Text Detection and classification in Camera-Captured Documents using CRNN + CRF Deep Learning Techniques

[1] Athapuram Divya, Assistant Professor, Dept of IT, Malla Reddy Engineering College and Management Sciences, Hyderabad
[2] J.Sri Latha, Assistant Professor, Dept of CSE, G. Narayanamma Institute of Technology & Science, Hyderabad
[3]B.Mamatha, Asst.Prof , Dept of Computer Science and engineering, CMR Engineering College, Hyderabad
[4] E Raju, Assistant Professor, Dept of CSE, Guru Nanak institute of technology, Hyderabad

Abstract –A major goal in the field of document image analysis (DIA) is to convert image data into a format that can be easily interpreted by machines. Within DIA-based systems, layout analysis plays a key role in preprocessing to identify and extract accurate and error-free text segments. However, as far as Pashto is concerned, document images are yet to be explored. Recognizing Pashto text in documents captured with a camera can be difficult due to variations in image quality, lighting conditions, complex backgrounds, unavailability of labeled documents, cursive writing, form and context dependence, and multiple scripts per image. , and language-specific layouts make it a difficult task. The proposed OCR system is developed using a CRNN+CRF-based deep learning model. The applicability of the proposed model is verified using a decision tree (DT) classification tool based on zonal feature extraction techniques and invariant moments approach. For his OCR system based on CRNN + CRF, the overall accuracy rate is calculated as 97%.

Keywords: DT, CRNN, CRF, OCR and DIA

I. 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.

After studying and analyzing 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 neighbors, 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.

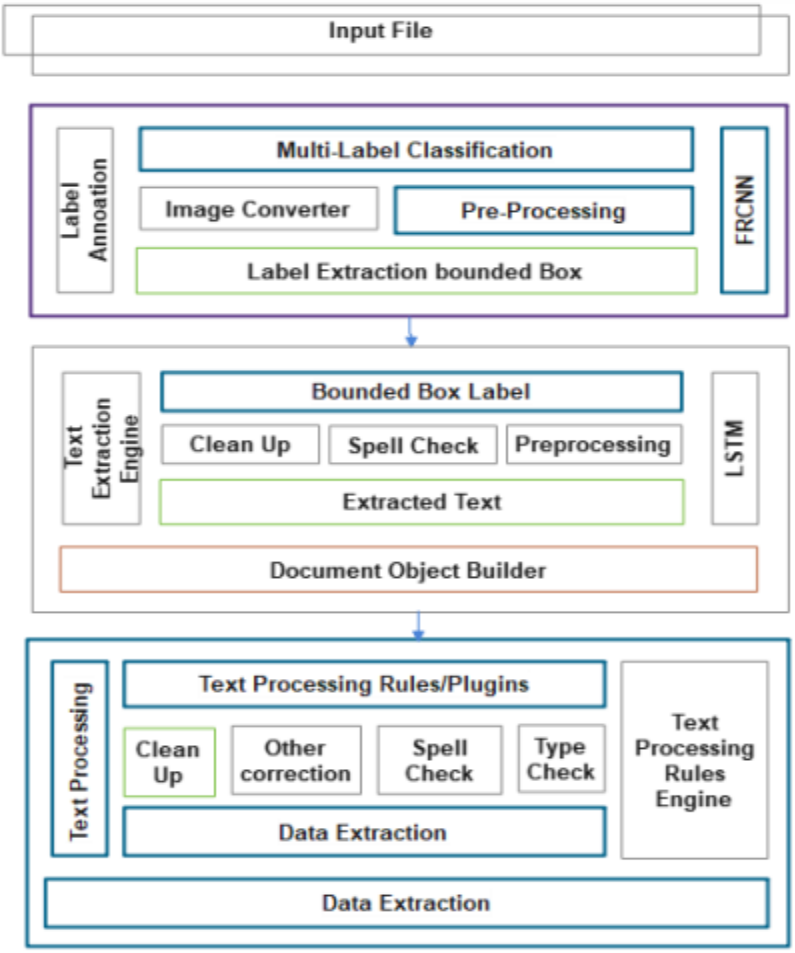

Fig 1: a generic text data identification process

Bukhari et al. [14] introduced a system for the layout analysis achieving a well organized and robust text and nontext segmentation, text-line extraction, and reading order determination methods for Urdu and Arabic document images. Tuan-Anh Tran et al. [15] proposed a system for the analysis of the textual and non-textual elements in document images. Their method is a mixture of white-space analysis method with multi-layer uniform areas. The system was

validated on page segmentation competition held by ICDAR2009 [16]. They achieved above 90% accuracy for text detection, non-text detection, text region detection, and non-text region detection. Ahmad et al. [17] introduced a text-line extraction method for the extraction of titles and large headings in cursive script i.e. Pashto [18], [19]. Their method is based on Horizontal Projection Pro-file (HPP) [20] and Hanning window smoothing technique [21]. They obtained an accuracy of 99.30%. However, the system needs de-skewed images for a better performance.

This research highlights the problem of layout analysis and classification considering the Pashto language, further it analyses the very basic layouts regarding the DIA system. However, regarding the Pashto language, there is a little work in the area of DIA system. The reasons are language-specific complexities, which include; writing direction, availability of different languages per document and language-specific layouts etc. For example, several Pashto documents contain Arabic as well as Pashto text in a single document (as shown in Figure 2). Ignoring this specific pattern will lead to the extraction of textual blocks that contain either Arabic text or Pashto text or mix of both. As a result, it becomes a multilanguage case for OCR, and could not be handled easily on a single OCR system. The major contributions of this research are given below. 1) Creation of a new dataset based on camera captured images of Pashto documents. 2) Fine tuning of CRNN + CRF Based deep learning models including SSD, Yolov5 and Yolov7 for baseline evaluation.

II. Literature Survey

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.

Simon et al. [5] introduced a bottom-up approach for the layout analysis of document using Kruskal's algorithm [6] and to build the structure of the real page through the utilization of a particular distance metric between the segments. Their algorithm is more limited according to the computational complexity, because of its linear structure regarding the amount of the relevant elements [4]. Thomas M. Breuel [7] introduced many novel algorithms and statistical methods for layout analysis. The methods consist of (1) to find rectangles of tall white spaces and assess

them as candidates for the channels, separators of column, etc (2) to find the text-lines concerning to the column-structure of the document, (3) to recognize paragraphs, headings and titles based on spacing, size and indentation, etc. and (4) to determine the reading order by using geometric and linguistic information. These algorithms are also applicable to Cursive Script. Kevin Laven et al. [8] presented an algorithm using statistical patterns like grammar-based and rule-based techniques. They first introduced a unique software for the manual segmentation and labeling of the page.

Their dataset contains a 932 pages as images from academic journals1 . Shafait et al. [9] introduced a system for the layout analysis of the cursive script. Their specified scheme experimented on 25 scanned images taken from various sources like magazines, newspapers, and books. Their algorithm obtained 90% precision in line detection, while in case of newspaper images 72% precision achieved.

Shafait et al. [10] also proposed a system for the classification and layout analysis of Breuel(Roman script text-line model) [11] to Nastaliq script and introduced a text-line extraction modeled for the Urdu text line. M. Sezer Erkilinc et al. [12] introduced an algorithm for document classification and page layout analysis. They tested a module for text detection which is based on wavelet analysis and Run Length Encoding (RLE) [13] method, and a second module to detect the image and graphic sections in the input document.

III. Text Detection Techniques

A) Tesseract OCR

Tesseract OCR combines character recognition, pattern matching, and contextual analysis to identify and interpret characters within images. It segments the image into individual characters or words, analyzes the shapes and features of these characters, and then matches them against its trained character set to produce machine-readable text. Tesseract's open-source nature, broad language support, and accuracy make it a reliable choice for various text recognition tasks. It can handle diverse fonts, sizes, and styles, making it suitable for digitizing printed materials like books, documents, and signage.

B) CRNN: Merging Convolution and Recurrent Networks for Text Recognition

CRNN combines convolutional layers for feature extraction from local regions of an image and recurrent layers to capture sequential dependencies among characters. This dual architecture enables it to recognize text within images with varying layouts and sequences. CRNN's ability to handle sequences of varying lengths and contextual understanding make it valuable for recognizing paragraphs, sentences, and single characters. It excels in scenarios involving handwritten notes, documents with variable text sizes, and text extraction from images with complex backgrounds.

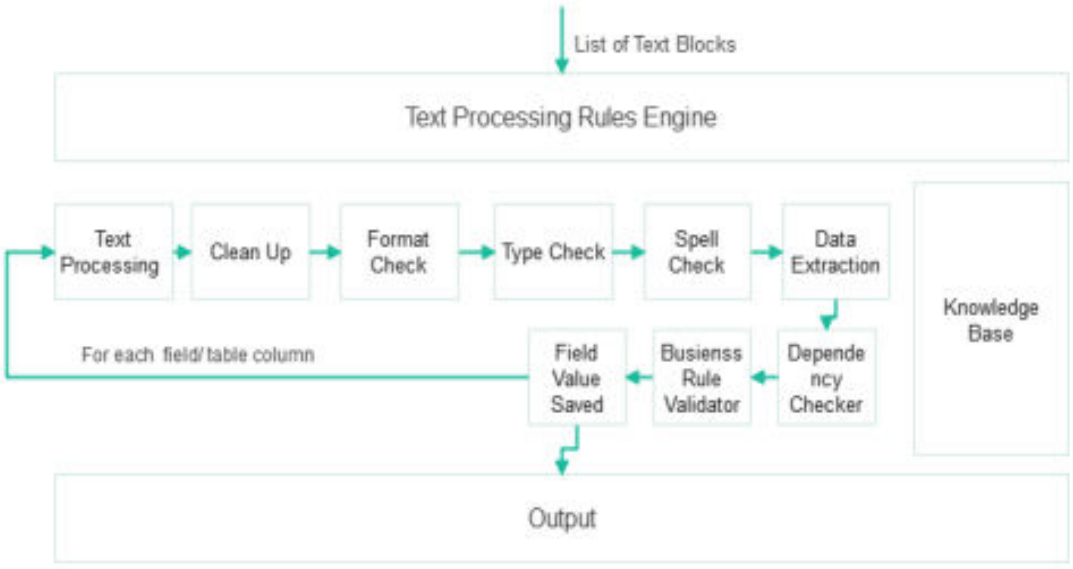

Fig 2: Text Extraction Flow

C) EAST: Efficient Text Detection in Natural Scenes

EAST employs a two-stage architecture for text detection. The first stage generates coarse text region proposals, and the second stage refines these proposals for accurate localization. It analyzes the image using convolutional layers to detect text regions based on their distinctive characteristics. EAST is known for its speed and efficiency, making it well-suited for real-time applications. Its ability to detect text in images with varying orientations, fonts, and backgrounds sets it apart from other algorithms.

D) CTPN: Accurate Text Proposal Generation

CTPN utilizes a neural network to generate text proposals and identify regions in the image that likely contain text. Subsequently, it refines these proposals to detect and segment text regions accurately. CTPN's neural network analyzes the spatial characteristics of the image to identify potential text regions. CTPN's accuracy in detecting text regions, including curved and rotated text, proves invaluable for scenarios that demand precise text localization. It commonly finds use in tasks involving document layout analysis and content extraction.

E) CRF: Contextual Understanding for Improved Recognition

Conditional Random Fields (CRF) refines text recognition results by considering the context of neighboring characters or words. It factors in contextual information to correct recognition errors caused by ambiguous characters or variations in text appearance. CRF enhances accuracy by improving contextual understanding of the text recognition process. It's precious in scenarios with noisy or degraded input images.

F) LSTM: Long Short-Term Memory for Sequence Recognition

Long Short-Term Memory (LSTM) is a recurrent neural network that captures long-range dependencies within sequences. It processes input sequences step by step, retaining information relevant for recognizing entire lines of text, sentences, or series of characters. LSTM's ability to handle sequences of varying lengths makes it helpful in recognizing text in structured documents, extracting data from invoices and forms, and converting handwritten notes into digital text.

IV. Proposed CRNN + CRF: Seamless Scene Text Recognition

The images are acquired via handheld camera. We choose two books (Tafseer-ul-Quran and Meshkat-sharif) that contain plenty of pages. However, we have captured only those pages where Pashto and Arabic text per page were notable. Additionally, while capturing the images, it was insured to avoid the skew and perspective distortion. However, blurriness and shadow effects are present due to various lighting condition in the acquired images.

In supervised learning, data must be transcribed or annotated with suitable labels. In our case, it is important to label/ transcribe the Arabic and Pashto text blocks separately. We use a tool created by MIT named LabelMe2 for the annotation. The annotation of each textual block is done by considering its contour/edges by taking polygons. The relevant annotation for each image is stored in a separate .json file. The prefix of the image filename is same as annotation or json file.

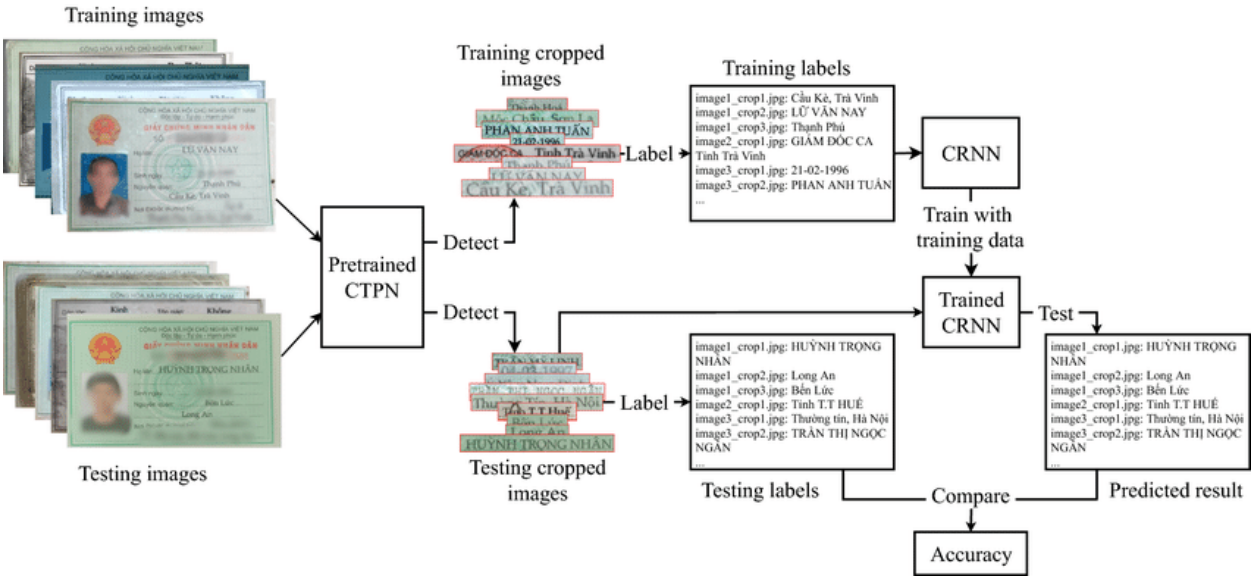

Fig 3: CRNN text extraction

Combining Convolutional Recurrent Neural Networks (CRNN) with Connectionist Temporal Classification (CTC) allows CRNN to handle variable-length sequences without requiring an explicit alignment between input images and text outputs. CTC assists in decoding the variable-length results. CRNN + CTC is decisive in recognizing scene text, including text with varying layouts and lengths. It's valuable for street sign reading, scene text translation, and text-based geolocation tasks.

Each text recognition algorithm discussed brings unique strengths, catering to specific use cases and requirements. By understanding these algorithms' working principles and advantages, businesses and researchers can choose the most suitable solution to enhance OCR and IDP capabilities across various domains.

V. Results Analysis

The Proposed Algorithm Results Were Compared With Existing Algorithms And The Results Tabulated Here.

Table 1: Performance reports of various algorithms

	Accuracy	Recal	Precision	F1 score	Time Taken to Extract	Overall efficiency %
SVM	86.35	78.65	78.65	86.32	60 sec	93
CRNN	87.63	74.24	81.55	85.97	59 sec	92
CRF	87.54	76.98	84.52	88.78	62 sec	92
LSTM	92.54	77.85	83.41	92.36	58 sec	94
CRNN + CRF	93.54	79.25	88.25	94.11	56 sec	97

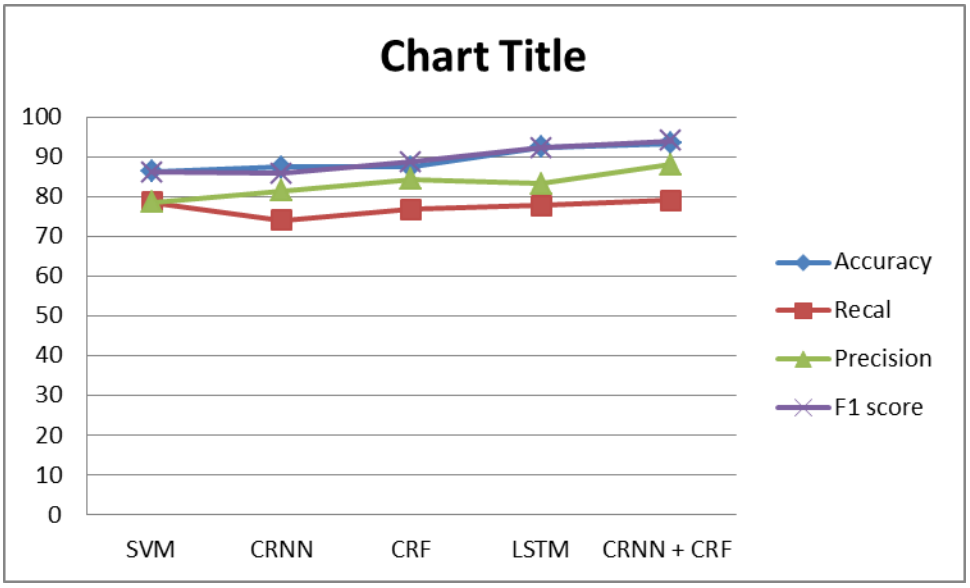

Fig 4:Comparison of Accuracy,Recal,Precision,F1 score of all algorithms

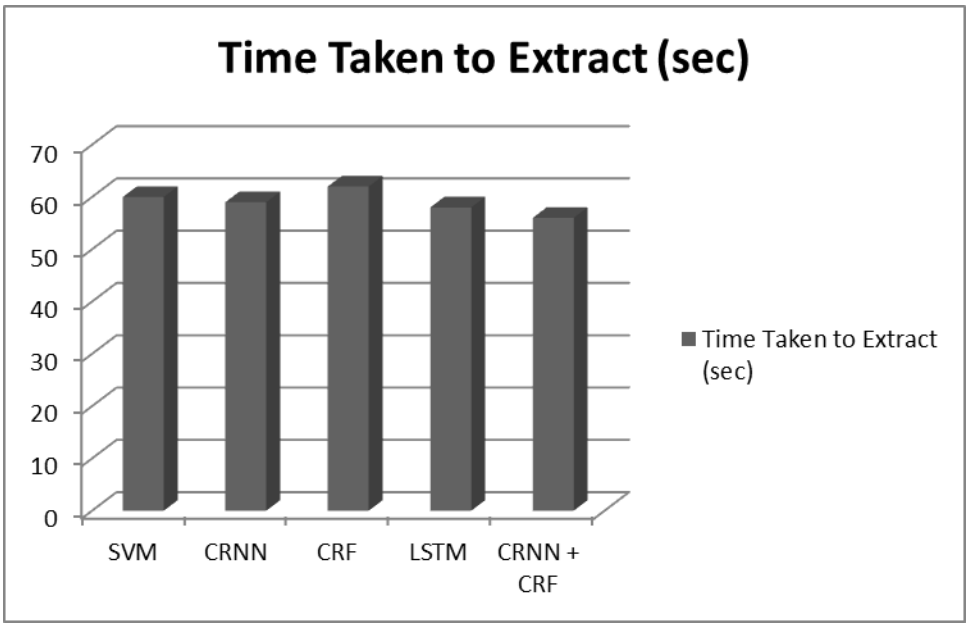

Fig 5:Time taken by all algorithms

VI. Conclusion

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 97% is calculated for the CRNN + CRF based OCR system, while SVM-based recognition rates of 86.35% is achieved for invariant moments-based feature map. Developed the model for Telugu text extraction and recognition using convolutional and recurrent neural networks .The model developed is confined to developed dataset. By using additional CRNN + CRF networks this model can be extended to predict at sentence level. It can also be extended to other languages by training on their datasets.

References

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[6] 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).

[7] 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).

[8] M.-K. Hu. Visual pattern recognition by moment invariants. IRE transactions on information theory, 8: 179-187 (1962).

[9] 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).

[10] 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).

[11] 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).

[12] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206.

[13] Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[14] P. Naresh, S. V. N. Pavan, A. R. Mohammed, N. Chanti and M. Tharun, "Comparative Study of Machine Learning Algorithms for Fake Review Detection with Emphasis on SVM," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 170-176, doi: 10.1109/ICSCSS57650.2023.10169190.

[15] Hussan, M.I. & Reddy, G. & Anitha, P. & Kanagaraj, A. & Pannangi, Naresh. (2023). DDoS attack detection in IoT environment using optimized Elman recurrent neural networks based on chaotic bacterial colony optimization. Cluster Computing. 1-22. 10.1007/s10586-023-04187-4.

[16] P. Naresh, P. Srinath, K. Akshit, M. S. S. Raju and P. VenkataTeja, "Decoding Network Anomalies using Supervised Machine Learning and Deep Learning Approaches," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 1598-1603, doi: 10.1109/ICACRS58579.2023.10404866.

[17] 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).

[18] 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).

[19] 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)..

[20] Nagesh, C., Chaganti, K.R. , Chaganti, S. , Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. *International Journal on Recent and Innovation Trends in Computing and Communication*. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010.

[21] S. Khaleelullah, P. Marry, P. Naresh, P. Srilatha, G. Sirisha and C. Nagesh, "A Framework for Design and Development of Message sharing using Open-Source Software," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 639-646, doi: 10.1109/ICSCDS56580.2023.1010467

[22] M. I. Thariq Hussan, D. Saidulu, P. T. Anitha, A. Manikandan and P. Naresh (2022), Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. IJEER 10(2), 80-86. DOI: 10.37391/IJEER.100205.

[23] 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).