Proceedings of the Pakistan Academy of Sciences: A: Physical and Computational Sciences 58(3): 49-58 (2021)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
https://doi.org/10.53560/PPASA(58-3)743

Pakistan Academy of Sciences

Research Article

Deep Learning Based Pashto Characters Recognition

Sulaiman Khan*, and Shah Nazir

Department of Computer Science, University of Swabi, Pakistan

Abstract: In artificial intelligence, text identification and analysis that are based on images play a vital role in the text retrieving process. Automatic text recognition system development is a difficult task in machine learning, but in the case of cursive languages, it poses a big challenge to the research community due to slight changes in character's shapes and the unavailability of a standard dataset. While this recognition task becomes more challenging in the case of Pashto language due to a large number of characters in its dataset than other similar cursive languages (Persian, Urdu, Arabic) and a slight change in character's shape. This paper aims to address accept these challenges by developing an optimal optical character recognition (OCR) system to recognise isolated handwritten Pashto characters. The proposed OCR system is developed using multiple long short-term memory (LSTM) based deep learning model. The applicability of the proposed model is validated by using the decision trees (DT) classification tool based on the zoning feature extraction technique and the invariant moment approaches. An overall accuracy rate of 89.03% is calculated for the multiple LSTM-based OCR system while DT-based recognition rate of 72.9% is achieved using zoning feature vector and 74.56% is achieved for invariant moments-based feature map. Applicability of the system is evaluated using different performance metrics of accuracy, f-score, specificity, and varying training and test sets.

Keywords: Optical Characters Recognition, Decision Trees, Deep Learning, Pashto, Zoning Technique, Invariant Moments, Long Short Term Memory.

1. INTRODUCTION

Handwriting not only varies from person to person but sometimes, most people cannot even read and understand their own handwritten notes. Handwritten letters are vague in nature as the handwritten letters have no perfectly sharp straight lines or sharp curves like printed letters. Furthermore, the handwritten letters are not only drawn in different font sizes and styles but they are often written in different positions in the specified location (defined cell); for example, sometimes some people write text in the centre, some in the bottom, some in the right and some in the left position of the cell, that is also a challenging task in recognition problem. Since from the early days of Computer pattern processing is a natural way of communication between the computer and the human being, that's why it is the most interesting and challenging field of research in the field of machine learning and pattern recognition with a large number of applications.

From all over the world, the researcher takes a keen interest in the development of handwritten characters recognition models in different languages around the world. Some of the researchers developed a novel approach for the character recognition models like Sara *et al*., [1] proposed spatial-temporal based features for the recognition of cursive text in Arabic/Persian languages. Rafeeq *et al*., [2] and Khan *et al*., [3] proposed the concepts of a deep neural network for the recognition of Urdu ligatures. While some researchers like Hussain *et al*., [4] and Tagougui *et al*., [5] worked on the existing techniques and presented survey papers for addressing the limitations in the available studies. To address the trade-offs in the existing research work, tremendous advances are made to the existing intelligence algorithms to provide a simple and most efficient tool for developing intelligent character recognition models. The arrival of deep learning-based models has revolutionised many fields such as healthcare [6], network security [7, 8], pervasive computing

Received: April 2021; Accepted: September 2021
*Corresponding Author: Sulaiman Khan <engr.sulaiman88@gmail.com>

[9], and many other fields. Deep learning has gained significant attention from the research community for different research problems due to its automatic feature extraction capabilities. Especially, in the optical character recognition development process, many researchers around the globe have proposed neural network approaches such as Naz *et al*., [10, 11] proposed a multi-dimensional recurrent neural network and convolutional recursive deep learning approach for the automatic recognition of Urdu text. ElAdil *et al*., [12] proposed a trained convolution neural network for the recognition of Arabic text using beta filters. A multi-step hybrid approach is suggested by Jabbar *et al*., [13] for Urdu text mining and stemming purposes. Jehangir *et al*., [14] proposed linear discriminant analysis for the automatic recognition of the handwritten Pashto characters using Zernike moments as a feature extractor, while Huang *et al*., [15] proposed zoning and histogram of oriented gradients for the recognition of the Pashto characters.

After studying and analysing the existing research work reported in the domain of cursive text recognition, it was concluded that currently, most of the researchers proposes deep neural networks for text classification and recognition purposes in cursive languages due to high recognition abilities compared to the traditional shallow architectures (support vector machine, Naïve Bayes, K nearest neighbours, random forest, and other generic classification techniques). The main contributions of the proposed research work are to present an optimal OCR model for the recognition of isolated handwritten Pashto characters. This model consists of zoning techniques and invariant moments for feature extraction, and multiple LSTM-based architectures are used for classification and recognition purposes.

The applicability of the proposed OCR system is tested by validating its results with generic decision tree recognition results. Other performance metrics such as specificity, f-score, error-rate, varying training and test sets, time consumption are also used to check the applicability of the proposed deep learning-based OCR system.

This benchmark work will encourage other researchers to further explore the proposed field (recognition of cursive text in the Pashto language) by applying classification and recognition tools. Also, they can extend their work to other cursive languages such as Arabic, Persian and Urdu.

The rest of the paper is organised as follows. Section 2 represents the Pashto language. Section 3 explains the proposed methodology. Section 4 provides details about the results of the experiments reported for the identification of the handwritten Pashto characters using both deep learning and shallow architectures based on zoning and invariant moments feature maps. It also outlines the applicability of the proposed OCR system by evaluating the model using different performance metrics. Section 6 concludes the proposed research work in detail.

2. PASHTO LANGUAGE

Pashto is the official language of Afghanistan and a major language in northern areas (Khyber Pakhtunkhwa) in Pakistan. According to the census 2008 and 2009, it was estimated that around 60 million people around the world could understand this language [16]. It has borrowed characters from Arabic, Urdu and Persian languages. It has encapsulated all the characters of Arabic (27 characters), Persian (32 characters) and Urdu (38 characters) languages with some modification Urdu language follows Nasta'liq writing style. Pashto has encapsulated all the 38 characters of Urdu language with the modification of Urdu specific characters, as shown in Table 1. Table 2 represents the Pashto specific characters. Pashto follows Naskh writing style while Nasta'liq writing style is to an extent.

After borrowing the 37 characters from Urdu language the Pashto language add seven more special characters to the borrowed character to develop a character dataset of 44 characters as shown in (Figure 1). Slight change in between

Table 1. Urdu characters modified in the Pashto language.

Urdu characters	ٹ	ڈ	ڑ	گ	ے
Pashto Equivalent	ټ	ډ	ړ	ګ	ی

Table 2. Pashto specific characters

څ	څ	ږ	ښ	ې	ې	ئ

different character's shape make the recognition task more challenging. The unavailability of a standard dataset for the proposed language impedes the researchers in developing an optimal OCR system for the Pashto language.

Research work on developing an OCR model is reported in several languages (Chines, Japanese, English, Urdu, Persian, Arabic and many more) around the globe, but there is no work reported for the recognition of this language. A little work is reported for printed letters in Pashto language like Nasir *et al*., [17] proposed neural network for Pashto optical characters recognition. On the other side, many researchers developed state of the art techniques for the recognition of Arabic languages like ElAdil *et al*., [12] presented the concept of a trained convolution neural network for the recognition of Arabic letters based on selected beta filters. Das *et al*., [18] proposed the concepts of extreme learning machine (ELM) for the recognition of handwritten characters in the Arabic language.

3. METHODOLOGY

The experimental setup of the proposed OCR system is depicted in Figure 2. This system consists of three building blocks: data collection block, feature extraction block, and the classification and recognition block. Data collection is the first and core step of any recognition model. No OCR system can be developed without data (input). To address this problem, a handwritten Pashto characters database developed by Khan *et al*., [16] is used for the simulation and experimental work. While zoning techniques and invariant moments are suggested as feature extraction tools in the proposed research work. LSTM-based deep learning architecture is used to recognise the handwritten Pashto characters in this research work. The applicability of the system is tested using a generic classification tool named decision trees. Other performance metrics are also used for evaluation and validation purposes. These metrics include specificity, accuracy, f-score, time consumption, varying training and tests.

High recognition rates compared to the traditional DT-based classification model reflects the applicability of the proposed OCR system for the identification of the handwritten Pashto characters. As explained in section 2 that Pashto has accumulated characters from Arabic, Persian and Urdu languages with some modifications in a few characters shape, so we can say that the proposed system can also be applied for the recognition of

Pashto Character Set

S/No	Alphabet	S/No	Alphabet	Name	Alphabet
1	ا	16	ر	31	ق
2	ب	17	ړ	32	ک
3	پ	18	ز	33	ګ
4	ت	19	ژ	34	ل
5	ټ	20	ږ	35	م
6	ث	21	س	36	ن
7	ج	22	ش	37	ڼ
8	څ	23	ښ	38	و
9	چ	24	ص	39	ه
10	ځ	25	ض	40	ي
11	ح	26	ط	41	ې
12	خ	27	ظ	42	ی
13	د	28	ع	43	ۍ
14	ډ	29	غ	44	ئ
15	ذ	30	ف		

Fig 1. Pashto characters set

the handwritten characters in these languages.

3.1 Data Description

For the experimental and simulation work the handwritten Pashto characters database developed by Khan *et al*., [16] is used that consists of 4488 handwritten Pashto samples. These samples are accumulated from different people (both male and female), of different ages and with varying styles of writing. It consists of 102 samples for each 44 characters in the Pashto language. Details of the dataset used for the training and testing purposes is shown in Table 3. 65% of the data is used for training purposes, while the remaining data is used for testing purposes. Where 66 of each character's samples are contributed to the training phase while the remaining 36 characters contributed in the testing phase.

3.2 Feature Extraction

Feature extraction is the most important step in any OCR development process because the recognition capabilities of any OCR system is solely dependent on the most accurate features of a certain problem to make decisions. But the selection of an accurate and optimal feature extraction technique is the most difficult task because the overall recognition task solely depends upon the feature extraction map and ultimately, the high validity of a feature extraction algorithms promises the high classification results. In the proposed research work two different feature extraction techniques are used for the accumulation of astute features of the handwritten Pashto characters.

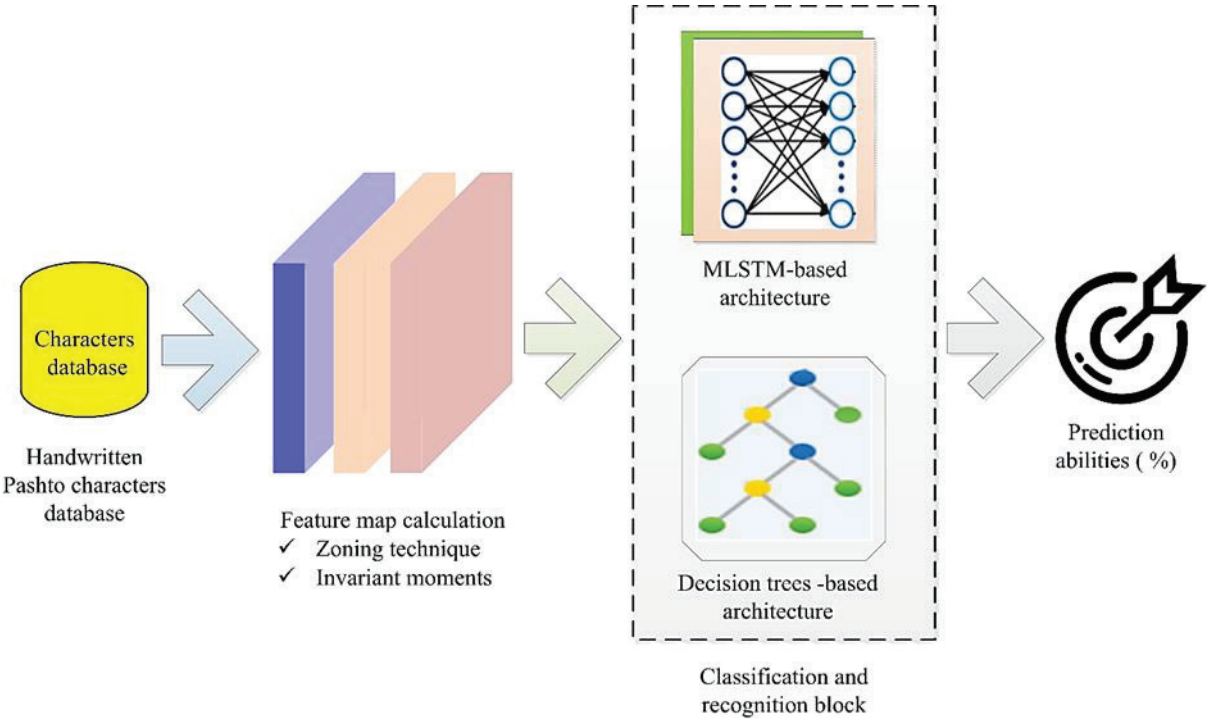

Fig. 2. Proposed methodology

Table 3. Description of the samples used in this research work.

Total number of samples	Samples used for training purposes	Samples used for testing purposes
4488	2917	1571
Individual character's contribution	66	36

3.2.1 Zoning Technique

The zoning technique works by superimposing a static or dynamic grid/zones over the interested image and calculates features using geometrical calculations. These calculations include averaging, summation, and many other types. In our case, we calculated zoning features by counting the number of character's pixels in a particular grid. This process is continued for the 64 (8×8) grids. Many researchers like Naz *et al*., [19] and Khan *et al*., [16, 20] suggested zoning techniques for feature extraction in Urdu and Pashto languages, respectively.

After applying this technique, a feature map is developed that is further used in the recognition phase. An overall accuracy rate of 72.9% is achieved for the decision tree-based classifier using a zoning feature vector.

3.2.2 Invariant Moment Technique

Invariant moments are up to 3rd order normalised spatial moments defined by HU in 1962 [21]. These features remained unchanged under image rotation, scaling and translation. Mathematically it can be represented by equation 1.

$$\eta_{p,q} = \frac{\mu_{p,q}}{(\mu_{0,0})^{\gamma}} \quad (1)$$

Where $\gamma = \frac{P+q+2}{2}$

In equation (1) "p" and "q" are the normalised moments in x and y axes while η_(p,q) is the spatial central moment of order (p, q). The central moments can be represented using equation (2).

$$\eta_{p,q} = \sum_{m=0}^{row-1} \sum_{n=0}^{col-1} ((n - C_n)^p . (m - C_m)^q . I_{m,n}) \quad (2)$$

Where "row" and "col" represents the number of rows and columns in the image "I", and C_n,C_m represents the centre of the interested image.

Many researchers have proposed this technique in many pattern recognition and text recognition problems such as Michael *et al*., [22] proposed moments for 2D flow detection purposes while Chen *et al*., [23] performed a comparative analysis of the Fourier descriptor and HU moments for image recognition purposes.

3.3 Classification Techniques

Classification is the core heart of any OCR system in cursive text recognition. Researchers around the globe suggest multiple classification techniques to achieve high recognition abilities. These techniques include both deep and shallow architectures. Mouhcine *et al*., [24] proposed a hidden Markova model (HMM) for the recognition of handwritten cursive Arabic text. Marie-Sainte and Alyani used the Firefly algorithm for the classification of Arabic text [25]. After analysing the existing research work reported (2008 – 2020 (a section of 2020 is included)) in the cursive text recognition domain, it was concluded that after 2010 the deep learning-based recognition models gained significant attention of the research community in many research problems like traffic prediction [26], object detection [27], and characters recognition [28], network security [29] as depicted in (Figure 3). This significant attention is due to automatic feature extraction capabilities and achieving high recognition rates for deep learning architectures in many pattern recognition problems especially in cursive text recognition problems. Elleuch *et al*., [30] proposed a deep belief network-based regularisation method for handwritten Arabic script recognition. Humaidi and Kadhim proposed a spiking neural network for the recognition of Arabic characters [31].

This research work has proposed a deep learning-based architecture MLSTM for the OCR development process. The performance of the MLSTM-based classifier is validated by testing it with the generic DT-based OCR system recognition capabilities.

4. PERFORMANCE EVALUATION AND ANALYSIS

The proposed OCR system is tested for the handwritten Pashto characters as specified in (Table 3). An overall accuracy rate of 89.03% is calculated for the MLSTM-based OCR model. This 18.97% error rate is because of a similar character's shape for many characters in the Pashto language that caused the miss-classification rate. As most of the characters in the Pashto language share the same basic shape only differs by secondary shape (diacritics/number of dots below or above a certain character) such as س (seen) and ښ (heen), څ (seem)

and چ (che). These characters are only differed by the number of dots or dots position. The results of the MLSTM-based OCR system are depicted in Figure 4.

The performance of the MLSTM-based classifier is validated by testing it with the generic DT-based OCR system recognition capabilities. After testing it was concluded that the MLSTM generates accuracy rate of 89.03% much better than DT-based recognition rates of 72.9% and 74.56% for zoning feature map and invariant moments-based feature map respectively as depicted in Figure 5 and Figure 6.

From the above graph, it is concluded that due to slight in character's shape makes the recognition task is restricted. The results of DT-based classifier using invariant moments feature map is depicted in (Figure 6). An overall accuracy rate of 74.56% is achieved. A miss-classification rate of 15.44% is reported for the DT-based OCR model. It concludes that the decision tree-based model fails in classifying similar characters. In Pashto language, most of the characters such as (ج, څ, چ, څ, ح), (س, ش, ښ) and many other characters have same basic shape only differs by secondary parts (diacritics) that cause miss-classification. The DT-based OCR system confuse these characters and treat them as the same characters.

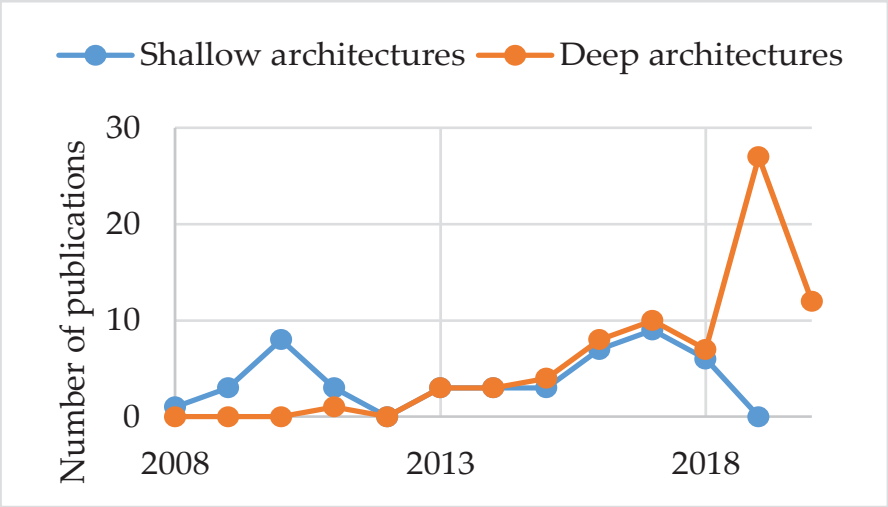

Fig. 3. Annual work reported for deep and shallow architectures for cursive text (Urdu, Arabic, and Persian) recognition.

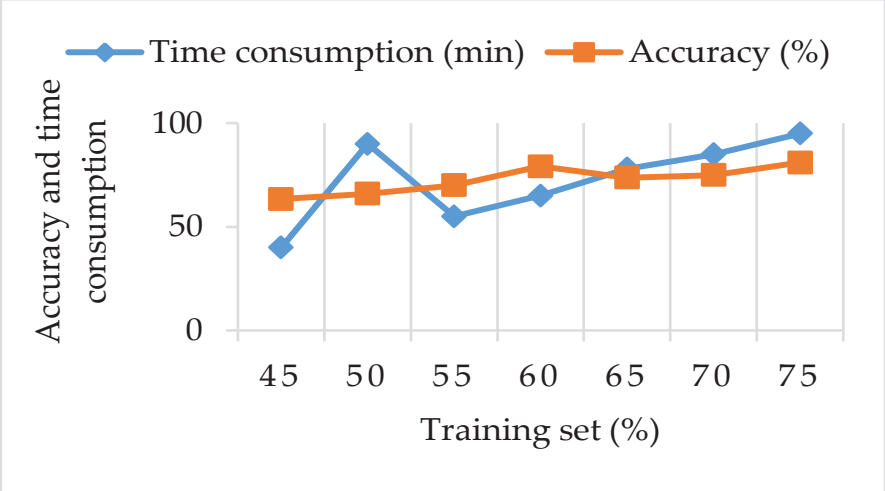

Fig. 4. MLSTM-based OCR results

After analysing the performance metrics of accuracy, recall, precision, F-score and misclassification rate based on true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) rates, it was concluded that the model outperforms by generating a recognition rate of 89.03% for the detection of handwritten Pashto characters. The results generated by MLSTM-based OCR system are shown in Figure 7.

The applicability of the proposed system is validated by comparing the proposed model identification capabilities with the decision trees-based OCR model. The performance results of the decision trees-based classifier are depicted in Figure 8 and Figure 9, respectively.

From Figure 8 and Figure 9, it is concluded that an overall accuracy rate of 72.9% is generated for the decision tree-based classification technique using zoning feature map, while 74.56% is generated for the decision tree-based classification technique using invariant moments feature map. A recognition rate of 89.03% is generated for MLSTM-based classification technique. The high recognition rate of 89.03% shows the applicability of the proposed deep learning-based model for the accurate identification of handwritten Pashto characters.

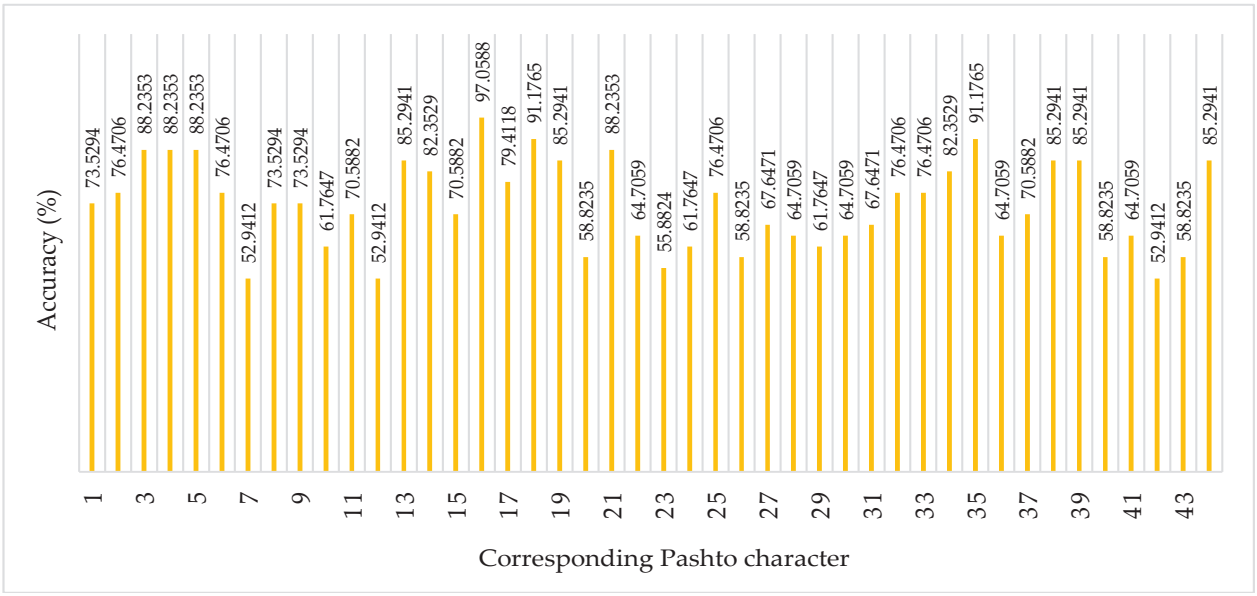

Fig. 5. DCT-based recognition results of the proposed OCR system using zoning feature map.

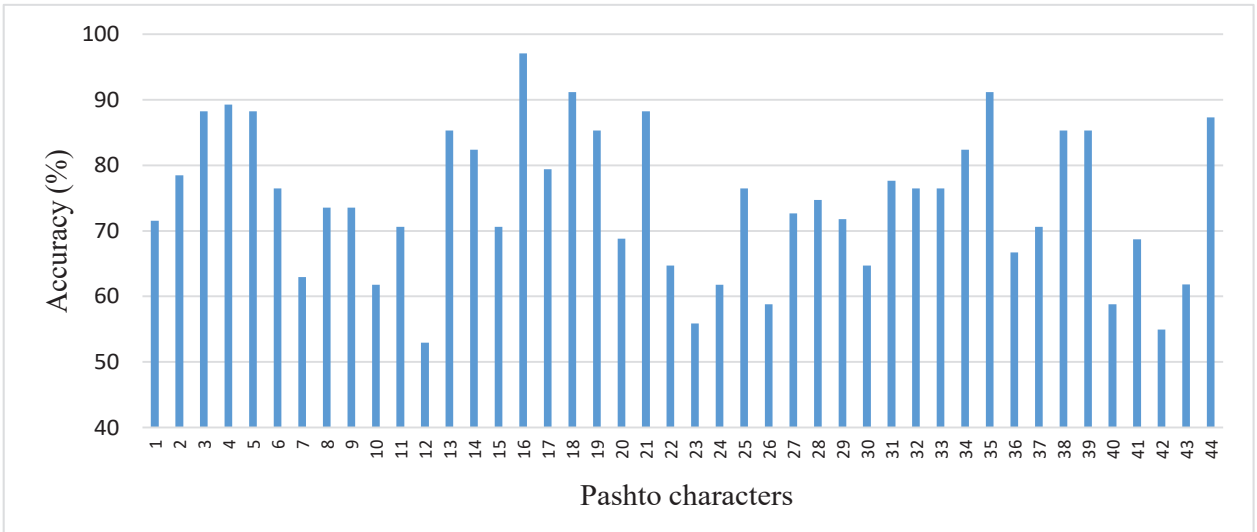

Fig. 6. DCT-based recognition results of the proposed OCR system using invariant moments feature

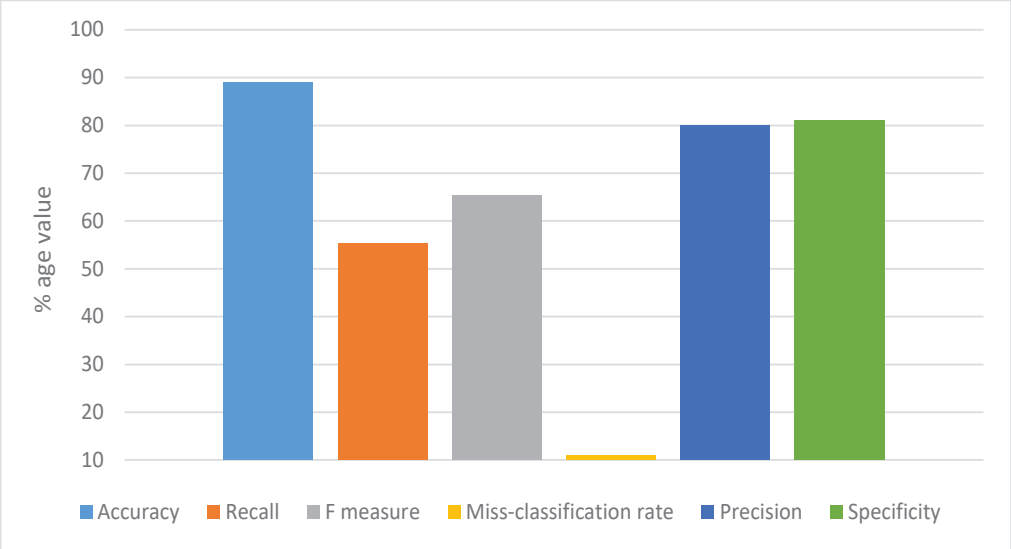

Fig. 7. MLSTM-based performance results

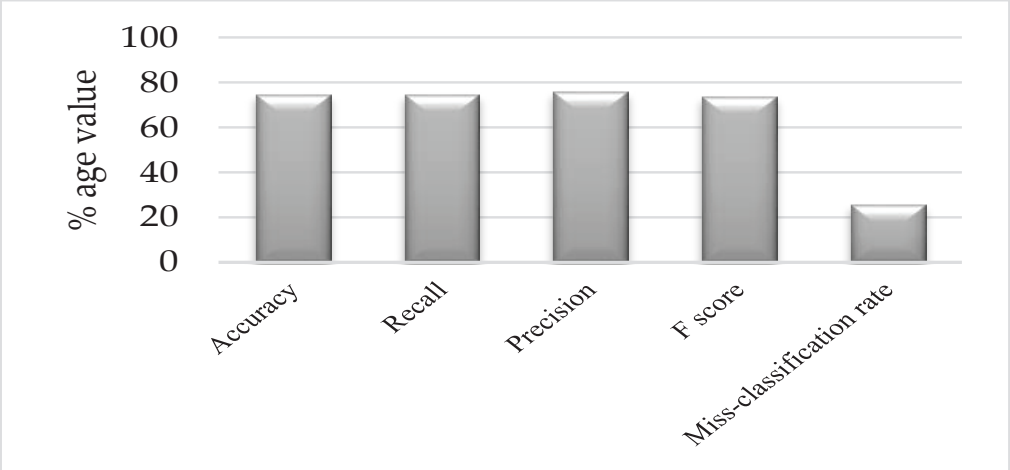

Fig. 8. DT-based performance results for invariant moments feature map.

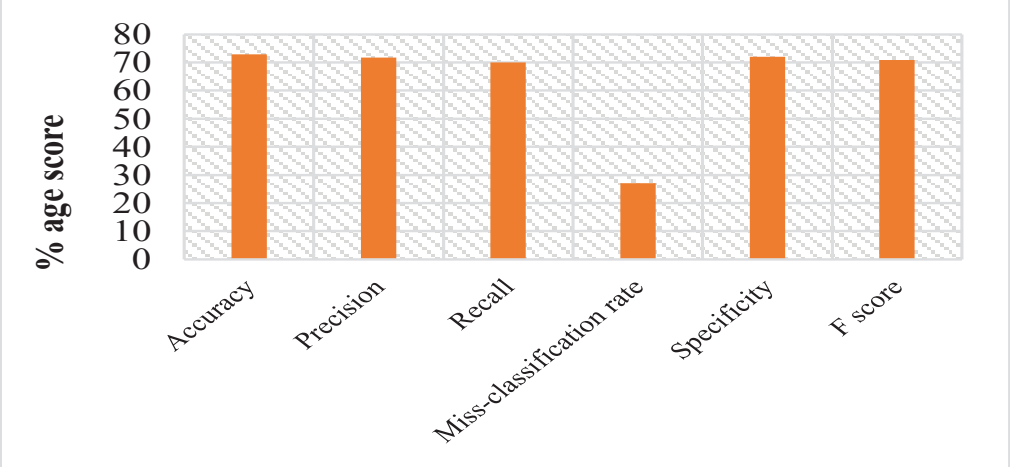

Fig. 9. Decision trees-based performance results using zoning feature map.

5. CONCLUSION

Cursive text recognition is considered the most challenging task in the fields of machine learning and pattern recognition, but in the case of Pashto language, it poses more hurdles to the research community due to slight changes in character's shape and large number of characters in its character's database. The proposed research work presents the development of an optimum OCR system for the recognition of isolated handwritten Pashto characters using MLSTM-based deep learning approach. Applicability of the proposed model is validated by using the decision trees classification tool based on zoning feature extraction technique and invariant moments-based approaches. An overall accuracy rate of 89.03% is calculated for the MLSTM-based OCR system, while DT-based recognition rates of 72.9% are achieved using zoning feature vector and 74.56% is achieved for invariant moments-based feature map. Applicability of the system is evaluated using different performance metrics of accuracy, f-score, specificity, and varying training and test sets. The contribution of this research work includes the providence of a benchmark simulated results for the recognition of the Pashto languages.

In future, we want to extend the proposed work for the automatic recognition of the cursive Pashto text. Also, we want to test the applicability of the proposed deep learning-based model for other similar cursive languages.

6. ACKNOWLEDGMENT

This research work is supported by Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar AND Department of Computer Science, University of Swabi, KP, Pakistan.

7. CONFLICT OF INTEREST

Author declares no conflict of interest.

8. REFERENCES

1. S. Valikhani, F. Abdali-Mohammadi, and A. Fathi. Online continuous multi-stroke Persian/Arabic character recognition by novel spatio-temporal features for digitiser pen devices. Neural Computing and Applications p. 1-20 (2019).
2. M. J. Rafeeq, Z. ur Rehman, A. Khan, I. A. Khan, and W. Jadoon. Ligature categorisation based Nastaliq Urdu recognition using deep neural networks. Computational and Mathematical Organisation Theory 25: 184-195 (2019).
3. W. Khan, A. Daud, F. Alotaibi, N. Aljohani, and S. Arafat. Deep recurrent neural networks with word embeddings for Urdu named entity recognition. ETRI Journal (2019).
4. R. Hussain, A. Raza, I. Siddiqi, K. Khurshid, and C. Djeddi. A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation. EURASIP Journal on Image and Video Processing, 2015: 46, (2015).
5. N. Tagougui, M. Kherallah, and A. M. Alimi. Online Arabic handwriting recognition: a survey. International Journal on Document Analysis and Recognition (IJDAR) 16: 209-226 (2013).
6. H. Chen, S. Khan, B. Kou, S. Nazir, W. Liu, and A. Hussain. A Smart Machine Learning Model for the Detection of Brain Hemorrhage Diagnosis Based Internet of Things in Smart Cities. Complexity 2020: 3047869 (2020).
7. Z. Gu, S. Nazir, C. Hong, and S. Khan. Convolution Neural Network-Based Higher Accurate Intrusion Identification System for the Network Security and Communication. Security and Communication Networks 2020: 8830903 (2020).
8. Y. He, S. Nazir, B. Nie, S. Khan, and J. Zhang. Developing an Efficient Deep Learning-Based Trusted Model for Pervasive Computing Using an LSTM-Based Classification Model. Complexity 2020: 4579495 (2020).
9. S. Wang, S. Khan, C. Xu, S. Nazir, and A. Hafeez, Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers. Complexity 2020: 8694796 (2020).
10. S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, and M. I. Razzak. Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. Neural computing and applications 28: 219-231 (2017).
11. S. Naz, A. I. Umar, R. Ahmad, I. Siddiqi, S. B. Ahmed, M. I. Razzak, *et al*. Urdu Nastaliq recognition using convolutional–recursive deep learning. Neurocomputing, 243: 80-87 (2017).
12. A. ElAdel, M. Zaied, and C. Ben Amar. Trained

convolutional neural network based on selected beta filters for Arabic letter recognition. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9: e1250 (2019).
13. A. Jabbar, S. Iqbal, A. Akhunzada, and Q. Abbas. An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach. Journal of Experimental & Theoretical Artificial Intelligence, 30: 703-723 (2018).
14. S. Jehangir, S. Khan, S. Khan, S. Nazir, and A. Hussain. Zernike moments based handwritten Pashto character recognition using linear discriminant analysis. Mehran University Research Journal Of Engineering & Technology, 40(1), 152-159 (2021).
15. J. Huang, I. U. Haq, C. Dai, S. Khan, S. Nazir, and M. Imtiaz. Isolated Handwritten Pashto Character Recognition Using a K-NN Classification Tool based on Zoning and HOG Feature Extraction Techniques. Complexity, 2021: 5558373 (2021).
16. S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez. KNN and ANN-based Recognition of Handwritten Pashto Letters using Zoning Features. Machine learning 9: 570-577 (2018).
17. N. Ahmad, M. Naeem, S. A. R. Abid, and A. Gul. Pashto optical character recognition using neural network. Journal of engineering and applied sciences (37) (2018).
18. D. Das, D. R. Nayak, R. Dash, and B. Majhi. An empirical evaluation of extreme learning machine: application to handwritten character recognition. Multimedia Tools and Applications, p. 1-29 (2019).
19. S. Naz, S. B. Ahmed, R. Ahmad, and M. I. Razzak. Zoning features and 2DLSTM for Urdu text-line recognition. Procedia Computer Science 96: 16-22, (2016).
20. M.-K. Hu. Visual pattern recognition by moment invariants. IRE transactions on information theory, 8: 179-187 (1962).
21. S. Khan, A. Hafeez, H. Ali, S. Nazir, A. Hussain. Pioneer dataset and recognition of Handwritten Pashto characters using Convolution Neural Networks. Measurement and Control (2020).
22. M. Schlemmer, M. Heringer, F. Morr, I. Hotz, M. Hering-Bertram, C. Garth, *et al*. Moment invariants for the analysis of 2D flow fields. IEEE Transactions on Visualization and Computer Graphics, 13: 1743-1750 (2007).
23. Q. Chen, E. Petriu, and X. Yang. A comparative study of Fourier descriptors and Hu's seven moment invariants for image recognition. in Canadian conference on electrical and computer engineering 2004 (IEEE Cat. No. 04CH37513), p. 103-106 (2004).
24. R. Mouhcine, A. Mustapha, and M. Zouhir. Recognition of cursive Arabic handwritten text using embedded training based on HMMs. Journal of Electrical Systems and Information Technology, 5: 245-251 (2018).
25. S. L. Marie-Sainte and N. Alalyani. Firefly algorithm based feature selection for Arabic text classification. Journal of King Saud University-Computer and Information Sciences (2018).
26. S. Khan, S. Nazir, I. García-Magariño, A. Hussain. Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion. Computers & Electrical Engineering, 89: 106906, (2021).
27. S. Khan, S. Nazir, H. U. Khan. Smart object detection and home appliances control system in smart cities. Computers, Materials and Continua, 67: 895-915 (2021).
28. S. Khan, S. Nazir, H. U. Khan, and A. Hussain. Pashto Characters recognition using multi-class enabled support vector machine. CMC-Computers Materials & Continua, 67(3), 2831-2844 (2021).
29. X. Liao, S. Nazir, J. Shen, B. Shen, and S. Khan. Rough Set Approach toward Data Modelling and User Knowledge for Extracting Insights. Complexity 2021: 7815418 (2021).
30. M. Elleuch, N. Tagougui, and M. Kherallah. Optimisation of DBN using regularisation methods applied for recognising arabic handwritten script. Procedia Computer Science 108: 2292-2297 (2017).
31. A. J. Humaidi and T. M. Kadhim. Recognition of Arabic Characters using Spiking Neural Networks. in 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), p. 7-11 (2017).