Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

Research Comparative Analysis of OCR Models for Urdu Language Characters Recognition

Muhammad Murad[1], Muhammad Shahzad[2], Naheeda Fareed[3], Dr. Kashif Saghar[4]
[1,2,3]School of Computing, Alhamd Islamic University Quetta-Pakistan.
CESAT, Islamabad, Pakistan.
muhammadmurad445@gmail.com, zadbaloch123@gamil.com, kashif.saghar.v@nu.edu.pk

DOI: 10.5281/zenodo.14028816

ABSTRACT

There have been many research works to digitalize Urdu Characters through machine learning algorithms. The algorithms that were already used for Urdu Optical Character Recognition [OCR] are Convolutional Neural Network [CNN], Recurrent Neural Network [RNN], and Transformer etc. There are also many machine learning algorithms that have not been used for Urdu OCR e.g Support Vector Machine, Graph Neural Network etc. This research paper proposes a comparative study between the performances of the already implemented Urdu OCR on some of following algorithms like Convolutional Neural Network/ Transformer Model it also proposed a new implemented Urdu OCR using on Support Vector Machine algorithm.

Keywords: Support Vector Machin (SVM), Convolutional Neural Network, Artificial Neural Network, Recurrent Neural Network (RNN), Optical Character Recognition, Transformer Model.

Cite as: Muhammad Murad, Muhammad Shahzad, & Naheeda Fareed. (2024). Research Comparative Analysis of OCR Models for Urdu Language Characters Recognition. *LC International Journal of STEM, 5*(3), 55–63. **https://doi.org/10.5281/zenodo.14028816**

INTRODUCTION

With progress of technology and Artificial Intelligence: the natural language processing [NLP], optical character recognition or handwritten recognition, speech or voice recognition, and machine vision are also growing fast. Normally research is based on character recognition because now days paper materials are getting out of people sight or being burden for offices to keep. Plethora amount of paper materials is to be transformed to computer editable format. In order to process hundred or thousand papers into digital format is quite time consuming and much expensive. So optical character recognition provides an easy and inexpensive way to convert handwritten paper materials into digital format. Optical character recognition is fascinating and challenging area for researcher due cursive nature of languages. However, in optical character recognition and machine learning the researchers have to focus on development of the new algorithms that make computers to learn from the given input data and then makes decisions. There is some more common machine learning algorithm that include Convolutional Neural Network [CNN], Recurrent Neural Network [RNN], and Support Vector Machine [SVM] and these models have been commonly used for different languages OCR. However, Convolutional Neural Network have been used for most of the language's character recognition including Urdu language. On the other hand, Support Vector Machine [SVM] Model, although is commonly used for the other languages, but it has not been used for the Urdu optical character recognition yet. In this research paper

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

we first used the support vector machine algorithm for the Urdu language by giving it Urdu dataset we then trained and tested the SVM algorithm for the given Urdu dataset. Finally, a comparative study is made between the performance of the already made Urdu OCR CNN, Transformer Model and our new Urdu OCR on SVM Model. There are a lot of work on Urdu OCR using deep learning model like CNN, ANN RNN etc and this research paper uses supervised learning model SVM for the purpose of Urdu OCR. Finally, comparison will be made in term of accuracy, efficiency, and adoptability of these models. The findings of this research will not only contribute to the advancement of Urdu OCR technology but also provide valuable insights for the development of OCR systems for other languages.

LITERATURE REVIEW

Character recognition is growing field in research. Researchers are using different machine learning algorithms with feature extraction techniques and trained them with different datasets the aim is to get maximum accuracy, less computational time and cost. But the work of researchers is inadequate to implement OCR for the local languages using supervise learning model. Therefore, it is important to use and analyze the result of supervised learning model for the local languages like Urdu. Pal, et at [1] and his team used deep learning model CNN with EMINST dataset and got 93% accuracy for English Language. Gulzar Ahmed et al [2] has got the highest level of accuracy 98% for Urdu Language using CNN. Whereas in [9] the CNN is also used for recognition of mathematical equation and to solve them. Mohammad Daniyal et al and his team [7] used Transformers Models but they get below 25% accuracy rate for Urdu text recognition they still didn't use supervised leaning technique. SVM supervise learning model used for English Character recognition [3] by Dewi Nasian et al, has got accuracy of 73%. In Urdu Nastaliq recognition [4] combine the CNN and RNN for better feature extraction. They achieve a good accuracy of 98.12 %, although in [8] the researcher combines Support Vector Machine (SVM) and Artificial Neural Network for recognition of English language they got 94.43% accuracy. In [5] CNN with two datasets MNIST and Koggle and got very high accuracy more than 99 % for digit and English. Abu Sayeed Ahsanul, et al [6] amalgamate three different models KNN SVM and SRC for Arabic character and digit recognition, the have got average good result of 90% accuracy rate. In this research we used supervise learning model SVM for the local language like Urdu and got much better result.

METHODOLOGY

Data

In this paper we are using Urdu Handwritten text dataset (Uhat-dataset) that contains more than 700 grayscale sample images for each Urdu character with resolution 28x28 Pixel for training and testing purpose. Each Urdu character's samples are organized in separate sub-directories. The Style of Uhat-dataset is same as MINST dataset which is very popular for English. Uhat-dataset is freely available for researcher to work on Urdu text recognition [https://www.kaggle.com/datasets/hazrat/uhat-urdu-handwritten-text-dataset].

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

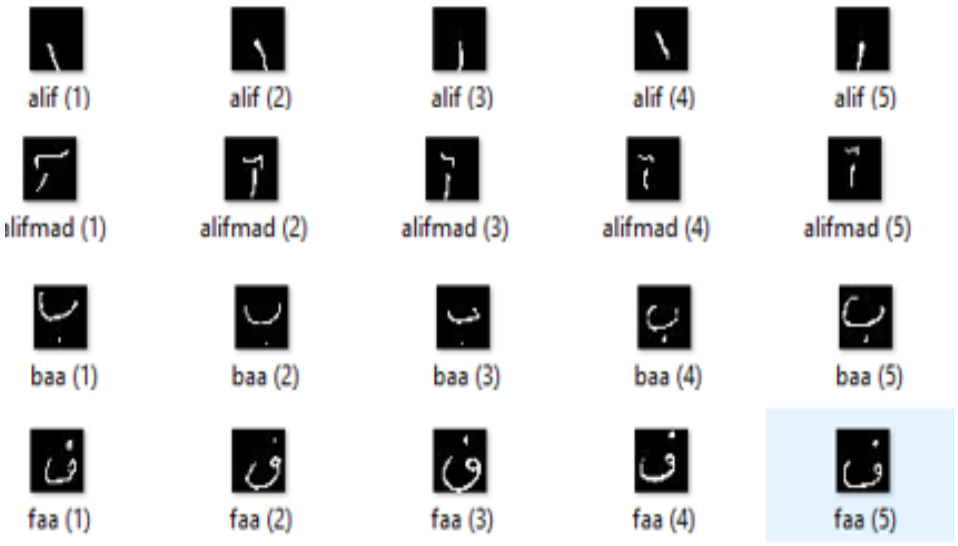


Figure 1: Few samples of Uhat_dataset

However, this research is based on implementation of Urdu optical character recognition using Support Vector Machine algorithm. This is very popular machine learning classifier suitable for linear and non-linear data it can also use for face recognition, voice recognition and written text recognition. Support Vector Machine (SVM) works on the principle of finding maximum separating hyperplane among the different class of data. For the Urdu text recognition first Support Vector Machine (SVM) is loaded with Urdu Handwritten text dataset by using python instructions. In the second step data preprocessing is performed in which images are resized, labeled and converted into grayscale. In the Third step all images of dataset are organized into list of single arrays for better manipulation, while loading; the datasets were stored into different directories. In the fourth step, normalization of image and splitting of dataset into 2 parts for training and testing purpose: 80% for training and 20% for testing is performed. In the fifth step Principal Component Analysis (PCA) technique is used for feature extraction from training data. In the sixth and last step SVM is initialize and trained using python code for future prediction.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

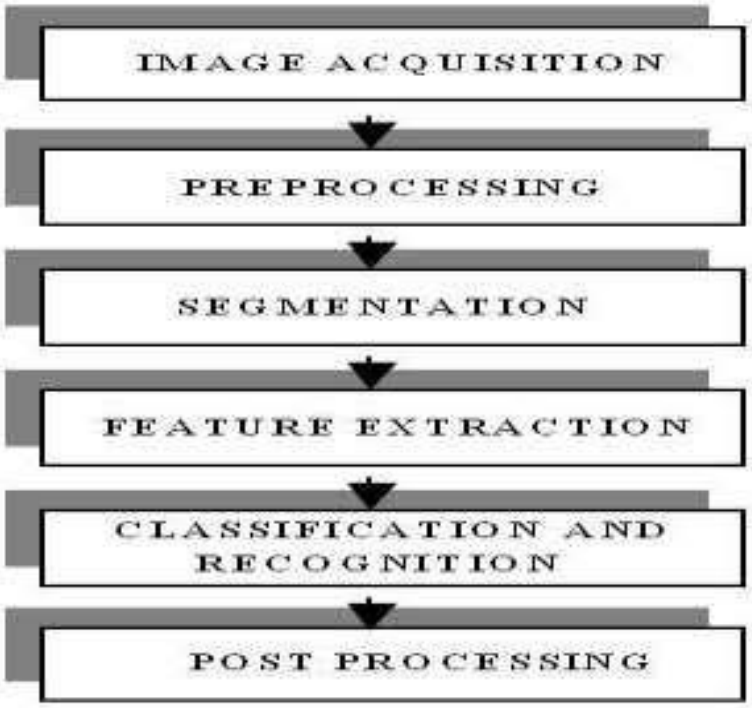


Figure 2: Depict how SVM works.

Model Development

As we have discussed above there are many machines learning models line CNN, RNN, Transformer Model for developing of this research we have used one of the supervised learning Model the Support Vector Machine SVM.

Method

We have used Python Programming language and its IDE for overall implementation of Urdu Optical character recognition. Through python code we have load dataset and preprocessed them for the training and test.

DATA ANALYSIS AND RESULTS

Results

Although, it is first time we are using support vector machine (SVM) algorithm for making the Urdu language OCR. We have trained and tested the model with 29,230 Urdu characters samples, these samples images were well preprocessed before load into model for getting better result in less time. The System can take 10 to 15 minutes for training of SVM model and finally we achieved an accuracy of 78.42% which may be considered to be a better result, using this supervise learning model for the cursive Urdu language. For evaluating the accuracy based on prediction by SVM model we have used Accuracy matrix that is very simple and easy technique for calculation of accuracy on the base the number of correct predictions divided by Total number of prediction (correct and in-correct prediction).

$$Accuracy = \frac{(Number\ of\ Correct\ Prediction)}{Total\ Number\ of\ Prediction}$$

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

We used two feature vectors 'y-test' and 'y-predict, y-test contains all labels that used for test and y-predict contains only all predicted labels from the test labels.

Example: y-test = [1 1 0 1 0]

y-predict = [1 0 0 1 1]

Total prediction = 5

Correct prediction = 3

Accuracy = 3/5 = 0.6 or (60%)

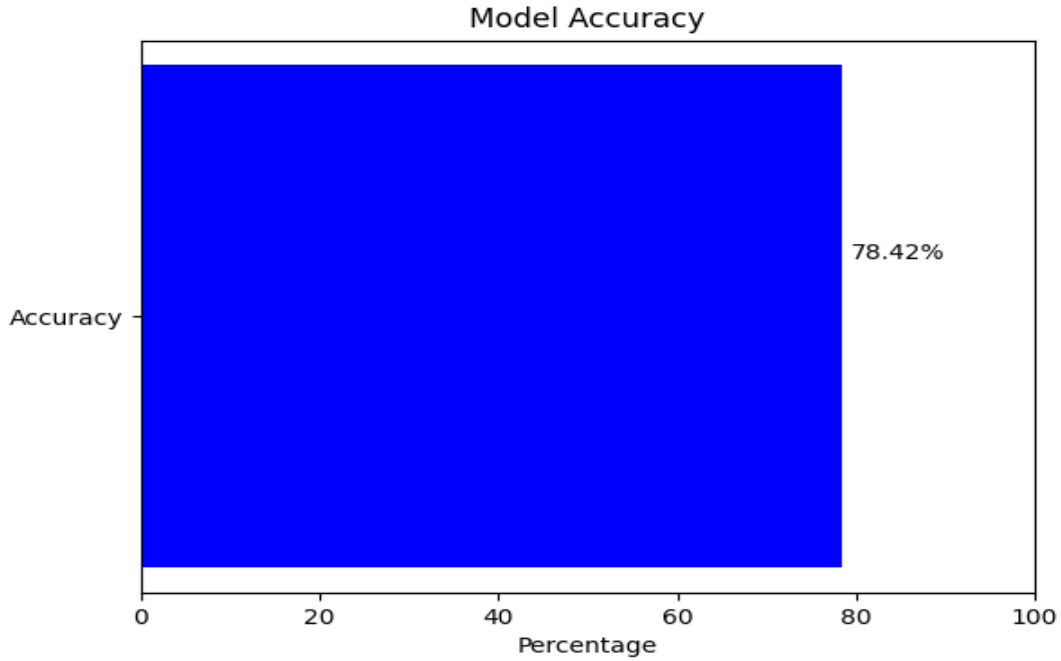


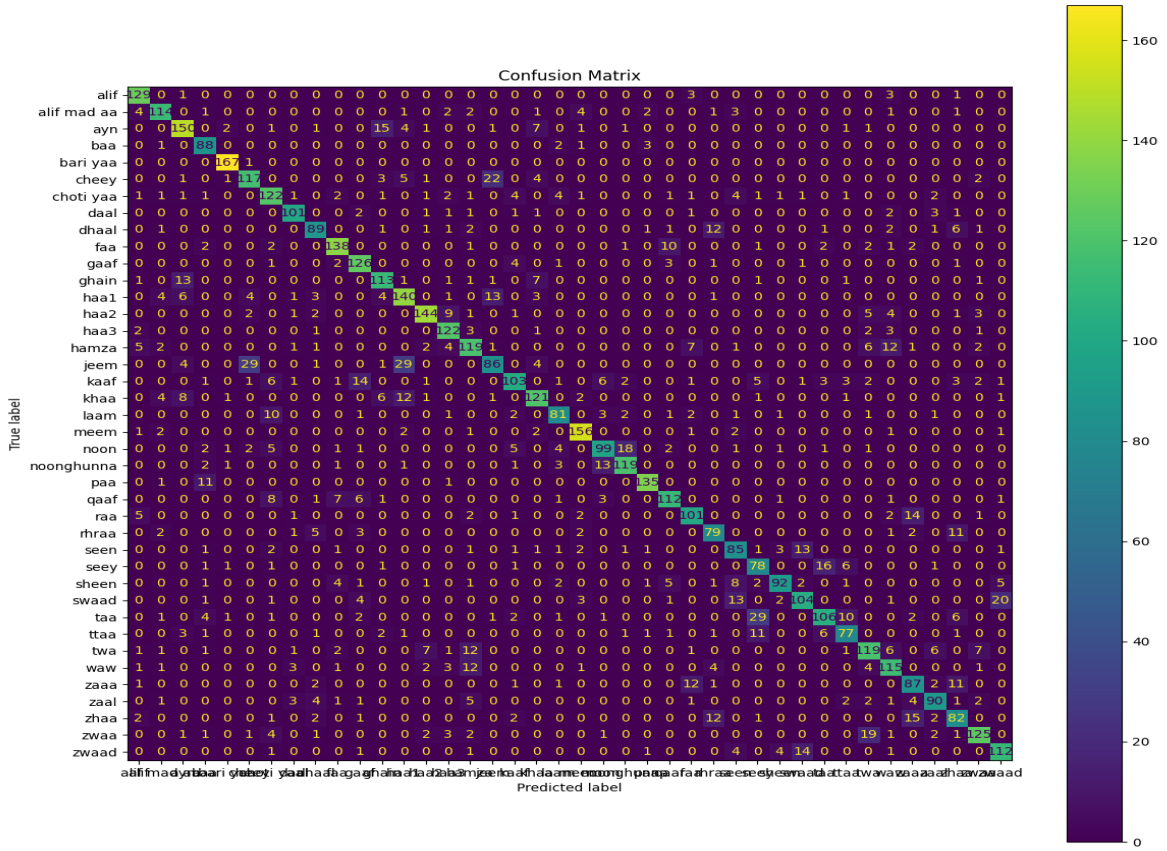
Figure 3: Accuracy matrix result for SVM

Additionally, we also used confusion matrix for calculation the performance of Support Vector Machine. The confusion matrix is table based which measurement the performance of machine learning algorithm on the basis of its two parameters (Predicted and Parameter). Further each table contained the following values, True Positive (TF): are the values or predictions that are correctly classified and are true, True Negative (TN): are the values or predictions that are classified as true but are false, False Positive (FP) are the values or predictions which wrongly classified as true but are false and False Negative (FN): are the values which wrongly classified as false, but are True.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Accuracy by this technique can be calculated by adding of True positive and true negative then divided by total prediction (Ture positive, true negative, False Positive, and False negative). On the basis of these calculation the result of confusion can be determined.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024



Here the result of confusion matrix in x-axis we have predicted values and in y-axis we have the actual or the true value. In the diagonal of the matrix is the value of true prediction of each sample. We have given 5840 instances of actual valued or test valued and the SVM model made 4581 true or accurate prediction.

Analysis

Here the result of confusion matrix in x-axis we have predicted values and in y-axis we have the actual or the true value. In the diagonal of the matrix is the value of true prediction of each sample. We have given 5840 instances of actual valued or test valued and the SVM model made 4581 true or accurate prediction.

Finally, we have compared the result of research our OCR with Urdu the already existing Urdu OCR that were made using Convolutional Neural Network and Transformer Model. The accuracy achieved by this research paper 78.42% accuracy rate while OCR using CNN was achieved is 98% and Transformer Model used for Urdu language achieved blow 25% accuracy rate.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

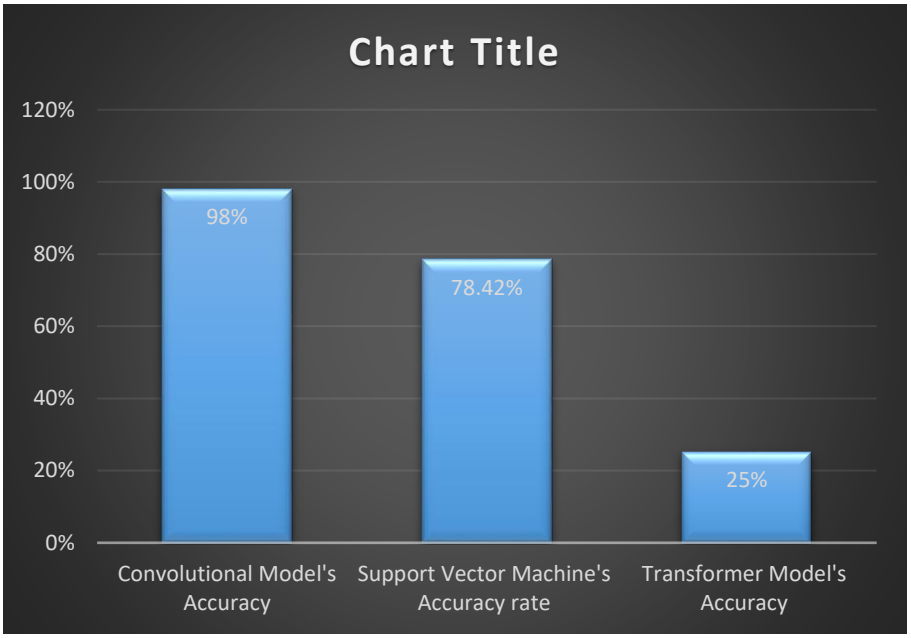


Figure 5: Comparison of result of Three Models

CONCLUSION AND RECOMMENDATIONS

The objective of this research is to make Urdu Optical character recognizer using Support Vector Machine algorithm and in the second step, we have compared the accuracy rate of SVM, CNN and transformer model. Algorithm itself matters in result but less, there are the following reasons: the first thing is dataset, it should be large enough for getting a good accuracy as in at el [2] they collected 38000 samples to Convolutional Neural Network and achieved 98% accuracy and we have given 29230 samples. The second thing is technique of feature extraction from these samples for prediction. By using the best technique like Histogram of Gradient is very common in SVM algorithm for feature extraction but we have Principal Component Analysis (PCA) technique for the said purpose because it is very easy to implement. And the last reason of their getting good result is model or algorithm: the classifier Convolution Neural Network, they used to be very common and adoptable for every type of natural language because of its best features.

Although, the result of our proposed model can be improved by adopted the professional well preprocessed and large enough dataset of Urdu language. For prediction and best result feature extraction technique.

DATA AVAILABILITY

This research will be available on LC International Journal of Stem.

CONFLICT OF INTEREST

We did not have any conflict of interest including any financial, personal or other relationships with other people.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

FUNDING SOURCE

None.

ACKNOWLEDGMENT

I have no words at my command to express my deepest sense of gratitude and thanks to the most beneficent, the most compassionate, and the most gracious Almighty Allah, and I thank Him with utmost gratitude for giving me blessings, opportunity, determination, and strength to do and complete my research. His continuous grace and mercy were with me throughout my life and even more during the tenure of my research.

We would like to extend our sincere gratitude to our HoD **Dr. Kashif Saghar** for his patience, motivation, enthusiasm, and vast knowledge who supported us wholeheartedly, we are also thankful to all the internal and external reviews of my thesis for their valuable scholarly comments to make the thesis further perfect.

REFERENCES

[1] A. Pal, and S. Dayashankar , "Handwritten English character recognition using neural network," *International Journal of Computer Science & Communication*, no. 1, pp. 141-144, 2010.

[2] Ahmed, Gulzar, et al. "Recognition of Urdu Handwritten Alphabet Using "Recognition of Urdu Handwritten Alphabet Using Convolutional Neural Network (CNN)," *Computers, Materials & Continua* , no. 2, pp. 2967-2984, 2022.

[3] Nasien, Dewi, Habibollah Haron, and Siti Sophiayati Yuhaniz "Support Vector Machine (SVM) for English handwritten character recognition.," *2010 Second international conference on computer engineering and applications*, vol. Vol. 1. IEEE, no. 3, pp. 249-252, 2010.

[4] Naz, S, Umar, A. I., Ahmad, R., Siddiq, B., Razzak, M and & Shafait, F, "Urdu Nastaliq recognition using convolutional–recursive deep learning," *Neurocomputing*, vol. 243, no. 4, pp. 80-87, 2017.

[5] Saqib, Nazmus, et al. "Convolutional-neural-network-based handwritten character recognition: an approach with massive multisource data." Algorithms 15.4 (2022): 129.

[6] Huque, Abu Sayeed Ahsanul, et al. "Comparative Study of KNN, SVM and SR Classifiers in Recognizing Arabic Handwritten Characters Employing Feature Fusion." *Signal and Image Processing Letters* 1.2 (2019): 41-49.

[7] Daniyal Shaiq, Mohammad, Musa Dildar Ahmed Cheema, and Ali Kamal. "Transformer based Urdu Handwritten Text Optical Character Reader." *arXiv e-prints* (2022): arXiv-2206.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-05 | Issue-03 | September-2024

[8] Phangtriastu, "Comparison between neural network and support vector machine in optical character recognition," *Procedia computer scienc*e, vol. 116, no. 5, pp. 351-357, 2017.
[9] Navaneetha Krishnan M, "Comparative Analysis of Convolutional Neural Neural Network and Character Recognition Techniques for Handwritten Mathematical Equation Solver," *Journal of Survey in Fisheries Science*s, no. 9, pp. 1609-1632, 2023

AUTHORS PROFILE (All author profiles are mandatory)

Authoor-1 Muhammad Murad, completed MCS from University of Balochistan University Quetta, in 2018. Currently serving as lecturer at Balochistan College Department. Additionally, I am doing MSCS from Al-Hamad Islamic University, Quetta. For competition of MSCS degree we are required to perform research. For the said purpose we haven chosen the topic" Comparative analysis of OCR for Urdu language" in which we had to make new Urdu OCR then it is to compare with already existing Urdu OCR. While performing this research the following areas were completed by me: Writing synopsis, Abstract introduction, Introduction, obtaining result and implementing OCR through python programming language and getting result from it and I also work in the area of conclusion of this research with help of our author-3. Email id : Muhammadmurad445@gmail.com

Authoor-2 Muhammad Shahzad, completed MCS from University of Balochistan Quetta, in 2008. Currently serving as Deputy Director in Balochistan University of Information Technology and Management Sciences Quetta. I am doing MSCS from Al-Hamad Islamic University, Quetta. For competition of MSCS degree we are required to perform research. For the said purpose we have chosen the topic" Comparative analysis of OCR for Urdu language" in which we had to make new Urdu OCR then it is to compare with already existing Urdu OCR. While performing this research I worked in the area of literature review, review whole the work and communicate the progress with supervisor. Email id:zadbaloch123@gamil.com

Authoor-3 Naheeda Fareed, competed BSIT from University of Balochistan University Quetta, in 2018. Currently serving as lecturer at Balochistan College Department. Additionally, I am doing MSCS from Al-Hamad Islamic University, Quetta. For competition of MSCS degree we are required to perform research. For the said purpose we have chosen the topic" Comparative analysis of OCR for Urdu language" in which we had to make new Urdu OCR then it is to compare with already existing Urdu OCR. While performing this research the following area of conclusion of this research was completed by me.