Enhancing Text Extraction: The Evolution and Future Potential of EasyOCR

Mr. ch. Ravi Kishore Reddy
Assistant professor,
Dept. of CSE
Vignan's Foundation for Science, Technology & Research Vadlamudi, AndhraPradesh-522213,India

Gullipalli Mohan
Dept of CSE
Vignan's Foundation for Science, Technology & Research Vadlamudi, AndhraPradesh-522213,India

Akula Yeshwanth
Dept of CSE
Vignan's Foundation for Science, Technology & Research Vadlamudi, AndhraPradesh-522213,India

Rama Krishna Chintapalli
Dept of CSE
Vignan's Foundation for Science, Technology & Research Vadlamudi, AndhraPradesh-522213,India

Abstract-This paper dive into the applications of Machine learning and Deeplearning by solving the challenges of Text detection and extraction.OCR encompasses new methodologies and also ensures the accuracy of the models on the recognition. Retrieving unaltered text from the sources plays a vital role in the system. These models ensures the quality of retrieval of data from the text irrespective of font, language and pattern. Hence this technology is a integration AI, ML and DL It not only focuses on just the recognition but also augments the quality of recognition, supports for multiple languages, diverse fonts and sources. Concentrating on all these pivotal tasks the proposed system enhances and make advancement on overall performance of the model.

Index Terms: Machine Learning, Deep Learning, Accuracy, Text recognition.

I. INTRODUCTION

In today's digital era, the sheer volume of information generated in various forms presents in printed documents, handwritten notes, and other physical media. We cannotuse the text directly which are present in other formats like above.The only way to getthe text present in these non-editable sources is OCR. In the previous models the observed performance is low compared to our OCR project. We've looked over various projects which presented before but there's a inadequacy in the models. Handwritten_OCR, Scene Text Telescope, MaskOCR, From Text Image, are some ofthe projects which have done before for the recognition and having unique functioning in performance of model like retrieving text from images, blurred scene, retrieving through decoding, Multi-language detection are the key features of these projects. Wehave all of these in various segments but having all them in a single project can be beimmensely beneficial. This trail also had done before but it did not reached expectations of performance. For an instance if we consider extracting text from a image the previous model of OCR didn't performed well in retrieving the data from the source. So we came up with a advancement which overcomes the past issues and fulfills the holes of inadequacy.

Our OCR bought up a unique pattern by incorporating different technologies like deep learning, machine learning and also libraries of Machine learning and tools of computer vision. A deep diving into the neural networks caused a major impact in the functioning of the model. We a proven significance that Ml and Dl algorithms like LST, RNN, CNN, GRU works on different tasks collectively to produce effective results in terms of Text retrieval. OCR encompasses a range of sophisticated algorithms and techniques that analyze the patterns and shapes of characters within an image, effectively 'reading' and interpreting the text therein. By harnessing the power of Libraries and frameworks of image processing and machine learning live OpenCV and CNN, OCR systems have become increasingly adept at accurately recognizing textin various languages, fonts, and writing styles.It's Significance extends across diverse domains and industries, offering myriad benefits and applications. This sets the stage for exploring the intricacies of OCR technology, delving into its underlying principles, functionalities, and real-world applications. By understanding the capabilities and potential, we can unlock new possibilities for data-driven innovation, productivity gains, and improved accessibility in an increasingly digitized world.

Extracting text from images in Optical Character Recognition involves different types to accurately capture, analyze, and convert text data. This technology evolved it's potential for improved detection and retrival such that it can ensure all the contents will be retrivedfrom the sources.By incorporating all the features this OCR provides flexibility and can be helpful in extraction from Text, Handwriting, Table, Barcode, Metadata, Structured data, Key-value pair, Entity Extraction. Each type of these plays a crucial role in enhancing the accuracy and efficiency of OCR systems, catering to a wide range of document types and formats for diverseapplications.

I. LITERATURE REVIEW

"Historical Review of OCR Research and Developmental" by Shunji Mori, Ching Y. Suen, and Kazuhiko Yamamoto (1992): This paper provides a historical review of research, outlining the evolution and its milestones, and advancements up to the early 1990s.[1]

"Scene Text Telescope: Text-Focused Scene Image Super-Resolution" by Jingye Chen, Bin Li, Xiangyang Xue (Year not specified): This paper focuses on enhancing the resolution of scene text images to improve accuracy, particularly in challenging real-world scenarios.[2]

"Datasets in OCR for Handwriting Recognition: A Survey" by Prarthana Dutta and Naresh Babu Muppalaneni (2023): This survey paper explores datasets used in research for handwriting recognition, providing insights into their characteristics, availability, and relevance for training and evaluating systems.[3]

"Survey of Post-OCR Processing Approaches" by Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, Antoine Doucet (2021): This survey covers various post-processing methods applied to output, such as error correction and data normalization, to enhance performance and usability.[4]

"Mask OCR: Text Recognition with Masked Encoder-Decoder Pretraining" by Pengyuan Lyu, Chengquan Zhang, ShanShan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, Jingdong Wang (2023): This paper introduces a novel approach to text recognition using masked encoder-decoder pretraining, showcasing advancements in deep learning techniques.[5]
"Handwritten text recognition: A review" by K. Simistira, N. Tsipas, and G. Economou (2016): This review paper provides an overview of handwritten text recognition techniques, discussing various approaches and their application.[6]

"Gradient-based learning applied to document recognition" by Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner (1998): [7]This paper discusses the application of gradient-based learning techniques, particularly in neural networks, for document recognition tasks, including Optical Character Recognition.

"Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks" by A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber (2006): This paper introduces the Connectionist Temporal Classification (CTC) algorithm, which is widely used for labeling unsegmented sequence data.[8]

"An experimental study of convolutional neural networks for handwritten digit recognition" by J. L. G. Pires and D. M. de Matos (2019): This paper presents an experimental study of convolutional neural networks (CNNs) for handwritten digit recognition, showcasing the effectiveness of CNNs in tasks.[9]

"ICDAR 2013 robust reading competition" by D. Karatzas et al. (2013): This paper discusses the results and findings of the ICDAR 2013 robust reading competition, which focuses on evaluating OCR systems' robustness to various challenges in real-world scenarios.[10]

Processing. Their study advances the field of image processing by offering a robust method for analyzing and classifying textures in images, a technique that holds potential for enhancing the detection of text from multiple sources.

"An analysis of single-layer networks in unsupervised feature learning" by A. Coates, A. Ng, and H. Lee (2011): This paper investigates the effectiveness of single-layer networks in unsupervised feature learning, which can be beneficial for feature extraction.[11]

"Structural pattern recognition" by T. Pavlidis (1977): This book covers the principles and techniques of structural pattern recognition, providing foundational knowledge applicable to Optical Character Recognition.[13]

"A top-down approach to document recognition" by R. K. Srihari (1988): This paper presents a top-down approach to document recognition, discussing methods for segmenting and recognizing text within documents.[14]

"A mathematical approach to character recognition" by R. D. Kent (1968): This paper presents a mathematical approach to character recognition, discussing methods for analyzing and recognizing patterns in characters.[15]

"Error-correcting output codes: A general method for improving multiclass inductive learning programs" by H. Bunke (1998): This paper introduces error-correcting output codes (ECOC), which is a method for improving the performance of multiclass classification tasks.[16]

"Offline handwritten character recognition using SVM and HMM" by K. V. Prathyusha and K. Latha (2018) [17]: This paper discusses the application of Support Vector Machines(SVM) and Hidden Markov Models (HMM) for offline handwritten character recognition.

"A review of neural network applications in handwritten Hindi character recognition" by S. Shah, M. S. Anand, and N. N. Patil (2020): This review paper discusses the application of neural networks in handwritten Hindi character recognition, highlighting recent advancements and challenges.[18]

"Handwritten Devanagari character recognition using deep learning techniques: A review" by S. Ojha and K. K. Tyagi (2021)[19]: This review paper focuses on the application of deep learning techniques for handwritten Devanagari character recognition, providing insights into recent developments in the field.

"A study of Bangla handwritten character recognition system using deep neural network" by P. K. Bharati and S. K. Jena (2020): This paper presents a study on Bangla handwritten character recognition using deep neural networks, showcasing the effectiveness of deep learning techniques for specific languages.[20]

II. ABOUT DATASET

The dataset which pursued for this project are multiple because here we not only extracting the text from the pdf or some external file but also from the images, multiple language text extraction, from various websites etc. so we need to train the model with several datasets so that it can work properly. The datasets we considered for the training are “Text Extraction Of OCR”, “ Standard OCR dataset”, “Water Meters Dataset”, “ Captcha Data”, “ Receipts Text-Detection”.Text Extraction consists of XML files and images.The XML files contain the extracted data from the image of the invoices, name of text and file is kept the same for clarity. Users of the dataset should extract entities like invoice no, invoice data, company name (invoice from company1 to company2/person), telephone number of the company, address e.t.c. IT consists a total of 1560 text files which are used for training and testing of model.Standard dataset contains Various Fonts and Style. Font and style differentiates the character. So we have to train by using the font and style so that it do not mislead into recognition. It also comprises of 2 folders. One for training another for testing. Each folder consists of 36 other files.

Water Meter Dataset comprises of 3 folders : collage, images, masks. A diverse collection of water meter images along with corresponding segmentation masks and labels for the meter readings.

collage - images of water meters with bounding boxes.
images - original images of water meters
masks - segmentation masks for images

Captcha dataset is used to make a custom OCR. This consists of 3 folders sample, train , val.

Receipts Text-Detection The Grocery Store Receipts Dataset is a collection of photos captured from various grocery store receipts.This dataset is specifically designed for tasks related to Optical Character Recognition and is useful for retail. Each image in the dataset is accompanied by bounding box annotations, indicating the precise locations of specific text segments on the receipts. The text segments are categorized into four classes: item, store, date, time and total.

III. PROPOSED METHODOLOGY

The proposed methodology of the Optical Character Recognition is shown briefly in the data flow diagram. A sample set of documents or images will be given to the pre-processor and upon it the feature and segmentation extraction will be done through CNN and RNN and the retrived result will be given to train the model.By training the model with all the sources and storing it all in the model file will be useful for the recognition. So for the end of the model training model file will be familiar with all the fonts, shapes, patterns of the text in multiple languages. By using CNN and RNN OCR model created a tremendous impact in the text recognition. Irrespective of the language, font, size, pattern

Detection will be at ease.Libraries and framewroks used to to build the model are OpenCV or PIL using TensorFlow, PyTorch, or Keras Tesseract OCR, OCRopus, or EasyOCR PyTesseract Google Cloud Vision API AWS Textract Microsoft Azure Computer Vision API.

A. Data Acquisition and Preprocessing

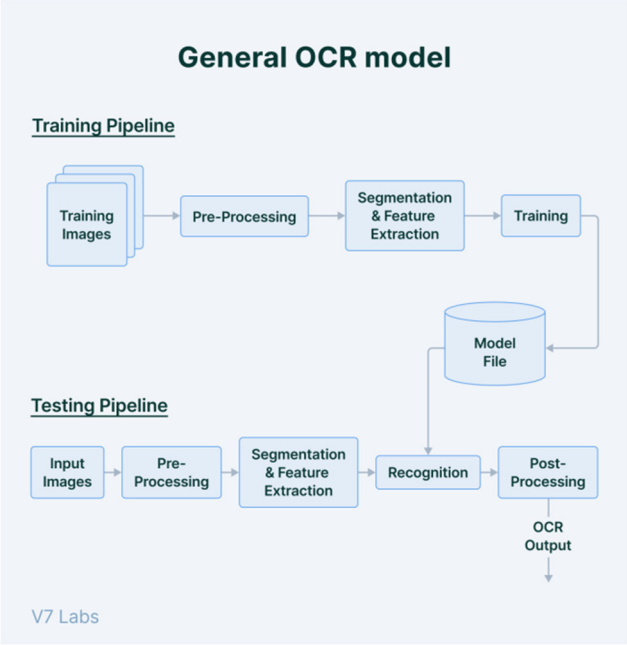

Fig 1 : Block Diagram of Proposed Model

Deep learning algorithms like CNN,RNN created a great impact in the character recognition. As we know that both of the above mentioned algorithms have a vital role in image detection, sequence detections in usage of DL.

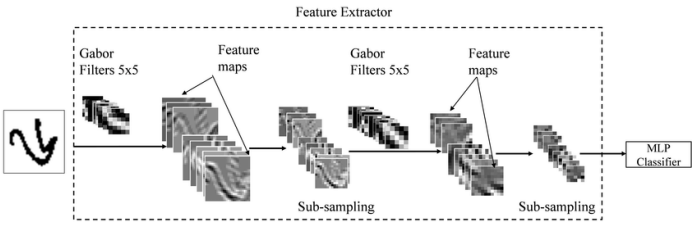

Fig 2: Feature extraction of a character using CNN

CNN is used for feature extraction from the images. Multiple convolutional layers progressively learn abstract features, enabling robust representation of characters. Trained CNN models classify individual characters or character regions, recognizing them from a predefined character set and can handle variations in text size, orientation, and font, improving OCR robustness.

Algorithms of CNN like Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) is used here. These handcrafted feature extraction algorithms were widely used before the advent of deep learning-based approaches like CNNs.

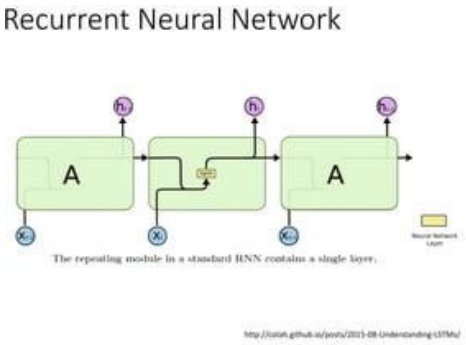

Fig 3: RNN Sequence extraction

Effective for sequential data processing, RNNs maintain a hidden (h1,h2,…hn) state that captures information from previous inputs, making them suitable for tasks like natural language processing and time series prediction.A type of RNN designed to overcome the vanishing gradient problem and capture long-term dependencies in sequential data.Similar to LSTM but with a simpler architecture, GRUs also address the vanishing gradient problem and are often used in sequence modeling tasks. By using DL algorithms features will be extracted but now the API and SDK together creates a interface to show up the obtained output from the DL algos.So we can say that the proposed methodology goes like the first step is to fed the images to the CNN or RNN and the obtained output will be shown to the user through user interface.

IV. RESULTS AND DISCUSSION

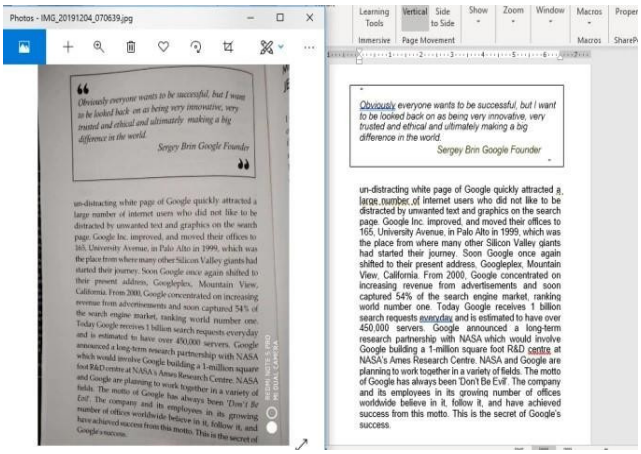

Fig 4 : Text from Picture

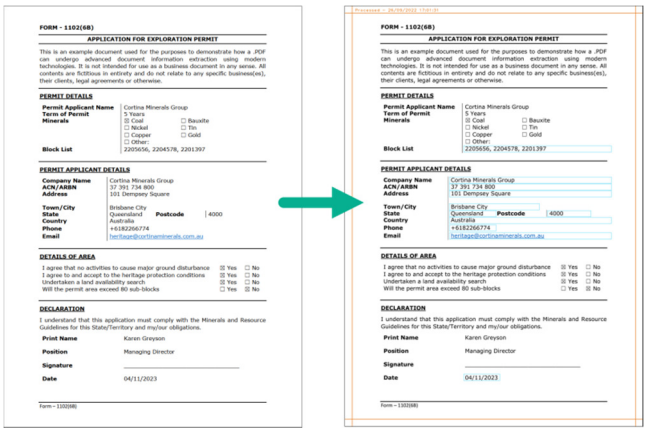

Fig 5 : Text from pdf

Fig 8 : Text from Image

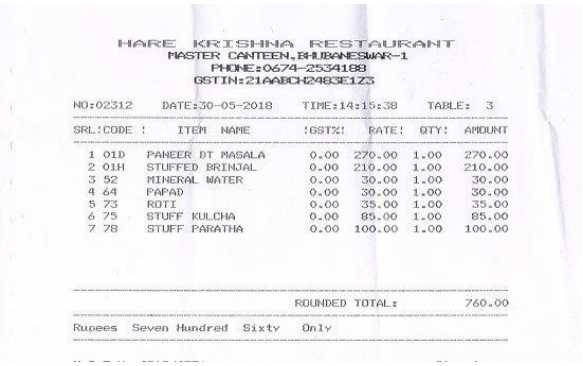

Fig 6: Receipt of ordered food from a hotel

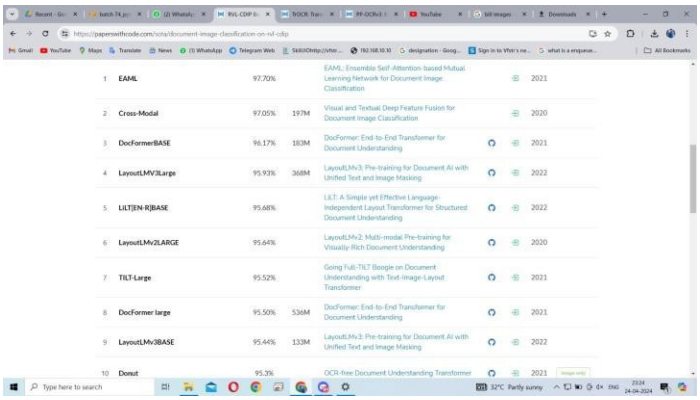

Fig 9: Extraction from various files.

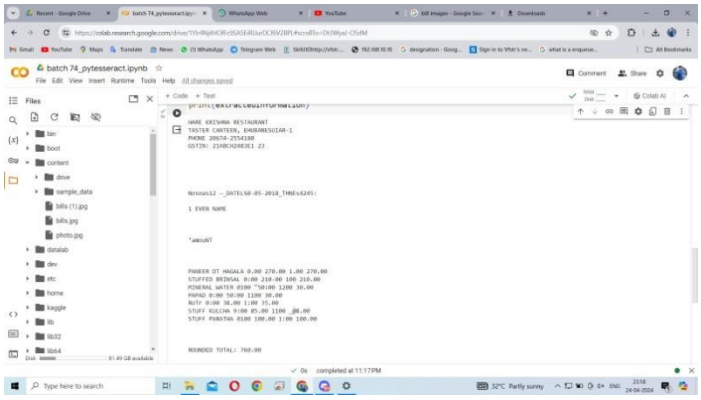

Fig 7: Text from the receipt.

The above picture depicts that the text which is taken from the receipt.when the machine learning algorithm which trained with Grocery Store Receipts Dataset gives us this output .likewise we get the other formats when we trained with specific datasets.

Service provider	Amazon Textract	Microsoft Azure Computer Vision	Google cloud Vision	Nanonets
Accuracy	50%	**97%**	61%	**97%**
Speed	**5 seconds per page**	6 seconds per page	6 seconds per page	16 seconds per page
Price	£1.28 per 1000 pages	£1.20 per 1000 pages	**£1.08 per 1000 pages**	£365 per month
Ease of implementation	Textract offers an intuitive CLI which is useful if you can download it.	**An Alteryx tool from the gallery allows for easy and fast integration.**	Cloud vision offers a CLI which can be implemented into Alteryx but requires installation.	Requires models to be created on training data. Then using API calls to train and score your data.
Other features	**Human review, confidence score, offers table and form detection**	Offers an on-premises solution, and some interesting spatial analysis	Facial recognition	**Has pre-built models such as passports, invoices which allow for immediate results.**

Fig 10: Displaying Accuracy for various platforms

V. CONCLUSION

In conclusion, Optical Character Recognition (technologies have evolved significantly, offering a range of solutions for text extraction from images. Leveraging libraries like OpenCV or PIL combined with deep learning frameworks such as TensorFlow, PyTorch, or Keras provides robust and customizable OCR pipelines.Tesseract OCR, OCRopus, and EasyOCR offer accessible options with varying levels of accuracy and ease of use. PyTesseract streamlines Tesseract's integration into Python workflows, simplifying implementation. Cloud-based OCR services like Google Cloud Vision API, AWS

Textract, and Microsoft Azure Computer Vision API offer scalable, reliable solutions with additional features like language detection and document analysis.The choice between these OCR methods depends on factors like accuracy requirements, resource constraints, and integration complexity, empowering users to select the most suitable solution for their specific needs.

Additionally, the flexibility of deep learning frameworks allows for the development of advanced OCR models tailored to specific domains or languages. Integration with cloud-based OCR services offers the advantage of offloading computational tasks and accessing continuous updates and improvements. Ultimately, the synergy between these technologies fosters the development of efficient, accurate, and scalable OCR solutions for various applications and industries.

VI. REFERENCES

1. SHUNJI MORI, MEMBER, IEEE, CHING Y. SUEN, FELLOW, IEEE, AND KAZUHIKO YAMAMOTO, MEMBER,“Historical Review of OCR Research and Developmental” PROCEEDINGS OF THE IEEE. VOL 80. NO 7. JULY 1092, 0018-9219/92$03.00 0 1992 IEEE.
2. Jingye Chen, Bin Li*, Xiangyang Xue Shanghai Key Laboratory of Intelligent Information Processing School of Computer Science, Fudan University,“Scene Text Telescope: Text-Focused Scene Image Super-Resolution”S. in IEEE.
3. Prarthana Dutta, Naresh Babu Muppalaneni, “Datasets in OCR for Handwriting Recognition: A Survey” 2023 4th International Conference on Computing and Communication Systems (I3CS) | 979-8-3503-2377-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/I3CS58314.2023.10127285.
4. THI TUYET HAI NGUYEN, L3i, University of La Rochelle ADAM JATOWT, University of Innsbruck MICKAEL COUSTATY and ANTOINE DOUCET, L3i, University of La Rochelle. “Survey of Post-OCR Processing Approaches” ACM Computing Surveys, Vol. 54, No. 6, Article 124. Publication date: July 2021.
5. Pengyuan Lyu, Chengquan Zhang, ShanShan Liu, Meina Qiao, Yangliu Xu, Liang Wu Kun Yao, Junyu Han, Errui Ding, Jingdong Wang * VIS, Baidu Inc. “MaskOCR: Text Recognition with Masked Encoder-Decoder Pretraining” arXiv:2206.00311v3 [cs.CV] 10 Oct 2023.
6. K. Simistira, N. Tsipas, and G. Economou, "Handwritten text recognition: A review," in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 2016.
7. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, 1998.
8. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning (ICML), 2006.
9. J. L. G. Pires and D. M. de Matos, "An experimental study of convolutional neural networks for handwritten digit recognition," in 2019 IEEE Symposium Series on Computational Intelligence (SSCI), 2019.
10. D. Karatzas et al., "ICDAR 2013 robust reading competition," in 2013 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.
11. A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2011.
12. L. Breiman, "Random forests," Machine Learning, 2001.
13. T. Pavlidis, "Structural pattern recognition," in Springer-Verlag New York Inc., 1977.
14. R. K. Srihari, "A top-down approach to document

recognition," in Pattern Recognition, 1988.

15. R. D. Kent, "A mathematical approach to character recognition," in Journal of the ACM, 1968.
16. H. Bunke, "Error-correcting output codes: A general method for improving multiclass inductive learning programs," in 1998 Springer-Verlag Berlin Heidelberg, 1998.
17. H. Bunke, "Error-correcting output codes: A general method for improving multiclass inductive learning programs," in 1998 Springer-Verlag Berlin Heidelberg, 1998.
18. K. V. Prathyusha and K. Latha, "Offline handwritten character recognition using SVM and HMM," in Procedia Computer Science, 2018.
19. S. Shah, M. S. Anand, and N. N. Patil, "A review of neural network applications in handwritten Hindi character recognition," in Materials Today: Proceedings, 2020.
20. S. Ojha and K. K. Tyagi, "Handwritten Devanagari character recognition using deep learning techniques: A review," in Materials Today: Proceedings, 2021.
21. P. K. Bharati and S. K. Jena, "A study of Bangla handwritten character recognition system using deep neural network," in Procedia Computer Science, 2020.
22. A. Thakur, S. Singla, and R. Juneja, "An efficient approach for OCR of printed Gurumukhi script," in Procedia Computer Science, 2018.