IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

PREDICTION OF CEREBROVASCULAR ACCIDENT SEVERITY USING MACHINE LEARNING APPROACH

Monica J[1], Nischitha G[2], Poornima M[3], Mohammed Hidayath[4], Girish S C[5]

NIE Institute of Technology Mysore, Karnataka[1-4]

Corresponding Author, Assistant professor, NIE Institute of Technology, Mysore, Karnataka[5]

Abstract: Stroke is a medical condition in which the blood vessels in the brain rupture, causing brain damage. If the brain supply of blood and other nutrients is compromised, symptoms could develop. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. To forecast the likelihood of a stroke happening in the brain, many machine learning (ML) models have been developed. This study uses a variety of physiological characteristics and machine learning methods to train four different models for reliable prediction, including Decision Tree (DT) Classification, Random Forest (RF) Classification, SVM, K-Neighbor classifiers. The datasets downloaded from Kaggle website was used in the development of the approach. The accuracy of the models employed in this study is substantially greater than in earlier studies, showing that the models utilized in this study are more trustworthy. The scheme may be determined from the study analysis, which has been proven by numerous model comparisons.

Keywords: DT, RF, SVM, K-Neighbors, Resnet-34, Vgg-16, Densenet-121.

I. INTRODUCTION

Stroke is a significant health issue nowadays. It is also referred to as a cerebrovascular accident and is a neurological condition that can be brought on by brain ischemia (which may be temporary) or brain artery hemorrhage. It typically results in a range of functionally compromised motor and cognitive abilities. There are three different types of strokes: ischemic, which happens when blood clots form in the blood vessels, hemorrhagic, which occurs when a brain artery ruptures, and transient-ischemic, sometimes known as a "mini-stroke," which occurs when blood flow to the brain is temporarily interrupted. If someone has specific risk factors, their chance of getting a stroke rises. Blood pressure, WBC and RBC counts, heart disease, diabetes, smoking, alcohol use, and other risk factors for stroke can all be altered or treated. The severity of a hemorrhagic stroke lesion can be classified as mild (15–30 percent of the total blood volume), moderate (30–40 percent of the circulating blood volume), or severe (>40 percent of the circulating blood is lost).

Stroke causes a significant number of fatalities, and the rate is rising in emerging nations. It can significantly contribute to stroke prevention and early treatment if these risk factor predictions are properly established early and precisely. The relationship between these risk factors and different types of strokes can be better understood with the aid of predictive algorithms. Through early detection and treatment, the machine learning algorithm can enhance patients' health. With the use of a number of machine learning algorithms, we were able to determine from a patient's clinical report what kind of stroke the patient may have or has already experienced. We have obtained a dataset from Kaggle that consists of 13485 data records and 18 different features (risk factors), of which 12 characteristics are changeable risk factors.

On the other hand, a physical examination and medical imaging tests like a CT or MRI are often used to diagnose strokes. However, because it is so good at getting more precise images, an MRI scan is regarded as superior medical imaging. The majority of hospitals throughout the world still rely on doctors to diagnose medical images, which is expensive for the radiologist and hinders the patients' access to therapy and recovery opportunities. It is now possible to interpret MRI data intelligently thanks to a new study area called deep learning, which extracts characteristic values from training data samples to provide more sophisticated abstract features for classification, detection, and segmentation. We have obtained a dataset from Kaggle that includes a normal MRI scan with 524 data points and 3 stages of hemorrhagic stroke.

Following is our primary contribution to this project:

• Dataset that includes 18 risk variables

1. The pre-processing of the obtained text information involves using median values and the interpolate method to fill in any missing values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

2. Four different models, including the Random Forest Classifier, Decision Tree Classifier, SVM, and K-Neighbors Model, have been trained by our team.
3. To validate the models, tests and statistical accuracy calculations were made.
4. Examining how various risk factors affect a specific type of stroke.

• MRI images in the dataset

1. The number of data in the dataset is increased by applying augmentation techniques such as flipping.
2. Used three deep learning models to analyse augmented and non-augmented data.
3. Validated the models by testing their accuracy using augmented and nonaugmented datasets.

II. SYSTEM METHODOLOGY

Different datasets from Kaggle were taken into consideration in order to move on with the implementation. A suitable dataset was selected for model construction from all of the available datasets.The process of preparing the dataset once it has been collected is what must be done in order for the machine to understand it. Data pre-processing refers to this stage. In this step, the text data set's label encoding is performed along with feature extraction and management of missing values. Pre-processing in the instance of an MRI scan dataset entails augmentation, the removal of damaged images, and scaling to a dimension of 226 x 226.The data is available for model development after pre-processing. Pre-processed datasets and machine learning and deep learning algorithms are needed for the model creation. For the text dataset, classifiers including Random Forest Classifier, Decision Tree, SVM, and N-neighbors are employed, while the MRI scan data set uses pretrained deep learning models like Resnet-34, vgg-16, and Densenet (augmented and non-augmented). Following the development of machine learning models, the Accuracy Score, Precision Score, Recall Score, and F1 Score accuracy metrics are used to compare the models.
The proposed system's methodology's flow chart is shown in Fig.2. 1.

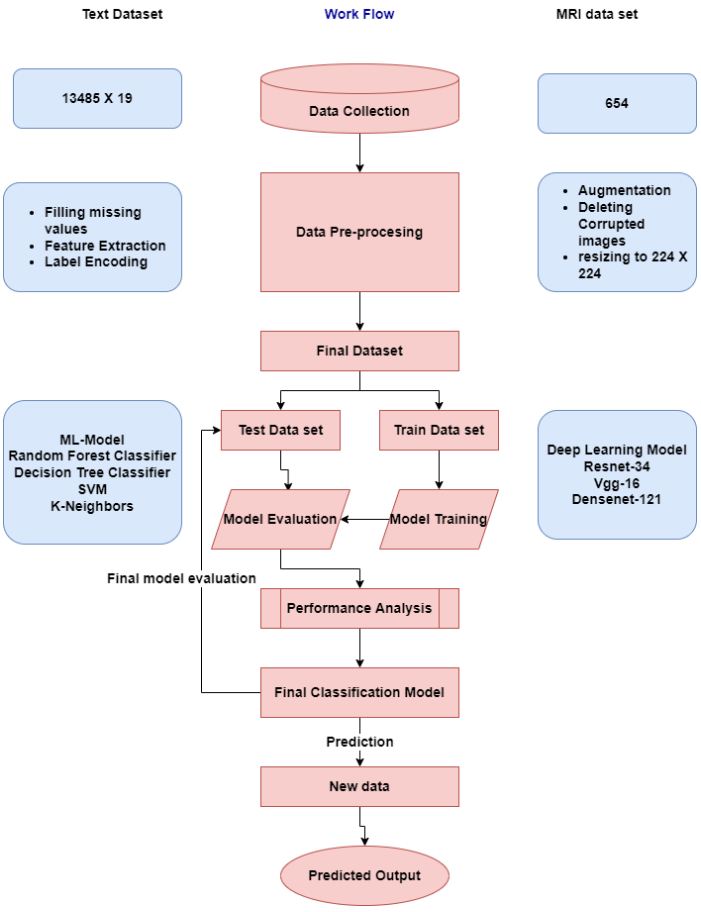

Fig. 2.1. Proposed system work flow

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

III. IMPLEMENTATION

A. DATA COLLECTION

A.1. TEXT DATASET

The Kaggle dataset is used to predict the kind of stroke. This particular dataset has 19 columns and 13485 rows, including the results. The result column's value might be "0," which denotes the absence of a stroke, "1," which denotes a transient ischemic stroke, "2," which denotes a hemorrhagic stroke, or "3," which denotes an ischemic stroke.

A.2 Dataset with MRI images:

We have obtained the MRI scanned image dataset from Kaggle, which contains 4 sets of data, i.e., normal, mild, moderate, and severe classes of hemorrhagic stroke, in order to determine the kind of hemorrhagic stroke. The dataset's sample may be shown in Fig 1. Fig. 2depicts the percentage of type of stroke samples.

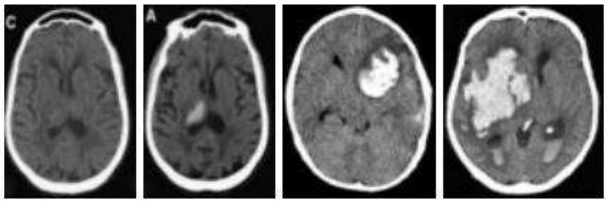

Fig 3.1: Sample MRI images

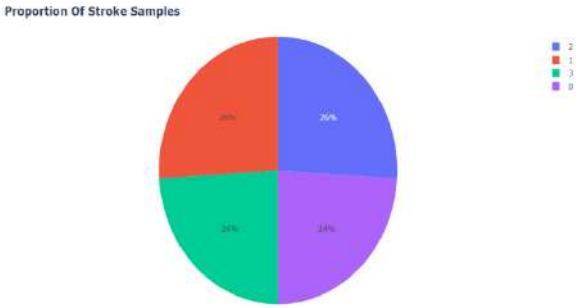

Fig 3. 2: Percentage of type of stroke data 24%(0), 26%(1),26%(2),24%(3)

B.PRE-PROCESSING OF THE DATA

Before developing a model, pre-processing is necessary to remove the dataset's undesired noise and outliers, which can cause the model to train improperly. This stage deals with anything preventing the model from operating more effectively. The process of cleaning the data and ensuring that it is prepared for model development comes after selecting the right dataset has been collected.

B.1. Image data set:

The dataset that was taken had the 19 characteristics listed in Table 1. The taken dataset is examined for empty fields. In this instance, as shown in Fig.4, the columns smoking status, white blood cell count, and red blood cell count all have null values. The interpolate method, which is essentially a function used to fill Null values in the data frame, is used to fill the missing values for the smoking status. The median approach, which fills the median of the column data, is used to fill in the missing values in the WBC and RBC counts. Fig 5 summarizes the missing values after handling missing values. For the machine to interpret the string literals, the label encoding transforms them into integer values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

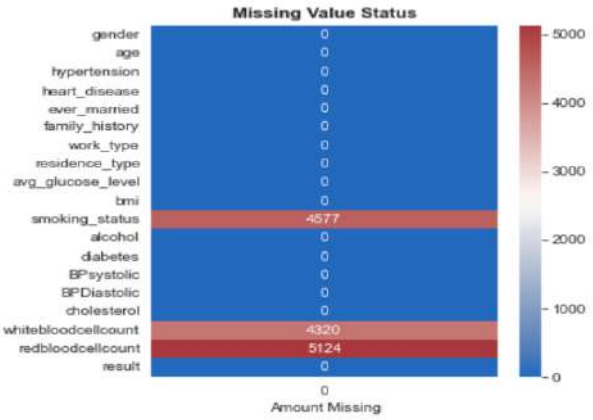

Fig 3. 3: Missing values before handling

Fig 3. 4: Missing values after handling

B.2 MRI scan dataset:

654 MRI pictures were gathered for the MRI scan dataset. Data augmentation is a technique that greatly broadens the variety of data that is accessible for training models. In order to train neural networks, techniques for enhancing the data are frequently used, such as cropping, padding, and flipping. The existing image dataset is expanded by this procedure from 654 to 1600 images. This allowed us to compare the performance of the two sets of data, one of which was supplemented and the other which was not, after the model had been trained. The corrupted photos are removed from the dataset of images used to train the model, and they are resized to have a dimension of 224 * 224. In Fig. 3.5, we can see a sample of.

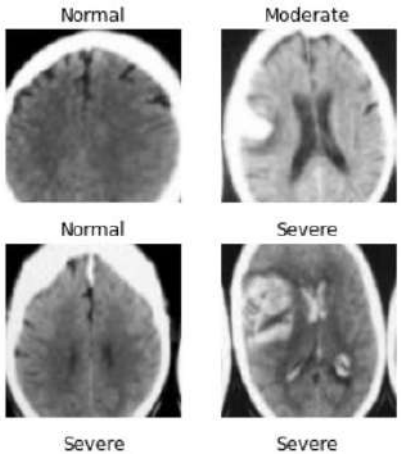

Fig 3.5: Samples of MRI Images

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

C.MODEL BUILDING

C.1 Data splitting:

The next step is developing the model after handling the imbalanced dataset and finishing data pre-processing. To improve accuracy and efficiency for this task, the under-sampled data is divided into training and testing data with the ratio remaining at 80% training data and 20% testing data. After splitting, the model is trained using a variety of classification algorithms. K-Neighbors classifier, Decision tree, Random Forest, and SVM are the classification algorithms employed for the text data set. The MRI scan(image) dataset uses Resnet-34, Vgg-16, and Densenet-121.

C.2 Classification Algorithm:

1. Random Forest: The k records in the data set are first split into n random records for Random Forest. A different decision tree is supplied to be constructed for each sample. Each decision tree results in a certain outcome. Averaging or majority voting are used to assess the final outcome for data classification and regression. The accuracy achieved after training this model is 100 percent. Utilizing many other accuracy indicators, such as precision score, recall score, and F1 score, one may also determine how effective the algorithm is. Each accuracy metric's value is equal to 100 percent in this score. The same, together with confusion matrices, can be shown in Fig.3.6.

```
Training Accuracy of Random Forest Classifier is 1.0
Test Accuracy of Random Forest Classifier is 1.0

Confusion Matrix :-
[[632   0   0   0]
 [  0 653   0   0]
 [  0   0 748   0]
 [  0   0   0 664]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       632
           1       1.00      1.00      1.00       653
           2       1.00      1.00      1.00       748
           3       1.00      1.00      1.00       664

    accuracy                           1.00      2697
   macro avg       1.00      1.00      1.00      2697
weighted avg       1.00      1.00      1.00      2697
```

Fig 3.6: Confusion metrics and various accuracy metrics score of Random Forest

2. Decision Tree: Because the root node of the tree contains the complete dataset, S suggests beginning there. Find the top attribute in the dataset using the Attribute Selection Measure (ASM). Divide S into subsets in order to identify possible values for the best attribute. Find the decision tree node with the best attribute and produce it. using the subgroups generated by the dataset, new decision trees were iteratively created. Continue in this manner until you are unable to further classify the nodes and must refer to the final node as a leaf node. The accuracy, precision, recall, and F1 score values for the used dataset were all 100% for this algorithm, as shown in Fig.3. 7.

```
Training Accuracy of Decision Tree Classifier is 1.0
Test Accuracy of Decision Tree Classifier is 1.0

Confusion Matrix :-
[[632   0   0   0]
 [  0 653   0   0]
 [  0   0 748   0]
 [  0   0   0 664]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       632
           1       1.00      1.00      1.00       653
           2       1.00      1.00      1.00       748
           3       1.00      1.00      1.00       664

    accuracy                           1.00      2697
   macro avg       1.00      1.00      1.00      2697
weighted avg       1.00      1.00      1.00      2697
```

Fig 3.7: Confusion metrics and various accuracy metrics score of Decision tree

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

3. SVM: By transforming the data into a high-dimensional feature space, SVM categorizes data points even when they cannot be separated linearly in other ways. The data are altered to enable the portrayal of the separator as a hyperplane once a separator between the categories has been found. Additionally, this algorithm achieved a 100% F1 Score, accuracy, precision, and recall as shown in the Fig 3.8.

```
Training Accuracy of SVM is 1.0
Test Accuracy of SVM is 1.0

Confusion Matrix :-
[[632   0   0   0]
 [  0 653   0   0]
 [  0   0 748   0]
 [  0   0   0 664]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       632
           1       1.00      1.00      1.00       653
           2       1.00      1.00      1.00       748
           3       1.00      1.00      1.00       664

    accuracy                           1.00      2697
   macro avg       1.00      1.00      1.00      2697
weighted avg       1.00      1.00      1.00      2697
```

Fig 3.8: Confusion metrics and various accuracy metrics score of SVM

4. K-Neighbor Classifier: Choose the neighbor in the Kth slot under the category "K - Neighbor." Determine the Euclidean separation between K neighbors Pick the K neighbors that are the closest according to the Euclidean distance calculation. Determine the number of data points in each category among these k neighbors. The category with the largest neighbor count should receive the additional data points. Fig.9 demonstrates that the obtained accuracy score, precision score, recall score, and F1 score are all 100%.

```
Training Accuracy of K-Neighbor classifier is 1.0
Test Accuracy of K-Neighbor classifier is 1.0

Confusion Matrix :-
[[632   0   0   0]
 [  0 653   0   0]
 [  0   0 748   0]
 [  0   0   0 664]]

Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       632
           1       1.00      1.00      1.00       653
           2       1.00      1.00      1.00       748
           3       1.00      1.00      1.00       664

    accuracy                           1.00      2697
   macro avg       1.00      1.00      1.00      2697
weighted avg       1.00      1.00      1.00      2697
```

Fig 3. 9: Confusion metrics and various accuracy metrics score of Decision tree

5.Resnet-34: Convolutional neural network with 34 layers is called Resnet34. It can be used as a model for picture classification. This model was previously trained using the ImageNet dataset, a sizable classification dataset. With almost 23 million trainable parameters, the model offers a deep architecture that improves image recognition. It is simple to train networks with several layers—even thousands—without raising the training error percentage. When we train the resnet-34 model with the non-augmented data, and augmented data, the obtained accuracy is 64% and 99.7% respectively. Fig 3.10 and Fig 3. 11 summarize the same.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.808019	1.885977	0.461538	0.538462	03:24
1	1.346459	1.290876	0.638462	0.361538	03:00
2	1.060441	1.081626	0.646154	0.353846	03:00

Fig 3. 10: Training resnet-34 with non-augmented data

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.491601	1.117839	0.620370	0.379630	06:29
1	0.896975	0.331567	0.925926	0.074074	06:34
2	0.603336	0.051703	0.987654	0.012346	06:33

Fig 3.11: training resnet-34 with augmented data

6.Vgg-16: VGG16 is the name given to 16 layers with weights. Any of the network configurations is thought of as having as its input a fixed 224 by 224 images with three channels-R, G, and B. The only pre-processing carried out is the normalization of each pixel's RGB values. After the image has been transmitted, the padding and convolution stride are both fixed at 1 pixel. The output activation map's size and spatial resolution are both maintained in this setup. In comparison to the accuracy of 80% with non-augmented data as shown in Figs. 3.12 and 3.13, we obtained the highest accuracy of 93 percent after training the Vgg-16 model.

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	2.096730	1.544774	0.376923	0.623077	10:25
1	1.634502	0.860445	0.684615	0.315385	10:31
2	1.294742	0.603686	0.800000	0.200000	10:33

Fig 3. 12: Training Vgg-16 with non-augmented data

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.539041	0.701021	0.709877	0.290123	28:03
1	0.953762	0.214099	0.910494	0.089506	28:09
2	0.662319	0.167777	0.938272	0.061728	28:16

Fig 3.13: Training Vgg-16 with Augmented data

7. Densenet 121: Dense Net is a CNN in which every layer is connected to every layer below it. It promotes the reuse of features. The number of parameters is significantly down. All layers are related to one another directly by dense blocks. fixes the vanishing gradient issue. When this model is trained, the non-supplemented data yields an accuracy of just 72% compared to the 98.4% accuracy obtained from the augmented data of MRI scans. Figs. 3. 14 and 3.15 show the model's accuracy, error rate, train loss, etc.

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.941158	2.129205	0.384615	0.615385	04:37
1	1.289455	1.099026	0.546154	0.453846	04:42
2	0.970828	0.807137	0.723077	0.276923	04:38

Fig 3.14: Training Densenet-121 with non-augmented data

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.339776	1.063902	0.666667	0.333333	09:58
1	0.720915	0.332542	0.907407	0.092593	09:56
2	0.446950	0.044996	0.984568	0.015432	09:53

Fig 3.15: Training Densenet-121 with augmented data

IV. RESULTS AND DISCUSSION

All three deep learning models outperform the non-augmented dataset when using the augmented dataset after training on the MRI scan picture dataset. The best accuracy in non-augmented data is provided by Resnet 34. However, because we used a static dataset for the text dataset, the accuracy of each model is the same. Therefore, we are unable to choose the optimum model. Tables 2 and 3 provide an overview of the accuracy results from several deep learning and machine learning models

TABLE 2

Model Name	Accuracy(%) of Augmented dataset	Accuracy(%) Of Non-Augmented dataset	
Resnet-34	98.7	64	
Vgg-16	93.9	80	
Densenet-121	98.4	72	

TABLE 3

ML model	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
RF	100	100	100	100
DT	100	100	100	100
SVM	100	100	100	100
K-N	100	100	100	100

V. CONCLUSION

Stroke is a serious medical illness that needs to be addressed right away. The development of deep learning and machine learning models can aid in the early detection of stroke subtypes and help to lessen the severity of future effects. The success of several machine learning and deep learning models in successfully predicting the stroke subtype based on various physiological issues and by using MRI scan images is demonstrated in this paper. The samples of the prediction are depicted in Figures below.

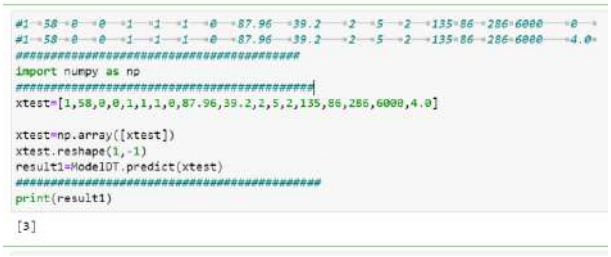

Fig 5.1: Testing result

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116133

Fig 5.2: Selecting file for testing

```
pred,idx,outputs = learn.predict(img)
print(pred)
result=int(pred)
print(result)
```

```
moderate
3
```

Fig 5.3: Testing Result

REFERENCES

[1] Analyzing the Performance of Stroke Prediction using ML Classification Algorithms (2021)
[2] Prediction of Stroke Lesion at 90-Day Follow-Up by Fusing Raw DSC-MRI With Para Parametric Maps Using Deep Learning (2021)
[3] Intracranial Hemorrhage Detection in Head CT Using Double-Branch Convolutional Neural Network, Support Vector Machine, and Random Forest (2020)
[4] Stroke Detection and Dating from FLAIR MRI Scans (2021)
[5] Combining unsupervised and supervised learning for predicting the final stroke lesion (2021)
[6] Image Thresholding Improves 3-Dimensional Convolutional Neural Network Diagnosis of Different Acute Brain Hemorrhages on Computed Tomography Scans (2019).