Extracting Meaningful Data from Education Credentials using OCR and Image Processing

Jagdish Kharatmol
REDX Innovation Lab
Prin. L. N. Welingkar Institute of Management Development and Research (PGDM)
Mumbai, India
jagdish.kharatmol@welingkar.org

Nishil Visawadia
REDX Innovation Lab
Prin. L. N. Welingkar Institute of Management Development and Research (PGDM)
Mumbai, India
nishil.visawadia@welingkar.org

Tanmay Gurav
REDX Innovation Lab
Prin. L. N. Welingkar Institute of Management Development and Research (PGDM)
Mumbai, India
tanmay.gurav@welingkar.org

Sonal Daulatkar
REDX Innovation Lab
Prin. L. N. Welingkar Institute of Management Development and Research (PGDM)
Mumbai, India
sonal.dualatkar@welingkar.org

Kaustubh Dhargalkar
REDX Innovation Lab
Prin. L. N. Welingkar Institute of Management Development and Research (PGDM)
Mumbai, India
kaustubh.dhargalkar@welingkar.org

***Abstract*—To apply for higher education and job opportunities, a student's marksheet serves as a reference document. The conventional way of manually extracting meaningful information for companies and colleges is time-consuming, error-prone, and non-automatic. Optical Character Recognition (OCR) helps to facilitate the procedure. The authors have employed OpenCV template matching method using Python, noise removal using the custom autoencoder model, and OCR to provide solutions. The authors have also compared three OCR models using Python: Tesseract OCR, Keras OCR, and Easy OCR. The conclusions are based on observations made about error rate and effectiveness in handling noisy data. In recruiting process, organizations need to extract data from education credentials. The proposed method offers a streamlined and effective option that can help automate the process and enhance accuracy.**

Keywords—Image processing, education credentials, noise removal using a custom autoencoder model, OCR

I. Introduction

The future is going to be data and analytics. Considering India's huge population [1], we are generating a large amount of data in the present and will continue to in the future as well. The data needs to be collected, organized, and stored in a scalable and efficient manner. Maintaining physical copies of documents is not a feasible solution in today's digital world. The time and effort required to organize and search physical documents is high. Keeping a digital record seems to be correct. But it also has a few shortcomings; storing a good resolution of a digital record would take around 4 MB to 10 MB which increases the storage cost. High-end data servers require high electricity [2] and water requirements [3] for computing and cooling it down respectively, which is sustainably not acceptable. A better solution would be to store important data points on a digital database.

In India, for the year 2019-2020, a total of 3.85 crore students enrolled in various graduation courses [4]. The students have to submit academic credentials in terms of tenth standard, twelfth standard, diploma, and degree marksheets to the college and later the college verifies their marksheets.

They manually check the subject and marks scored for the subject respectively which is laborious, tedious, and error prone. Between the years 2014 to 2022, the government job applications were 22.05 crore [5]. Verifying each one's academic credentials is a laborious job.

In view of the above, there is a need of a reliable and computationally efficient system for extracting meaningful data from education credentials using Optical Character Recognition (OCR) as it is becoming increasingly important in educational institutions, recruitment agencies, and other sectors that handle large volumes of academic records.

Education credentials such as diplomas, transcripts, and certificates contain critical information, but manual data entry is time-consuming, prone to errors, and inefficient. OCR technology has the potential to automate this process by converting scanned or photographed documents into machine-readable text. However, the challenge lies in accurately extracting relevant data such as names, grades, and degree information from documents that may have varying layouts, fonts, and quality.

OCR (Optical Character Recognition) is a technology which happens to extract text from images. The extracted text gives the capability for searching and editing data. OCR is extensively applied in many different areas like legal industry, banking, captcha reading, institutional repositories, digital libraries, optical music recognition, automatic number recognition, and healthcare [6]. Despite OCR technology's extensive acceptance in other areas, the education and job-seeking sectors have yet to capitalize on its transformative potential completely.

Copyright © INDIACom-2025; ISSN 0973-7529; ISBN 978-93-80544-58-8

The major steps in any given OCR system are to take pictures, preprocess them via binarization, divide them into segments, and then use matching and recognition to determine the content [7]. To put it plainly, A physical document is first scanned, transformed into an image, and then converted into a binary pixel format. After that, the image is then segmented to break it into different parts for easier recognition. Line segmentation, word segmentation, and character segmentation are all done in chronological order. A classification model is used to recognize the text within these parts and provide the corresponding text output.

The purpose of this research is to explore and develop an efficient system for accurately extracting structured data from education credentials, improving the reliability and speed of academic record processing, and reducing human errors in data handling.

In the paper, the author has provided an OCR based system for extracting key data points from education credentials after experimentation with three well-known pre-trained OCR Models namely - Easy OCR, Keras OCR, and Pytesseract OCR. A custom denoising autoencoder model has been used to lower image noise before submitting it to a pre-trained OCR model to make it efficient in terms of text prediction accuracy [8].

II. Literature Review

Some authors have tried to compare different OCR models in various domains. Authors in [9] compared pre-trained OCR models called Pytesseract, EasyOCR, and KerasOCR on license plate recognition. The authors examined how these three pre-trained OCR models handle the challenges of recognizing closed characters and dealing with blurred text. The dataset used for evaluating the models consists of Malaysian license plates. The image was processed before passing it to OCR Models. On the image, operations like erosion, dilation, opening, and closing were applied in an attempt to improve the recognition accuracy. On the same note authors used Tesseract, EasyOCR, and DocTR OCR models for digitalizing medical reports [10].

Various image pre-processing techniques such as image binarization and brightness transformations were used with different OCR models and examine how these pre-processing methods influence OCR accuracy. The results demonstrated that OCR accuracy was significantly improved when pre-processing methods were applied, compared to cases where no pre-processing was used. On similar grounds, the experiment needs to be carried out on marksheets images with varying lighting conditions, blur and noise for getting a real time solution with acceptable accuracy.

Advanced deep learning models such as Mask R-CNN, and OCR technology for creating a powerful framework for automatic extraction of student performance from gazette assessment data have been used [11]. The main point the authors solved is the ability to extract the data accurately from tabular format. For creating training data Dataturks annotation tool was used which returned annotations in COCO format with bounding boxes around document fields such as Name, Roll Number, and others. Mask R-CNN with ResNeXt101-32d backbone and feature Pyramid Network has been used. It detects and recognizes various fields in exam result documents. The model workflow consists of two stages. In the first stage backbone network uses CNN to generate regions of interest with anchor boxes. The region proposal network provides region proposals based on scores from binary classifiers. In the second stage of Mask R-CNN, the fixed-dimension features obtained from the first stage are classified using a fully connected layer and softmax function. At the same time, a Mask classifier generates masks for each region of interest. OpenCV is used to crop the bounding box and then it is sent to the Pytesseract OCR model for predicting text which is further stored in the CSV file. Complex deep learning models were used which require heavy computing resources and might be a challenge to scale the solution. Getting the text prediction in real time without delay would be difficult.

The system for extracting information from digital marksheets has been described by the author in [12]. Image processing techniques, particularly binarization, and thresholding have been used in preprocessing. After preprocessing the image is given input to the OCR model. The OCR model has been used by the authors to convert images to text and store them in the database. The purpose of the system is to automate the process of converting physical marksheets into a database of marksheets. The paper also throws light on different methods for binarization, including Otsu's method and Adaptive smart binarization method which are used for performing automatic image thresholding. However, the authors would have stored key data points which would have used less storage space as compared to storing a whole marksheet.

A system of managing education data through an OCR system by automating the extraction of text from scanned or photographed marksheets reducing manual work and making the process efficient has been used [13]. However, the specific OCR models have not been discussed.

A solution for automating the data entry and validation process of marksheet while registration has been discussed wherein the authors claim that their proposed solution shows an improvement in processing speed and error reduction. At the core of the solution, data entered from the input field by the user is matched with extracted data from the marksheet using Pytesseract OCR with some level of preprocessing for extracting data from the marksheet. Once extracted, the data is stored inside a database. [14]

Over time, technology has advanced, bringing with it OCR models like EasyOCR and KerasOCR that perform better under certain circumstances and this paper uses these advanced technologies.

Thus, after seeing various studies of different OCR models and pre-processing techniques for image processing yet gaps remain, particularly in the areas of marksheet dataset diversity under varied conditions with text prediction using different OCR models. The study to find a solution for extracting key data points from a marksheet which is scalable, technologically feasible and economically and sustainable acceptable needs to be carried out.

III. Methodology

To create primary data for testing the OCR system and training the custom denoising autoencoder model, the authors have created a Google form accepting name and marksheet image data. Thus, real-world data has been collected wherein images of different quality were uploaded by respondents. A total of 25 marksheet images were received as a response. All the images were captured using mobile phones. The images varied in resolution, noise, lighting condition, size, and format. Some images were blurry and had bright spots caused by the flash. The diverse representations of data helped the authors build a robust model.

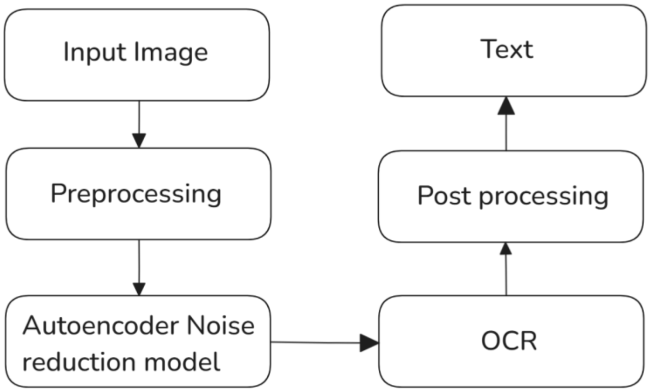


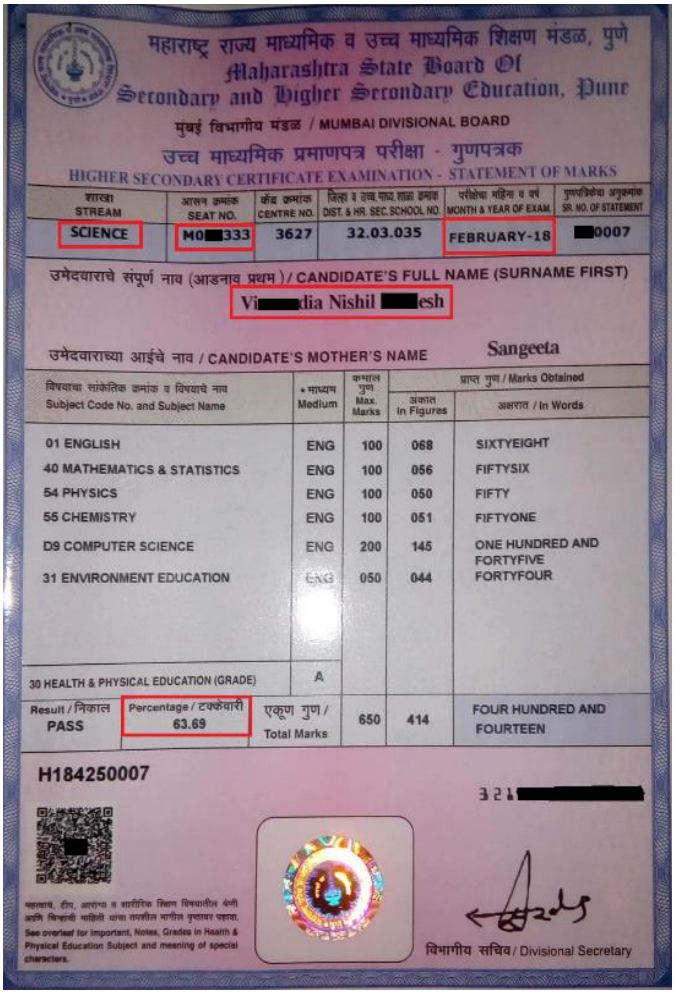
Fig. 1. Flow of Proposed OCR System

Fig. 1 shows the flow used by the Proposed OCR system built and summarized in this paper for text extraction from the image:

1) Users provide image to our system using a web-based platform.
2) Preprocessing of the image is done by performing resizing, gray scaling, thresholding, and template matching.
3) For processing, a Custom Autoencoder model has been introduced in the OCR system for noise reduction.
4) Pass the processed image to OCR for prediction of text.
5) Post-processing identifies the highlighted data in Fig 2.
6) Return text to the web page.

Fig. 2 displays the marksheet with highlighted data points to be extracted. Details including the student's stream, seat number, year of examination, name, and percentage are extracted by using OCR technology and stored in the database.

Fig. 3 specifies steps for preprocessing - resizing image, grayscaling, thresholding and template matching. Once the input image is given, it is adjusted to specific dimensions. The authors experimented with different dimensions but finalized on 1200 x 800-pixel dimensions. The grayscale method in OpenCV maps the image pixels to the range of 0 to 255, after which thresholding is applied to binarize the image, turning the pixels either white or black.



महाराष्ट्र राज्य माध्यमिक व उच्च माध्यमिक शिक्षण मंडळ, पुणे
Maharashtra State Board Of
Secondary and Higher Secondary Education, Pune

मुंबई विभागीय मंडळ / MUMBAI DIVISIONAL BOARD

उच्च माध्यमिक प्रमाणपत्र परीक्षा - गुणपत्रक
HIGHER SECONDARY CERTIFICATE EXAMINATION - STATEMENT OF MARKS

शाखा STREAM	आसन क्रमांक SEAT NO.	केंद्र क्रमांक CENTRE NO.	जिल्हा व उच्च माध्य. शाळा क्रमांक DIST. & HR. SEC. SCHOOL NO.	परीक्षेचा महिना व वर्ष MONTH & YEAR OF EXAM.	गुणपत्रिकेचा अनुक्रमांक SR. NO. OF STATEMENT
SCIENCE	M0[redacted]333	3627	32.03.035	FEBRUARY-18	[redacted]0007

उमेदवाराचे संपूर्ण नाव (आडनाव प्रथम) / CANDIDATE'S FULL NAME (SURNAME FIRST)
Vi[redacted]dia Nishil [redacted]esh

उमेदवाराच्या आईचे नाव / CANDIDATE'S MOTHER'S NAME Sangeeta

विषयाचा सांकेतिक क्रमांक व विषयाचे नाव Subject Code No. and Subject Name	• माध्यम Medium	कमाल गुण Max. Marks	प्राप्त गुण / Marks Obtained अंकात In Figures	 अक्षरात / In Words
01 ENGLISH	ENG	100	068	SIXTYEIGHT
40 MATHEMATICS & STATISTICS	ENG	100	056	FIFTYSIX
54 PHYSICS	ENG	100	050	FIFTY
55 CHEMISTRY	ENG	100	051	FIFTYONE
D9 COMPUTER SCIENCE	ENG	200	145	ONE HUNDRED AND FORTYFIVE
31 ENVIRONMENT EDUCATION	ENG	050	044	FORTYFOUR
30 HEALTH & PHYSICAL EDUCATION (GRADE)	A			
Result / निकाल PASS | Percentage / टक्केवारी 63.69	एकूण गुण / Total Marks	650	414	FOUR HUNDRED AND FOURTEEN

H184250007

321[redacted]

महत्वाचे, टीप, आरोग्य व शारीरिक शिक्षण विषयातील श्रेणी आणि विशेष चिन्हांची माहिती यांचा तपशील मागील पृष्ठावर पहावा.
See overleaf for important, Notes, Grades in Health & Physical Education Subject and meaning of special characters.

विभागीय सचिव / Divisional Secretary

Fig. 2. Displays marksheet with highlighted data

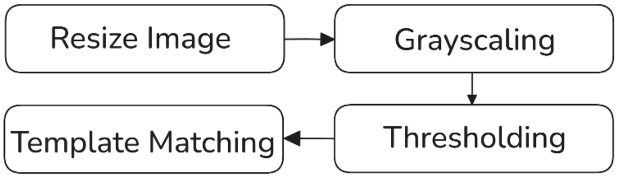


Fig. 3. Preprocessing Steps

Template matching has been performed to target the region of interest. It is a process for finding the location of a template image in a bigger image. OpenCV comes with a function cv.matchTemplate() [15] for this purpose. The template for matching is chosen in such a way that they have a lot of text in common with the bigger image to increase the accuracy of the matched template. The authors could have implemented a deep learning approach as demonstrated in [10] but they chose template matching instead. This decision was driven by the need to balance processing time with accuracy, as template matching provided a more efficient solution given their specific requirements.

Fig. 4 shows the output after the template matching process. Images after preprocessing and template matching still have noise which needs removal to improve the OCR performance. A custom autoencoder model using Python language has been trained for noise removal. Fig. 6 displays the clean image after it has been passed through the custom autoencoder noise removal model.

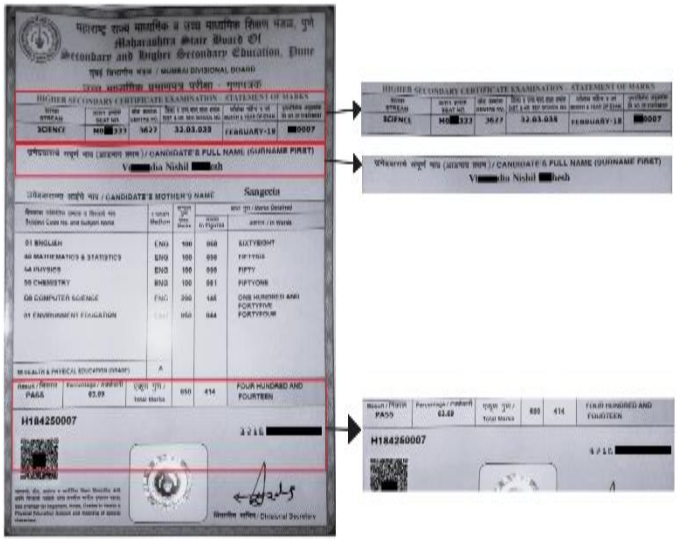

Maharashtra State Board Of
Secondary and Higher Secondary Education, Pune
MUMBAI DIVISIONAL BOARD
HIGHER SECONDARY CERTIFICATE EXAMINATION - STATEMENT OF MARKS
STREAM
SEAT NO.
CENTRE NO.
SCIENCE
M0[redacted]333
3627
33.03.038
FEBRUARY-18
[redacted]0007
CANDIDATE'S FULL NAME (SURNAME FIRST)
V[redacted]da Nishil [redacted]sh
CANDIDATE'S MOTHER'S NAME
Sangeeta
01 ENGLISH
ENG
100
068
SIXTYEIGHT
MATHEMATICS & STATISTICS
PHYSICS
CHEMISTRY
COMPUTER SCIENCE
ENVIRONMENT EDUCATION
FIFTY
FIFTYONE
ONE HUNDRED AND FORTYFIVE
FORTYFOUR
HEALTH & PHYSICAL EDUCATION
PASS
Total Marks
414
FOUR HUNDRED AND FOURTEEN
H184250007
[redacted]
Divisional Secretary

Fig. 4. Output of template matching process

The entire project has been configured on Google Colab with 12.7 GB RAM and 107.7 GB storage space. Version 3 of Python has been used for running the code. The tensorflow framework has been utilized to train the custom autoencoder model.

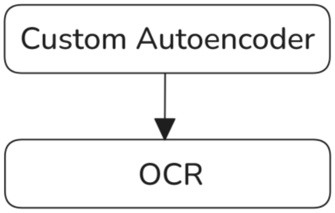


Fig. 5. Processing steps

Fig. 5 Shows the processing steps involved:

1) Use of an custom autoencoder model to remove noise

2) The output from the custom autoencoder model is provided to OCR engine to convert image to text.

The custom autoencoder model architecture consists of five dense layers starting with an input layer with 84846 neurons. Followed by three hidden layers with neuron counts 500, 300, and 100 respectively utilizing Rectified Linear Unit (ReLU) activation. The decoder reverses the encoder's structure to restore the image to its initial dimensions. The output layer matches the input layer's neuron count and uses the sigmoid activation function. The model is trained using mean squared error loss and Adam optimizer. For training the custom autoencoder model more marksheet images were collected from the internet in the public domain. A total of 95 images were used to train the model and 5 images were used for testing the model. The model has been trained for 100 epochs with a batch size of 5. This resulted in mean squared error loss to be low with a value of 2.

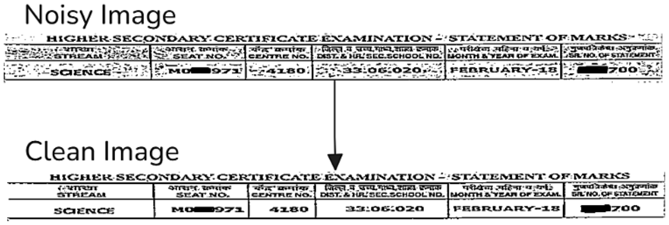

Noisy Image
HIGHER SECONDARY CERTIFICATE EXAMINATION – STATEMENT OF MARKS
STREAM
SEAT NO.
CENTRE NO.
DIST. & HR./SEC.SCHOOL NO.
MONTH & YEAR OF EXAM.
SR. NO. OF STATEMENT
SCIENCE
M0[redacted]971
4180
33:06:020
FEBRUARY-18
[redacted]700
Clean Image
HIGHER SECONDARY CERTIFICATE EXAMINATION – STATEMENT OF MARKS
STREAM
SEAT NO.
CENTRE NO.
DIST. & HR./SEC.SCHOOL NO.
MONTH & YEAR OF EXAM.
SR. NO. OF STATEMENT
SCIENCE
M0[redacted]971
4180
33:06:020
FEBRUARY-18
[redacted]700

Fig. 6. Output of custom autoencoder noise removal model

The authors experimented with three OCR models, namely Tesseract OCR [16], Keras OCR [17], and Easy OCR [18] and observations regarding their text prediction capabilities have been made. All three OCR models have different techniques of working. Pytesseract uses tesseract version 4 with the LSTM network for extracting printed text. Tesseract supports multiple languages with a focus on line detection. EasyOCR applies the Character Region Awareness for Text Detection (CRAFT) technique for text detection and a CRNN model with ResNet, Visual Geometry Group model (VGG), Long Short-Term Memory (LSTM), and Connectionist Temporal Classification (CTC) for recognition, and handling diverse text layouts. KerasOCR also uses the CRAFT technique for detecting text areas and employs bounding box detection through binary maps. For recognition, it utilizes Convolutional Recurrent Neural Network (CRNN) or spatial transformer networks for accurate text interpretation.

Fig. 7 shows the algorithm for OCR-based marksheet data extraction. The extracted text from the OCR has to be post-processed to get meaningful information like full name, seat number, percentage, board, year of passing, etc.

Algorithm 1 OCR-Based Marksheet Data Extraction

```
Begin
  Upload marksheet
  Apply preprocessing steps:
    Resize the image
    Convert the image to grayscale
    Apply thresholding
  Perform template matching to locate Region of Interest
  Pass matched image to Autoencoder model for noise removal
  Initialize OCR engine
  Pass the processed image to OCR model for text extraction
  Do postprocessing on the output text from OCR:
    Clean unwanted characters
  Save the processed image and output text:
    Processed image in a folder
    OCR output text in a CSV file
End
```

Fig. 7. Algorithm 1 – OCR-Based Marksheet Data Extraction

To find the percentage in predicted text, the authors filtered text by keeping words whose length is 5 or less because the percentage in the marksheet has been represented as DD.DD format where D is a digit between 0 to 9. After applying the word length filter we further narrowed down our search by screening those words that match the pattern i.e. two digits followed by two digits and a dot in between. If no match was found then we find the first four-digit number in predicted text and return that.

For the seat number, the authors looked into predicted text starting from M or N followed by 6 digits or 7-digit numbers which if matched it was returned to the webpage. The seat number had a general pattern of M followed by 6 digits in the twelfth marksheet. But sometimes M was predicted as N by OCR.

To fetch the academic branches (streams) from predicted text, a matrix has been created with columns labeled science, arts, and commerce and with predicted words arranged in rows. The cells of the matrix contained the corresponding edit distance values. The column name with the smallest edit distance values amongst the matrix has been considered as the expected stream.

Similarly, to get the year of passing from the twelfth marksheet, the authors have constructed a matrix with columns representing the months in a year and rows containing predicted words. After that, the edit distance of each cell in the matrix has been computed. The column with the lowest edit distance is identified as the year of passing.

IV. Results and Discussion

The results with corresponding analysis is as follows:

1) All the OCR models faced difficulty predicting characters like '8', 'h', and dot. Instead of the given characters, they predicted '0', 'b' and '' respectively.

2) There were a few instances where the OCR algorithms failed to accurately predict any letter in the text. The text had considerable noise and blur, but it was still readable by humans.

3) The edit distance between the predicted and actual words was used to compute the error percentage. The lowest error percentage was 41% for EasyOCR, followed by Pytesseract OCR and Keras OCR with 52% and 66%, respectively.

4) EasyOCR was found to be the most effective in handling noisy data, followed by Tesseract OCR and KerasOCR. Easy OCR has been used for implementing the solution due to its noise handling and accurate prediction of words.

5) In terms of speed, Pytesseract OCR was the fastest, followed by EasyOCR, with Keras OCR taking the longest. All models were executed on a CPU.

6) EasyOCR was the most effective in predicting characters and words, followed by Keras OCR and Pytesseract.

Based on the results we can say for applications that require high prediction accuracy, EasyOCR is a suitable choice, while Pytesseract OCR is better for applications that consider speed.

If we consider different languages for data extraction then Pytesseract OCR supports more than 100 languages making it the best option. In second place we have EasyOCR with 80 plus supported languages. Followed by Keras OCR which primarily works on English text but can be trained on different languages.

Future research could focus on proposing standardized formats for digital marksheets across the country which facilitates more efficient OCR processing. This could streamline the data management process and improve the consistency of information retrieval.

The study could also explore the use of advanced binarization techniques [19] to better handle issues like non-uniform illumination and low contrast in the marksheet images, aiming to further enhance OCR accuracy.

Usage of LLMs to analyze and interpret the text extracted from marksheets which could provide insights of student performance. Further, it can be used for generating context-rich reports. Integration of you only look once (YOLO) model to identify and locate key areas of marksheets, such as name, seat number, year of passing, and percentage. YOLOv7-B model consists of a Bi-directional feature pyramid network lowering the small objects' miss rate [20]. This would reduce preprocessing time and improve the recognition accuracy of different marksheet layouts. We should build a custom EasyOCR model by training on education credential image data to improve accuracy.

V. Conclusion

After conducting experimentation of extracting key data points using pre-trained OCR models, authors found EasyOCR outperforms Pytesseract OCR and Keras OCR. Given the results, EasyOCR can be used for custom training on marksheet image data for giving better results. The proposed solution uses less resources, scalable and economically acceptable.

An effective OCR system has been discovered that consists of the following steps: input image to system, preprocessing, running the processed image through a custom autoencoder model to remove noise, then passing the clean image to EasyOCR to convert it to text, and finally postprocessing to obtain the desired text.

References

[1] S. Galan, “Countries with the largest population 2024,” Statista, Feb. 10, 2025. [Online]. Available: https://www.statista.com/statistics/262879/countries-with-the-largest-population. [Accessed Feb. 17, 2025].

[2] India, Ministry of Power, Bureau of Energy Efficiency. Energy Efficiency Guidelines and Best Practices in Indian Datacenters. [Online]. Available: https://beeindia.gov.in/sites/default/files/datacenterbook.pdf [Accessed Feb. 17, 2025].

[3] D. Mytton, “Data centre water consumption,” npj Clean Water, vol. 4, no. 1, Feb. 2021. doi: https://doi.org/10.1038/s41545-021-00101-w.

[4] India, Ministry of Education, Department of Higher Education, New Delhi. All India Survey on Higher Education 2019-20. [Online]. Available:https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/aishe_eng.pdf. [Accessed Feb. 17, 2025].

[5] “Over 22 crore govt job applications in 2014-2022, 7.22 lakh recommended for appointment,” India Today, Jul. 28, 2022. [Online]. Available: https://www.indiatoday.in/education-today/news/story/over-22-crore-govt-job-applications-in-2014-2022-7-22-lakh-recommended-for-appointment-1981102-2022-07-28. [Accessed Feb. 17, 2025].

[6] A. Mir Asif, S. A. Hannan, Dr. Y. Perwej, and A. Mane, “An Overview and Applications of Optical Character Recognition,” International Journal of Advance Research in Science and Engineering, vol. 3, no. 7, Jul. 2014.

[7] S. Janvalkar, P. Manjrekar, S. Pawar, and Prof. L. Naik, “Text Recognition from an Image,” Journal of Engineering Research and Applications, vol. 4, no. 4, pp. 149–151, 2014.

[8] N. Alamsyah, M. N. Fauzan, A. G. Putrada, and S. F. Pane, “Autoencoder Image Denoising to Increase Optical Character Recognition Performance in Text Conversion,” in Proc. of the 2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS), pp. 1–6. doi: 10.1109/ICACNIS57039. 2022.10054885.

[9] H. Idrose, N. AlDahoul, H. A. Karim, R. Shahid, and M. K. Mishra, “An Evaluation of Various Pretrained Optical Character Recognition Models for Complex License Plates,” in Proc. of the Multimedia University Engineering Conference (MECON 2022), Atlantis Press, 2022, pp. 21–27. doi: 10.2991/9789464630824_4.

[10] P. Batra, N. Phalnikar, D. Kurmi, J. Tembhurne, P. Sahare, and T. Diwan, “OCR-MRD: performance analysis of different optical character recognition engines for medical report digitization,” International Journal of Information Technology, vol. 16, no. 1, pp. 447–455, Nov. 2023, doi: 10.1007/s41870-023-01610-2.

[11] R. Nikam, R. Pardeshi, Y. Patel, and E. Sarda, “Deep Learning based Automatic Extraction of Student Performance from Gazette Assessment Data,” ITM Web Conf., vol. 40, 2021.doi: 10.1051/itmconf/ 20214003022.

[12] R. Muke, S. Patil, J. Acharya, and S. Shiravale, “Marksheet Image Processing,” Current Trends in Technology and Science, vol. 3, no. 3, 2014.

[13] S. Kakade, T. Patle, R. Dable, Y. Nathe, P. Pranjali Bahalkar, “Marksheet Analysis using OCR,” Journal of Emerging Technologies and Innovative Research, vol. 11, 2024.

[14] N. Vinothkumar, S. Swathi, T. Pradeepika and K. Sarankumar, "Marksheet Data Collector," International Journal of Advance Research and Innovative Ideas in Education, vol. 10, no. 2, pp. 1265-1271, 2024.

[15] “OpenCV: Template Matching,” Opencv.org, 2022. [Online]. Available: https://docs.opencv.org/4.x/d4/dc6/tutorial_py_template_matching.html [Accessed Feb. 18 2025].

[16] M. Lee, “pytesseract: Python-tesseract is a python wrapper for Google’s Tesseract-OCR,” PyPI, Aug. 17, 2022. [Online]. Available: https://pypi.org/project/pytesseract/. [Accessed Feb. 18 2025].

[17] “keras-ocr — keras_ocr documentation,” Readthedocs.io, 2021. [Online]. Available: https://keras-ocr.readthedocs.io/en/latest. [Accessed Feb. 18, 2025].

[18] “Jaided AI: EasyOCR install,” Jaided.ai, 2025. [Online]. Available: https://www.jaided.ai/easyocr/install/. [Accessed Feb. 18, 2025].

[19] U. Rani, A. Kaur, and G. Josan, “A new binarization method for degraded document images,” International Journal of Information Technology, vol. 15, no. 2, pp. 1035–1053, 2023, doi: 10.1007/ s41870019003613.

[20] M. Yu and Y. Jia, “Improved YOLOv7 Small Object Detection Algorithm for Seaside Aerial Images,” Communications in computer and information science, pp. 483–491, Jan. 2024, doi: 10.1007/978-981-99-9109-9_46.