2021 24th International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2021

Bangla Optical Character Recognition (OCR) Using Deep Learning Based Image Classification Algorithms

Nadim Mahmud Dipu
Dept. of Electrical and Computer Engineering
North South University
nadim.dipu@northsouth.edu

Sifatul Alam Shohan
Dept. of Electrical and Computer Engineering
North South University
sifatul.shohan@northsouth.edu

K. M. A. Salam
Dept. of Electrical and Computer Engineering
North South University
kazi.salam@northsouth.edu

2021 24th International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-9435-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIT54785.2021.9689864

***Abstract*—Optical Character Recognition (OCR) refers to the process of converting images of printed, typed, or handwritten text into machine-readable text. OCR is one of the most widely researched topics in the field of computer vision. Furthermore, highly accurate, and sophisticated Optical Character Recognition systems have been built for most of the major languages of the world such as English, French, German, Mandarin, etc. However, despite having 300 million native speakers (4.00% of the world population) and being the 5th most spoken language of the world, the Bengali language still does not have a state-of-the-art OCR system. Moreover, most of the existing systems are not able to recognize compound letters. This study strives to resolve this issue by proposing three neural network based image classification models for Bangla OCR. These models are Inception V3, VGG-16, and Vision Transformer. These models have been trained on the BanglaLekha-Isolated dataset that contains 98,950 images of Bengali characters (vowels, consonants, digits, compound letters). The accuracy provided by the VGG-16, Inception V3, and Vision Transformer on the test set are 98.65%, 97.82%, and 96.88% respectively. Each of these models is much more accurate than the existing systems. Real-time implementation of these three models will be instrumental in building a state-of-the-art Bangla OCR system.**

***Keywords*—Deep Learning, Bangla OCR, Optical Character Recognition, OCR, CNN, Inception V3, VGG-16, Vision Transformer, Image Classification**

I. Introduction

Convolutional Neural Network based deep learning models have become quite popular in the past decade and these models have been extensively used in computer vision [1-3]. Furthermore, Object Character Recognition or OCR is a crucial topic in computer vision and deep learning. That is why a lot of research has been conducted on this topic and various cutting-edge OCR models have been proposed by those studies [4][5]. Several scientific studies have been conducted on the world's most popular and widely spoken languages such as English [6], Mandarin [7], Spanish, French [8], etc.

However, that is not the case when it comes to Bengali. Although a decent amount of research papers have been published on Bangla Optical Character Recognition, most of those studies were carried out on Bengali handwritten digits and simple vowels and consonants [10][11][19]. Most of those studies do not address the existence of compound letters and a lot of the existing models are not accurate enough for real-time use.

978-1-6654-9435-9/21/$31.00 ©2021 IEEE

These drawbacks are caused by the lack of a substantial dataset containing Bengali characters, the complexity and shape of the Bengali letters, and the geometric similarity of some of the characters. Nonetheless, these issues can be resolved due to the advent of cutting-edge, highly efficient image classification models like Inception V3 and Vision Transformer. Moreover, large-scale Bengali handwritten character datasets such as BanglaLekha-Isolated [12], Bangla.Ai, Ekush, CAMTERdb, ISI, etc. have been published in the past few years.

U. Garain et al. [20] proposed a character segmentation algorithm based on fuzzy multi-factorial analysis that is capable of distinguishing between touching characters. This study was done on Bengali and Devnagari.

P. Mahmud et al. [16] suggested using fuzzy logic in order to build an online OCR system. The goal of the proposed model is to recognize the curvaceous handwritten Bengali characters without using a lot of computational power.

M. A. Imran et al. [17] proposed a novel approach for Bangla OCR using the Outer Shape Detection Technique or OSDT that can detect the Bengali characters from scanned images based on outer shape of the character. This process involved image filtering, gray-scaling the images, binary conversion, segmentation, shape detection, and decision making.

T. Ahmed et al. [18] have used a Convolutional Neural Network based model that can detect and segment individual Bengali characters from images. The accuracy of the model was improved by using various techniques such as noise reduction and binarization.

Rabby et al. [19] developed a CNN-based optical character recognizer that was trained on three different datasets including BanglaLekha-Isolated, ISI, CAMTERdb. This model achieved $\mathbf{96.40}\%$ accuracy on the BanglaLekha dataset.

M. A. Hasnat et al. [21] built a Bengali optical character recognition system by integrating the open-source BanglaOCR software and Tesseract [13], which is a powerful OCR engine created and maintained by Google.

Although each and every one of these studies that we have discussed so far was innovative and they tried to create optical character recognition systems that can produce accurate results, they have certain limitations that prevent these models from being suitable for real-time use.

In order to resolve these issues and limitations, this proposes study proposes a slightly different approach than the traditional machine learning based or Convolutional Neural Network (CNN) based models. We have decided to utilize highly efficient and scalable, advanced neural networked based image classification algorithms that are trained on handwritten Bengali character images taken from the BanglaLekha-Isolated dataset. The objective of this study is to determine the performance of these algorithms in Bangla optical character recognition. In order to achieve this objective, we have utilized several techniques to reduce the computational time required for training the models and improving the accuracy of our algorithms. Furthermore, we have attempted to address several technical challenges such as lack of suitable training data, insufficient number of validation data, over-fitting, etc.

II. Materials and Method

This study is designed to investigate the performance of deep-learning based image classification frameworks/algorithms when they are used in Bangla OCR. The key features of this study are: (1) A robust and precise method for training Convolutional Neural Networks, (2) A suitable preprocessed classification dataset, (3) An implementation of these high-performance image classification algorithms on handwritten Bengali optical characters. The classification

Fig. 1. A look at some sample images of thN-Bengali alphabet and numerals taken from the BanglaLekha-Isolated dataset.

models used in this study are Inception V3 [14], VGG-16 [15], and Vision Transformer [9]. These three algorithms possess superior neural network architectures compared to traditional machine learning based or CNN-based models. In this section, we will examine the dataset as well as the image classification models used in this paper.

All three of these models were trained using the computational power of an NVIDIA GeForce GTX 1050 GPU. Since the models were run on a Graphics Processing Unit (GPU) instead of a Central Processing Unit (CPU), the training time was substantially reduced.

A. Dataset

As mentioned before we have used the BanglaLekha-Isolated dataset [12] in this study. This dataset comprises of 84

Fig. 2. Some sample images Bengali compound characters.

different characters that consist of 50 Bengali simple alphabet, 24 compound characters, and 10 digits. Each of these 84 characters have 2000 different samples associated with them. All of these samples were manually collected by taking handwriting samples from 2000 different individuals. These samples are preprocessed and digitized for being used in optical character recognition. In total this dataset contains 1,66,105 Bengali character images. This paper will demonstrate the performance of these different deep-learning methods when applied to handwritten Bengali optical characters that are contained in this dataset. Figure 1 illustrates some sample images of the Bengali vowels, consonants, and digits obtained from the BanglaLekha-Isolated dataset. And some sample images of the compound characters are shown in figure 2.

After obtaining the dataset, we applied image transformation techniques in order to preprocess the data. This was done in order to decrease the training time taken by our deep learning models and to enhance the performance of these models.

B. Dataset Preprocessing and Augmentation

Image augmentation refers to the process of artificially increasing the available data in the dataset in order to train a machine learning or deep learning model. The quality and accuracy of a deep learning based image classification model entirely depend on the size and diversity of the dataset that is used to train it. In order to enhance the generalizability of our model's performance, we went through various augmentation steps that added more learning examples for our model.

Each of the images was resized to 416x416 pixels and then they were augmented. The preprocessing and augmentation steps that we have taken are described below:

1) **Flip:** As the name suggests this technique flips the image horizontally and vertically in order to generate two new images. This process is done by flipping the NumPy array of the image from left to right (horizontal flip) or from up to down (vertical flip).

2) **45° Rotation:** The method rotates an image 45° to the clockwise direction and 45° to the counter-clockwise direction and generates two new augmented images.
3) **Exposure/Brightness:** We have generated new images by adjusting the gamma exposure of the image in order to make it brighter and darker.
4) **Blur:** This method utilizes the Gaussian Blur technique to generate a blurred version of the given image.
5) **Random Noise:** We altered $\mathbf{10}\%$ of the pixels in an image and changed it to black and white speckles.

All of these preprocessing and image augmentation steps were performed using the Keras deep learning library [23] of Python. Various image augmentation tasks such as flip, rotation, changing brightness, adding noise, etc. can be performed using the ImageDataGenerator class provided by the Keras library. An illustration of these preprocessing and augmentation methods can be seen in figure 3. After the new augmented images were generated, our dataset contained 498,315 images.

The pre-processed and augmented dataset was split into three parts of training, validation, and testing. We assigned $\mathbf{80}\%$, of the images to the training set, $\mathbf{20}\%$, to the validation set, and $\mathbf{10}\%$, to the test set.

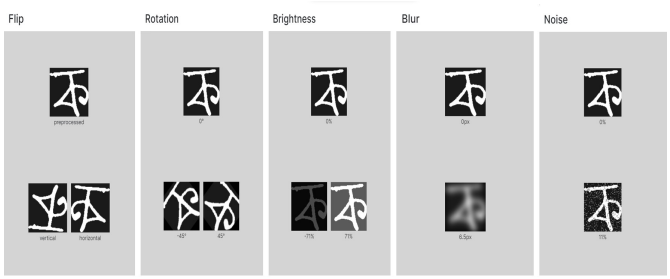

Fig. 3. Illustration of the data pre-processing and augmentation steps that were taken while preparing the dataset for training.

C. Inception V3

Inception v3 is one of the most widely-recognized image recognition algorithms in the world. This model was able to attain an accuracy of $\mathbf{78.1}\%$ on the ImageNet dataset. The model comprises convolutional layers, max pooling, concats, average pooling, fully connected neural networks, and dropouts. This model uses Softmax as its loss function and it applies Bathnorms to its activation layers.

Inception V3 utilizes a highly effective technique known as transfer learning where a pre-trained model is used to further train the model on a custom dataset. This method increases the accuracy of the model exponentially. In particular, the Inception V3 architecture is a pre-trained classification model made by Google AI that is trained on the benchmark ImageNet [22] dataset. This convolutional neural network based model is comprised of 48 layers. Moreover, this CNN model was already trained on more than 14 million images, divided into 1000 different object classes taken from the ImageNet dataset. That is why this is capable of recognizing the intricate features of different types of images.

The architecture of the Inception V3 is illustrated in figure number 4. It takes input images of size 299x299. The first part of the model is used to extract the features from the images and the second part is responsible for classifying the image based on the extracted features. We had to modify the Inception V3 model provided by the Keras library in order to train it on our dataset that contains 84 different classes.

At first, we traversed through the various layers of the Inception V3 model and set their trainable parameter to false. This was done so that the model becomes suitable for further training on our custom BanglaLekha-Isolated dataset. Next, we configured the model's parameters and metrics. For instance, we choose "categorical_entropy" to be the loss function, and we set the optimizer to be the "Adam" optimizer. Finally, we set the metrics for evaluation of the model's performance to be "accuracy".

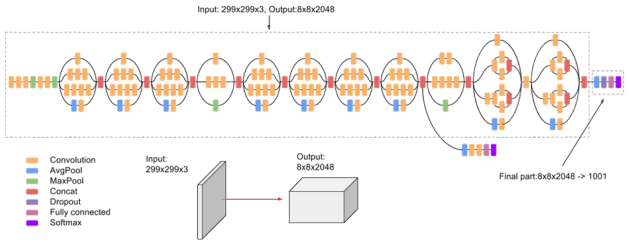

Fig. 4. The architecture of the Inception V3 model.

Since a transfer learning model such as Inception V3 reuses the features extracted from the original dataset, our classification model did not have to spend a lot of time extracting features from our input images. That is why this model did not require that much computational time and resources.

D. VGG-16

Just like Inception V3, VGG-16 is also a convolutional neural network based model that was proposed by the paper titled "Very Deep Convolutional Networks for Large-Scale Image Recognition" [15]. The model was able to achieve $\mathbf{92.70}\%$ accuracy on the ImageNet dataset. This model is considered to be an improvement over the AlexNet model as it replaces large kernel-shaped filters with multiple, tiny 3×3 kernel-shaped filters that are connected with each other. The VGG-16 model has been per-trained for multiple weeks using Titan Black GPUs made by NVIDIA. This model has

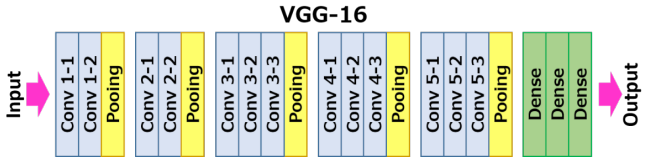

Fig. 5. Illustration of the architecture of the VGG-16 model.

a complex architecture and it was further modified in order to make it ideal for being used as a Bangla optical character classifier. The input layer of the VGG-16 model is composed

of a cov1 layer that only takes 224x224 RGB type images. This model has a complex architecture and it was further modified in order to make it ideal for being used as a Bangla optical character classifier. The most interesting aspect of the VGG-16 model is that it does not use a huge number of hyper-parameter, it utilizes multiple 3x3 and max-pool layers of 2x2 filter. Such an arrangement is consistent throughout the model's structure. At the end of the architecture, there exist two fully connected layers. This model is called VGG-16 because it has 16 layers that contain weights. The whole architecture of the VGG-16 model is shown in figure 5. The only drawback of the VGG-16 model is that it requires more computational power to train than other conventional classification models. The reason for that is VGG-16 has a complex architecture and it has 138 million parameters. Despite of that disadvantage, this model performed remarkably well in Bangla optical character recognition.

E. Vision Transformer

The Vision Transformer refers to the classification model that is based on the natural language processing embeddings taken from the BERT architecture. It utilizes the powerful BERT architecture in order to classify images.

The steps of implementing the Vision Transformer model on the BanglaLekha dataset is described below:

1) At first, the model splits the images into smaller patches (fixed-sized).
2) After that, the two-dimensional images get flattened into linear 1D embeddings. This is done using a fully connected layer.
3) Next, the positional embedding gets mixed with the patch embedding that the model had received from the previous step.
4) The encoded vectors are then passed to the Transformer Encoder as its input.
5) The Transformer Encoder is comprised of multiple layers of MLP blocks and self-attention layers. These layers get connected to LN or LayerNorm.
6) The classification task is performed by the MLP head and a linear layer is used for fine-tuning the results.

This whole process is illustrated in figure number 6.

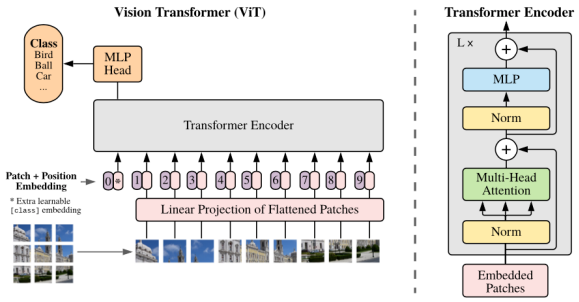

Fig. 6. Vision Transformer model architecture.

III. Experimental Results

The principal goal of this study is to detect, classify and recognize Bangla optical characters from images. The three models that we have discussed in the previous section were trained on the training set that we had extracted from the BanglaLekha dataset. After the successful completion of the training phase, we ran our model on the previously unseen images of the test set.

In order to perform a qualitative evaluation of the three image classification models, we used "accuracy" to be the defining metric of the model's performance. This metric was also used to compare our models with the models that had been proposed in previous studies.

$$Accuracy = \frac{NumberofCorrectPrediction}{TotalNumberofPredictions} \quad (1)$$

A. Determining Performance of Models

TABLE I
Performance of the Classification Models

Model	Training Accuracy	Test Accuracy
Inception V3	**98.77%**	**97.82%**
VGG-16	**99.23%**	**98.65%**
Vision Transformer	**97.56%**	**96.88%**

The performance of the model on the train and test sets are illustrated in table I. Each of these models was trained for 250 epochs on the same set of images in order to determine which one was the best. And as we can clearly see from the results the VGG-16 model was the most accurate as it received an accuracy of **98.65%** on the test set. In contrast, the Vision Transformer model had the worst performance out of the three algorithms, with an accuracy score of **96.88%**. Each

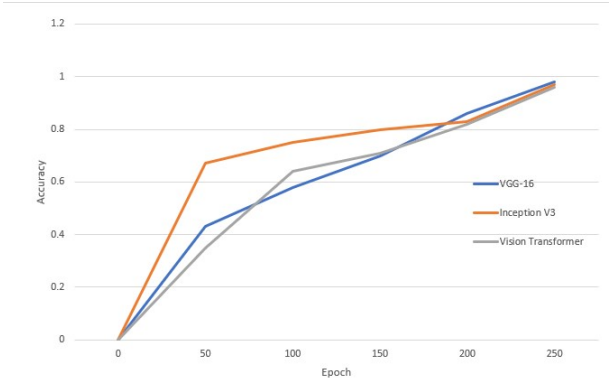

Fig. 7. Accuracy Vs. Epoch graph of the three image classification models.

of these models were trained for 250 epochs. The "Accuracy Vs. Epoch" graph illustrates how the training accuracy of the Inception V3, VGG-16, and Vision Transformer increases with the number of epochs.

Table II illustrates a comparison between the accuracy of our best model which is VGG-16 and the accuracy of the models

and algorithms that were proposed by previous studies. As we can clearly observe from the table, our VGG-16 model outperformed the previously established models.

TABLE II
COMPARISON BETWEEN OUR VGG-16 MODEL AND A FEW OTHER STUDIES

Author	Methodology	Accuracy
Rahman et al. [24]	CNN	$\mathbf{85.96}\%$
Halima Begum et al. [25]	Gabor Filter and ANN	$\mathbf{79.40}\%$
P. Mahmud [16]	Fuzzy Logic	$\mathbf{72.00}\%$
Rabby et al. [19]	Novel CNN	$\mathbf{95.71}\%$
Our Best Model	VGG-16	$\mathbf{98.65}\%$

IV. CONCLUSION

This study presented three different convolutional neural network based image classification models that were trained and tested on the BanglaLekha-Isolated dataset (divided into 84 classes) with the sole purpose of accurately recognizing Bangla optical characters in images. Each of the classification models is exceptionally fast and accurate. Out of the three deep learning models VGG-16 gained the highest accuracy of $\mathbf{98.65}\%$, and it will be an excellent alternative to the traditional machine learning and CNN or ANN models. Furthermore, this model will play a vital role in the construction of a stat- of-the-art, highly efficient Bangla OCR system. Such a system will have many useful applications such as scanning text from handwritten documents, archiving old Bengali books, reading license plates written in Bengali from vehicle images.

The reason behind the success of this study is that we augmented our dataset and generated new instances of the handwritten Bengali optical characters. Moreover, we were able to build three exceptionally fast models that did not require a lot of computation power because we had implemented a transfer learning technique by utilizing pre-trained deep learning models.

REFERENCES

[1] Xu, L., Ren, J. S., Liu, C., & Jia, J. (2014). Deep convolutional neural network for image deconvolution. Advances in neural information processing systems, 27, 1790-1798.

[2] Guo, T., Dong, J., Li, H., & Gao, Y. (2017, March). Simple convolutional neural network on image classification. In 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA) (pp. 721-724). IEEE.

[3] Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2018). A new deep convolutional neural network for fast hyperspectral image classification. ISPRS journal of photogrammetry and remote sensing, 145, 120-147.

[4] Wick, M., Ross, M., & Learned-Miller, E. (2007, September). Context-sensitive error correction: Using topic models to improve OCR. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 1168-1172). IEEE.

[5] Smith, R. (2011, September). Limits on the application of frequency-based language models to OCR. In 2011 International Conference on Document Analysis and Recognition (pp. 538-542). IEEE.

[6] Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013, August). High-performance OCR for printed English and Fraktur using LSTM networks. In 2013 12th international conference on document analysis and recognition (pp. 683-687). IEEE.

[7] Huo, Q., Ge, Y., & Feng, Z. D. (2001, May). High performance Chinese OCR based on Gabor features, discriminative feature extraction and model training. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 3, pp. 1517-1520). IEEE.

[8] Clematide, S., Furrer, L., & Volk, M. (2016, May). Crowdsourcing an OCR gold standard for a German and French heritage corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 975-982).

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.

[10] Pramanik, R., & Bag, S. (2018). Shape decomposition-based handwritten compound character recognition for Bangla OCR. Journal of Visual Communication and Image Representation, 50, 123-134.

[11] Omee, F. Y., Himel, S. S., Bikas, M., & Naser, A. (2012). A complete workflow for development of Bangla OCR. arXiv preprint arXiv:1204.1198.

[12] Biswas, M., Islam, R., Shom, G. K., Shopon, M., Mohammed, N., Momen, S., & Abedin, A. (2017). Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters. Data in brief, 12, 103-107.

[13] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

[14] Xia, X., Xu, C., & Nan, B. (2017, June). Inception-v3 for flower classification. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 783-787). IEEE.

[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[16] P. Mahmud, M. R. Rahman, M. J. Islam, R. M. Rahman and M. A. Matin, "Ankur: Bangla online character recognition," 5th Brunei International Conference on Engineering and Technology (BICET 2014), 2014, pp. 1-6, doi: 10.1049/cp.2014.1116.

[17] M. A. Imran, J. Hossain, T. Dey, B. K. Debroy and A. H. Abir, "OSDT: Outer Shape Detection Technique for Recognition of Bangla Optical Character," 2009 12th International Conference on Computers and Information Technology, 2009, pp. 384-389, doi: 10.1109/ICCIT.2009.5407268.

[18] T. Ahmed, M. N. Raihan, R. Kushol and M. S. Salekin, "A Complete Bangla Optical Character Recognition System: An Effective Approach," 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1-7, doi: 10.1109/ICCIT48885.2019.9038551.

[19] Rabby, A. S., Haque, S., Islam, S., Abujar, S., amp; Hossain, S. A. (2018). BornoNet: Bangla handwritten characters recognition using convolutional neural network. Procedia Computer Science, 143, 528–535. https://doi.org/10.1016/j.procs.2018.10.426

[20] Garain, U., & Chaudhuri, B. B. (2002). Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 32(4), 449-459.

[21] Hasnat, M. A., Chowdhury, M. R., & Khan, M. (2009, July). An open source tesseract based optical character recognizer for bangla script. In 2009 10th International Conference on Document Analysis and Recognition (pp. 671-675). IEEE.

[22] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[23] Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd. Begum, H., Islam, M. (2017). Recognition of Handwritten Bangla Characters using Gabor Filter and Artificial Neural Network. et al. [25] SegNet

[24] Rahman, Md Akhand, M. A. H. Islam, Shahidul Shill, Pintu Rahman, M. M.. (2015). Bangla Handwritten Character Recognition using Convolutional Neural Network. International Journal of Image, Graphics and Signal Processing (IJIGSP). 7. 42-49. 10.5815/ijigsp.2015.08.05.

[25] Begum, H., Islam, M. (2017). Recognition of Handwritten Bangla Characters using Gabor Filter and Artificial Neural Network.