INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Handwritten Urdu character recognition via images using different machine learning and deep learning techniques

M Ameen Chhajro[1]*, **Hadeeb Khan**[1], **Farrukh Khan**[1], **Kamlesh Kumar**[1], **Asif Ali Wagan**[1], **Sadaf Solangi**[2]

1 Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan
2 Department of Computer Science and Technology, Nanjing University of Science and Technology, China

OPEN ACCESS

Received: 02-04-2020
Accepted: 08-05-2020
Published: 12-06-2020

Editor: Dr. Natarajan Gajendran

Citation: Chhajro MA, Khan H, Khan F, Kumar K, Wagan AA, Solangi S (2020) Handwritten Urdu character recognition via images using different machine learning and deep learning techniques. Indian Journal of Science and Technology 13(17): 1746-1754. https://doi.org/10.17485/IJST/v13i17.113

***Corresponding author**.
M Ameen Chhajro
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan
ameen.chhajro@smiu.edu.pk

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

Abstract

Objectives: This research presents a model for Urdu Handwritten Character Recognition via images using various Machine Learning and Deep Learning Techniques. The main objective of this research is to provide comparative study on Urdu Handwritten Characters from images dataset. **Methods/Statistical analysis**: In this research paper, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN) algorithm, Multi-Layer Perceptron (MLP), Concurrent Neural Network (CNN), Recurrent Neural Network (RNN) and Random Forest Algorithm (RF) have been implemented in order to evaluate most suitable technique for Urdu Handwritten Characters Recognition via images. **Findings**: Ample amount of research work has been carried out on English Language but it is clearly shown through the conducted literature review that very lesser amount of work has been done on Urdu Handwritten Characters Recognition using images. Furthermore, It has been analyzed from this research that CNN models are most efficient compared to RF, SVM and MLP as to produce reliable results in terms of optimal accuracy. Therefore, using the CNN model is a viable choice to recognize Urdu handwritten characters from the images. And proposed study provides significant contribution in automatic learning of Urdu handwritten Characters.

Keywords: Urdu Handwritten Characters; Machine Learning; Deep Learning; Urdu Character Recognition

1 Introduction

The automatic handwritten text recognition via images is considered one of the most difficult tasks in pattern recognition research areas. The challenge is to identify the shape or pattern from the handwritten text. As every word in handwritten text has different shape because of different style of writing which varies person to person. As Urdu is one of the cursive language which is spoken and written in different regions and also known as the national language of Pakistan.

It is being spoken in other countries like Afghanistan, India, Bangladesh or some other languages include the literature of Urdu. When it comes to particularly Urdu handwritten text so the research has been carried out to a very small extent and the very first script was published in 2004 which is Optical Character Recognition. Urdu handwritten text recognition using image is so called ICR (Intelligent Character Recognition)[1–7]. However, no robust system is available to date which could produce reliable results. And for printed text which is called OCR (Optical Character Recognition) so there are few available systems which can be used but only for printed text and not for handwritten. Several machine learning and deep learning algorithms are applied on handwritten Urdu text where we have used the labels from 0 to 39 in counting and not Urdu labels were used. Machine learning algorithms used are Support Vector Machine (SVM) and K-nearest neighbor algorithm whereas the deep learning algorithms used are Multi-Layer Perceptron (MLP), Random Forest, Concurrent/Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)[1,8].

All the above mentioned algorithms have been analyzed and compared to check the final results of these algorithms that which of the algorithm provides better accuracy which can be considered reliable. All the results which have been obtained after applying the above mentioned algorithms are discussed in the paper which shows the difference between the outputs and accuracy through them. The deep learning approaches have been considered more accurate while getting better accuracy as compared to other conventional techniques[9]. It is also recommended by some researches to use the deep learning algorithms for extraction and identification of text.

The Urdu Language has quite complex structure of numbers therefore; it becomes much difficult to recognize the handwritten words. As every person has different writing style, in this regard to detect the word accurately is quite a big challenge in itself. Urdu Language script is written form right to left and it is also a cursive language[10]. The proposed system can be used to provide benefits to the people who are not familiar with Urdu but wants to learn to write Urdu so the person can match its handwritten word in the system that whether he/she has written down that particular alphabet right or not.

Literature review

The background study reveals that very limited amount of work has been carried out on Urdu language from with the perspective of images as compared to other languages. Instead of millions of people speak Urdu but not such commendable system is made for Urdu handwritten character's recognition using images. There is a model present for roman digit recognition which provides 97% accuracy and that model was built using support vector machine by Gorgevik et al[11]. The model extracts four attributes from the given picture and namely those attributes are ring-zone, contour profiles, histogram projections and features of kirsch. They used Principal Component Analysis (PCA) to convert 256 dimensional features into 128 directional features set. The simulations carried out can be used in various applications such as classification and recognition of pattern matching and alphabets recognition. The approach of this model can be used for recognition of different language characters such as Urdu, Sindhi etc. For hand written digit recognition, this model used back propagation technique minimal pre-processing technique was used for data. The model was provided with images of each digit separately and while testing 1% error and 9% rejection rate was shown.

As discussed above that handwritten Urdu character recognition using images is one of the most challenging part due to variations in handwriting and it has been observed through different studies that ANN (Artificial Neural Network) is widely used for character recognition. The ANN model creates bunch of nodes linked together and the neurons which are connected with each other passes the signal and information from one node to the other. The problem which can be faced by ANN that what of information is provided to the model because it takes an input and then trains itself to provide a labeled output. It is also not possible to develop a system which can be used for multiple languages because every language has different style and characters of writing respectively.

A methodology which was based on open mining algorithms and classification survey was presented by Kaur Harpreet et al[12]. There are various approaches present for sentimental analysis such as random forest, SVM, naïve Bayes and many more. The study was carried out to make analysis of IMBD, Amazon and Flipkart dataset to train the model for sentiment analysis of words and to also find the contextual meaning of those words. Another research was performed by Chen Mei-Hua to find the emotions of various text words[13]. The work over emotions of words is found at lower side and yet there is great room for exploration in this field. The main purpose of the study was to introduce a system which can be utilized for educational needs as it will help the people who are learning new languages to find the emotion which a words contains in that particular language. The results of this research provided reliable outputs with good accuracy.

2 Proposed Methodology

Urdu handwritten characters were collected from different people. Every box for writing a single character had equal breadth and width. All the people were asked to write in their own style so that the model can be best trained over versatile samples of handwritten characters. The data was collected from 100 people. All the images are separated and OpenCV filters were used to eliminate the noise from each image and some other techniques were deployed to attain high quality gray scale images which could be used for training and testing the model. The total collection of dataset is 4668 with dimensions 50*50 which is divided into two parts (training and testing). For training the model we used 3734 samples of images and 934 samples of images are used for testing the model.

2.1 Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a feed forward and deep artificial neural network which is used to train and supervise the learning problems to minimize the errors. In MLP, the supervised learning techniques which is used, known as backpropagation for training. Multi-layer perceptron contain three layers:

2.1.1 Input layer

It is used by the network to access the data

2.1.2 Intermediate layer

Used by network as computation machine to convert input into output and also known as hidden layers.

2.1.3 Output Layer

It shows the obtained results.

The multi-layers and non-linear activation is used to differentiate MLP from linear perceptron. Input layers are also known as multiple nonlinear layers whereas intermediate layers as hidden layers. The activation function is used to specify neurons in a network and all the neurons are connected with each other. MLP model also separates the data which is linearly inseparable. Furthermore, the communication between input and output layer is established and managed by MLP model.

2.2 Support Vector Machine (SVM)

Support vector machines are used to analyze, classify and perform linear regression over provided data. It is a supervised learning algorithm. An SVM algorithm creates a model which breaks data into categories and assigns newly created categories to each set of data which makes SVM a non-probabilistic binary linear classifier[14].

In SVM, all the categories of data are represented through mapping it with spaces to make the data plotting and understanding clearly. The newly created one are also mapped through the same technique by considering the gaps between them. Support Vector Machine can also be used to deal with non-linear classification using a trick called kernel trick by mapping inputs into feature spaces having high dimensions.

Support Vector Machine only works for labeled data but for unlabeled data, support vector clustering technique is used which works over unsupervised learning by making clusters of the data. This model was developed to extend SVM for tackling unlabeled data and is widely used for industrial applications.

K-NN algorithm

(K-Nearest Neighbors Algorithm) is used to identify the dataset for regression and classification. It is a non-parametric method where in feature space, it contains examples of k closest training sets. K-NN output mainly rely that whether it is used for classification and regression.

If K-NN is used for regression so it provides the average of all k values and that output is considered as property value for object and if the K-NN is used for classification purpose so it produces a class membership as output[15] .

Until the classification is not done so the whole computation process in KNN is delayed because it is kind of lazy learning or instance based learning approach. Furthermore, a meaningful method which is applied on both classification and regression is to assign the weights to each neighbor. The neighbors who are nearest would put more to the average value than the far ones. Neighbor selection is made either from class which his K-NN classification or from object property value which is K-NN regression. If dealing with local structure of the data so in that case, K-NN algorithm is sensitive.

2.3 Convolution Neural Network (CNN)

It is most commonly used to make analysis of visual images and it is a deep neural networks class. Applications of CNN can be widely found in the fields of image classification, image analysis, image recognition, image prediction, recommendation system, text prediction, video recognition, natural language processing and many more[16]. Moreover, CNN is multilayer perceptron regularized version but in multilayer each neuron connected to every other neuron which makes it fully connected but CNN opts the other technique by regularizing the data. It makes use of simple and small patterns found in hierarchical structure of data therefore, CNN is lesser connected than multilayer perceptron.

CNN inspiration was driven from biological process which deals with the connectivity of neurons to other parts in a body. The pre-processing in CNN is at lower side as compared to other image classification algorithms which proves that CNN learns many things more effectively than other traditional algorithms which consumed more time for data processing. Major advantage of using CNN is that it is independent of human efforts required with other algorithms. CNN is also known as space invariant artificial neural networks (SIANN).

2.4 Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is used for handwritten recognition, speech recognition, unsegmented recognition and it is related to artificial neural networks. A directed graph alongside temporal sequence is created between nodes when connection is established. Input sequences in RNN are processed through utilizing its internal memory. Recurrent Neural Network (RNN) is classified into parts, which are finite and infinite impulse which are directed acyclic graph and directed cyclic graph respectively. It is possible to replace or unroll finite impulse by following feedforward neural network strictly but unrolling and replacing cannot be done in infinite impulse.

In addition, finite and infinite impulse has the ability to store additional states but neural networks possess the control for this storage. If the stored state has time delays or loops in the feedback then it can be replaced with some other networks. The controlled states are known as gated state or memory which are part of gated recurrent units and LSTM (Long Short Term Memory) Networks. It is also named as feedback neural networks[17].

2.5 Random Forest

Random Forest algorithm works on the regression and classification of data. It constructs a decision tree at training portion. For classification and regression, it places classes and mean prediction from each individual tree respectively. Tin Kam Ho was the first person to construct an algorithm for Random Decision Forest which used stochastic discrimination approach for the classification of data. Leo Breiman and Adele Cutler together constructed and registered Random Forest Algorithm as trademark in 2019. It was extended to be able to create decision tree collections for variance control[18].

3 Results and Discussion

3.1 Description of dataset

The dataset has been created our own, the procedure of dataset collection started by writing hand written urdu characters on white papers. Data was written in such a way that each row consists of same hand written character written by different people along with one column that is considered as the label. Label was there for each hand written characters because machine learning supervised models has been applied for training and testing purpose in this research study. In order to create big corpus of hand written characters the same task of was communicated to our class students. The same approach of data collection was distributed throughout the university students. The idea was to cover maximum students so that we can have variety of hand written characters so that machine learning models can be trained well on different handwritten characters. The Figure 1 shows the Urdu digits data distribution. The dataset consist of 4668 total samples with each having 50*50 dimension of gray scale image. The dataset is labeled from 0 to 39 and also divided into two parts (training and testing). From the dataset 3734 sample images are utilized for training and 934 sample images for testing purpose respectively.

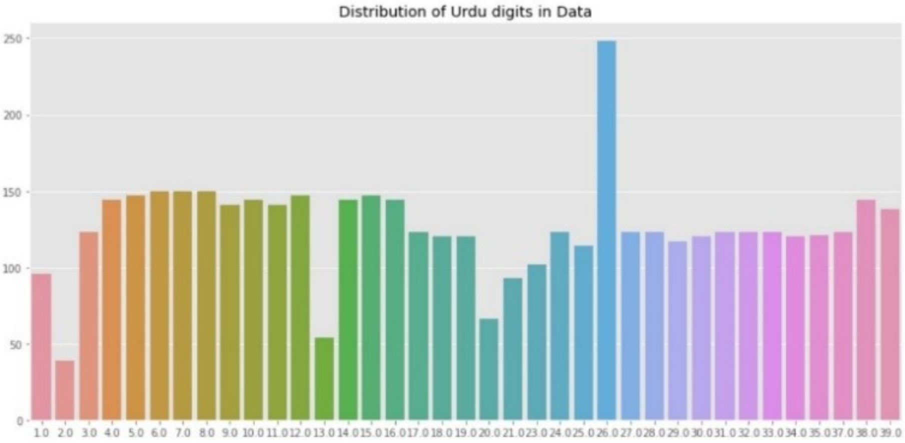

Fig 1. Distribution of Urdu digits in data

3.2 Accuracy evaluation

The accuracy of different algorithms used for the recognition of Urdu Handwritten text from images by using machine learning and deep learning techniques as shown in Table 1. Furthermore, the conclusive accuracy results of various Machine learning and Deep learning techniques have been plotted in Figure 2.

Table 1. Accuracy of different algorithms

S.no	Algorithms	Accuracy
1	Support Vector Machine (SVM)	97 %
2	K-Nearest Neighbor (K-NN)	38 %
3	Random Forest	97 %
4	Concurrent Neural Network (CNN)	99 %
5	Recurrent Neural Network (RNN)	80 %
6	Multi-Layer Perceptron (MLP)	98 %

It can be clearly seen from the table1 that Support Vector Machine and Random Forest both provided 97% accuracy over the given dataset of 4668 sample which had 3734 samples for training and remaining 934 is used

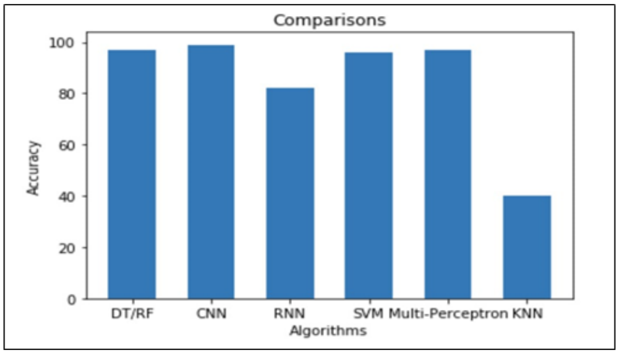

Fig 2. Accuracy comparison graphs of algorithms

for testing. Furthermore, the digit recognition results of both techniques from the handwritten characters from the image's dataset has been shown in Figure 3 and Figure 4 respectively.

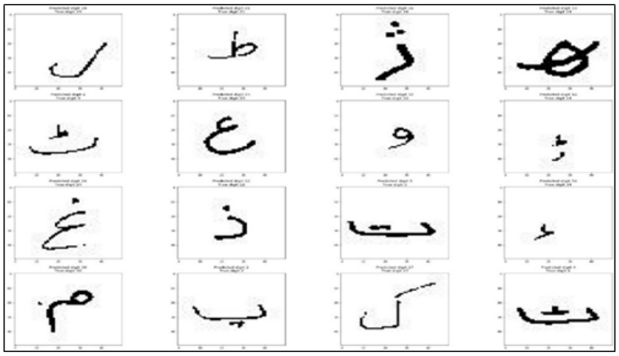

Fig 3. Support Vector Machine (SVM)

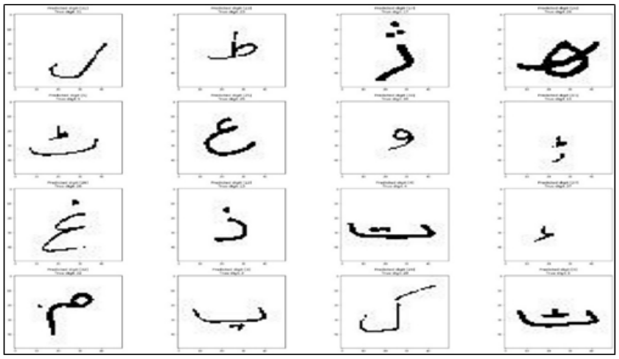

Fig 4. Random Forest (RF)

The K-Nearest Neighbor proved to be the worst algorithm for recognition of handwritten Urdu text from images because it attained the accuracy of only 38% which can be observed from the Table 1. In order to analyze the character recognition result of the KNN it is shown in Figure 5 **.** Whereas, Multi-Layer Perceptron is one of the best algorithms which can also be used for Urdu handwritten text recognition from images as it gets the accuracy of 98%.MLP result of character recognition is shown in the Figure 6.

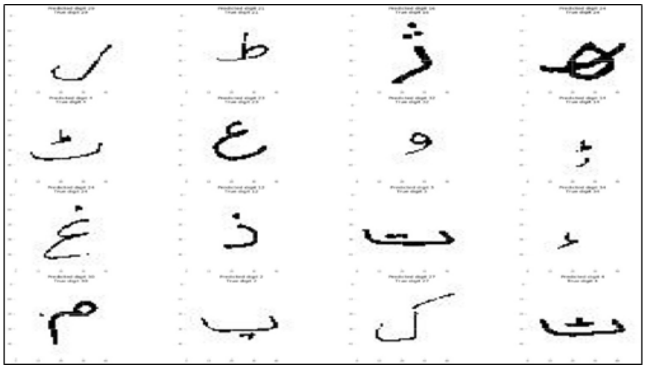

Fig 5. K-Nearest Neighbor (KNN)

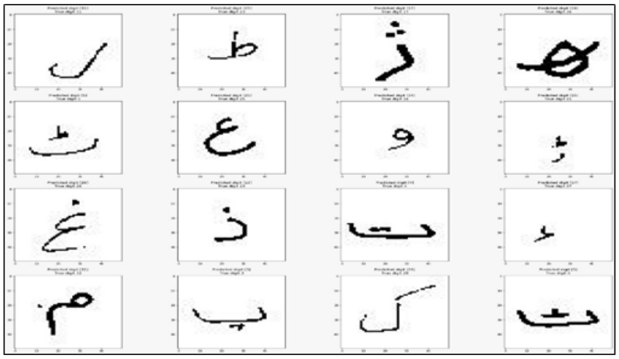

Fig 6. Multi-Layer Perceptron(MLP)

Concurrent Neural Network is the best algorithm to attain the best accuracy for Urdu handwritten text recognition from images because we get the accuracy of 98% from this technique in table1 along with its recognition result which can be seen in Figure 7. We also implemented Recurrent Neural Network to check whether it gets the reliable accuracy or not but it only attained 80% which is somewhat average accuracy. We can refer Figure 8 for its result of character recognition from the images.

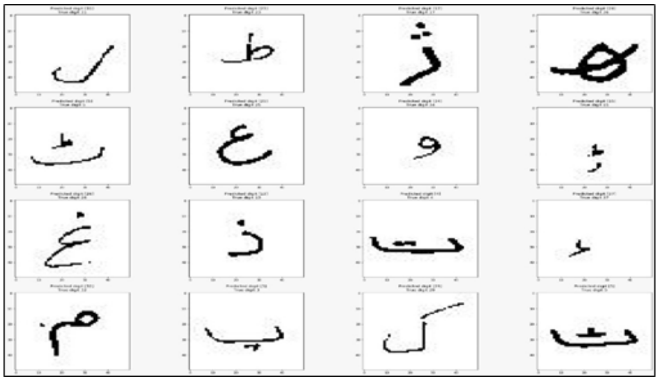

Fig 7. Concurrent NeuralNetwork (CNN)

Amongst the entire algorithms of this research, Multi-Layer Perceptron and Concurrent Neural Network are the best technique which could be used for Urdu handwritten text recognition from images as these two algorithms provide best and reliable results because Deep Learning algorithms work more effectively over text recognition from

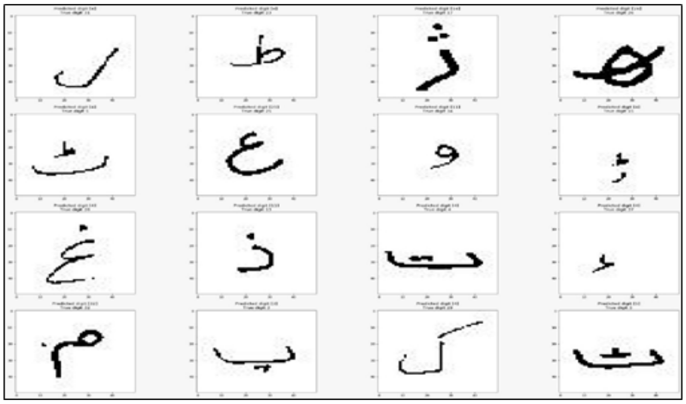

Fig 8. Recurrent NeuralNetwork (RNN)

images as compared to Machine Learning algorithms.

4 Conclusion

In this paper, the research was conducted on handwritten optical character recognition via images. For this purpose, we implemented different algorithms namely: Random Forest, Support Vector Machine (SVM), Concurrent Neural Network (CNN), Recurrent Neural Network (RNN), Multi-Layer Perceptron and K-Nearest Neighbor (KNN). The data set was divided into 20% and 80% for testing and training respectively.

The obtained results showed that CNN and MLP attained accuracy of 99% and 98% respectively during the handwritten characters recognition via same image dataset which is higher compared to 97%, 97%, 80%, and 38% obtained by Random forest, Support Vector Machines, Recurrent Neural Network and K-Nearest Neighbor respectively. These figures show that our study provides significant contribution in automatic optical character recognition of Urdu phonetics.

5 Future work

The proposed model works well for given input image; however, due to versatile style of writing, this research work has some limitation on hand written character recognition. For this purpose, in future we plan to extend existing techniques on more sophisticated Urdu language phonetics.

References

1) Husnain M, Missen MMS, Mumtaz S, Jhanidr MZ, Coustaty M, Luqman MM, et al. Recognition of Urdu Handwritten Characters Using Convolutional Neural Network. *Applied Sciences*. 2019;9(13):1–21. doi:10.3390/app9132758.
2) Khan NH, Adnan A. Urdu Optical Character Recognition Systems: Present Contributions and Future Directions. *IEEE Access*. 2018;6:46019–46046. doi:10.1109/access.2018.2865532.
3) Naz S, Umar AI, Ahmad R, Siddiqi I, Ahmed SB, Razzak MI, et al. Urdu Nastaliq recognition using convolutional–recursive deep learning. *Neurocomputing*. 2017;243:80–87. doi:10.1016/j.neucom.2017.02.081.
4) Rizvi SSR, Sagheer A, Adnan K, Muhammad A. Optical Character Recognition System for Nastalique Urdu-Like Script Languages Using Supervised Learning. *International Journal of Pattern Recognition and Artificial Intelligence*. 2019;33(10). doi:10.1142/s0218001419530045.
5) Sardar S, Wahab A, IEEE. Optical character recognition system for Urdu. In: and others, editor. International Conference on Information and Emerging Technologies. 2010;p. 1–5. Available from: 10.1109/ICIET.2010.5625694.
6) Naz S, Hayat K, Razzak MI, Anwar MW, Madani SA, Khan SU. The optical character recognition of Urdu-like cursive scripts. *Pattern Recognition*. 2014;47(3):1229–1248. doi:10.1016/j.patcog.2013.09.037.
7) Choudhary P, Nain N, Ahmed M. A Structure for Annotation and Ground-truthing of Urdu Handwritten Text Image Corpus. *Procedia - Social and Behavioral Sciences*. 2015;198:84–88. doi:10.1016/j.sbspro.2015.07.422.

8) Mukhtar N, Khan MA, Chiragh N. Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. *Telematics and Informatics*. 2018;35(8):2173–2183. doi:10.1016/j.tele.2018.08.003.

9) Jabbar A, ul Islam S, Hussain S, Akhunzada A, Ilahi M. A comparative review of Urdu stemmers: Approaches and challenges. *Computer Science Review*. 2019;34. doi:10.1016/j.cosrev.2019.100195.

10) Mahmood Z, Safder I, Nawab RMA, Bukhari F, Nawaz R, Alfakeeh AS, et al. Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model. *Information Processing & Management*. 2020;57(4):102233. doi:10.1016/j.ipm.2020.102233.

11) Gorgevik D, Cakmakov D, Radevski V. Handwritten digit recognition by combining support vector machines using rule-based reasoning. In: and others, editor. In Proceedings of the 23rd International Conference on Information Technology Interfaces. IEEE. 2001;p. 139–144. Available from: doi:10.1109/ITI.2001.938010.

12) Kaur H, Mangat V. A survey of sentiment analysis techniques.IoT in Social, Mobile, Analytics and Cloud, (I-SMAC). *IEEE*. 2017;p. 921–925. doi:10.1109/I-SMAC.2017.8058315.

13) Chen MH, Chen WF, Ku LW. Application of Sentiment Analysis to Language Learning. *IEEE*. 2018;6:24433–24442. doi:10.1109/access.2018.2832137.

14) Shehab A, Elhenway I. A drugs classifier system based on machine learning algorithms. *Indian Journal of Science and Technology*. 2020;13(09):1046–1056.

15) Bilal M, Israr H, Shahid M, Khan A. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences*. 2016;28(3):330–344. doi:10.1016/j.jksuci.2015.11.003.

16) Rafeeq MJ, ur Rehman Z, Khan A, Khan IA, Jadoon W. Ligature categorization based Nastaliq Urdu recognition using deep neural networks. *Computational and Mathematical Organization Theory*. 2019;25:184–195. doi:10.1007/s10588-018-9271-y.

17) Ali I, Ali I, Subhash, Khan A, Raza SA, Hassan B, et al. Sindhi Handwritten-Digits Recognition Using Machine Learning Techniques". *International Journal of Computer Network and Information Security*. 2019;9(5):195–201. Available from: http://paper.ijcsns.org/07_book/201905/20190526.pdf.

18) Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5–32. doi:doi.org/10.1023/A.