iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55164

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Review on Isolated Urdu Character Recognition: Offline Handwritten Approach

Sayma Shafeeque A. W. Siddiqui[1], Rajashri G. Kanke[2], Ramnath M. Gaikwad[3], Manasi R. Baheti[4]
[1]*Research scholar, Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431005, Maharashtra, India*
[2, 3, 4]*Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431005, Maharashtra, India*

Abstract: ***This paper summarizes a system for recognizing isolated Urdu characters using advanced machine learning algorithms. The system analyzes visual features of Urdu characters, like strokes and curves, to train models such as CNN, SVM, ANN, and MLP. With a large dataset, the system can accurately predict unseen characters. It can be integrated into various applications for real-time character recognition tasks like OCR (Optical Character Recognition) and handwriting recognition. This literature survey explores research papers focused on character recognition in languages like Urdu, Arabic, Persian, and Sindhi, proposing various techniques like feature extraction, deep learning, and machine learning to enhance character recognition technology. The survey highlights specific studies with high accuracy and discusses recognition systems for Arabic characters as well.***

Keywords: Convolution Neural Network (CNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Multilayer Perception Model (MLP), Hidden Markov Model (HMM), Hidden Markov Model Toolkit (HTK)

I. INTRODUCTION

Urdu isolated characters recognition system is a technology designed to identify and recognize individual characters of the Urdu script. Urdu is a widely spoken language primarily used in Pakistan and India, and it is written in a modified form of the Arabic script. The isolated characters recognition system aims to automate the process of character recognition, which is important for various applications such as optical character recognition (OCR), handwriting recognition, and machine translation. By accurately identifying and classifying individual Urdu characters, the system enables efficient and reliable processing of Urdu text in digital form. The system typically uses advanced machine learning algorithms and techniques to analyze the visual features of isolated Urdu characters. These algorithms can extract various aspects of the characters, such as stroke patterns, curves, and angles. The extracted features are then used to train a machine learning model, such as a convolution Neural Network (CNN) or a Support Vector Machine (SVM), Artificial Neural Networks (ANN), Multilayer Perception Model (MLP) to recognize and classify different Urdu language characters. The training process involves feeding the model with a large dataset of labeled Urdu characters, where each character is associated with its corresponding label or class. The model learns to identify patterns and relationships among different characters, enabling it to make accurate predictions on unseen characters. Once trained, the Urdu isolated characters recognition system can be combined into various applications and systems to perform real-time character recognition tasks. It can process scanned documents, handwritten text, or even recognize characters from live video streams, providing valuable support for tasks like document digitization, automated translation, and text-to-speech conversion.

Overall, an Urdu isolated characters recognition system plays an important role in advancing the digitization and automation of Urdu language processing, making it easier to handle and analyze Urdu text in various areas.

II. PROCEDURES FOR PAPER SUBMISSION

This literature survey provides a brief overview of various research papers focused on the recognition of isolated handwritten characters in languages such as Urdu, Arabic, Persian, and Sindhi. Each paper proposes different techniques and approaches for character recognition, utilizing methods such as feature extraction, machine learning algorithms, and deep learning architectures. These papers looks into a different aspects of character recognition, including the use of invariant moments, B - Spline Curve approximation, hybrid features combining structural and statistical features, wavelet transform, and hidden Markov models. Overall, these studies contribute to the advancement of character recognition technology for various languages.

©IJRASET: All Rights are Reserved

In 2009 Nabeel Shahzad et al Offers a technique for identifying solitary, handwritten Urdu characters that have been drawn on a tablet computer. They analyzed the defining characteristics and introduced a set of features based on the primary and secondary strokes that make up each Urdu character. The accuracy of preliminary results was found to be 92.8% for native Urdu writers, while a non-native participant only achieved 73% accuracy. This paper present recognition of isolated printed Urdu characters only [1].

In 2012 IMRAN KHAN PATHAN et al provides a method based on Invariant Moments for the recognition of offline handwritten solitary Urdu characters. This work attempts to apply a feature extraction method based on moment invariants, followed by primary and secondary component separation. SVM is adopted for classification and position of secondary component (Above, Below and middle) is considered for recognition. Overall performance rate for all offline handwritten isolated Urdu characters is found to be 93.59%.It is possible to enhance the accuracy of system by combining more structural and statistical features [2].

In 2010 Junaid Tariq et al introduces a "soft-converter" that uses a database instead of a neural network to create OCR for isolated letters in the Urdu language (or right to left writing). This study indicates that NN is not necessary to implement OCR.. The accuracy rate of soft converter is 97.43%.But this paper present recognition of only isolated printed Urdu characters [3].

In 2017 Mohd Jameel and Sanjay Kumar present a detailed survey of Urdu HCR techniques with respect to feature extraction along with their efficiency and accuracy. This paper also presents a new proposed B-Spline Curve approximation approach for feature extraction of offline isolated Urdu handwritten characters. Various characteristics such as feature extraction, preprocessing, segmentation and recognition techniques have been used and reported different accuracy levels but the use of B-Spline Curve has not been found for handwritten Urdu script to form the feature vector inspire of their robustness. We have proposed a technique in this regard in order to enhance the accuracy and efficiency of Urdu HCR [4].

In 2019 Osamah Abdulrahman Almansari et al focuses on utilizing the Python programming language to create a model of a deep learning architecture employing multi layer perception (MLP) and convolution neural networks (CNN). They analyzes the performance of a public database which is Arabic Handwritten Characters Dataset (AHCD).However, training this database with CNN model has achieved a test accuracy of 95.27% while training it with MLP model achieved 72.08% [5].

In 2019 Mustafa Salam and Alia Abdul Hassan proposed a new architecture for Offline Isolated Arabic Handwriting Character Recognition System Based on SVM (OIAHCR). An Arabic handwriting dataset is also offered for training and testing the suggested system.Although half of the dataset used for training the Support Vector Machine (SVM) and the second half used for testing, the system achieved high performance with less training data. Besides, the system achieved best recognition accuracy 99.64% based on several feature extraction methods and SVM classifier. The results show that the linear kernel of SVM is convergent and more accurate for recognition than other SVM kernels [6].

In 2016 Shaina and Harpreet Kaur Bajaj proposes a method for Urdu language isolated text Character recognition is obtained by OCR. The efficacy of characters using SVM Classifier utilizing Hierarchical method is represented in this study. This work was completed using the Sindhi Character Set. This experiment achieves a recognition rate of 93.0481% [7].

In 2015 JAWAD H ALKHATEEB presents an effective system for off-line Arabic handwritten character recognition. They focuses on employing hybrid features, where the structure features are combined with the statistical features. This system includes four main phases. The first stage entails acquiring images and extracting characters. The second phase involves the pre-processing methods. The third phase entails feature extraction, in which two types of characteristics are applied and integrated. His discrete cosine transform (DCT) coefficients are extracted; hence the structure features such as the number of dots, the position of dots, and the number of holes. All the structure features are extracted and merged together with the DCT coefficients. Finally, the classification phase takes place using Neural Networks. This proposed system has been tested using different sets of characters and gives a very good accuracy [8].

Ahmed Subhi Abdalkafor suggested an Arabic off-line handwritten isolated recognition system in 2017 that was built on unique feature extraction techniques and a back propagation artificial neural network as the classification phase. This work is implemented and tested via the CENPARMI database. The recognition accuracy 96.14% has been achieved. But this recognition accuracy of this research is highly dependable on many factors related to the used database, namely, CENPARMI. These factors include the way in which the letters are written by the writers who were involved in the process of creating this database, such as if the letters either badly written or drawn in an uncommon way or employing a distinct writing style, such as the AL-Rukaa style, and because CENBARMI comprises the majority of those writing flaws, we predicted the low accuracy of overall performance [9].

Amir Mowlaei and Karim Faez proposed an isolated handwritten Persian/Arabic character and numeric recognition system in 2003. Wavelet transform has been used for feature extraction in this system. They use wavelet. The Support Vector Machine (SVM), which is a new learning machine with very good generalization ability used in pattern recognition and regression estimation, uses as classifier in this system. This technique was used to recognize isolated handwritten postal addresses that included city names and zip codes.. This system yields the recognition rate of 98.96% for these postal addresses [10].

In 2016 Ahmed Subhi Abdalkafo and Sadeq AlHamouz proposed Arabic off-line isolated character handwritten recognition system based on novel feature extraction techniques and MLP neural network as classification engine, this work is implemented and tested via CENPARMI database. Competitive recognition accuracy reached up to (94.75%) has been achieved [11].

In 2015 Jawad H Alkhateeb presents a recognition system for Arabic handwritten isolated characters. This recognition system is based on hidden Markov model (HMM). The entire system is capable of recognizing the Arabic handwritten characters to begin, the technique eliminates any diversity in the character pictures. Second, Features are extracted using the sliding window technique with HMM. Then, the HMM is used for recognition and classification process. The Hidden Markov Model Toolkit (HTK) is used for the classification process. HTK is a special toolkit used for the purpose of speech recognition. This proposed system is applied to a database for Arabic characters only [12].

III. ARTIFICIAL NEURAL NETWORKS

The capacity to learn complicated nonlinear input-output connections, apply sequential training techniques, and adapt to data are the primary properties of neural networks. The most commonly used family of neural networks for pattern classification tasks is the feed-forward network, which includes multilayer perception and Radial-Basis Function (RBF) networks. [13] Another popular network is the Self-Organizing Map (SOM), or Kohonen-Network [14], which is mainly used for data clustering and feature mapping. The learning process entails modifying network design and connection weights in order for a network to accomplish a certain classification/clustering job efficiently. The increasing popularity of neural network models to solve pattern recognition problems has been primarily due to their seemingly low dependence on domain-specific knowledge and due to the availability of efficient learning algorithms for practitioners to use. Artificial neural networks (ANNs) provide a new suite of nonlinear algorithms for feature extraction (using hidden layers) and classification (e.g., multilayer perceptions Furthermore, pre-existing feature extraction and classification techniques may be transferred onto neural network topologies for efficient (hardware) implementation. An artificial neural network (ANN) is a data processing paradigm inspired by the way organic nervous systems, such as the brain, process information. The unique structure of the information processing system is a crucial component of this paradigm. It is made up of a huge number of highly linked processing components (neurons) that work together to solve issues. Through a learning process, an ANN is trained for a specific application, such as pattern recognition or data categorization. In biological systems, learning entails changes to the synaptic connections that occur between neurons. [15]

A. Support Vector Machine

The Support Vector Machine (SVM) is a state-of-the-art classification technique introduced in 1992 by Boser, Guyon and Vapnik. [16] The SVM classifier, which based on the statistical theory, is widely used in multiple disciplines due to its high accuracy and ability to deal with high-dimensional data [17]. SVM can be described as a set of related supervised learning methods used for classification and regression [18] SVM is a type of generalized linear classifier. SVM is a very useful tool that can be used to solve problems of pattern recognition and function estimation such as in handwritten text recognition. SVM has been applied to a wide range of pattern recognition applications such as face detection in image, verification and recognition, object detection and recognition, speaker identification and isolated handwritten digit recognition. [19][22] Support Vector Machine is a classification and regression prediction technique that uses the theory of machine learning to maximize predictive accuracy. The applications of SVM to pattern recognition are state-of-the-art performance Support Vector Machines, on the other hand, are systems that employ the hypothesis space of a linear function in a high dimensional feature space and are trained with an optimization theory learning algorithm that applies a statistical learning theory learning bias. Solving a restricted quadratic optimization issue is used to train SVM. SVM uses a collection of nonlinear basis functions to map inputs into a high-dimensional space. SVM may be used to train a number of representations, such as neural networks, splines, polynomial estimators, and so on, but each choice of representation has a unique optimum solution. The SVM parameters the mathematical formulas of SVM are available in [19] [23].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical model used to analyze sequential data with hidden states. It assumes that observed data is generated by an underlying process with unobservable states. HMMs consist of hidden states forming a Markov chain and observed data generated based on the current hidden state. The objective is to infer the hidden state sequence or estimate model parameters. HMMs find applications in speech recognition, natural language processing, bioinformatics, and other fields where hidden states can be inferred from observed data.

C. Multilayer Preceptor

A multi layer Preceptor (MLP) is a form of artificial neural network made up of many layers of linked nodes, or neurons.. It is one of the fundamental architectures used in deep learning. MLPs are feed forward neural networks, meaning that information flows in one direction, from the input layer through the hidden layers to the output layer, without any loops or feedback connections. Each neuron in an MLP performs a weighted sum of its inputs, applies an activation function to the result, and produces an output. The output of one layer serves as the input to the next layer, forming a hierarchical representation of the input data. The hidden layers between the input and Output layers allow the MLP to learn complex, nonlinear relationships in the data. MLPs are trained using a method called back propagation, which involves iteratively adjusting the weights of the connections between neurons to minimize the difference between the network's predicted output and the desired output. This training process involves computing gradients and updating the weights using optimization algorithms like gradient descent. MLPs have been successfully applied to various tasks, such as pattern recognition, classification, regression, and function approximation. They have shown remarkable capability in learning from large amounts of data and can be used in a wide range of applications, including image and speech recognition, natural language processing, and recommendation systems.

D. Convolution Neural Networks

Convolution Neural Networks (CNNs) are specialized artificial neural networks designed for analyzing visual data. They are inspired by the structure of the human visual cortex and are widely used in computer vision tasks such as image classification, object detection, and image segmentation. CNNs consist of interconnected layers of neurons organized hierarchically to extract meaningful features from input images.

The key component is the convolution layer, which applies filters to the input image, enabling the network to learn and detect low-level visual patterns. Pooling layers in CNNs down sample feature maps to reduce spatial dimensions, while fully connected layers perform classification or regression tasks based on extracted features. Activation functions like ReLU introduce non-linearity to the model. CNNs excel at capturing spatial hierarchies of features. With multiple layers and increasing receptive fields, they learn complex and abstract features at higher levels, making them effective for image recognition. CNNs are trained through back propagation, adjusting internal parameters to minimize the difference between predicted and actual outputs. Training often requires a large labeled dataset for optimal performance.

Overall, CNNs have transformed computer vision, enabling advancements in autonomous vehicles, medical imaging, facial recognition, and other applications.

IV. CONCLUSION

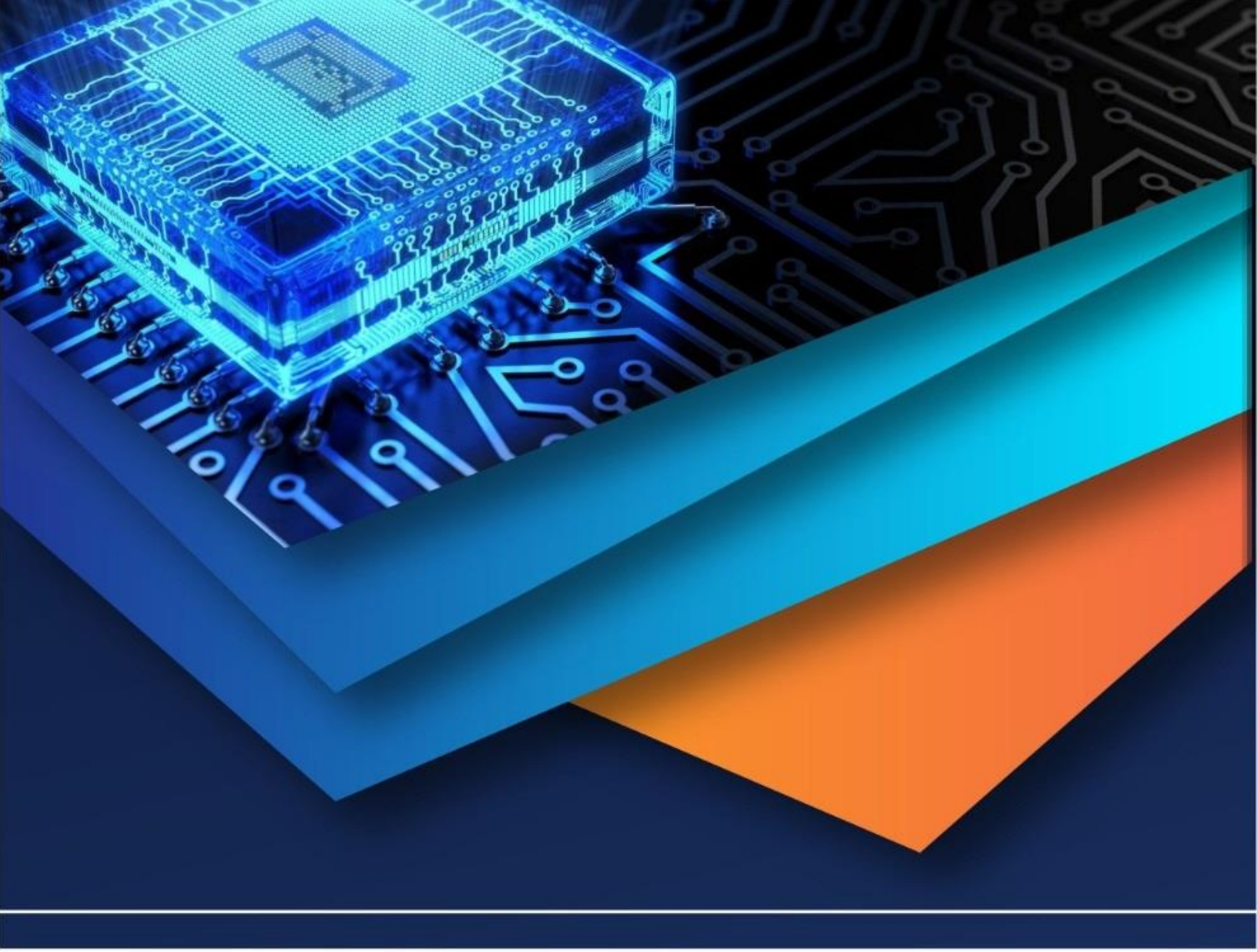
This paper highlights various techniques and approaches for recognizing isolated handwritten characters in Urdu and Arabic languages. These include feature extraction techniques, machine learning algorithms (such as SVM and MLP), deep learning architectures, and the use of hybrid features. The reported recognition accuracies range from 72.08% to 99.64%, depending on the specific approach and dataset used. Further research can explore combining different techniques and features to improve accuracy and efficiency in isolated character recognition tasks.

V. ACKNOWLEDGMENT

I would like to express my deepest gratitude to my guide Dr. Manasi R. Baheti for their invaluable guidance and support throughout this research. Their expertise and encouragement were instrumental in its successful completion. Additionally, I extend my appreciation to my family and friends for their unwavering support during this endeavor.

REFERENCES

[1] M Shahzad, N.; Paulson, B.; Hammond, T. 'Urdu Qaeda: Recognition system for isolated Urdu characters', In Proceedings of the IUI Workshop on Sketch Recognition, Sanibel Island, FL, USA, 8–11 February 2009.
[2] I, K. Pathan, A.A. Ali and R.R. J, 'Recognition of offline handwritten isolated Urdu characters', Advances in Computational Research, 2012.
[3] Tariq, J., Nauman, U. and Naru, M.U., 2010, April. 'Soft converter: A novel approach to construct OCR for printed Urdu isolated characters'. 2nd International Conference on Computer Engineering and Technology (Vol. 3, pp. V3-495). IEEE, 2010.
[4] Jameel, M. and Kumar, S., 2017. 'Offline recognition of handwritten Urdu characters using b spline curves: A survey'. International Journal of Computer Applications, 157(1), pp.28-34. 2017.
[5] Almansari, O.A. and Hashim, N.N.W.N., 2019, October. 'Recognition of isolated handwritten Arabic characters'. 7th International Conference on Mechatronics Engineering (ICOM) (pp. 1-5). IEEE, 2019.
[6] Salam, M. and Hassan, A.A., 2019. 'Offline isolated Arabic handwriting character recognition system based on SVM'. Int. Arab J. Inf. Technol., 16(3), pp.467-472, 2019.
[7] Shaina, Harpreet Kaur Bajaj. (2016). 'ISOLATED CHARACTER RECOGNITION USING HIERARCHICAL APPROACH WITH SVM CLASSIFIER'. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 5(9), 570–575, 2016.
[8] Alkhateeb, J.H., 2015. 'Off-Line Arabic Handwritten Isolated Character Recognition'. Int. J. Eng. Sci. Technol, 7, pp.251-257, 2015.
[9] Abdalkafor, A.S., 2017. 'Designing offline Arabic handwritten isolated character recognition system using artificial neura lnetwork approach'. International Journal of Technology, 8(3), pp.528-538, 2017.
[10] Mowlaei, A. and Faez, K., 2003, September. 'Recognition of isolated handwritten Persian/Arabic characters and numerals using support vector machines'. IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No. 03TH8718) (pp. 547-554). IEEE, 2003.
[11] Abdalkafor, A.S. and Sadeq, A., 2016. 'Arabic offline handwritten isolated character recognition system using neural network'. International Journal of Business and ICT, 2(3), pp.41-50, 2016.
[12] Alkhateeb, J.H., 2015. Off-line 'Arabic Handwritten Isolated Character Recognition using Hidden Markov Models'. Journal of Emerging Trends in Computing and Information Sciences, 6(8), pp.415-420,2015.
[13] A.K. Jain, J. Mao, and K.M. Mohiuddin, "Artificial Neural Networks: A Tutorial", Computer, pp. 31-44, Mar, 1996.
[14] T.Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, Berlin, vol.30,1995.
[15] Basu, J.K., Bhattacharyya, D. and Kim, T.H., 2010. 'Use of artificial neural network in pattern recognition'. International journal of software engineering and its applications, 4(2), 2010.
[16] B. E. Boser, I. M. Guyon and V. N. Vapnik, 'A Training Algorithm for Optimal Margin Classifiers', 5th Annual ACM Workshop on COLT, pp 144-152, ACM Press, 1992.
[17] B. Scholkopf, K. Tsuda, and J. P. Vert, editors, 'Kernel Methods in Computational Biology', MIT Press series on Computational Molecular Biology, MIT Press, 2004.
[18] Vikramaditya Jakkula, 'Tutorial on SVM', School of EECS, Washington State University, Pullman. C. J. C. BURGES, "A Tutorial on Support Vector Machin



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)