Sindhi Handwritten-Digits Recognition Using Machine Learning Techniques

Irfan Ali[1], Insaf Ali[2], Subhash[1], Asif Khan[1], Syed Ahmed Raza[3] ,Basit Hassan[1], Priha Bhatti[4]
[1]Sindh Madressatul Islam University, Karachi, Pakistan
[2]Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan
[3]Suleman Roshan College of Physiotherapy & Rehabilitation Science, Tando Adam, Sindh, Pakistan
[4]Sir Syed University of Engineering & Technology,Karachi, Pakista

Abstract

This study presents Sindhi Handwritten Digits Recognition using machine learning approaches. The purpose of this study is to explore the different ML techniques to identify its significance in the field of Sindhi handwritten digits recognition. Handwritten digits recognition is one of the important fields in Computer Science. In the past, a lot of research work has been carried out regarding the recognition of digits as well as the characters in various languages like Urdu, English, Chinese, and Arabic. The literature review suggests that limited work has been done on the Sindhi language. In this study, the model is designed for the recognition of Sindhi handwritten digits using K Nearest Neighbor, Decision Tree, Multilayer Perception and Random Forest Classifier. It is found that the performance of the Random Forest Classifier and Decision Tree on Sindhi Digits are more effective as compared to other approaches. The study helps for improving the automatic learning for Sindhi handwritten digits. It is recommended that RF and DT Classifiers should be used in Sindhi handwritten digits recognition. In the future, this research will pave the way to recognize Sindhi characters through deep learning models.

Keywords — Sindhi digit Recogniation, , Numerals recognition, Handwriten digits , Sindhi handwriten-digits, Machine learning techniques

I. INTRODUCTION

The Digit Recognition is art for identifying, detecting and segmenting the digits from the image. The basic phenomenon of the digit recognition system is recognizing and detecting the digits from the input image then convert that image into machine readable code or ASCII format. However, this automation is more helpful for processing the information between the human and machine. Since the last decade, the scope of digit recognition is increasing day by day due to the usage of a wide range of applications in the computer science field. The handwriting digit recognition to understand complex and intelligible inputs through the machine. In the field of pattern recognition, the interest of the research community high, for that reason the applications and its applicability of recognition tools in different fields are increasing day by day. Hence, the more flexible input devices and efficient data organization techniques and processing methods are used. Instant applications of the digit recognition techniques include several fields and it may probably use for checks recognition in banks and recognizing numbers of the car's plates, and furthermore, its used in postal mail sorting for ascending and descending order, and last, the automatically address reading and mail routing, etc.

Optical character recognition (OCR) is one of the most successful pattern recognition applications and handwritten digit recognition [3]. And its active topic in research for OCR learning and classification. In past years, various methods have been implemented to analysis or test case the theories of many patterns' detection and machine learning techniques. Furthermore, to promote the research culture in machine learning and pattern detection, there are many standard databases have designed in which digits of handwritten are processed, with normalization and segmentation. Therefore, many researchers have compared the results with respect to common factors. For the research purpose, one of the most famous and free MNIST handwritten digits database available for fast testing the machine learning algorithms. In different digit recognition, the major task is the variance of class. Sometimes, this variance comes through a single and different person. Because the person does not write the same digits always in the same way in the different time slot it may have format changes. So, therefore, various feature extraction methods have been designed for suggesting and describing shape invariance of class. And enhancing the discrimination ability within in class. [1] As the experiment depicted that to identifying the local structure, feature direction and moreover curvature features are more important. The performance related to accuracy and efficiency has been improved for the usage of these features [1]. Sindhi is an Indo-Aryan and ancient Indus valley language having 100 years old history and its official language of the province of Sindh, Pakistan. It is a widely spoken language, especially in Pakistan and India. As per the researchers [1-2], this language is spoken round about 40 million peoples of India

Manuscript received May 5, 2019
Manuscript revised May 20, 2019

within the different states and Pakistan in the province of Sindh. And however, this language is taught as a basic language in primary schools at public and private sectors in the province of Sindh, Pakistan. [2]

The Sindhi Language has a complex structure of numbers and characters, therefore, it's very much difficult to extract the exact features from online Sindhi handwritten. The writing style of Sindhi digits is the same just like Farsi and Arabic Numerals. Writing with a digital pen or stylus tool it's easy to write the different numbers of touch screen or touchpad. As scripts of Arabic and Farsi are just like the same, but some of the digits are written in the form of old traditional Arabic style. Some Sindhi numbers are such as '۴' and '۵' are written in different formats. And moreover, these digits are also similar just like Hindi and Urdu digits but there is a small difference in writing style.

II. RELATED WORK

Limited work has been done on Sindh language even the speakers of this language are million, although more research work has been other languages. Gorgevik at el [7] designed the handwritten digit recognition system for roman numbers using a support vector machine. They extracted 4 different features from the input image, namely contour profiles, projection histograms, ring-zone, and last kirsch features. The model showed 97.27%-digit detection accuracy on the national digits recognition database NIST [8]. The framework max-min posterior has been designed by Chen et al [9]. This model used posterior pseudo probabilities for the detection of handwritten digits from the input image. They pull out (256) directional dimension features from the image pattern, and moreover, the extracted features have converted into 128 set by implementing Principal component analysis (PCA). With the NIST dataset, the model reported the accuracy of recognition of 98.76% [8]. The spare coding method is suggested by labusch et al [10]. The model extracted the important features using (SVM) support vector machine classifier. They reported the model recognition accuracy of 99.41% on the modified or improved version of the NIST handwritten database (Modified-NIST) [11]. The research work described in [12] merged three different recognizers by high voting, dimensionality reduction of PCA and the other one is four orientations based on kirsch gradient, and the last one is classification through SVM. The model reported the accuracy of recognition 95.05% and with error 0.93% on 10,000 test data samples of the MNIST dataset. Mane et al [13]. They proposed an elastic image matching technique for handwritten digits recognition-based Eigen deformation. Therefore, the selection of the automatic deformation of elastic matching is based on the principal component analysis.The results are reported the overall accuracy of 94% on self-collected dataset individuals of the profession experiments. The presented work Cruz et al. [14] evaluated the six different features using ensemble classifiers for the handwritten digit detection system. The extracted features are modified edge maps, structural characteristics, Multi zoning, projections, concavities measurements, and last one gradient directional. The model accuracy and rate of recognition is 99.68% on training and testing, 60,000 images samples of training and 10,000 for testing of Modified NIST dataset [4].

Kapoor et al [5] presented model used a neural network for deeply analysis the handwritten digit recognition. And that model chose the main concept of the back propagation for selection of a suitable model of the simulation model, which basically recognizes the different digits. And moreover. The simulation of the back-propagation model achieved good accuracy compared to the other different techniques. The model used a handwritten input image with the help of image editor or paintbrush which checked against testing and training ready set. The model simulation also used in different applications like visual classification and different entities possible variations also an example of signature pattern matching and alphabet handwritten recognition. And moreover, this simulation can further extend to understand different characters of various languages (characters, Urdu). Fatahi at el [6] proposed the SNN model. There no dataset available for evaluation of the performance of the SNN model. The suggested model used Event-driven MNIST (evt_MNIST). This dataset consists of all testing and training sets. The dataset is designed Matlab file format and it's used by the brain [20] and for the MATLAB, related to the single image we have a collection of the 100 ms spikes to training collective all the pixels that trained is 1KHZ. The results show the generated spike sequences have Poisson distribution. And therefore, the model basically satisfactory for the simulation spikes that is a loophole in the cortex. This model presents a back propagation to handwritten digit recognition and for the data, we have required minimal preprocessing, but the network structure was fully constrained and highly focus on designed based tasks. The model used an input collection of various normalized images of separated digits. And it shows a 1% error rate and 9% of the rejection of the zip codes provided by US postal serves. [15]

III. METHODLOGY

.

In this section Sindhi handwritten numerals samples were collected from the difference students on A-4 size paper. Every samples are written grid form of equal width and height. During the sample collection student can write the samples in any style or size. The obtained samples collected from the 400 students, and secondly inputs are process with

noise. These noise features are added during the writing of samples in different size or invariant size of input image. The OpenCV image denoising filters are used for removing the noise from the images. Some preprocessing approaches are performed in order improving the quality of grayscale input image by removing the unwanted pattern and irreverent bit patterns from the image which is basically case of the noise.

A. K-NN Algorithm

KNN (K Nearest Neighbor) algorithm is a simplistic approach for learning purposes. Its non-parametric technique mostly implements for regression and classification. Somehow this method is called a lazy learner method .and furthermore, the functionality of KNN is based on the feature similarity function.
This model trained on training samples once it's fully trained then it will identify the similar samples in test data [18]. The KNN working mechanism set based on the K points in the handwritten digit recognition dataset. So, therefore, the method predicts those digits they are most close to K and it estimates the majority votes of the closest points. The core suggestion of the KNN is basically assigned a specific class. And however, that specific class more deeply represents the closest neighbor. The model KNN computes the similarity with help distance equations and that distance related to the single test examples z= (x', y') and complete training example. (x, y) D to decide the most closed neighbor DZ.

Algorithm 01: **K Nearest Neighbor**

Procedure of KNN

1 *K = nearest neighbor and D represent size samples*

2 *Repeat test sample z=(x',y') do*

3 *Compute the distance d(x',x) and (x, y) ∈D.*

4 *Dz ⊆ D the set of K nearest training to z*

5 End **Procedure**

B. Decision Tree Classifier

The decision tree classifier is part of supervised learning. Like other different supervised learning algorithms exits in machine learning and other fields. The approach most used for solving the classification and regression problems. The main objective is creating the training model which basically predicts the class or targeted values of learning decision rules that inferred from training and past data. The core object of the model is creating a small tree that can classify the unknown class or instance by determining the list of the rules the handwritten digits dataset. Multiple different tools are used by DT classifier to extract information, digit recognition, text predictive and machine learning [17]. The research study attention on the hunt's algorithm to create multiple decision trees based on sprevious existing DT's simulation algorithm with C4.5. The Weka DM tool is called J48 is an open source tool for employing the C4.5 algorithm [18]. The working procedure of the algorithm, first, picks the parameters for the root node and builds a branch for a single attribute. The approach preferred best split selection from the dataset by assessing uncleanness for each child nodes. The equations for GINI, entropy and child nodes for computing the uncleanness degree, normally C4.5 approach is used entropy.

Algorithm 02: **Decision Tree Classifier**

S_1 where set of classified instances

S ≠ Null number of attributes > 0

1 **Procedure** of Decision tree

2 ***Repeat***

3 *gainMax←0*

4 *split←null*

5 *e←Entropy of attributes*

6 *Compute for all attributes α in S (store in A)*

7 *Split(S, new A)*

8 ***Till all trees partition***

9 End **Procedure**

C. Multi-layer Perceptron Classifier

A multi-layer perceptron is a multilayered and feed forward supervised neural network used for the text or data classification. The MLP neural network consists of multiple nonlinear layers which are the input layer and required one or more hidden and the output layers. The model MLP used activation function on the hidden layer for creating nonlinear in data such as tan-sigmoid and sigmoid. The activation function characterized every neuron on the network and moreover, all neuron relates to the next layer [19]. Furthermore, every single neuron is connected and characterized by synaptic weight and its simple weight factor. The structure and configuration MLP are based on artificial neural network technique. The L resented by several layers in the network, the first input layer, and output and last layer output. L-1th hidden layers from to -1, Furthermore, the network shows are fully interconnected allowing every neuron in the network relates to the next

network layer. The model manages the forward communication between input and output layer with support possible hidden layers. [19] And furthermore the signals are moves, the inputs move onward direction with the support of the network.

Algorithm 03: **Multilayer Perceptron Classifier**

Dataset of Training
Dataset of Testing
Procedure of Decision tree
1 **Input Layer**
L MLP classifier
Training dataset of S;
Labels Y
2 ***Repeat***
3 *Train Classier S^k via L*
4 ***End repeat***
L (output)
5 End **Procedure**

D. Random Forest Classifier

The random forest is an ensemble learning method used for regression and classification. This method combines Breiman's bagging sample methodology, and selection some random features that introduced by independently. In general terms RF is combination of the L tree-structured classifiers {h(x, Θk)} such as k=1 to L, where x is input and Θk independent identically distributed vectors, with the respect of definition context, we can say that RF is the family of the different methods which basically exist in different algorithms. And this method uses a bagging approach for the selection of some random features. [16] The main stage of this model is based on training which consists of the multiple trees and, for addition, every segment is trained on bootstrap samples. This principle shows a bagging method combine with CART - like induction. Somehow this induction method called is RF where working mechanism based on CART based Algorithm and that updated the feature selection on every node.

Algorithm 04: **Random Forest Classifier**

Training samples
Testing samples
1 **Procedure** of Decision tree
2 ***Repeat***
3 *Bootstrap Z* the size of N*
4 *Tb min node n_{min}*
5 *m random variables, p variables*
6 *Split the m*
7 *$\{Tb\}^B$*
8 ***Till output trees***
9 End **Procedure**

IV. RESULT AND DISCUSSION

Sindhi handwritten digit dataset contains 30,000 records with dimension 32* 32 grayscale images in the distribution of 10 classes (0 – 9). The size of each class varies. Figure 1 shows the overall classes and their size of the digits.

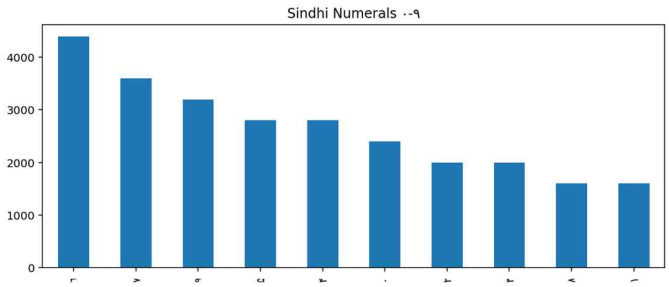

Figure. 1 Size of Dataset with labels

The whole dataset is divided into two parts (training and testing). There are 23000 samples of training images and 7000 test images. And furthermore, every single digit image consists of 1024 features that represent grayscale properties. The size of digit images is normalized and centered with the fixed size. The digit image of Sindhi handwritten is centered in dimension 32 by 32-pixel image by computing the center mass of the various pixels and moves the position of the pixel in the center of the 32 x 32 field. The focus of this research study is to find similarities between the shapes of Sindhi numerals. Figure 2 shows some random samples of handwritten digits from the dataset.

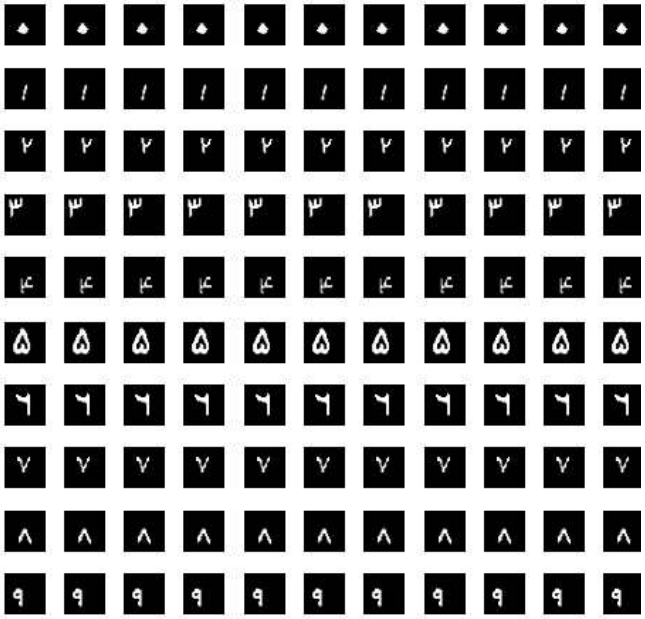

Fig. 2 Random Samples of handwritten digits from dataset

The handwritten digits are not always of the same size, width, orientation and justified to margins as they differ from writing of person to person, so the general problem would be while classifying the digits due to the similarity between digits such as ١ and ٧, ٥ and ٦, ٣ and ٨, ٢ and ٥, ٢ and ٧, etc. This problem is faced more when many people write a single digit with a variety of different handwritings. Lastly, the uniqueness and variety in the handwriting of different individuals also influence the formation and appearance of the digits. In the content of pixel to pixel comparison of two or more images. They are a lot of difference between the same image's pixels, but in the human eye, somehow both are same images. Hence, we find some new approaches that correctly identified the labels.

In this section, the performance of proposed model on Sindhi handwritten digits approaches is evaluated experimentally using 8500 test samples. The Recognition model is implemented on google colab laboratory with GPU configuration. And the results are the performed-on grayscale images with range (0-255) for black and white color. The size of images is 32 * 32 (width, height). The accuracy rate is used as assessment criteria of measuring the recognition performance of the proposed model. The below graphs are depicted the results of the following approaches: K Nearest Neighbor, Decision Tree, Multilayer Perception and Random Forest Classifier. Furthermore, it can be observed from Figure 6 that the accuracy of tested models for digit recognition also varies with different machine learning techniques. However, obtained accuracies are 74%, 95%, 82% and 96% for classifier KNN, DT, MLP, and Random forest. Comparatively, The Random forest classifier and Decision tree provide high accuracy or recognition rate for Sindhi handwritten numerals as compared to other ML techniques. it was noticed that the testing accuracy at 8500 samples of model recognizing the digits are more effectively. The correct predictions of the handwritten digits are shown in Figure 4. Figure 3 shows the accuracy of all classes on the Random forest classifier, the model on the diagonal side shows the actual label or correct prediction of classes from (٠-٩). These recognition accuracies depend on the size of samples. As Figure 5 represents the accuracy of all classes on the RF classifier. The size of the samples is varied in class to class. Some asses' samples are high, and some are low. However, the size of class six and seven are greater than the other eight classes. Therefore, the prediction accuracy of class six and seven are better than other class, comparatively.

Sindhi Digits	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
٠	**80.88**	00.1	00.11	00.16	00.22	00.22	00.22	00.22	00.22	00.21
١	00.04	**77.33**	00.18	00.17	00.22	00.22	00.22	00.22	00.22	00.11
٢	00.08	00.30	**80.88**	00.32	00.22	00.22	00.22	00.22	00.22	00.12
٣	00.10	00.15	00.17	**85.32**	00.22	00.22	00.22	00.22	00.22	00.17
٤	00.15	00.11	00.20	00.12	**85.38**	00.22	00.22	00.22	00.22	00.15
٥	0.25	00.09	00.15	00.17	00.22	**85.33**	00.22	00.22	00.22	00.27
٦	00.16	00.15	00.33	00.25	00.22	00.22	**95.32**	00.22	00.22	00.20
٧	00.14	00.11	00.16	00.19	00.13	00.11	00.16	**91.24**	00.22	00.23
٨	00.13	00.10	00.14	00.23	00.32	00.25	00.23	00.26	**77.44**	00.21
٩	03.50	00.16	00.18	00.19	00.44	00.21	00.17	00.11	00.19	**87.88**

Figure 3. Recognition rate of (٠-٩) Classes on Random Forest Classifier

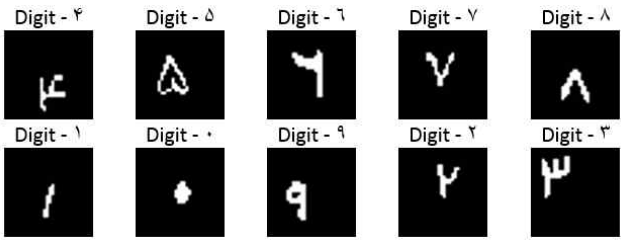

Figure 4. Random Forest Classifier generate results

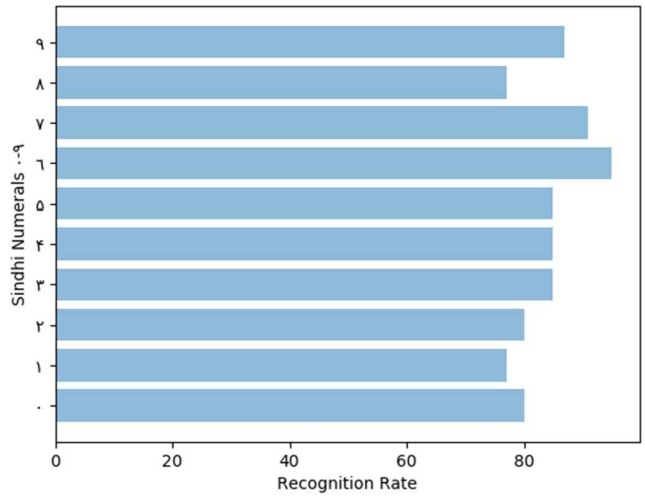

Figure 5. The Accuracy of (۰-۹) classes on Random Forest Classifier

.

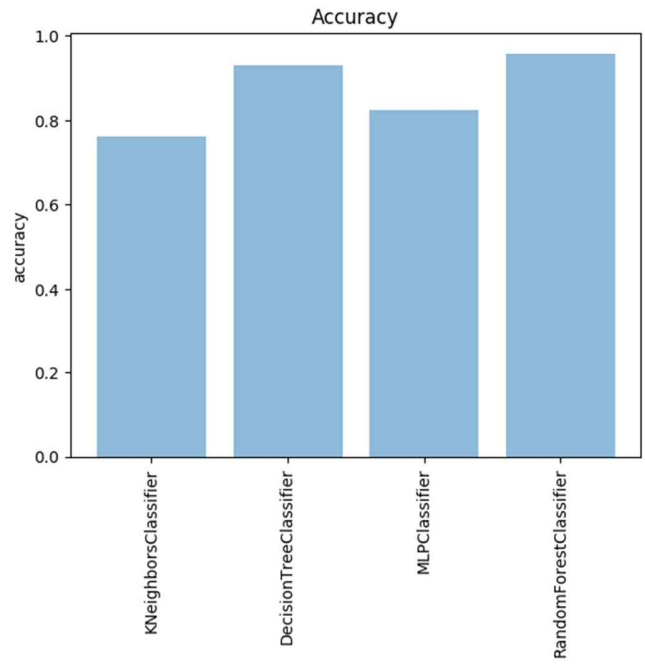

Figure 6. The Accuracy of Sindhi Handwritten digits on K Nearest Neighbor, Decision Tree, Multilayer Perception and Random Forest Classifier

V. CONCLUSION

The Sindh Handwritten Digits Recognition System (SHDS) database gives a relatively simple static classification task for researchers and students to explore machine learning and pattern recognition techniques. Sindhi digits are based on Arabic and Urdu digits. However, the writing style and direction of Sindhi digits are different from Arabic and Urdu digits. The total 30000 digits samples are collected from the students of different institutions. The samples were taken on A4 size paper by dividing the page in the grid form and furthermore each student could write a sample in any style and size. The dataset divides into training and testing which comprises of 30% for testing and 70% for training. The proposed model implemented on K Nearest Neighbor, Decision tree, Multilayer perception, and Random Forest Classifier. The results show that the Accuracy of KNN (0.74%), DT (0.95%), MLP (0.82%) and Random forest (0.96%). The RF and DT classifier provides a high accuracy recognition rate for Sindhi handwritten numerals as compared to other ML techniques. However, the recognition rate for style and shape matching digits was low. There are many issues in Sindhi scripts because that language is complex in syntax structures.

REFERENCES

[1] Liu, Cheng-Lin, et al. "Handwritten digit recognition: benchmarking of state-of-the-art techniques." Pattern recognition 36.10 (2003): 2271-2285.

[2] Chandio, Asghar Ali, et al. "Multi-Digit Handwritten Sindhi Numerals Recognition using SOM Neural Network." Mehran University Research Journal of Engineering and Technology 36.4 (2017): 8.

[3] Sadri, Javad, Ching Y. Suen, and Tien D. Bui. "Application of support vector machines for recognition of handwritten Arabic/Persian digits." Proceedings of Second Iranian Conference on Machine Vision and Image Processing. Vol. 1. 2003.

[4] Deng, Li. "The MNIST database of handwritten digit images for machine learning research [best of the web]." IEEE Signal Processing Magazine 29.6 (2012): 141-142.

[5] Kapoor, V., and Priyanka Gupta. "Digit Recognition System by using Back Propagation Algorithm." International Journal of Computer Applications 83.8 (2013).

[6] Fatahi, Mazdak, et al. "evt_MNIST: A spike based version of traditional MNIST." arXiv preprint arXiv:1604.06751 (2016).

[7] D. Gorgevik and D. Cakmakov, “Handwritten digit recognition by combining SVM classifiers,” in Proceedings of the International Conference on Computer as a Tool (EUROCON '05), vol. 2, pp. 1393–1396, Belgrade, Serbia, November 2005

[8] M. D. Garris, J. L. Blue, and G. T. Candela, NIST Form-Based Handprint Recognition System, NIST, 1997.

[9] X. Chen, X. Liu, and Y. Jia, “Learning handwritten digit recognition by the max-min posterior pseudo-probabilities method,” in Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR '07), pp. 342–346, Parana, Brazil, September 2007.

[10] K. Labusch, E. Barth, and T. Martinetz, “Simple method for high-performance digit recognition based on sparse coding,” IEEE Transactions on Neural Networks, vol. 19, no. 11, pp. 1985–1989, 2008.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] Y. Wen, Y. Lu, and P. Shi, “Handwritten Bangla numeral recognition system and its application to postal automation,” Pattern Recognition, vol. 40, no. 1, pp. 99–107, 2007. View at Publisher ·

[13] V. Mane and L. Ragha, “Handwritten character recognition using elastic matching and PCA,” in Proceedings of the International Conference on Advances in Computing, Communication and Control, pp. 410–415, ACM, Mumbai, India, January 2009.

[14] R. M. O. Cruz, G. D. C. Cavalcanti, and T. I. Ren, “Handwritten digit recognition using multiple feature extraction techniques and classifier ensemble,” in Proceedings of the 17th International Conference on Systems, Signals and Image Processing, pp. 215–218, Rio de Janeiro, Brazil, June 2010.

[15] LeCun, Yann, et al. "Handwritten digit recognition with a back-propagation network." Advances in neural information processing systems. 1990.

[16] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[17] Breiman, Leo. Classification and regression trees. Routledge, 2017.

[18] AL-Behadili, Hayder Naser Khraibet. "Classification Algorithms for Determining Handwritten Digit." Iraqi Journal for Electrical And Electronic Engineering 12.1 (2016): 96-102.

[19] Frias-Martinez, Enrique, Angel Sanchez, and Jose Velez. "Support vector machines versus multi-layer perceptrons for efficient off-line signature recognition." Engineering Applications of Artificial Intelligence 19.6 (2006): 693-704

Mr. Irfan obtained BS in Computer Science from the University of Sindh in 2011. He is pursuing MS (CS) from Muhammad Ali Jinnah University. He has served for more than four years as Lecturer and Software Engineer. His areas of interest are Machine Learning,Computer Vision, Image Processing, Natural language processing and Database Designing.

Dr.Insaf Ali siming is founder Chairman department of English QUEST Nawabshah . Dr Siming has did his masters from Shah Abdul Latif University Khairpur and earned his PhD degree from UTHM Malaysia. He has immense experience in Research and has published great number of research publication in well reputed journals .Dr siming has presented his research work globly in national and in international conferences ,including malaysia, singapore, thailand and pakistan. Dr. Siming is active researcher .he supervises good number of research scholars and visits different univeristies as external examiners and expert.He is currently working as Foundation chairman Department of English at QUEST. Nawabshah.

Dr. Subhash has done PhD in Development Studies from Sindh Development Studies Centre, University of Sindh, Jamshoro. His PhD dissertation was focused on microfinance and poverty in Sindh. He has published three research articles and attended several workshops and conferences at national and international level. He has served as a visiting guest lecturer at Berlin School of Economics and Law in Germany. He has also worked in development sector and microfinance industry for more than 12 years on different management positions.

Mr. Asif obtained ME in Electronics Engineering from Mehran University of Engineering & Technology Jamshoro. He has two years of industrial and more than 3 years of teaching experience. His particular interests are in Control System Design, Digital Logic Design, Fuzzy Logic, Computer Architecture and Electronics Hardware.

Syed Ahmed Raza is a seasoned Electronics Engineer, and holds a master’s Degree in Electronics Engineering with over 8 years of experience in project management and cross-functional leadership for structural analysis, quality engineering, and systems engineering projects. He is presently lecturer and lab in-charge in Suleman Roshan College of Physiotherapy &

Rehabilitation Science, Tando Adam Pakistan. His area of interest is in Automation, Control System and Instrumentation.

Mr. Basit Hassan holds a Master of Information Systems from the University of Melbourne, Australia, and a Master of Science in Computer Science from Iqra University, Karachi. He has more than two years of experience as lecturer and also served as Software Engineer specializing in Enterprise Resource Planning Systems design and implementation. His areas of interest include Software Engineering, Data Modeling and System Engineering.

Ms.Priha Bhatti she earned her undersgradute degree in computer science engineering from mehran university and Technology. She also hold Master's degree in software engineering from Sir Syed University of Engineering and Technology.she is currently working as Lecturer at department of computer science ,Sir Syed University of Engineering & Technology