E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2022; 3(2): 59-68
Received: 16-04-2022
Accepted: 26-06-2022

Jhankar Moolchandani
Research Scholar, Department of Computer Science, Tantia University Sri. Ganganagar, Rajasthan, India

Dr. Kulvinder Hundal
Research Supervisor, Tantia University, Sri Ganganagar, Rajasthan, India

Corresponding Author:
Jhankar Moolchandani
Research Scholar, Department of Computer Science, Tantia University Sri. Ganganagar, Rajasthan, India

A review of English language from the scanned document with the use of pattern recognition: Detailed study on methodology and working of optical recognition system

Jhankar Moolchandani and Dr. Kulvinder Hundal

DOI: https://doi.org/10.33545/27076636.2022.v3.i2a.71

Abstract
Optical Character Recognition (OCR) is the procedure through which a machine analyses scanned printed or handwritten documents. OCR is a new study topic for Indian scripts that has the potential to replace the habit of manually completing data entries for diverse documents from printed paper records. The primary benefit of this digital textual material is that it can be altered, something that is not feasible with digitized files.

Keywords: Pattern recognition, machine learning, document analysis recognition, optical character recognition, diabetes retinopathy

1. Introduction
Classification, pre-processing, and feature extraction make up the fundamentals of the OCR system. The process of modifying and enhancing the final product Use features like skew detection and response, quantization, text classification, and flattening in the pre-processing stage to help improve the final image. The OCR procedure is divided into four basic stages such as, retrieval of features before disposal, classifying and modifying the results. Among the methods used in pre-processing are those for identifying and correcting skew, banalisation, object classification, and flattening, among others.
Feature extraction is the process of dissecting a bunch of letters into their separate images [39]. A feature set is derived and character-specific from these segmented characters. One of the most important step in OCR production is character recognition. OCR performance relies heavily on character recognition. Among other things, the enormous range of text types, skew unpredictability, and poor document integrity make it difficult to distinguish cursive script.
Numerous ways for segmenting lines, words, and characters in printed Indian scripts are described. The scholars#39; approaches to segmenting the Hindi and Bangla languages are collected.

2. Segmentation of Characters
The diverse ways of the segmentation of characters are shown below:

- To execute character segmentation, thinning words [41] are first subjected to polygonal estimation. As derived from a digital curve, Digital Straight-Line Segments (DSS) are described by this method. A little amount of data is needed to describe the word, which necessitates a small number of segments. A rough form for the word is generated by combining the approximate points of segments acquired by approximation.
- Only those junction points are represented based on the estimated positions. There are connection locations where more than two additional positions surround them. It selected header junction points from all intersections (header line junction points). Green points represent junctions, while magenta points represent header junctions.
- A graph structure that is explored used a graph traversal method to find the segmentation points, when all the estimated points are added together. Found to dynamically classify data [25].
- To locate the segmentation areas, a graph search method is used to connect all estimated locations. The sentence is segmented into unique characters when it has obtained the segmentation positions; the editable words, on either side, are not segmented. The next section goes into detail about both the higher and lower modifiers.

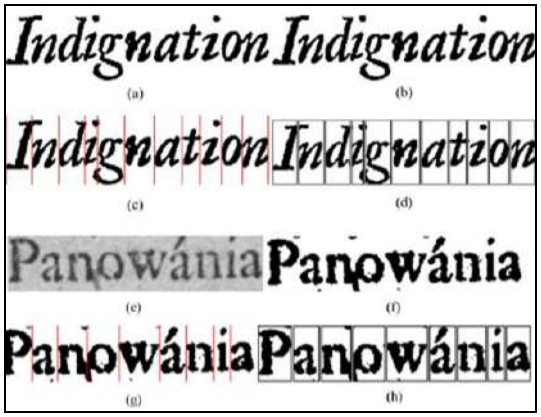


Fig 1: Shows the segmentation of the characters. [40].

3. Vector-Based Feature Extraction

Learning mechanisms can be improved by considering prior knowledge of the basic qualities and characteristics of handwritten and printed characters. Two instances of these qualities in the existence of linear and/or structures in characters, as well as the use of OCR, should be noted and the relevance of character information being conveyed using boundaries. Linear and circular shapes make up the characters in both handwritten and printed text. For example, feature extractors can search for these forms in character images. Character prototypes and ellipsoids can be found using the Hough transform, as can the lines in the Bengali script. An alternate method involves supplying the neural network with intentionally created patterns to push the neural Networks' location of specified forms during training. Non Cognitive training used this method to learn feature maps in convolutional networks. First, a neural network is constructed with 12 cell planes, one for each line direction, to train it to detect patterns of lines. Complex (e.g., circular) patterns are learned in the subsequent layers. Forcing each cell plane to recognize a predefined type of pattern is a necessary part of the learning process [42].

4. Advantages of OCR

OCR software has a slew of benefits, including the ability to save institutions and individuals both time and cost. A few advantages of OCR involve:

- Text transcribed into the set of words processing document can be edited with OCR
- Software, which also makes it possible to perform keyword or phrase searches to find specific documents.
- OCR automatically categorizes documents, improving efficiency and lowering employment costs.
- The data's precision is improved because it is extracted, vetted, and verified in a matter of seconds with a minimum of human intervention.
- Businesses reap the benefits of reducing human data entry.
- Paper documents can be saved on storage space by scanning them instead of storing the originals. A single Compact Disc can hold a cabinet full of files.

Businesses can increase their productivity and maximize their return on investment when their OCR system is more stable and performs better [43]. OCR is a process for classifying an image's optical patterns represented digitally. Feature extraction, categorization, and segmentation is used to accomplish character recognition. The author suggested fundamental concepts of OCR. The chapter begins with a brief overview of OCR systems and their history.

Several OCR scanning, pre-processing and extraction of features and training and recognition as well as post-processing be discussed. The many applications of OCR systems are discussed, then OCR technologies are discussed [44].

OCR and handwritten character identification based on Neural Networks (NNs). Experiments demonstrate that the strategy improves optical character recognition and handwritten character recognition accuracy. Each algorithm is very self-explanatory. It has described new approaches to handwritten character recognition, which is a widely popular and computationally demanding activity. The goal of the investigation is to make the most comparisons possible. Significant existing procedures and systematize them according to their characteristic considerations. The results in the algorithm's behavior approaching the expected similarity [45].

5. Disadvantages of optical character recognition

The following are some disadvantages of OCR given below:

- Handwritten text cannot be read by OCR software because it is too slow. The computer must be taught how to write by hand.
- A large amount of room is needed to accommodate the image.
- However, this can result in a decrease in the quality of the image.
- The quality of the final image is directly related to the quality of the initial image.
- High-quality final images are dependent on high-quality initial images.
- Each document must be meticulously examined and then manually corrected.
- Some errors can be made throughout the process.
- Not worth the effort for small quantities of text [46].

6. Classification of Optical Character Recognition

OCR is a technology that identifies and classifies every graphic and character in the typewritten text. One word at a time, optical word recognition identifies entered text in a computer. OCR is typically performed offline, which means that it does not require a computer. It read material that isn't being moved by the user. Some cloud-based providers offer an OCR Application Programming Interface (API) accessible through the web Handwriting recognition systems, for example, benefitting from an investigation of the handwriting movement. Rather than simply recording words and glyphs, this technique allows us to capture behavior such as the order in which components are drawn, as well as their placement and elevation. Accuracy could be improved because of this new knowledge. Along with online character recognition, this system is sometimes mentioned as, real-time character recognition, intelligent character recognition, and dynamic character recognition [47].

The Royal Botanic Garden Edinburgh (RBGE) has studied the use of OCR to aid in the digitization process. The was validated using a two-stage data input approach for herbarium specimens. According to categorizing by collector and/or Country of origin, batch processing of records was used to add data. The OCR text was used to get the information. The specimen records were supplemented

with new data by a group of six digitizers. using photographs of the specimens. A series of tests comparing the efficiency of data input across batches of specimens sorted and unsorted were done to establish the role of OCR data in the digitization process. A poll was conducted to confirm the digitizing staff's perceptions of the various sorting alternatives. Total specimens processed amounted to 7,200. OCR-sorted samples outperformed a control group that had been randomly selected. Took significantly less time to digitize. It was found that the most efficient method was one that needed only a few fields of data to be entered and filtered records by collector and Country. Additionally, a vast variety of OCR products are commercially accessible. The literature review covers the history of OCR and the many strategies used to build OCR in chronological sequence. Further improvements must be made to the process of producing OCRs for printed and handwritten materials on low-quality paper. Additionally, OCRs must be produced for regional languages, even though works of literature for some languages, such as Hindi, Kannada, and so on, are available. Additionally, study might be conducted to produce OCRs capable of identifying multi-font or multiscript characters [49].

The OCR literature with an emphasis on Urdu-inspired cursive scripts. The Urdu, Pushto, and Sindhi languages are examined in detail, with the Nasta and Naskh scripts being highlighted. Before discussing the OCR works, the characteristics of Urdu-like scripts are discussed, followed by a discussion of the existing text image databases. The numerous approaches have been divided into three categories: (Printing, handwriting, and online character recognition are all examples of methods for capturing characters. Each section analyses the works in comparison to a convention when it comes to OCR, there is a lot that goes into preprocessing and segmentation [50]. The texts serve as a significant representation of language.

Due to the volume of text generated and the historical significance of particular documents, generated texts must be read by computers and made editable and searchable. One of the primary goals of pattern recognition study is to replicate artificial systems with human-like perceptual capacities such as papers. After decades of study and advancements in computing powers, machine intelligence is still far from matching humans' capacity to understand typed or handwritten text. More complicated scripts or handwritten messages, particularly ones found on business cards and letters, challenge the present state of OCR technology. Additionally, many existing OCR systems are language dependent. As a result, advancements in OCR technology have been inconsistent between languages. Despite the requirement for processing many Persian historical texts and the widespread usage of OCR in a range of applications, few Persian OCR systems achieve a high recognition rate. As a result, the challenge of reading Persian typed documents mechanically with near-human performance remains an outstanding problem and the primary subject of the dissertation. It presented a new strategy for two critical pre-processing phases in any OCR system: skew detection and page segmentation. Then, character segmentation is customarily suggested by segmenting Persian manuscripts into sub-words. Sub-word segmentation was chosen to circumvent the difficulties associated with segmenting heavily cursive Persian texts into separate isolated characters. It has offered a hybrid technique for feature extraction that combines three widely used methods and then applies a nonparametric classification method. Numerous studies and patents show near-100 percent recognition rates. Such assertions provide the impression that automation issues have been resolved. OCR is extensively used; however, its accuracy is still significantly lower than that of a child's reading ability. The failure of several real-world applications demonstrates that performance issues with composite and degraded documents persist and that there is still room for improvement [51].

6.1 Pre-processing

OCR software commonly pre-processes pictures to increase the likelihood of effective recognition. The following are a few methods to consider:

- It can be necessary to rotate the page by a few degrees counter clock wise or clockwise to align text lines horizontally or vertically if the document was not correctly scanned.
- Color or monochromatic photographs can be transformed into black then white images (called a binary image because there are two colors). Binary images are the most common type used in commercial image recognition systems because it was simple and easy to process. This process does have a profound effect on the performance of the personality recognition phase, and prudent decisions are made considering the method to be used for a particular input image type. This is the excellence of the technique used to obtain a number value which is dependent upon the type of source images (historical degraded, scene text image, scanned document, etc.)
- The font in international texts might vary down to individual words, it's indeed necessary to determine the script before applying the proper OCR. It's indeed necessary to determine the script before applying the proper OCR, because the font in international texts might vary down to individual words.

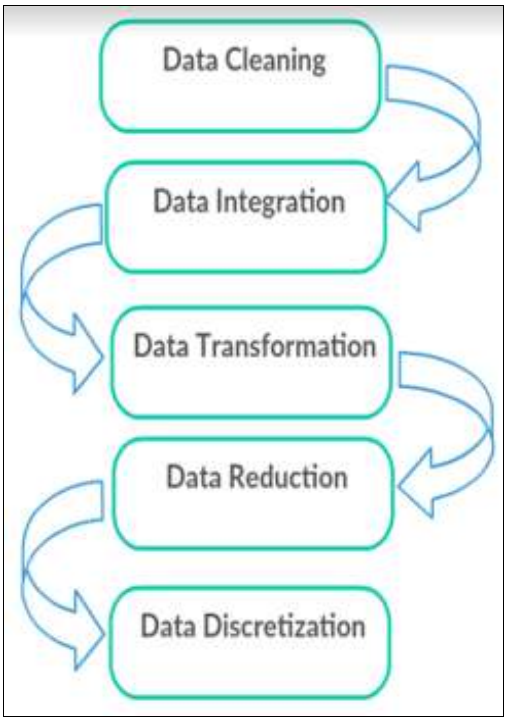


Fig 2: Shows the steps of pre-processing [52]

6.2 Text Recognition

Prioritized character lists can be generated using either arithmetic or logical core OCR methods,with the former being the most common.There are other names for this technique, including pattern recognition, pattern matching, and image correlation. Matrix matching is the process of comparing pixels by pixels in a picture to those in a previously recorded glyph. The input character must be correctly isolated from the main of the picture, and the stored glyph must be in the same typeface and height as the input glyph for it to be identified. Users who are unfamiliar with a font's style cannot even be able to make use of this strategy. During the early days of OCR, actual cluster light was intended to recognize characters [54].

- Feature extraction is the process of breaking down glyphs into features such as lines,shuttered loops, line vision, and intersections of lines. The extraction characteristics lower the depth of the representation, allowing the recognition procedure to be performed more effectively computationally. A comparison is made between these characteristics and a character represented as an abstract vector-like recognition that could consist of just one or several character prototypes. With intelligent OCR, as is common in intelligent OCR and most modern OCR systems, the general feature detection algorithms used in ML can be applied. For example, using the k-nearest relatives' technique, images can be compared to recorded glyph features to determine which glyph has the closest visual appearance [55].
- Text line recognition and localization is a critical step in whole-page document analysis, it is nevertheless hampered by the variability of real-world documents. The author described a novel technique for full-page text recognition that is both efficient and accurate. Using regressions as contextual layers, the background estimating approach is used to localize the text lines in the document. It is just the location of the text lines on the screen's left side tha is anticipated to maximize the efficacy of the localization approach. The text recognizer is then responsible for anticipating when the text to recognize and getting the conclusion of the document. On the diverse Maurdor dataset, the technique has demonstrated promising results in terms of whole-page text recognition [56].
- The transcribing platform i.e., identification and retrieval of digitized historical writings using machine learning. The primary user interface is supplied by an open desktop program that includes tools for segmenting document pictures, adding transcriptions, and tagging items included inside them, among other things. The desktop programmer can communicate with the documents that can be stored in the platform's document management system. Using a set of document image analysis tools, including the evaluation of the design as well as the automated identification of handwritten and typed text features other users must be granted access to documents that have been submitted to the platform [57].

6.3 Post-Processing

The precision of OCR by restricting the result to terms that only occur in the lexicon is effective.The scope of this dictionary could range from the whole English language to a more tightly focused technical lexicon, depending on the industry. While the output stream can be a plain text stream or a file with characters in certain circumstances, more sophisticated OCR algorithms could preserve the page's original layout and provide files such as an augmented PDF that comprises both the actual photo of the pages and a searching textual representation. The located close analysis carmaker of founder frequency can be estimated by observing how frequently certain words are in the vicinity. The actual world, curve-oriented text, appears nowhere in well-accepted scene text datasets the International Conference on Document Analysis and Recognition (ICDAR) is curved text in addition to the usual horizontal and multi-oriented types. More than half of the photos in the total text have more than two orientations, making it one of the most diverse collections of images. Recently, new generations of systems treat text. Segmentation as a challenge in identification has shown its efficacy in dealing with multiple-oriented texts [58].

The worldwide Document Image Binarization Competition (DIBCO 2017) to see a significant rise in the number of supervised techniques. The overall goal of the competition is to discover current improvements in machine-printed and handwritten document images binarized in the process. Utilizing performance evaluation metrics that are inspired by document image analysis and recognition criteria. 26 techniques were submitted to the competition, which is described along with the assessment metrics utilized and the performance of each approach. For the study community working on enhancing both machine-printed and handwritten document image binarization approaches, various conclusions must be taken from the methods and their performance. DIBCO 2017 saw a significant rise in the number of supervised techniques that occupied the top spots in the rankings. The strategy that received the best score was a supervised

approach. It's worth mentioning that the DIBCO datasets from prior years were used to train the supervised techniques document images [59].

The methods that seemed to be highly performant in prior years'; DIBCO challenges, either as a whole or through the usage of a specific component of the algorithm that was previously utilized. It is also worth noting that the majority of the methods make use of an explicit post-processing stage. Demonstrating the success of the binarization process is highly dependent on the success of those stages in which a preprocessing stage is part of the binarization pipeline.It should also be noted that the submitted techniques tended to gravitate toward a strategy that relied on phases that were dominant in the study on document image binarization, which should be taken into consideration. Such steps are concerned with Background estimation methods, such as the stroke width transform [60].

The CNN introduce a novel Handwritten Chinese character recognition (HCCR) in its entirety, which has shown promising results in many computers vision challenges, similar to their remarkable success in tackling many other computer vision problems. The study demonstrated that a deeper architecture can greatly help HCCR in terms of performance while also allowing for the construction of systems with fewer parameters. The conventional method of using techniques like Gabor or gradient feature maps to extract features continue to be beneficial in enhancing the performance of CNN as demonstrated by the results of the

study. With the help of Google Net, it has been used to classify images and has a fairly complex design. It created a simplified version of the algorithm for HCCR (denoted as HCCR-Google Net). The HCCR Google Net that employed has 19 layers and just 7.26 million parameters, which is a small number in comparison to other networks. It has been demonstrated that the single and ensemble HCCR-Google Net models suggested here reach newly developed state-of-the-art recognition accuracy of 96% and 97%when traditional directional feature maps are properly incorporated, outperforming the previous best result by a significant margin [61].

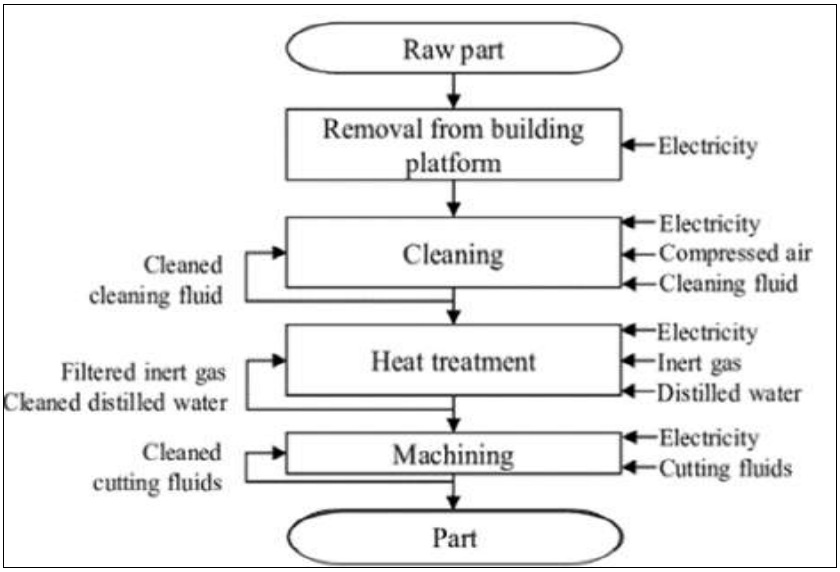


Fig 3: Shows the processing of the post-processing [62]

6.4 Digitization

The input for the off-line OCR would be either in printed or handwritten form on paper. The digitization process is the conversion of a paper-based document to a digital one. The conversion into electronic form is accomplished by scanning the original document and producing an electronic representation in the form of a bitmap image. As a matrix of dots, this imaging method records variations in light intensity reflected from the original paper. Each dot's light/color value(s) is encoded in binary digits. In a binary scan, each dot would need one bit, but a color scan can take up to 32 bits per dot. Normally the document image is scanned in gray tone mode, as the gray-tone image can be easily converted into binary form i.e., black, and white form for further processing. This digitized image document is then fed to the pre-processing phase.

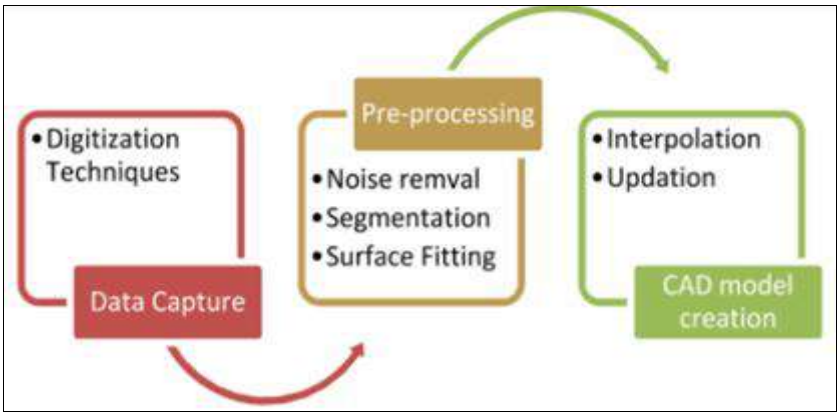
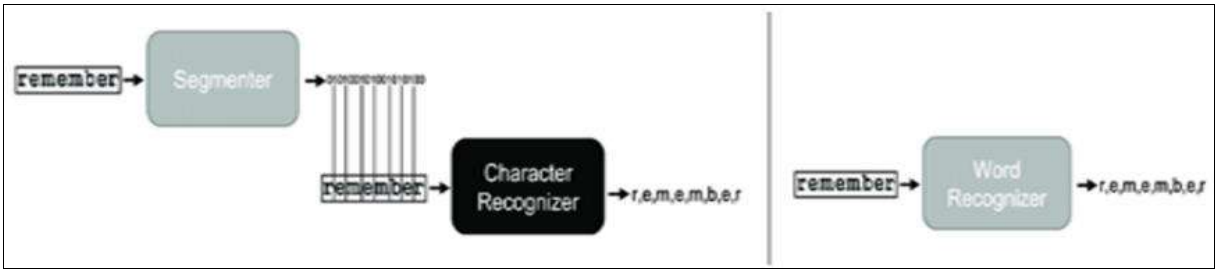


Fig 4: Digitization Process [66]

OCR became a powerful tool for most digital world applications. There are a wide variety of languages and script styles throughout the world, but in countries like America, Russia, and Europe, the script is nearly identical or similar, and study began several decades ago with the development of efficient algorithms for any printed document or content on an image or handwritten document. The beauty of India is that it has a single basic language, Devanagari (script), and many languages such as Hindi, Telugu, Malayalam, Tamil, and Kannada are spoken in various parts of the country and use similar character styles from the Devanagari character set. Due to the similarity in character styles, existing algorithms designed for foreign languages have difficulty identifying incorrect letters in printed or handwritten documents. It has listed the most widely used techniques or algorithms in optical character recognition. Pattern recognition is a subfield of image processing based on the findings, there is no single algorithm or OCR model that can accurately recognize the characters of the Devanagari language now. However, existing algorithms with classifiers based on convolutional neural networks that have the highest accuracy must invest significantly more time and effort in the subject to improve the accuracy of the Devanagari script algorithm [65].

6.5 Segmentation

Clean documents are produced following the pre-processing stage. Segmentation would be the next step. The next step is to break the document down into its constituent parts. It distinguishes between the many logical components, such as text and images, a paragraph's line alphabetical arrangement of a letter. In OCR, segmentation is an essential phase. since it can affect the script's recognition rate by separating words, lines, or characters. Proper segmentation is essential for accurate recognition.



7. Literature Survey

Pattern recognition and machine learning are used in the work to analyze language in scanned documents. The following are the explanations provided by various researchers and authors of related works:

Rabby *et al*., (2021) [22] explain that Bangladeshi is the world's most widely spoken language. To develop an OCR system, it is critical to identify execute character recognition and segmentation modules in a certain language and print style. Utilizing numerous deep learning models, the author suggested a novel technique for automatic identification of each document's language and printing style, including both printed and handwritten, in terms of a script (Bangla or English). Moreover, they could classify text as printed or handwritten, a classification challenge that obtained greater than 99 percent test accuracy at the character level. Additionally, under the suggested model, Only Bangla and English can now be detected, and there are only printed and handwritten forms of the written word that can be detected at the moment. But they are working to expand their detection capabilities to include more languages and writing styles such as letterpress characters and font families.

Wei *et al*., (2021) [23] stated that relaxometry based on nuclear magnetic resonance (NMR) is widely utilized in a variety of disciplines of research due to its advantages over metabolomics techniques, including ease of preparation, ease of use, and low price. However, there are no publications on metabolic mixes that can be evaluated by T2 relaxation curves, which are commonly used in metabolomics research such as determining their geographical origin and extracting features via Data mining and pattern recognition. The author would go through the data mining technique for relaxing metric data in that work (i.e., relaxometry learning). Analysis and a machine-learning method are the foundations of the approach, which is well-suited for studying relaxation curves.

Rani, N. Shobha., (2020) [24] suggested that the handwritten Kannada character identification based on Devanagari handwriting recognition technology can be used to impart knowledge. The enormous Devanagari recognition system data corpus will be used as training material to recognize traditional Kannada characters written by hand, albeit with a smaller database.VGG19 NET uses deep learning network architecture to transfer knowledge for recognition.

A hidden output layer, two tightly connected ones, and five blocks of hidden layers make up the VGG19 NET architecture. Each block has layers of convolution and a layer of maximum pooling except the block. A total of 92000 photos with 46 classes make up the Devanagari character set in the suggested classification framework, whereas the 81654 training pictures and 9401 testing photos total 188 classes with 200-500 sample images in each. VGG19 NET uses 1,23,654 data samples in its training. They employed 9401 samples with an accuracy of close to 90% for experiments with 188 classes made up of 40-100 samples apiece. 73.51 percent accuracy after 10 epochs of evaluation with the VGG19 NET, with a loss of 16.18 percent.

Alyoubi *et al*., (2020) [25] planned that Diabetic Retinopathy (DR) causes abnormalities in the retina that impede vision when they're not well planned. Due to the
irreversibility of DR, therapy can only sustain eyesight, not restore it. It is possible to drastically lower the risk of visual loss with early detection and treatment of DR. Ophthalmologists' use of DR retina fundus images for diagnosis is labor-intensive, expensive, and prone to error compared to computer-aided evaluation. Recently, deep learning has evolved as one of the most important technologies widely used approaches for optimization in a range of sectors, most notably medical picture analysis and classification. Convolutional neural networks (CNN) are becoming more extensively employed as a method for deep learning in medicinal copy processing due to their effectiveness. The analyses of recent ways of recognition that are state-of-the-art and categorization of Deep learning methods are used to DR fundus pictures. Table 1 summarizes the related work.

Wang *et al*., (2020) [26] planned that computers can now understand human languages through Natural Language Processing (NLP). An important function of natural language processing (NLP) is to segment words for deeper grammatical and semantic analysis. Multimodal neural networks (MNN) are suggested. There is a multilayer sub-neural network for each mode, and each one has its distinct structure. It is a tool for converting features from one model to another. An English word recognition system based on a network model approach is developed to address the problem word segmentation approaches cannot guarantee the long-term dependability of text semantics and extended training prediction time. Shortens network training and prediction times by utilizing the Conditional Random Field (CRF) model to annotate several phrases at once. Bi-direction Gated Recurrent Unit (BI-GRU). According to the results of the experiments, as it relates to word segmentation, however, this technique performs comparably to the BI LSTM-CRF model, however, the average

estimated processing rapidity is 1.94 times quicker boosting word segmentation processing efficiency.

Paliwal *et al*., (2019) [27] analyzed that scanners and mobile phones are making it increasingly difficult to extract information from unstructured document pictures like receipts, insurance claim forms, and financial bills. Data extraction from photographs containing tabular sub-images presents a distinct set of challenges, compounding the difficulty of the task.

It involves accurately detecting identifying and extracting information from the rows and columns of the specified table in a picture. Detection of tables has come a long way, but extracting the data from them remains a challenge since it requires a more precise recognition of the table structure (rows and columns). There have been several previous attempts that used two different models to tackle the table detection and structure recognition issues. An end-to-end deep learning network for the identification and recognition of tables and structures is presented as Table Net. The approach relies on the interdependence between the two objectives of identifying tables and recognizing their structures to separate the table and column areas.

International Conference on Document Analysis and Recognition (ICDAR) 2013 and Marmot Table, two publicly available datasets, were used for these experiments and yielded. The suggested model and extraction technique produce state-of-the-art results.

Sánchez *et al*., (2019) [28] intended the handwritten text recognition to be utilized to access the worldwide collection of historical materials housed in archives and libraries. Automated Handwritten Text Recognition (HTR) can be a difficult problem to solve because Feature extraction, image processing, and document image analysis are among the sophisticated Pattern Recognition techniques. They must be used in conjunction with one another. This work provides the HTR benchmarks that increase in complexity from various perspectives, based on historical documents provided in English and German during the 2013 through 2017 ICFHR and ICAR conferences' open contest. There is a suggested system for each benchmark that improves upon the previous work that has been done under similar circumstances. The goal of the study is to set new standards and benchmarks for HTR technology progress by presenting fresh challenges and illuminating current state-of-the-art outcomes together with the datasets and all of the software tools necessary to build the most basic systems accessible for free.

Lu *et al*., (2019) [29] analyzed optical coherence tomography (OCT) as minimally invasive imaging technology. It may produce micrometer-resolution three dimensional pictures of retinal constructions. These pictures can aid in identifying virus-connected changes beneath the retinal surface, the existence of edema or fluid accumulation can impair vision, and are indicative of retinal vascular abnormalities.

The goal of the author is to present Multiclass Fluid detection (MFD) and segmentation in OCT pictures of the retina as a new framework. A neural network with all of its connections convolutional was trained to distinguish and classify by a graph cut technique, fluid pixels can be produced based on OCT pixel intensity and segmentation of the retinal layers. Random forest classification was used to identify and remove the incorrectly identified fluid areas from the segmented regions of fluid. The suggested framework takes first place in Segmentation difficulty in detecting (mean dice: 0.7667) (mean AUC: 1.00).

Rashid *et al*., (2018) [30] suggested that tables are a convenient way to express data structurally. Recognizing tables is critical for extracting the information from document images. Typically, current OCR algorithms deliver textual data extracted from tables without understanding the table's real structure. Recognizing the table structure is critical for deriving the content's contextual significance. Recognizing table structures in diverse texts is difficult because of the wide variety of table layouts. It becomes more difficult when there are no physical rulings on the table. In a model of a pre-trained neural network, the textual content of documents is categorized as a table or non-table elements. The system was trained on a portion of the photos for UW3 images and demonstrated greater than 97 percent table and non-table detection components on a test set.

Chen *et al*., (2015) [3] stated that OCR methods cannot be directly utilized. As far as Recognition systems are concerned, they are designed to work with only one language and a single orientation. Therefore, they could only handle those types of texts. Many non-character-based ways of recognition have been developed to address the issue. These approaches did not perform as well as more advanced OCR systems. As a result, it is preferable to identify the linguistic type and position before performing OCR. Moreover, it is quite difficult to extract consistent information for recognition. Since the forms of the letters in different languages are significantly confusing. Convolutional neural networks (CNN) have recently demonstrated remarkable effectiveness in pattern recognition tasks. As a result, CNN is an excellent choice for such demanding assignments. The author started a CNN to acknowledge text attributes. There is indeed a new sliding window voting method suggested to reduce the size of the network. The technique demonstrated a very high recognition rate in the experiments. The results validated the suggested strategy, which may also be used to develop a document interpretation system using OCR technology. There is a wide range of authors who used the technique and presented their discoveries, as can be seen in table 1.

Table 1: Summary of related work

S. no	Author's	Techniques	Outcome
1.	Rabby *et al*., (2021) [22]	Conventional neural network (CNN)	1. CNN was able to identify the document's characteristics. 2. The new voting method with a sliding window reduces the size of the network while maximizing the use of the text line's content. 3. The method achieved a very high success percentage in the tests.
2.	Wei *et al*., (2021 [23]	Optical character recognition (OCR)	1. Due to its ability to distinguish between handwritten and printed texts and the high degree of accuracy achieved in testing. 2. Optical character recognition (OCR) is a critical component in character recognition and segmentation systems.
3.	Rani, N. Shobha., (2020) [24]	Optical character recognition (OCR)	1. OCR is employed to establish tables and extract data from document images by table recognition. 2. Test sets yielded a 97% accuracy rate in the ability to distinguish between components on and off a table. 3. Understanding the context of the content requires an understanding of the table structure.
4.	Alyoubi *et al*., (2020) [25]	TableNet and deep learning	1. Techniques are employed in the identification and recognition of tables and their structures. 2. It comprises accurately detecting the tabular region inside a picture and then recognizing and extracting data from the selected table's rows and columns.
5.	Wang *et al*., (2020) [26]	NLP (Natural Language Processing)	1. An important function of natural language processing (NLP) is to segment words for deeper grammatical and semantic analysis. 2. Tool for converting features from one model to another. There are issues with word segmentation processing.
6.	Paliwal *et al*., (2019) [27]	Handwritten Text Recognition (HTR)	1. It is utilized to identify the textual material. They are aimed at becoming new challenges for HTR technologies to spur future innovation. 2. Baseline systems are implemented by publicly available datasets and software tools.
7.	Sánchez *et al*., (2019) [28]	Deep Learning Network (DLN)	1. A novel model for handwritten Kannada character recognition is described that uses handwritten Kannada characters. 2. The Devanagari handwritten recognition system's training data was used.
8.	Lu *et al*., (2019) [29]	Machine learning	Automated learning technique and analytic framework specifically designed to analyze relaxation curves.
9.	Rashid *et al*., (2018) [30]	Optical coherence tomography (OCT)	A graph-cut algorithm can extract the intensity of OCT pictures and retinal layer segmentation to produce 3D scans of retinal structures with micrometer resolution.
10.	Chen *et al*., (2015) [3]	Diabetic Retinopathy (DR)	In comparison to computer-aided diagnosis techniques, Doctors' manual inspections of DR retina fundus pictures are time-consuming, labor intensive, expensive, and prone to misdiagnosis.

8. Comparative Analysis

This section of the study offers a comparative analysis of the pattern recognition abilities of several deep learning approaches. It is the most widely used technique for recognizing patterns. CNN Table Net, CNN Simple, Deep Learning Network (DLN), and Machine Learning (ML) are only a few of the approaches employed (ML). An OCR scanner has a 99 percent success rate. First, OCR has an accuracy rate of 97%. Table 2 displays the accuracy comparison. Figure 5 provides a graph comparing the accuracy of the results:

Table 2: Comparison based on Accuracy 120% 100% 80% 60% 40% 20% 0%

Technique	Accuracy (%)
OCR [24]	99%
OCR [32]	97%
DLN [25]	73.51%
DR [27]	94.5%

Accuracy (%) OCR [27] OCR [28] DLN [32] DR [35]

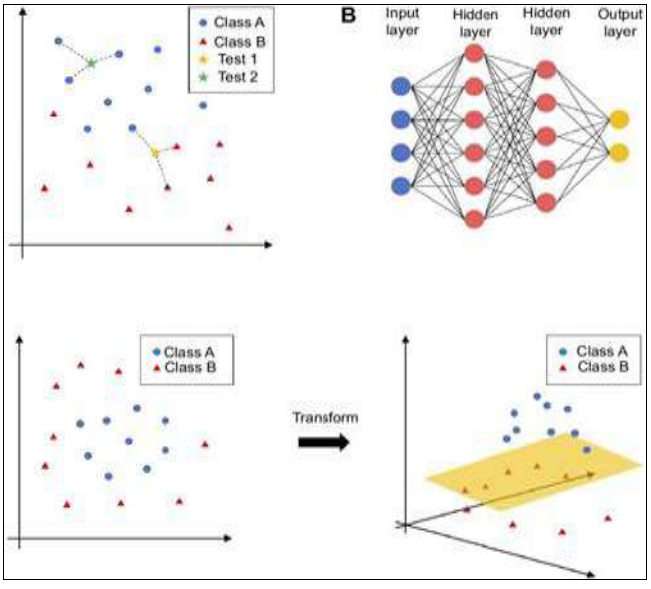


Fig 5: Comparison graph based on Accuracy

9. Conclusion and future scope

The author recommends a CNN model for the identification of orientation and document language. The Convolutional Neural Network (CNN) provides substantial improvement when used in conjunction with the handwritten document character recognition approach. To categorize document terms into a table or non-table categories, a neural network is trained on these contextual cues and then tested. The suggested framework investigated the difficulties associated with the classification of characters in the Devanagari Dataset Both the quality of the training data and the effectiveness of the machine learning procedures are necessary for a successful model. Additionally, machine learning methodologies were discussed to illustrate their applicability for a variety of real-world issues across a variety of major application fields. In the future, another method of improving the character segmentation model would be to move beyond a greedy search for the most likely solution. To implement a robust technique providing more accuracy and less error rate in the future.

10. References

1. Tanaka Hiroshi. X-BROT: prototyping of compatibility testing tool for web application based on document analysis technology. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), IEEE. 2019;7:18-21.
2. Elsayed, Omnia, Khaled Mahar, Mohamed Kholief, Hatem A. Khater. Automatic detection of the pulmonary nodules from CT images. In 2015 SAI Intelligent Systems Conference (IntelliSys). IEEE; c2015. p. 742-746.
3. Javed, Mohammed P, Nagabhushan, Bidyut B Chaudhuri. A review on document image analysis techniques directly in the compressed domain. Artificial Intelligence Review. 2018;50(4):539-568.
4. Saeed, Sana, Saeeda Naz, Muhammad Imran Razzak. An application of deep learning in character recognition: an overview. Handbook of Deep Learning Applications; c2019. p. 53-81.
5. Rajalakshmi M, Saranya P, Shanmugavadivu P. Pattern recognition-recognition of a handwritten document using convolutional neural networks. In 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), IEEE; c2019. p. 1-7.
6. Li Junchan, Yu Wang, Pengfei Wang, Qing Bai, Yan Gao, Hongjuan Zhang, *et al*. Pattern recognition for distributed optical fiber vibration sensing: A review. IEEE Sensors Journal; c2021.
7. Chen Li, Song Wang, Wei Fan, Jun Sun, Naoi Satoshi. Deep learning-based language and orientation recognition in document analysis. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE; c2015. p. 436-440.
8. Jebadurai, Jebaveerasingh, Immanuel Johnraja Jebadurai, Getzi Jeba Leelipushpam Paulraj, Sushen Vallabh Vangeepuram. Handwritten Text Recognition and Conversion Using Convolutional Neural Network (CNN) Based Deep Learning Model. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA); IEEE; c2021. p. 1037-1042.
9. Valenzuela Sergio E, Juan B Calabrese, Josue Ortiz-Medina, Claudia N Sánchez. Convolutional neural networks for detection of hand-written drawings. In 2020 IEEE Andescon, IEEE; c2020. p. 1-5.
10. Berry, Michael W, Azlinah Mohamed, Bee Wah Yap. eds. Supervised and unsupervised learning for data science. Springer Nature; c2019.
11. Sarker, Iqbal H. Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science. 2022;3(2):1-20.
12. Sarker, Iqbal H. Machine learning: Algorithms, real-world applications, and research directions. SN Computer Science. 2021;2(3):1-21.
13. Alzubi, Jafar, Anand Nayyar, Akshi Kumar. Machine learning from theory to algorithms: an overview. In Journal of physics: conference series. IOP Publishing. 2018; 1142(1):012012.
14. Mahesh Batta. Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet]. 2020;9:381-386.
15. Singh, Amanpreet, Narina Thakur, Aakanksha Sharma. A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (India.Com), Ieee; 2016. p.1310-1315.
16. Livieris, Ioannis E, Konstantina Drakopoulou, Vassilis T Tampakas, Tassos A Mikropoulos, Panagiotis Pintelas. Predicting secondary school students' performance utilizing a semi-supervised learning approach. Jthenal of educational computing research. 2019;57(2):448-470.
17. Ray Susmita. A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE; c2019. p. 35-39.
18. Dhall, Devanshi, Ravinder Kaur, Mamta Juneja. Machine learning: a review of the algorithms and its applications. Proceedings of ICRIC. 2019(2020):47-63.
19. Sharma, Priyanka, Manavjeet Kaur. Classification in pattern recognition: A review. International Journal of Advanced Research in Computer Science and Software Engineering; c2013, 3(4).
20. Martinez-Luengo, Maria, Athanasios Kolios, Lin Wang. Structural health monitoring of offshore wind turbines: A review through the Statistical Pattern Recognition Paradigm. Renewable and Sustainable Energy Reviews. 2016;64:91-105.
21. Silva, Jonathan A, Elaine R Faria, Rodrigo C Barros, Eduardo R Hruschka, André CPLF de Carvalho, *et al*. Data stream clustering: A survey. ACM Computing Surveys (CSUR). 2013;461:1-31.
22. Rabby, AKM Shahariar Azad, Md Majedul Islam, Nazmul Hasan, Jebun Nahar, Fuad Rahman. A Novel Deep Learning Character Level Solution to Detect Language and Printing Style from a Bilingual Scanned Document. In 2020 IEEE International Conference on Big Data (Big Data), IEEE; c2020. p. 5218-5226.
23. Wei, Feifei, Yuuri Tsuboi, Kengo Ito, Kenji Sakata, Jun Kikuchi. Relaxometric learning: A pattern recognition method for T2 relaxation curves based on machine learning supported by an analytical framework. BMC chemistry. 2021;15(1):1-8.

24. Rani N, Shobha AC, Subramani, Akshay Kumar, Pushpa BR. Deep learning network architecture based Kannada hand written character recognition. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE; c2020. p. 213-220.
25. Alyoubi, Wejdan L, Wafaa M Shalash, Maysoon F Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. Informatics in Medicine Unlocked. 2020;20:100377.
26. Wang, Dongyang, Junli Su, Hongbin Yu. Feature extraction and analysis of natural language processing for deep learning English language. IEEE Access. 2020;8:46335-46345.
27. Paliwal, Shubham Singh, Vishwanath D, Rohit Rahul, Monika Sharma, Lovekesh Vig. Tablenet: Deep learning model for end to-end table detection and tabular data extraction from scanned document images. In 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE; c2019. p. 128-133.
28. Sánchez, Joan Andreu, Verónica Romero, Alejandro H Toselli, Mauricio Villegas, Enrique Vidal. A set of benchmarks for handwritten text recognition on historical documents. Pattern Recognition. 2019;94:122-134.
29. Lu Donghuan, Morgan Heisler, Sieun Lee, Gavin Weiguang Ding, Eduardo Navajas, Marinko V Sarunic, *et al*. Deep learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. Medical image analysis. 2019;54:100-110.
30. Rashid, Sheikh Faisal, Abdullah Akmal, Muhammad Adnan, Ali Adnan Aslam, Andreas Dengel. Table recognition in heterogeneous documents using machine learning. In 2017 14th IAPR International conference on document analysis and recognition (ICDAR), IEEE. 2017;1:777-782.