke illness.

G. Vijayadeep *et al.* [2] proposed a hybrid feature extraction-based system to predict brain stroke applying a random forest classifier. K- cross-validation technique used for extracting the feature and the obtained accuracy was 98.23%.

A Real-time gait monitoring system for stroke prediction service referred by SJ. Park *et al.* [3] .Their work was conducted with Gait parameters of 63 stroke patients and 208 healthy patients. They used RT, CART-, C5.0, SVM, LR and LSVM classifier in their model and generate highest accuracy AUC: 0.995, gini: 0.993 with C5.0 classifier. They resulted in the lowest accuracy AUC: 0.908, Gini: 0.816 with LSVM classifier.

T. Badriyah *et al.* [4] build a classification model for predicting two sub-types of stroke disease, Ischemic stroke, and stroke hemorrhage depend on CT scan data those obtained 102 patients. They used eight machines learning algorithm to generate the accuracy. Their model resulted in 95.97% accuracy with random forest algorithm and lowest 71.18% accuracy with Naïve Bayes algorithm.

T. Liu *et al.* [5] proposed a hybrid machine learning approach to predict cerebral strokes. For their proposed method, they used HealthData.gov, which was utilized from the benchmark dataset from Kaggle. They counted the accuracy of KNN, LR, RF DECT depending on smoking status and SVM, RFR, BayRid, and GBM with BMI status variable and obtained 49.6% accuracy from the KNN classifier and 87.6% accuracy from the RFR classifier.

G. Fang *et al.* [6] proposed a machine learning approach for predicting stroke prognosis. They used an IST dataset that contained 19435 patient data of 467 hospitals. They used SVC, MLP, RF, and AdaBoost classifiers to predict RVISINF of acute stroke. They obtained the highest accuracy on the RVISINF model.

P. Govindarajan *et al.* [7] proposed a Machine learning model to detect stroke patients. Their work used 507 patient data collected from Sugam Multispecialty hospital, Kumbakonam. They used ANN, SVM, DT, LR, Boosting, and bagging techniques. ANN resulted in the highest accuracy for their approach that was 95.3%.

M. Emon *et al.* [8] proposed a machine learning approach for stroke prediction. In their research, they used 5110 people's data from the medical clinic of Bangladesh. They used LR, SGD, DT, AdaBoost, Gaussian, Quadratic Discriminant Analysis, MLP, KNN, and XGBoost Classifier for predicting the stroke. They obtained 97% accuracy from the Weighted Voting classifier and 65% accuracy from SGD.

A novel prediction model was introduced in the paper of T. Shoily *et al.* [9] with several general classifications and different combinations of features techniques. They used NB, J48, KNN, and RF Classifiers to build their model. Obtained accuracy of 99.8% with theJ48, KNN, RF classifier, and 85.6% with NB Classifier.

The essential risk factors were identified in the study by M. Amin *et al.* [10], and machine learning models (k-NN, DT, NB, LR, SVM, Neural Network, and a hybrid of voting with NB and LR) were used to undertake a comparative analysis. According to their findings, it was founded that when combined with the selected attributes, the hybrid model attained an accuracy of 87.41

For prediction, M. Ashraf et al *et al.* [11] employed individual learning algorithms and ensemble techniques such as Bayes Net, J48, KNN, MLP, NB, RT, and RF. J48 was the most accurate, with a score of 70.77 percent. They then used cutting-edge approaches, with KERAS achieving an accuracy rate of 80.

S. Saqlain *et al.* [12] proposed a technique that combined the mean Fisher score feature selection algorithm (MFSFSA) with the SVM classification model. They obtained the targeted feature subset using an SVM and utilized a validation method to calculate MCC. The combination of MFSFSA and SVM yielded an accuracy of 81.19 percent, a sensitivity of 72.92 percent, and a specificity of 88.68 percent.

From the above studies, it is observed that there is very little research on brain stroke prediction, and they adopted tiny datasets to implement their works.

III. Proposed Work

In this section, the working procedure to predict brain stroke, which is presented in Fig. 1, is described clearly. The methodology is divided into four sub-sections: data acquisition, data preprocessing, classifier description, and performance evaluation. The detailed description of the sub-section is as follows:

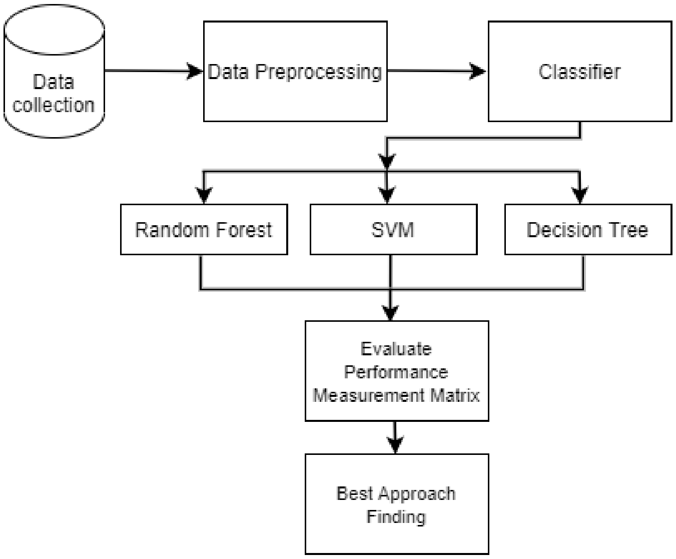

Fig. 1. The working procedure of proposed work

A. Data Collection

Data is the initial and most fundamental component of machine learning algorithms to produce accurate results. We used the ”Stroke prediction dataset,” collected from Kaggle [13], an established data source. There were 5110 occurrences and 12 attributes in this dataset. Eleven attributes are utilized as inputs, with one attribute functioning as an output.

B. Data Pre-processing

Data pre-processing is the most crucial stage of any research. Missing values and redundant data are dealt with in this stage. This work dealt with missing values and redundant data using image processing techniques. We utilized 80% of the data for training and 20% of the data to test our model.

C. Classifiers Description

In this sub-section, applied three machine learning algorithms to accomplish our proposed work have been depicted.

1) Decision Tree: One of the most potent and well-known prediction tools is the Decision Tree method, with only two num Classes. Every internal node in a Decision Tree structure represents a property being tested, every branch represents a test outcome, and each leaf node represents a separate class[14]. The tree develops from the root node by picking a 'Best Feature' or 'Best Attribute' from a list of available characteristics, then splitting.' To choose the 'Best feature,' two additional measures, 'Entropy' as shown in equ. (1) and 'Information Gain' as shown in equ. (2), are usually utilized (2).

$$E(D) = -P(positive)\log_2 P(positive) - P(negative)\log_2 P(negative) \tag{1}$$

Equ. 1 is used to determine the Entropy E of a dataset D with positive and negative 'Decision Attributes' (1).

$$Gain(AttributeX) = Entropy(DecisionAttributeY) - Entropy(X;Y) \tag{2}$$

2) Random Forest: An ensemble algorithm, the Random Forest (RF) classifier, comprises multiple algorithms. During the training portion, RF constructed an entire forest using numerous uncorrelated and random Decision Trees[15]. Ensemble learning approaches integrate many learning algorithms to generate the most excellent predictive model possible, which outperforms the predictions of any single model.

3) Support Vector Machine: SVM stands for support vector machine and is a linear classifier based on the margin maximization principle. They use structural risk minimization to increase the classifier's complexity and improve generalization [16]. The SVM addresses the classification issue by finding the hyperplane in a higher-dimensional space that best splits the data into two groups. It outperformed another classifier in terms of efficiency and accuracy.

D. Performance Measurement Metrics

Performance assessment metrics are used to assess the efficiency of an ML model. We have applied three classifiers to predict brain stroke. To evaluate the performance of utilized classifiers, we have calculated the performance evaluation metrics [17][18][19]. The equation for those metrics is as follows:

$$Accuracy = (\frac{TP+TN}{TP+FP+TN+FN}) \times 100\% \tag{3}$$

TABLE I
CONFUSION MATRIX FOR EACH CLASSIFIER

Classifier	Confusion Matrix	
Random Forest	TP 973	FN 3
	FP 45	TN 1
SVM	TP 934	FN 42
	FP 40	TN 6
Decision Tree	TP 907	FN 69
	FP 38	TN 8

$$Sensitivity = (\frac{TP}{TP+FN}) \times 100\% \tag{4}$$

$$Specificity = (\frac{TN}{TN+FP}) \times 100\% \tag{5}$$

$$FPR = (\frac{FP}{TN+FP}) \times 100\% \tag{6}$$

$$FNR = (\frac{FN}{TP+FN}) \times 100\% \tag{7}$$

IV. RESULT ANALYSIS

We used a "Brain stroke prediction dataset" to build our model. This dataset contained a total of 5110 patient observations and 12 attributes. The attributes were gradually gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose, BMI, smoking status, and stroke. Each attribute holds different symptoms, which helps to make a decision. Is the patient affected by stroke or not? We evaluated the correlation with each attribute. After considering correlation among whole attributes, we generated a heat-map. Fig. 2 shows the relationship between the same attribute, which consists of both input and output and is strongly correlated. Those attributes are used to train and test the classifier. After testing each classifier, the confusion matrix is generated. Table I shows the confusion matrix for each classifier.

The performance evaluation metrics for each classifier have been estimated and are presented in Table II. From Table II, it has been noticed that the Random Forest classifier gained 95.30% accuracy, 95.57% sensitivity, 25.00% specificity, 75.00% false-positive rate, 4.42% false-negative rate, 4.69% mean absolute error, 21.67% root mean square error, and 1.62% log loss, respectively. Besides the performance evaluation metrics such as accuracy, sensitivity, specificity,false-positive rate, false-negative rate, mean absolute error, root mean square error, and log loss for the support vector machine (SVM) classifier were successively 91.98%, 95.89%, 12.50%, 87.50%, 4.10%, 8.02%, 28.32%, and 2.77% respectively. On the other hand, the decision tree classifier achieved the lowest accuracy, 89.53% accuracy among the entire classifiers. Rather it obtained 95.97% sensitivity, 10.38%specificity, 89.61% false-positive rate, 4.02 false-negative rate,10.46%

TABLE II
PERFORMANCE EVALUATION METRICS FOR THREE CLASSIFIERS

Classifier	Accuracy	Sensitivity	Specificity	FPR	FNR	Mean Absolute Error	Root Mean Square Error	Log Loss
Random Forest	95.30%	95.57%	25.00%	75.00%	4.42%	4.69%	21.67%	1.62
SVM	91.98%	95.89%	12.50%	87.50%	4.10%	8.02%	28.32%	2.77
Decision Tree	89.53%	95.97%	10.38%	89.61%	4.02%	10.46%	32.35%	3.62

TABLE III
COMPARATIVE ANALYSIS BETWEEN EXISTING WORKS AND THIS WORK

Reference	Object	Data set Source	Dataset Size	Applied Classifier	Best classifier	Obtained Accuracy
This work	Stroke Prediction	Kaggle	5110	RF, DT, SVM	RF	95.30%
M. Ashraf et al [11]	Cardiovascular Disease	Stanford online healthcare repository	304	Bayes Net, J48, KNN, MLP, NB	KERAS	80.00%
S. Saqlain et al [12]	Heart disease diagnosis	4 UCI marge dataset	976	SVM	SVM	92.68%
P. Govindarajan et al [13]	Classification of stroke disease	Sugam Multispecialty Hospital, India	507	ANN, DT, SVM, LR	ANN	95.03%

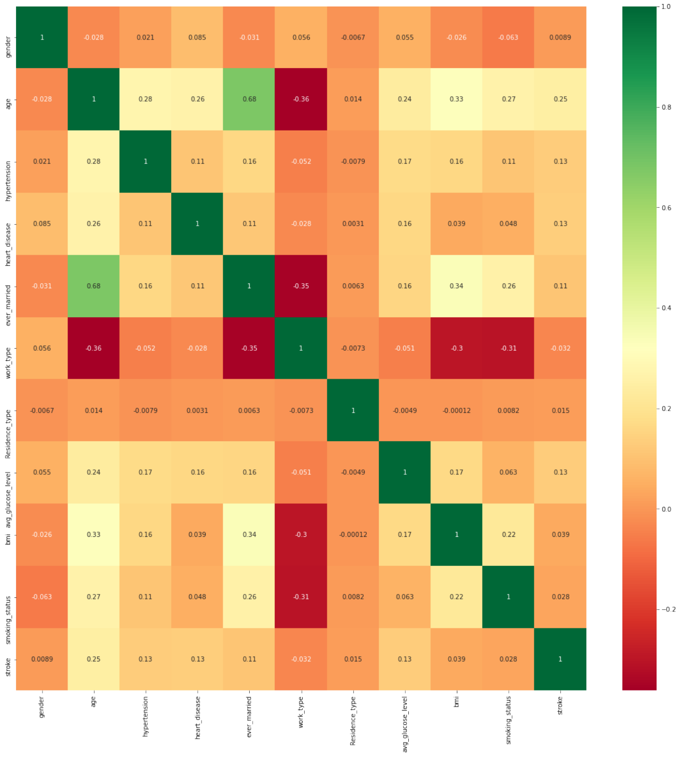

Fig. 2. Correlation among attributes

mean absolute error, 32.35% root mean square error,and log loss value 3.62%.

For clear visualization of the comparison accuracy among classifiers, Fig. 3 is presented. It can easily show that random forest achieved the highest accuracy, and decision tree achieved the lowest accuracy for our proposed model. On the other side, the visualization of the comparison of Sensitivity and Specificity values among the random forest, SVM, and decision tree classifiers is presented in Fig. 4. Here, The highest sensitivity is 95.97%, which is obtained by SVM. Instead, the greatest specificity is 25.00%, which is gained by a random forest classifier.

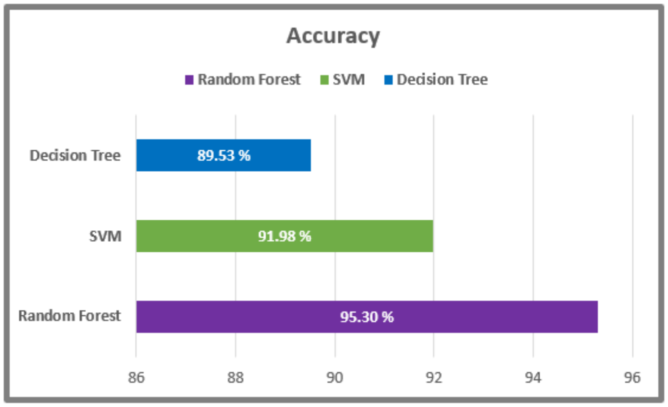

Fig. 3. Comparison accuracy among three classifiers

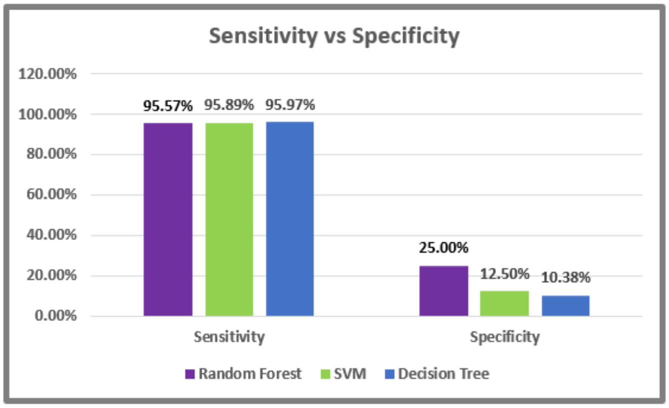

Fig. 4. Comparison Sensitivity & Specificity among three classifiers

Comparison [20] is necessary for the findings of the significance of a work. A comparison of earlier and current work is shown in Table III. In their proposed work, M. Ashraf et al. [10], S. Saqlain et al. [11], and P. Govindarajan et al. [12] employed different machine learning algorithms to identify cardiovascular disease, heart disease, and stroke disease, respectively. Even though their recommended models were executed effectively, their accuracy was inadequate. They also have a restricted dataset and only work with people in specific locations. On the other hand, our model exceeded theirs in terms of accuracy. We used a vast dataset with a wide range of people to create our model.

V. Conclusion

Identifying the risk of brain stroke with reasonable precision could significantly impact human long-term death rates, regardless of social or cultural background. Early detection is crucial to attaining that goal. Machine learning has already been used in various research to predict brain stroke. In this paper, we have followed a similar path, but with a better and more novel strategy and a larger dataset to train the model on. The utilized dataset consists of 5110 patients' observation issues with 12 attributes relevant to brain stroke. The image processing technique has been applied to make the dataset more adaptable to train and test the three classifiers. It can be observed that the Random Forest classifier performs exceptionally well with high-impact features and has a significantly higher accuracy which is 95.30% compared to other classifiers. In the future, We want to generalize the model with different feature selection algorithms and make it robust against datasets with a lot of missing data to improve the accuracy.

References

[1] World wide Stroke affect, WHO Available Link:http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html, Last Access:[6-10-2021].

[2] G. Vijayadeep and N. N. M. Rao, "A hybrid feature extraction based optimized random forest learning model for brain stroke prediction," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 11, no. 3, pp. 1152-1165, 2020.

[3] SJ. Park, I. Hussain, S. Hong, D. Kim, H. Park, and HC. Benjamin, "Real-time Gait Monitoring System for Consumer Stroke Prediction Service," IEEE International Conference on Consumer Electronics (ICCE), pp. 1-4, 2020.

[4] T. Badriyah, N. Sakinah, I. Syarif, and D. R. Syarif, "Machine Learning Algorithm for Stroke Disease Classification," International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1-5, 2020.

[5] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," Artificial intelligence in medicine, vol. 101, pp. 101723, 2019.

[6] G. Fang, W. Liu, and L. Wang, "A machine learning approach to select features important to stroke prognosis," Computational Biology and Chemistry, vol. 88, pp.107316, 2020.

[7] P. Govindarajan, RK. Soundarapandian, AH. Gandomi, R. Patan, P. Jayaraman, and R.Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817-828, 2020.

[8] M. Emon, M. Keya, T. Meghla, M. Rahman M, A. Mamun, and M. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, pp. 1464-1469, 2020.

[9] T. Shoily, T. Islam, and S. Jannat, "Detection of stroke disease using machine learning algorithms," 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-6, 2019.

[10] M. Amin, Y. Chiam and K. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol.36, pp.82-93, 2019.

[11] M. Ashraf, S. Ahmad, N. Ganai, R. Shah, M. Zamanand, and S. Khan, "Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies," An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS. International Conference on Innovative Computing and Communications, Springer, Singapore, pp. 239-255, 2021.

[12] S. Saqlain, M. Sher, F. Shah, I. Khan, M. Ashraf, M. Awais, and A. Ghani, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowledge and Information Systems, vol.58, no.1, pp.139-167, 2019.

[13] Dataset, Available Link:https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. Last Access:[06-10-2021].

[14] Decision Tree, Available Link:https://link.springer.com/book/10.1007/b107408. Last Access:[10-10-2021].

[15] Random Forest, Available Link:https://link.springer.com/article/10.1023/A:1010933404324 Last Access:[10-10-2021].

[16] Support Vector Machine, Available Link:https://link.springer.com/chapter/10.1007/0-387-25465-X_124 Last Access:[10-10-2021].

[17] M. M. Rahman, A. A. Biswas, A. Rajbongshi, and A. Majumder, "Recognition of Local Birds of Bangladesh using MobileNet and Inception-v3," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 309-316, 2020.

[18] A. Rajbongshi, T. Sarker, M. M. Ahamad, and M. M. Rahman, "Rose Diseases Recognition using MobileNet," In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-7. IEEE, 2020.

[19] J. I. Z. Chen, and P. Hengjinda. "Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study," Journal of Artificial Intelligence, vol. 3, no. 01, pp. 17-33, 2021.

[20] T. Vijayakumar,"Posed Inverse Problem Rectification Using Novel Deep Convolutional Neural Network," Journal of Innovative Image Processing (JIIP), vol. 2, no. 03, pp. 121-127, 2020.