Stacked Auto Encoder Training incorporating Printed Text Data for Handwritten Bangla Numeral Recognition

Mahtab Ahmed, Animesh Kumar Paul, M. A. H. Akhand
Dept. of Computer Science and Engineering
Khulna University of Engineering & Technology, Khulna, Bangladesh

Abstract—**Recognition of handwritten numerals has gained much interest in recent years due to its various application potentials. Bangla is a major language in Indian subcontinent and is the first language of Bangladesh; but unfortunately, study regarding handwritten Bangla numeral recognition (HBNR) is very few with respect to other major languages such as English, Roman etc. Some noteworthy research works have been conducted for recognition of Bangla handwritten numeral using artificial neural network (ANN) as ANN and its various updated models are found to be efficient for classification task. The aim of this study is to develop a better HBNR system and hence investigated deep architecture of stacked auto encoder (SAE) incorporating printed text (SAEPT) method. SAE is a variant of neural networks (NNs) and is applied efficiently for hierarchical feature extraction from its input. The proposed SAEPT contains the encoding of handwritten numeral into printed form in the course of pre-training and finally initializing a multi-layer perceptron (MLP) using these pre-trained weights. Unlike other methods, it does not employ any feature extraction technique. Benchmark dataset with 22000 hand written numerals with different shapes, sizes and variations are used in this study. The proposed method is shown to outperform other prominent existing methods achieving satisfactory recognition accuracy.**

Keywords—Bangla Handwritten Numeral; Stacked Auto Encoder; printed text; Deep Neural Networks.

I. INTRODUCTION

Handwritten numeral recognition is widely considered as a classical pattern recognition and artificial intelligence problem all around the globe. The researchers get motivations on this research due to its various applications such as automation of the postal system, processing bank cheque automatically, verification and identification of the writer, analysis of passports and documents and identification of number plate etc. Recognition techniques can be applied on machine printed numeral and handwritten numeral. Recognition of handwritten numeral is more difficult than printed numeral because handwritten numeral are non-uniform because of the inconsistency of the writing style of different persons. Despite its high challenges, various research has made impressive progress in Roman, Chinese and Arabic script [1, 2, 3]. But a few researches have been done in the field of Bangla (Indian scripts) although Bangla is the major language in the Indian subcontinent [4] and the national language in Bangladesh.

Some noteworthy research works have been conducted for recognition of Bangla handwritten numeral using artificial neural network (ANN). Khan et al. [5] utilized evolutionary approach while training ANN with handwritten numeral data in Bengali language. At first, they used boundary extraction for extracting numeral in a single window by horizontal and vertical scanning; and scaled the image into fixed sized matrix. Then Multi-Layer Perceptron (MLP) are developed for recognition. Basu et al. [3] used Dempster-Shafer (DS) technique where they combined the classification decisions of two MLP based classifiers for handwritten Bangla numeral using two different feature sets. They investigated two feature sets called shadow feature and centroid feature.

Bhattacharya and Chaudhuri [6] presented a multistage cascaded recognition scheme using MLP classifiers. In their scheme, they used a majority voting concept using multi revolutionary representative wavelet process and MLP classifiers. First they computed features using wavelet-filtered image at different resolutions and then used a cascade of three MLP classifiers. Pal et al. [2] proposed a technique which contains segmentation and recognition parts. In the segmentation part, they extract the feature using the technique of water overflow from the reservoir. Finally, they used binary tree classifier for recognition of digits.

Wen et al. [5] proposed a handwritten Bangla numeral recognition system for automatic letter sorting machine based on image reconstruction and feature extraction techniques with principal component analysis (PCA) and kernel PCA (KPCA) . Finally they used Support Vector Machine (SVM) for the classification purpose. Most recently, Nasir and Uddin [8] inaugurated a hybrid system of Bangla handwritten numeral recognition for automated postal system, which uses k-means clustering, Baye’s theorem and Maximum a Posteriori for feature extraction. Finally they employed SVM for the recognition purpose.

Recently, deep neural networks (DNNs) have been found to be very potential in pattern classification imitating the underlying features of the data through its deep architecture. Among them, deep recurrent neural network (RNN) has shown better numeral recognition accuracy than tradition NN based methods. Oval et al. [15] proposed a method for recognizing Devanagari words. At first they employed Stentiford algorithm as thinning techniques and then trained the network using the extracted feature of the image. Finally they assigned level to the words considering the bidirectional dependencies and employed recurrent neural network for classification.

A few works of Bangla handwritten numeral recognition are also available which uses different techniques rather than

ANN. Bashar et al. [10] proposed a histogram based technique for recognition of the Bangla numeral digit. At first they extracted features of a given input image using a windowing technique and generated histogram from the feature. Finally, they compares the recognizing histogram with the reference histogram for making decisions about the recognition of numeral. Recently, Wen and He [11] proposed a method to recognize handwritten Bangla numeral based on kernel and Bayesian Discriminant. Das et al. [15] proposed a technique in which they select the optimal feature set with discriminating features using Genetic algorithm (GA). Finally they use support vector machine (SVM) for recognition process on the Bangla handwritten digits. Suranta et al. [16] introduced a contour angular technique to capture the curvature of the handwritten image and finally used a nonlinear SVM classifier for classification.

The aim of this study is to develop a better Bangla handwritten numeral recognition system and hence investigated deep architecture of auto encoder (AE) based technique. AE is a simple feed forward NN except that the training label is exactly or noisy version of input. It is mainly used for reconstructing the original input in the course of extracting hierarchical features which are then used for training any classifier. Stacking a bunch of single AEs gives stacked AE (SAE), a deep architecture which subsequently extracts features within features and finally ends up reducing a high dimensional problem into a low one. The SAE concept used in this work consists of two NNs where the first one is trained with raw handwritten data having corresponding printed data as output and the second one is trained by printed data having corresponding label as output. Finally, these two NNs are combined. In order to highlight the significance of SAE incorporating printed text, performance is compared with traditional SAE and feed forward neural net tested on the same benchmark dataset. Experimental results reveal that the proposed method shows satisfactory results achieving higher accuracy in very less number of training steps without any fine-tuning and outperformed some other existing methods.

The rest of the paper is organized as follows. Section II explains proposed SAE based handwritten Bangla numeral recognition. For better understanding the section also gives brief description of traditional MLP and SAE. Section III presents experimental results of the proposed method and performance comparison with other related works. Finally, a brief conclusion of the work is given in Section IV.

II. SAE Training Incorporating Printed Text (SAEPT) for Handwritten Bangla Numeral Recognition

Handwritten numeral recognition (HNR) is a high-dimensional complex classification task. A number of neural network based techniques have been investigated for this purpose as it is found to be effective for classification task. Although NN may classify directly from pixel values, classification from extracted features is also investigated and found to be performed better in some cases. This section first explains the HBNR using traditional MLP and SAE and then presents proposed SAE training with printed text data.

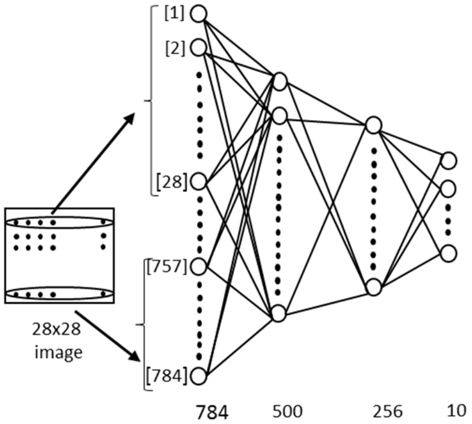


Fig.1. Structure of the NN considered in this study

A. Classification using MLP

MLP is the pioneer structure of NN which is found to be very effective for classification task. It has input layer, output layer and several layers of hidden layers. For HNR, input layer takes image pixels and outpu layer contains 10 neuron to identify class label. But this architecture suffers from vanishing gradient problem and consists of very large amount of network parameters to handle when the raw image pixels are directly fed to the network as input.

A NN architecture considered in this study as [784-500-256-10] is shown in Fig. 1 for classifying the Bangla handwritten numerals. It consists of an input layer, two hidden layers and an output layer. The input layer consists of 784 input units which takes the linear representation of the 28x28 sized handwritten image. For doing this, the first 28 units of input layer are the first row of the image, next 28 units are the second row and so on. The hidden layers consists of 500 and 256 hidden units respectively. Finally the output layer consists of 10 output units similar to the number of class labels. Additionally the network have feed forward connection between the layers and the layers are separated by sigmoid activation function. For classification purpose the output layer consists of softmax activation function. Finally during training, the network parameters are updated using the traditional voila back-propagation algorithm.

B. Classification using SAE

A SAE is formed by stacking a number of auto encoders (AEs) extracting features within features consecutively. AEs are generative networks, which can be used to learn the reconstruction of input data at the output layer. These networks use unsupervised learning algorithm, since, class labels are not present during training. Generally, an AE can be considered as having an input layer, a hidden layer, which learns the compressed representation of input layer attributes (i.e. as in encoding), and an output layer which contains expanded form of the hidden representation (i.e. as in decoding). The encoding is done using Eq. (1) as,

$$f_h = s(Wx + b) \qquad (1)$$

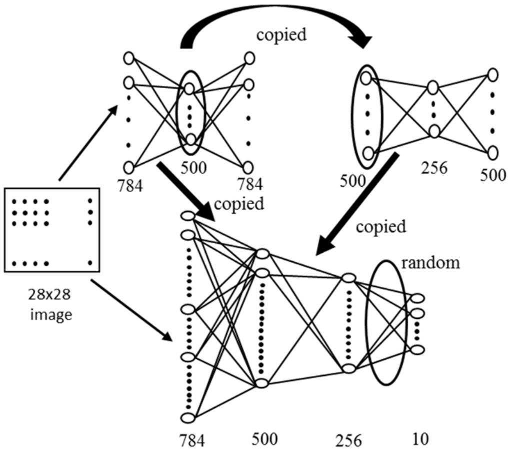


Fig.2. Structure of SAE considered in this study.

where f_h transforms an input vector into its hidden representation, W is the weight connecting input-hidden units and s is the activation function. This hidden representation is generally in low dimension compared to the input x. The resulting hidden representation is then mapped back to reconstruct high dimensional X and it is termed as decoder as shown in Eq. (2),

$$X = s(W'f_h + b') \qquad (2)$$

where W' is generally the same input-hidden tied weights except in the transpose form and same goes for bias b'. The reconstructed X may not be the exact representation as of original input x rather than a predictive one. This requires an associative reconstruction error to be minimized as shown in Eq. (3),

$$L(x, X) \propto -\log \mathrm{p}(x|X) \qquad (3)$$

Fig. 2 presents a SAE construction as [784-500-256-10] from two AEs. The first AE has the architecture of 784 input units, 500 hidden units and 784 output units. Similarly, the second AE has an architecture of 500 input units, 256 hidden units and again 500 output units. The first AE takes handwritten Bangla numeral as input and try to reconstructs the exact form of it at the output in the course of encoding the original input as 500 hidden unit features. Similarly, the second one of AE takes this 500 features as input, encodes it as 256 hidden units and finally reconstructs it. Taking inputs and trying to reconstruct it as above is called pre-training. Next a feed forward neural network is considered having two hidden layers of size 500 and 256 units respectively. Also it has 784 input units and 10 output units. The input-hidden weights (784x500) of this NN are initialized by the first AE, the hidden-hidden weights (500x256) are initialized by second AE and the hidden-output weights (256x10) are initialized randomly. Finally this NN is fine-tuned using the pre-trained weights from AEs for classification.

C. Classification using proposed SAEPT

SAEPT is a modified version of SAE which is efficient than MLP and SAE for handwritten numeral recognition. The traditional MLP and NN discussed above have similar architecture of (784-500-256-10). In terms of updating parameters, MLP has 784x500x256x10 weights along with bias terms. This leads to a very large computation of gradients for each of the parameters from above and if the input is sparse, most of the gradients becomes zero resulting the

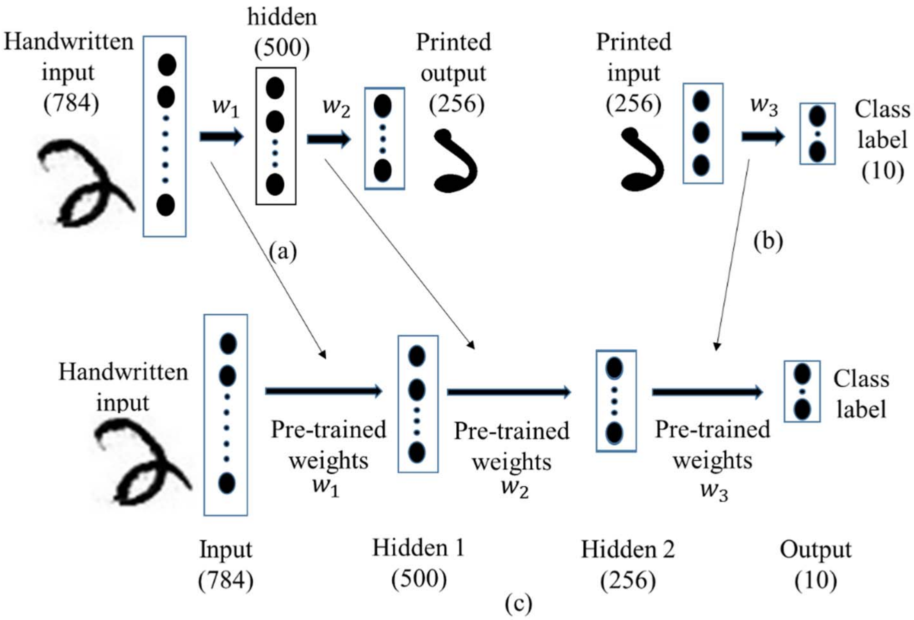


Fig. 3. Structure of the network for proposed SAEPT: (a) AE to encode handwritten Bangla numeral into printed form (b) NN to classify printed Bangla numeral (c) Final SAEPT structure using pre-trained weights from (a) and (b).

network to be suffered from vanishing gradient problem. Moreover the network needs very large time to converge. The SAE has the similar architecture (784-500-256-10) at the final stage except the weights are pre-trained from two AEs (784-500-784, 500-256-500). Thus the number of network parameters are more than the one in MLP. Calculating gradients and minimizing error signal of these large number of parameters requires huge computation. On the other hand if we use the concept of reconstructing input as a form of printed one which is smaller in size, the number of weights becomes smaller. As printed data is easier to map with the class labels if the reconstruction from handwritten to printed works well, the network would easily classify them into class label. Also the final MLP does not need any further fine-tuning.

Fig. 3 shows the proposed SAEPT structure for handwritten numeral recognition in which Fig. 3(a) is the AE for mapping handwritten Bangla numeral into printed from and Fig. 3(b) is for printed to class label mapping. The AE consists of 784 input units, followed by 500 hidden units and finally 256 output units. They are separated by feed-forward layers and *sigmoid* is used as the activation function. A 28x28 sized handwritten numeral image is directly fed as input to the network. Generally in AE, Input data is supplied at the input layer as well as in the output layer in the form of desired output. But in this case, keeping the above concept in mind, at the output layer, corresponding printed form of the handwritten numeral is provided instead of the exact handwritten input. This allows the network to reconstruct the handwritten numeral to corresponding printed form. On the other hand, there is another NN as shown in Figure 3(b) which consists of 256 input units and 10 output units. This NN is trained using 16x16 sized printed numeral images having corresponding class label at the output. Finally the above two networks are combined keeping all of the pre-trained weights from the above two networks as shown in Figure 3(c). The final network consists of a NN having the architecture of 784 input units, 500 1st hidden layer units, 256 2nd hidden layer units and 10 output layer units. The 784x500 and 500x256 weights are replicated from the AE whereas 256x10 weights come from the NN. The final network does not need any fine-tuning and directly gives the class label with only a forward pass.

In the backward pass for both the AE and NN, traditional Back-propagation (BP) calculates the derivative of the loss function (i.e. gradient) with respect to output layer activation for each point. Finally this gradient is applied to update the network parameters converging the network towards generalization.

III. Results and Discussions

Experimental results of the proposed recognition scheme have been collected on the benchmark dataset of ISI [17]. The dataset contains total 22000 images with extensive variation due to difference in writing styles of different people. The samples are divided into 18000 and 4000 samples, for training and testing, respectively. Training samples are evenly distributed over the underlying 10 classes. The recognition performances reported in this paper are based on the test set accuracies. In the test set, equal number of samples (i.e., 400) for each numeral were considered. In order to make the raw images appropriate to feed into classifiers, some preprocessing techniques are applied such as converting raw images into gray scale images, cropping the images to maintain uniformity and finally conversion from background numeral white to black and foreground changed to white.

We applied the three architectures as described in the previous section on resized and normalized grayscale image files without any feature extraction technique. The method is implemented in Matlab R2015a. The experiment has been conducted on MacBook Pro laptop machine (CPU: Intel Core i5 @ 2.70 GHz and RAM: 8.00 GB) in OS X Yosemite environment.

We have observed the classification accuracy of the proposed system for various epochs and it is observed that the method performs well at 260th epoch of pre-training AE. The NN with printed data as input was run for fixed 200 epochs. Fig. 4 and Fig. 5 show training set and test set recognition accuracy respectively of NN, traditional SAE and proposed SAE for different epochs with a fixed batch size of 50. It is observed from these figures that training set accuracy of the proposed method reaches at 100% whereas the test set cannot

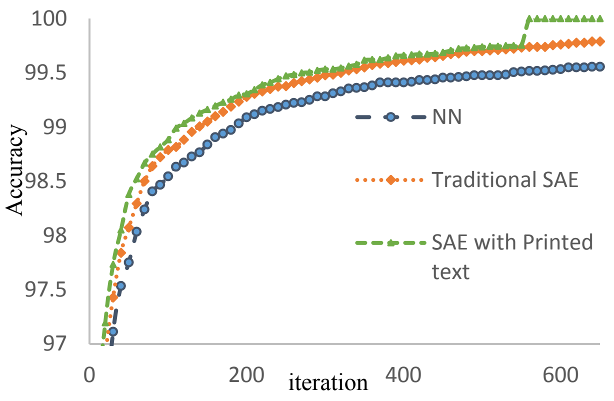


Fig.4. Training Set Recognition Accuracy with batch size 50.

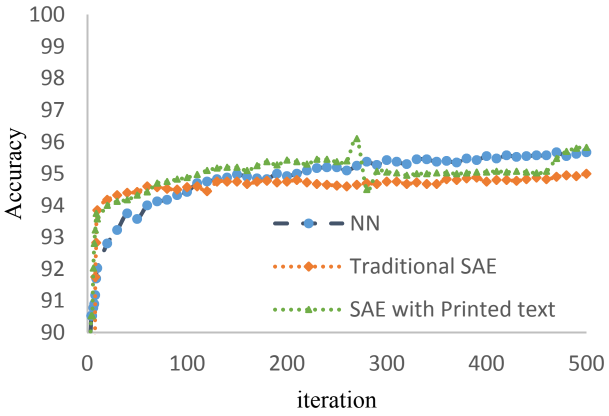


Fig. 5. Test Set Recognition Accuracy with batch size 50.

coincide with this as its patterns were unseen during the course of training achieving the accuracy of 96.30%. It is notable that the accuracy of the test set is more desirable as it indicates the generalization ability of the system. Although the other two networks achieved training and test set recognition accuracy closed to the proposed one but still they require large number of epochs to converge whereas the proposed one requires only pre-trained weights without any further fine-tuning. The pre-training epochs are 500 and 200 for the networks of fig 2(a) and 2(b) respectively.

Table I shows the confusion matrix of test set samples after fixed 500 iterations. It is clearly observed from the table that the proposed method worst performed for the numeral "১" and out of 400 test cases, classified it truly in 360 cases. In 21 cases this numeral classified as "৯". Both "১" and "৯" have ambiguity in their handwritten form. In Bangla handwritten numeral script, there is also a similarity between "৫" and "৬"; therefore in 12 cases "৬" classified as "৫". Similarly, the numeral "০" classified as "৩" and "৬" in eight and three cases, respectively; it is clearly observed that confusion in these handwritten numerals arises due to the diversity of different individuals writing. But the proposed method is shown to perform best for "২", correctly classifying all of the 400 test samples.

Table II shows some handwritten numeral images those are misclassified. Due to large variation in writing styles, such numeral images are difficult to classify even by human. Finally, the proposed SAEPT misclassified 148 test samples out of 4000 test cases achieving test set accuracy of 96.30%. On the other hand, the method didn't misclassify any of the 18000 training samples showing training set accuracy of 100%. This indicates that there is a chance to improve deep SAE training and get better performance with the proposed method.

TABLE II. SAMPLE HANDWRITTEN NUMERALS THAT ARE MISCLASSIFIED BY SAEPT.

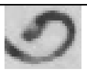
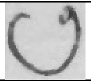
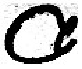
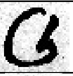
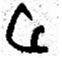
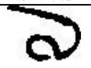
Handwritten Numeral Image	True Numeral	Image Classified as
	৩	০
	৩	০
	৫	০
	৬	৫
	৬	৫
	৯	১

Table III compares the outcome of the proposed method with other prominent works of Bangla handwritten numeral recognition on the basis of accuracy provided by corresponding authors. It also presents distinct features of individual methods. For better understanding the significance of the proposed method, traditional AE and feed forward NN having the same architecture is further considered and its outcome is also included in the table. It is notable that, proposed method did not employ any feature selection technique whereas existing methods use one or more feature selection methods. Without feature selection, proposed SAEPT method is shown to outperform the existing methods.

TABLE I. CONFUSION MATRIX PRODUCED FOR TEST SAMPLES OF BANGLA HANDWRITTEN NUMERALS. TOTAL SAMPLES ARE 4000 HAVING 400 FOR EACH NUMERAL.

English Numeral	Bangla Numeral	Total samples of a particular numeral classified as									
		০	১	২	৩	৪	৫	৬	৭	৮	৯
0	০	388	0	1	8	0	0	3	0	0	0
1	১	0	360	9	0	4	0	6	0	0	21
2	২	0	0	400	0	0	0	0	0	0	0
3	৩	4	0	0	392	0	1	3	0	0	0
4	৪	2	0	0	0	395	2	0	0	1	0
5	৫	0	0	0	1	3	392	4	0	0	0
6	৬	1	0	0	11	0	12	374	0	0	2
7	৭	3	0	1	0	0	3	0	393	0	0
8	৮	1	0	0	0	0	4	0	0	395	0
9	৯	5	18	3	5	2	0	4	0	0	363

TABLE III. A COMPARATIVE DESCRIPTION OF PROPOSED SAEPT WITH SOME CONTEMPORARY METHODS.

The work reference	Feature Selection	Classification	Recognition Accuracy
Pal et al. [2]	Water overflow from the reservoir based feature selection	Binary decision tree	92.80%
Wen et al. [7]	Principal component analysis (PCA) and Kernel PCA	SVM	95.05 %
Basu et al. [3]	Shadow feature and Centroid feature	MLPs with Dempster-Shafer technique	95.10%
Bhattacharya and Chaudhuri [6]	Wavelet filter at different resolutions	Four MLPs in two stages (three + one)	98.20%
Surtinta et al. [16]	Contour angular	SVM	86.80%
Feed forward NN	No	NN	95.67%
Traditional SAE	No	SAE	95.75%
Proposed SAEPT	No	SAE	**96.30%**

According to the table, SAEPT achieved test set accuracy of 96.30%. On the other hand, the test set accuracies are 92.80%, 95.10%, 95.05% and 86.8% for the works of [2], [3], [7] and [16] respectively. Test set accuracy of work [6] is higher than the proposed one, but it used a scaled up version of the original dataset which is 10 times larger than the one used in this study. Moreover, the two other networks of traditional AE and NN having the same architecture achieved test set accuracy of 95.75% and 95.67% respectively which is not better than the proposed one. Although performance comparison here with other prominent works are for different datasets, the efficacy of the proposed SAEPT is quite interesting and identified the ability of SAE with printed text based classifier for Bangla handwritten numeral recognition.

IV. CONCLUSIONS

This paper proposes a SAE architecture along with the printed text for handwritten Bangla numeral recognition. SAE has the ability to exactly or predictively reconstruct the original form of its input and this gives some of the hierarchical features that explains the data very well. SAE also extracts the visual patterns directly from pixel images with minimal preprocessing. Therefore, a SAE structure is investigated without any feature selection for handwritten Bangla numeral recognition in this study achieving promising result. Study also reveals that the deep networks need lesser data than shallow networks. The method has been tested on a large handwritten numeral dataset and outcome is compared with existing prominent works for Bangla. The proposed method is shown to outperform the existing methods on the basis of test set accuracy without scaling the data size. Moreover, the proposed scheme seems efficient in size and computation.

REFERENCES

[1] R. Plamondon and S. N. Srihari, “On-line and off-line handwritten recognition: A comprehensive survey,” *IEEE Trans. on PAMI,* vol. 22, pp. 62-84, 2000.

[2] U. Pal, C. B. B. Chaudhuri and A. Belaid, “A System for Bangla Handwritten Numeral Recognition,” *IETE Journal of Research, Institution of Electronics and Telecommunication Engineers*, vol. 52, no. 1, pp. 27-34, 2006.

[3] S. Basu, R. Sarkar, N. Das, M. Kundu, M. Nasipuri and D. K. Basu, “Handwritten BanglaDigit Recognition Using Classifier Combination Through DS Technique,” *LNCS*, vol. 3776, pp. 236–241, 2005

[4] List of languages by number of native users. Available: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_users

[5] M. M. R. Khan, S. M. A. Rahman and M. M. Alam, “Bangla Handwritten Digits Recognition using Evolutionary Artificial Neural Networks*” in Proc. of the 7th International Conference on Computer and Information Technology (ICCIT 2004),* 26-28 December, 2004, Dhaka, Bangladesh.

[6] U. Bhattacharya and B. B. Chaudhuri, “Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 444-457, 2009.

[7] Y. Wen, Y. Lu and P. Shi, “Handwritten Bangla numeral recognition system and its application to postal automation,” *Pattern Recognition*, vol. 40, pp. 99-107, 2007.

[8] M. K. Nasir and M. S. Uddin, “Hand Written Bangla Numerals Recognition for Automated Postal System,” *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 8, no. 6, pp. 43-48, 2013.

[9] S. G. Oval and S. Shirawale, “Recognizing handwritten devanagari words using recurrent neural network,” in *Proc. of International Conference on Frontiers of Intelligent Computing; Theory and Applications (FICTA)*, vol. 2, pp. 413-421, 2015.

[10] M. R. Bashar, M. A. F. M. R. Hasan, M. A. Hossain and D. Das, “Handwritten Bangla Numerical Digit Recognition using Histogram Technique,” *Asian Journal of Information Technology*, vol. 3, pp. 611-615, 2004.

[11] Y. Wen and L. He, “A classifier for Bangla handwritten numeral recognition,” *Expert Systems with Applications*, vol. 39, pp. 948-953, 2012.

[12] M. Ahmed et al., “Acoustic Modeling of Bangla Words using Deep Belief Network”, *International Journal of Image, Graphics and Signal Processing (IJIGSP),* vol. 7, no. 10, pp. 19-27, Sep. 2015.

[13] Vanishing gradient problem. Available: https://en.wikipedia.org/wiki/Vanishing_gradient_problem

[14] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.” *In A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

[15] N. Das, R.Sarkar, S. Basu, M. Kundu, M. Nasipuri and D. K. Basu, “A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application”, *Applied Soft Computing*, vol. 12, pp. 1592-1606, 2012

[16] O. Surinta, L. Schomaker, and M. Wiering, "A comparison of feature and pixel-based methods for recognizing handwritten bangla digits," in *Document Analysis and Recognition (ICDAR), 2013 12th Interational Conference on. IEEE*, 2013,pp. 165-169.

[17] Off-Line Handwritten Bangla Numeral Database, http://www.isical.ac.in/~ujjwal/download/database.html