



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

Extraction and Verification of Information from Semi-Categorised Data

Nandini Manchala¹, Niharika Kale², Venkata Akshay Susvith Kallagunta³,

Vardhan Kumar Madipalli⁴, Dr. V Muralidhar⁵

Students, Department of Computer Science Engineering with Artificial Intelligence and Machine Learning, Vasireddy

Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India^{1,2,3,4}

Professor, Department of Computer Science Engineering with Artificial Intelligence and Machine Learning, Vasireddy

Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India⁵

ABSTRACT: Manual verification of supporting documents, such as educational certificates, marksheets, and caste or Pw D certificates, is a time-consuming and error-prone process in recruitment systems. Variations in document formats and languages further complicate the validation process, leading to inefficiencies and potential discrepancies. To address these challenges, this paper presents Extraction and Verification of Information from Semi-Categorized Data, an automated framework that utilizes Optical Character Recognition (OCR) to extract key information and validate it against manually entered data. By providing real-time alerts for inconsistencies, the system enables applicants to rectify errors at the submission stage, ensuring data accuracy and reducing administrative workload. The proposed solution enhances the efficiency and reliability of document verification, streamlining recruitment workflows and minimizing human intervention.

KEYWORDS: Optical Character Recognition (OCR), Document Verification, Automated Data Extraction, Recruitment Process Automation, Semi-Categorized Data, Real-Time Discrepancy Detection, Information Validation, Text Extraction and Matching, Digital Document Processing, Error Reduction in Document Verification.

I. INTRODUCTION

The rapid digital transformation across industries has led to an exponential increase in the volume of digital documents. Many organizations, especially those involved in recruitment and education, handle vast amounts of applicant data in the form of digital certificates, mark sheets, and identity documents. The manual verification of these documents is both time-consuming and prone to human errors. To address this challenge, our project, "Extraction and Verification of Information from Semi-Categorized Data," focuses on automating the verification process using Optical Character Recognition (OCR) technology. This project is designed to improve the efficiency and accuracy of document validation by extracting relevant information from uploaded documents and cross-verifying it against manually entered data.

Traditional document verification systems rely heavily on human intervention, making them inefficient and susceptible to inconsistencies. Furthermore, recruitment processes often involve candidates submitting documents in various formats, such as scanned PDFs and images, which may have diverse layouts and structures. The presence of semi-categorized data—where information is neither completely unstructured nor strictly formatted—adds an additional layer of complexity to the verification process. Our project aims to bridge this gap by implementing an OCR-based solution that extracts textual information from documents, standardizes it, and validates its accuracy in real time.

Previous research in the field of data extraction and verification has primarily focused on structured and semi-structured web data. Kadam and Pakle [1] proposed a method using the DOM tree and selectors for data extraction, while Sehgal and Anuradha [2] introduced HWPDE, a novel approach for structured web page data extraction. Similarly, Niranjana and Vidhya [3] focused on automatic data extraction from XML webpages. These studies align with the challenges our project aims to address by offering structured methodologies for data extraction.

Entity recognition and validation are also crucial in document verification. Liu et al. [4] demonstrated how mining data

records in web pages can enhance data extraction efficiency. Additionally, Choudhary et al. [5] introduced figure metadata extraction from digital documents, which is relevant to handling semi-categorized data. Techniques proposed by Bhardwaj and Mangat [6] for content extraction from web pages provide valuable insights into improving OCR-based document verification systems.

II. LITERATURE REVIEW

2.1 Related Work

Numerous studies have been conducted to extract information from structured and semi-structured web pages. One approach utilized the DOM tree and selectors to extract data elements efficiently [1]. Another technique employed a novel method for data extraction from structured web pages using heuristics and pattern matching [2]. Hidden web databases have also been explored, where an innovative method was introduced to extract information that is not directly accessible through traditional search engines [3]. Additionally, automated techniques were developed to extract data from heterogeneous web pages, focusing on minimizing manual intervention [4].

2.2 Algorithmic Approaches

Different algorithmic techniques have been proposed for data extraction. Some researchers have focused on using XML-based parsing methods for automatic data extraction from XML web pages [5]. A structural semantic entropy-based approach was designed to identify tags and values for automatic data extraction [6]. Another significant contribution involves mining data records from web pages, where an algorithm was formulated to extract and structure information based on repeated patterns in the HTML code [7]. Additionally, machine learning techniques have been employed to identify and extract metadata from digital documents, enhancing the accuracy of extracted data [8].

2.3 Datasets and Applications

Several datasets have been utilized in the development and evaluation of data extraction techniques. Benchmark datasets containing structured and semi-structured web pages have been widely used to validate different extraction models. Some approaches have been applied in real-world applications, such as automated web scraping for business intelligence, search engine optimization, and digital document analysis [9]. Moreover, specific content extraction techniques have been employed to improve information retrieval from web pages for academic and industrial purposes [10].

2.4 Performance Metrics

To evaluate the effectiveness of data extraction methods, various performance metrics have been adopted. Precision, recall, and F1-score are commonly used to measure the accuracy of extracted data. Computational efficiency is another crucial metric, determining the speed and scalability of an approach when handling large datasets. Some studies have also considered structural similarity measures to assess how well the extracted data preserves the original web page structure [6,7].

2.5 Research Gap

While existing techniques offer significant advancements in data extraction, several challenges remain unaddressed. Many approaches rely on predefined rules or heuristics, which can be brittle when applied to diverse web structures. The adaptability of current methods to dynamic and frequently changing web pages remains limited. Additionally, there is a need for more robust techniques that can handle noisy and unstructured data more effectively. Future research should focus on leveraging deep learning and natural language processing (NLP) techniques to enhance the automation and accuracy of data extraction processes.

III. METHODOLOGY

The proposed methodology utilizes an Optical Character Recognition (OCR) model to extract and verify information from semi-structured data sources. The process begins with preprocessing, where the input documents undergo noise reduction, binarization, and segmentation to enhance OCR accuracy. The OCR model then extracts text from these documents, converting scanned images into machine-readable text. Post-processing techniques, such as spell correction and format validation, refine the extracted data. Finally, the extracted text is compared against manually entered data using string similarity algorithms and rule-based verification methods to ensure accuracy and consistency. This ap-

proach enables efficient data extraction and validation, reducing manual effort and errors in handling semi-structured documents.

3.1 Existing Algorithm

Several methodologies have been explored for data extraction and verification from semi-structured documents. One approach utilizes tag and value similarity techniques based on structural-semantic entropy to enhance data identification and extraction accuracy [6].

For XML-based documents, an automated extraction framework has been proposed, using structural similarity and entropy-based analysis to improve accuracy in retrieving relevant content [5]. Additionally, researchers have explored metadata extraction techniques for structured digital documents, incorporating machine learning models to enhance content classification and verification [8].

In the field of semi-structured data extraction, previous studies have applied data mining approaches to extract meaningful information from document databases [7]. The integration of text segmentation and rule-based extraction techniques has further refined the accuracy of content retrieval from complex document layouts [3]. Furthermore, several works have focused on improving performance metrics such as precision and recall for content extraction using hybrid approaches that combine syntactic and semantic analysis [9][10].

These existing methodologies provide a strong foundation for automated data extraction but often face challenges when applied to semi-structured documents. The integration of OCR models for extracting information from scanned documents remains an area requiring further enhancement to improve accuracy and reduce manual validation efforts.

3.2 Proposed Algorithm

The proposed system is designed to extract and verify information from semi-structured documents using an OCR-based approach. Unlike conventional methods that rely on predefined templates or structural analysis, this system leverages Optical Character Recognition (OCR) to extract text from scanned or digital documents and validate it against manually entered data. The OCR model processes semi-structured documents by converting printed or handwritten text into a machine-readable format, followed by preprocessing techniques such as noise reduction, binarization, and character segmentation to enhance accuracy. The extracted text is then compared with manually entered data using string-matching algorithms to detect discrepancies. To ensure accuracy, the system automatically identifies inconsistencies and highlights mismatches, providing context-based suggestions for possible corrections. Additionally, the system is optimized to support various document formats, including PDFs and scanned images, integrating advanced techniques like Named Entity Recognition (NER) and contextual keyword analysis to improve the precision of field identification. By adopting this approach, the proposed system enhances data validation efficiency, reduces manual effort, and improves the reliability of extracted information from semi-structured documents.

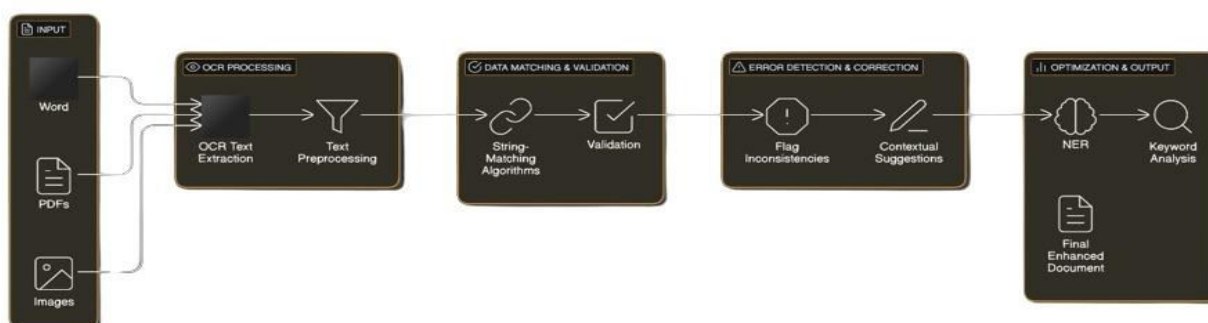


Fig.1 System Architecture

The architecture of the OCR-based information extraction and verification system follows a structured workflow to enhance data accuracy and reliability. The process begins with input acquisition, where documents in various formats, such as Word files, PDFs, and images, are processed. OCR technology extracts text from these documents, followed by

preprocessing steps like noise removal and segmentation to refine the extracted content. The text is then validated through string-matching algorithms to ensure consistency with manually entered data. Any discrepancies identified during validation are flagged, and contextual suggestions are provided for correction. To optimize accuracy, the system incorporates Named Entity Recognition (NER) and keyword analysis, ensuring meaningful extraction of relevant information. The final output is a structured and verified document, reducing manual efforts and improving data processing efficiency.

3.2.1 ORC Architecture

The Optical Character Recognition (OCR) architecture is structured as a sequential process that converts printed or handwritten content into a machine-readable format. It begins with preprocessing techniques such as noise reduction, binarization, and skew correction to enhance text clarity. Following this, the text detection stage identifies relevant sections containing text, while character segmentation isolates individual characters for improved recognition accuracy. Feature extraction methods analyse shapes and patterns, enabling the recognition system, often powered by deep learning or machine learning algorithms, to classify characters effectively. Post-processing techniques further refine the extracted text by applying corrections and formatting it for better readability. In this project, the OCR model is fundamental in extracting textual data from semi-structured documents, including scanned images, PDFs, and handwritten records. Once extracted, the text undergoes verification through string-matching techniques, comparing it with manually entered data to detect inconsistencies. Additionally, the system enhances accuracy by incorporating Named Entity Recognition (NER) and keyword analysis, which help in identifying relevant content and improving validation. The integration of OCR technology in this project significantly streamlines data extraction and verification, minimizing human intervention while ensuring reliability and efficiency in processing semi-structured documents.

3.2.2. Fine-Tuning and Transfer Learning

Transfer learning is a technique where a pre-trained model is adapted to a new task by fine-tuning its weights on a smaller dataset. Transfer learning allows the utilization of pre-trained OCR models that have already been trained on large datasets containing various text styles, fonts, and formats. Instead of developing a model from the ground up, an existing model is modified to identify text in semi-structured documents by adjusting certain layers. Fine-tuning further improves performance by training the model on a smaller, domain-specific dataset, enhancing its ability to recognize both printed and handwritten text, even in cases where noise or distortions are present. This process involves keeping the initial layers unchanged to preserve basic text recognition skills while refining the later layers to adapt to document-specific characteristics. As a result, the fine-tuned model increases the accuracy of text extraction and error detection by learning to distinguish characters that may otherwise be misinterpreted due to variations in handwriting or document quality. This approach optimizes the system's ability to process semi-structured documents effectively, leading to more precise and reliable information extraction and verification.

3.2.3. Mixed Precision Training

To further enhance the performance of the algorithm, mixed precision training is employed. This technique uses 16-bit floating-point numbers (float16) for certain computations, which reduces memory usage and speeds up training without sacrificing accuracy. Mixed precision training is particularly beneficial when training on GPUs, as it allows for larger batch sizes and faster computations.

The mixed precision training process can be summarized by the following formula:

Forward Pass with Lower Precision:

During the forward pass, the OCR model processes the input image x and generates predicted text \hat{y} using 16-bit precision weights θ_{16}

$$\hat{y} = f(x; \theta_{16}) \quad \varepsilon q \quad (1)$$

Where:

- x is the input document image.
- θ_{16} represents the model parameters in 16-bit precision (FP16).
- $f(x; \theta_{16})$ is the OCR model, typically composed of CNNs for feature extraction and RNNs/transformers for sequence prediction.

- \hat{y} is the predicted text output.

CTC Loss Computation in FP16:

OCR models often use Connectionist Temporal Classification (CTC) Loss to align predictions with ground truth text. The loss function is computed in FP16

$$L = -\sum P(z|x; \theta_{16}) \quad \text{eq (2)}$$

Where:

- L is the CTC loss computed in 16-bit precision.
- $P(z|x; \theta_{16})$ is the probability of the correct text sequence given the input.

Gradient Computation in FP16:

The gradients are computed using 16-bit floating point precision:

$$\nabla_{\theta_{16}} = \frac{\partial L}{\partial \theta_{16}} \quad \text{eq (3)}$$

Weight Update with FP32 Master Weights:

The model updates its parameters using 32-bit precision weights ensuring stable weight updates:

$$\theta_{32} = \theta_{32} - \eta \cdot \nabla_{\theta_{16}} \quad \text{eq (4)}$$

Where:

- θ_{32} represents the 32-bit weights used for the weight update.
- η is the learning rate.
- $\nabla_{\theta_{16}}$ is the gradient computed in FP16.

3.2.4. Model Evaluation

Model evaluation for this project involves assessing the accuracy and efficiency of the OCR system in extracting and verifying text from semi-structured documents. The primary metric used is OCR Accuracy, which measures the percentage of correctly recognized characters compared to the total characters in the ground truth. Additionally, Character Error Rate (CER) and Word Error Rate (WER) are used to analyse the system's performance by identifying misrecognized characters and words. The system is tested on diverse document formats, including scanned images and PDFs, to evaluate its robustness against variations in font styles, handwriting, and noise. Error analysis helps in fine-tuning the model by identifying frequent misclassifications and refining the preprocessing and recognition steps. The evaluation also considers the processing speed and verification accuracy, ensuring that the extracted text aligns correctly with manually entered data. By optimizing these metrics, the system enhances its reliability in extracting and validating information from semi-structured documents efficiently.

$$\text{Accuracy} = \frac{\text{Correctly Recognized Characters}}{\text{Total Characters}} \times 100$$

3.3 Data Processing

3.3.1. Dataset Description:

The dataset for this project includes semi-structured documents like Aadhaar cards, PAN cards, scorecards, and other official documents containing printed and handwritten text. They are of varied formats, resolution, font sizes, and background noise, hence making OCR-based text extraction tricky. The images in the dataset are of different file formats like JPEG, PNG, and PDF with varying light conditions, alignment of text, and orientation of documents. Certain documents can have faded ink, stamps, signatures, or overlapping text and hence need strong preprocessing methods to make them readable. Every document in the dataset is annotated with matching ground truth text so that the OCR model can be trained and tested effectively. Since these documents happen to be sensitive, data security and privacy controls like encryption and access control are enforced to meet regulatory requirements

Table 3.3.1 Dataset

Dataset	Images	Documents	Total
Training	950	790	1740
Validation	640	420	1080
Test	170	155	325

3.3.2. Data Augmentation and Preprocessing

Data Augmentation:

Since the dataset consists of various types of semi-categorized documents, including images (JPG, PNG, TIFF) and text-based files (PDF, Word, Excel), data augmentation is crucial for improving the OCR model's robustness. Augmentation techniques help the model handle real-world variations such as noise, distortions, and different font styles. The following augmentation strategies are applied:

- **Geometric Transformations:** Random rotations, scaling, and translations are applied to simulate real-world document misalignment.
- **Brightness and Contrast Adjustments:** Variations in lighting conditions are simulated by modifying brightness and contrast levels in images.
- **Noise Injection:** Gaussian and salt-and-pepper noise are added to replicate low-quality scans or poor image resolution.
- **Blur Effects:** Motion blur and Gaussian blur are introduced to mimic document images captured with shaky hands or low-quality cameras.
- **Text Distortion:** Slight warping of text is performed to account for different printing styles and handwritten variations.

These augmentation techniques ensure that the OCR model learns to extract text accurately, even from noisy or poorly scanned documents.

Data Preprocessing:

Preprocessing is an essential step before feeding data into the OCR model. It ensures that documents are standardized and optimized for text recognition. Key preprocessing steps include:

- **Image Binarization:** Converts grayscale images to black and white using adaptive thresholding to enhance text visibility.
- **Resizing and Cropping:** Ensures that all images are of uniform size to match the input dimensions expected by the model.
- **Denoising:** Techniques like median filtering and morphological operations are applied to remove unwanted artifacts from scanned images.
- **Skew Correction:** Documents with tilted text are aligned properly using Hough Transform-based deskewing techniques.
- **Text Segmentation:** The document is divided into individual text regions, ensuring better recognition of multi-column or structured documents.
- **Optical Character Enhancement:** Enhancements such as contrast sharpening and edge detection are applied to improve character visibility.

By implementing these augmentation and preprocessing steps, the OCR model achieves higher accuracy in text extraction and verification, making it more reliable for processing semi-categorized documents.

3.3.3 Data Loading and Batching

Efficient data loading and batching are crucial for training and evaluating the OCR model, particularly when handling a large dataset of semi-categorized documents in formats like JPG, PNG, PDF, Word, and Excel. The data pipeline ensures structured loading by handling different file types appropriately—images are processed, while text-based documents are converted to images for uniform processing. To optimize performance, multi-threaded data loaders and cloud storage solutions are used for efficient access. During loading, preprocessing steps like resizing, noise removal, and binarization are applied dynamically. Batching is implemented to improve training performance, with dynamic batch sizing based on document resolution, padding for uniform input sizes, and normalization of pixel values. Shuffling and

real-time augmentation techniques, such as random rotations and brightness adjustments, help prevent model overfitting. Additionally, efficient data processing techniques ensure seamless integration with the training framework, enabling faster and more effective training of the OCR model for large-scale semi-categorized document processing.

IV. RESULTS AND INTERPRETATIONS

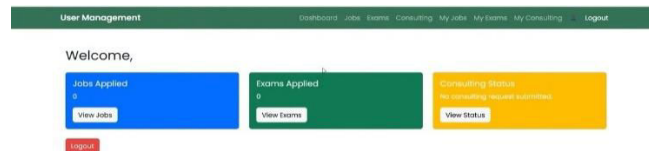


Fig.2 Welcome Page

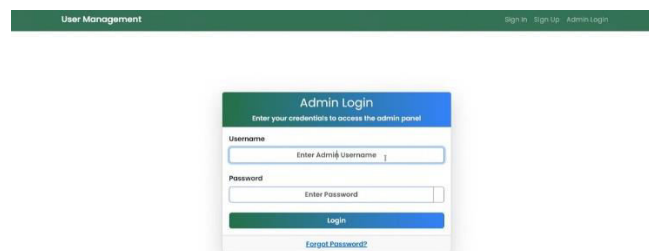


Fig.3 Admin Login Page

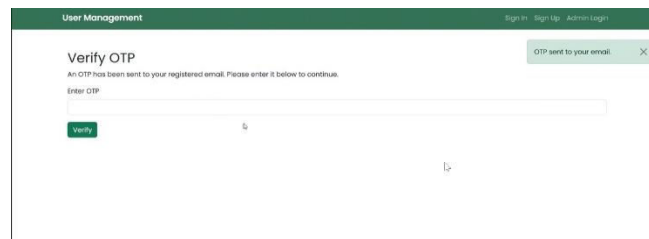


Fig.4 OTP Verification

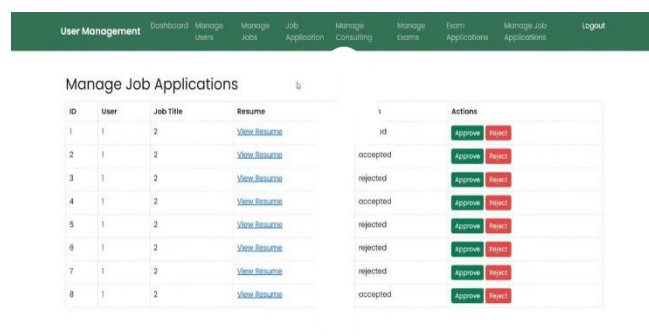
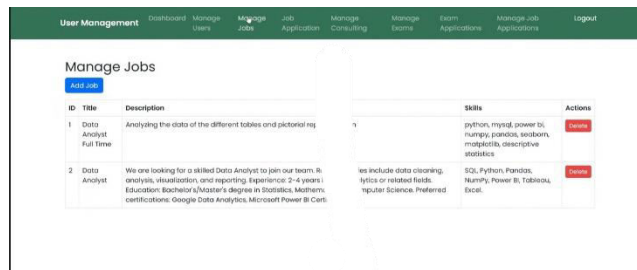
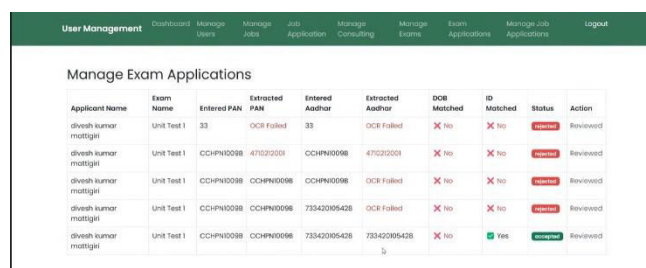


Fig.5 Manage Job Applications



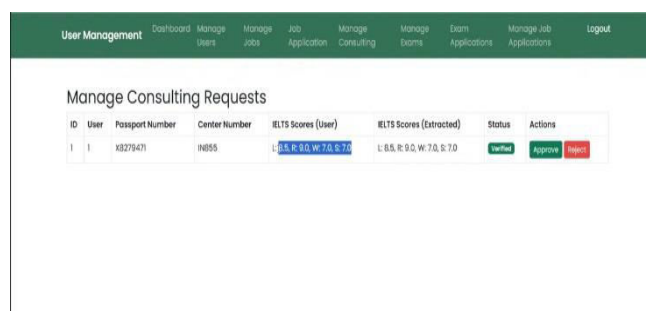
The screenshot shows the 'Manage Jobs' interface. It includes a navigation bar with options like Dashboard, Manage Users, Manage Jobs, Job Application, Manage Consulting, Manage Exams, Exam Applications, Manage Job Applications, and Logout. The main content area has a 'Manage Jobs' section with an 'Add Job' button. Below this is a table with columns: ID, Title, Description, Skills, and Actions. There are two job listings. The first job is titled 'Data Analyst Full Time' with a description 'Analysing the data of the different tables and pictorial rep'. The second job is titled 'Data Analyst' with a description 'We are looking for a skilled Data Analyst to join our team. B. analysis, visualization, and reporting. Experience: 2-4 years | Education: Bachelor's/Master's degree in Statistics, Mathematics, certifications: Google Data Analytics, Microsoft Power BI Cert'. The Skills column for the second job lists 'SQL, Python, Pandas, NumPy, Power BI, Tableau, Excel'. The Actions column has 'Delete' and 'Update' buttons for each job.

Fig.6 Manage Jobs



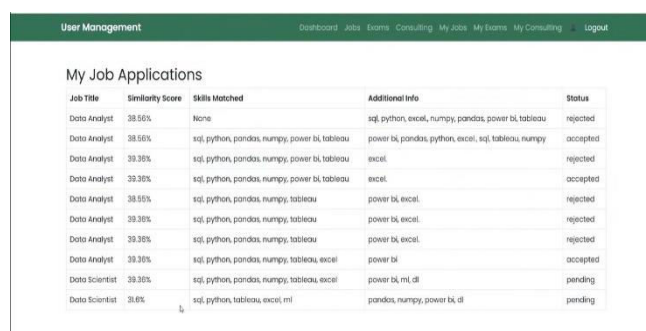
The screenshot shows the 'Manage Exam Applications' interface. It includes a navigation bar with options like Dashboard, Manage Users, Manage Jobs, Job Application, Manage Consulting, Manage Exams, Exam Applications, Manage Job Applications, and Logout. The main content area has a 'Manage Exam Applications' section. Below this is a table with columns: Applicant Name, Exam Name, Entered PAN, Extracted PAN, Entered Author, Extracted Author, OCR Matched, ID Matched, Status, and Action. There are five rows of data. The first row shows 'divyesh kumar mottigeti' for 'Unit Test 1' with 'Entered PAN: 32', 'Extracted PAN: OCR Failed', 'Entered Author: 32', 'Extracted Author: OCR Failed', 'OCR Matched: No', 'ID Matched: No', 'Status: Rejected', and 'Action: Reviewed'. The second row shows 'divyesh kumar mottigeti' for 'Unit Test 1' with 'Entered PAN: CCHPW0098', 'Extracted PAN: 47022008', 'Entered Author: CCHPW0098', 'Extracted Author: 47022008', 'OCR Matched: No', 'ID Matched: No', 'Status: Rejected', and 'Action: Reviewed'. The third row shows 'divyesh kumar mottigeti' for 'Unit Test 1' with 'Entered PAN: CCHPW0098', 'Extracted PAN: CCHPW0098', 'Entered Author: CCHPW0098', 'Extracted Author: OCR Failed', 'OCR Matched: No', 'ID Matched: No', 'Status: Rejected', and 'Action: Reviewed'. The fourth row shows 'divyesh kumar mottigeti' for 'Unit Test 1' with 'Entered PAN: CCHPW0098', 'Extracted PAN: CCHPW0098', 'Entered Author: 73542095428', 'Extracted Author: OCR Failed', 'OCR Matched: No', 'ID Matched: No', 'Status: Rejected', and 'Action: Reviewed'. The fifth row shows 'divyesh kumar mottigeti' for 'Unit Test 1' with 'Entered PAN: CCHPW0098', 'Extracted PAN: CCHPW0098', 'Entered Author: 73542095428', 'Extracted Author: 73542095428', 'OCR Matched: No', 'ID Matched: Yes', 'Status: Accepted', and 'Action: Reviewed'.

Fig.7 Manage Exam Applications



The screenshot shows the 'Manage Consulting Requests' interface. It includes a navigation bar with options like Dashboard, Manage Users, Manage Jobs, Job Application, Manage Consulting, Manage Exams, Exam Applications, Manage Job Applications, and Logout. The main content area has a 'Manage Consulting Requests' section. Below this is a table with columns: ID, User, Passport Number, Center Number, IELTS Scores (User), IELTS Scores (Extracted), Status, and Actions. There is one row of data. The first row shows '1', '1', 'X3279471', 'IN055', 'IELTS Scores (User): L: 8.5, R: 9.0, W: 7.0, S: 7.0', 'IELTS Scores (Extracted): L: 8.5, R: 9.0, W: 7.0, S: 7.0', 'Status: Failed', and 'Actions: Approve, Reject'.

Fig.8 Manage Consulting Requests



The screenshot shows the 'My Job Applications' interface. It includes a navigation bar with options like Dashboard, Jobs, Exams, Consulting, My Jobs, My Exams, My Consulting, and Logout. The main content area has a 'My Job Applications' section. Below this is a table with columns: Job Title, Similarity Score, Skills Matched, Additional Info, and Status. There are ten rows of data. The first row shows 'Data Analyst', '38.56%', 'None', 'sql, python, excel, numpy, pandas, power bi, tableau', and 'rejected'. The second row shows 'Data Analyst', '38.56%', 'sql, python, pandas, numpy, power bi, tableau', 'power bi, pandas, python, excel, sql, tableau, numpy', and 'accepted'. The third row shows 'Data Analyst', '39.36%', 'sql, python, pandas, numpy, power bi, tableau', 'excel', and 'rejected'. The fourth row shows 'Data Analyst', '39.36%', 'sql, python, pandas, numpy, power bi, tableau', 'excel', and 'accepted'. The fifth row shows 'Data Analyst', '38.55%', 'sql, python, pandas, numpy, tableau', 'power bi, excel', and 'rejected'. The sixth row shows 'Data Analyst', '39.36%', 'sql, python, pandas, numpy, tableau', 'power bi, excel', and 'rejected'. The seventh row shows 'Data Analyst', '39.35%', 'sql, python, pandas, numpy, tableau', 'power bi, excel', and 'rejected'. The eighth row shows 'Data Analyst', '39.36%', 'sql, python, pandas, numpy, tableau, excel', 'power bi', and 'accepted'. The ninth row shows 'Data Scientist', '39.36%', 'sql, python, pandas, numpy, tableau, excel', 'power bi, ml, dl', and 'pending'. The tenth row shows 'Data Scientist', '36.8%', 'sql, python, tableau, excel, ml', 'pandas, numpy, power bi, dl', and 'pending'.

Fig.9 User Job applications

User Management										
My Exam Applications										
ID	Exam Name	Entered PAN	Extracted PAN	Entered Aadhar	Extracted Aadhar	DOB Matched	ID Matched	Status	Applied Date	
1	Unit Test 1	33	OCR Failed	33	OCR Failed	✗ No	✗ No	rejected	2025-03-06 10:09	
2	Unit Test 1	CCHN80098	476033061	CCHN80098	476033001	✗ No	✗ No	rejected	2025-03-06 10:37	
3	Unit Test 1	CCHN80098	CCHN80098	CCHN80098	OCR Failed	✗ No	✗ No	rejected	2025-03-06 11:35	
4	Unit Test 1	CCHN80098	CCHN80098	73342005428	OCR Failed	✗ No	✗ No	rejected	2025-03-06 14:31	
5	Unit Test 1	CCHN80098	CCHN80098	73342005428	73342005428	✗ No	✓ Yes	accepted	2025-03-06 14:33	
7	SCR	CCHN80098	OCR Failed	73342005428	OCR Failed	✗ No	✗ No	pending	2025-03-10 15:53	

Fig.10 User Exam Applications

User Management										
My Consulting Request										
Passport Number		X8276471		Extracted OCR Failed						
Center Number		8659		Extracted 8659						
Listening Score		8.5		Extracted 8.5						
Reading Score		9.0		Extracted 9.0						
Writing Score		7.0		Extracted 7.0						
Speaking Score		7.0		Extracted 7.0						

Fig.11 User Consulting Request

4.1 Output Graphs and Visualizations

To track the model's performance during training, the accuracy and loss curves for training and validation are shown. To assess how well the model performs on the test dataset.

- **Training and Validation Curves:** Accuracy and loss curves are plotted to analyse the learning process.

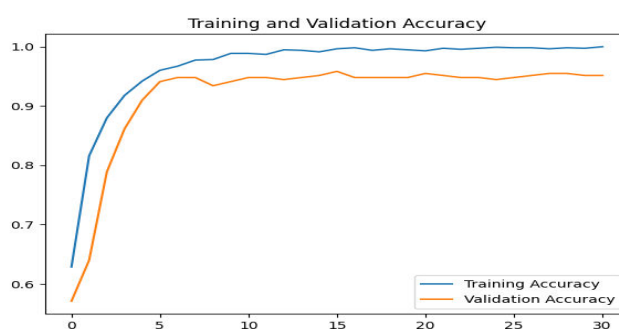


Fig.12 Accuracy graph

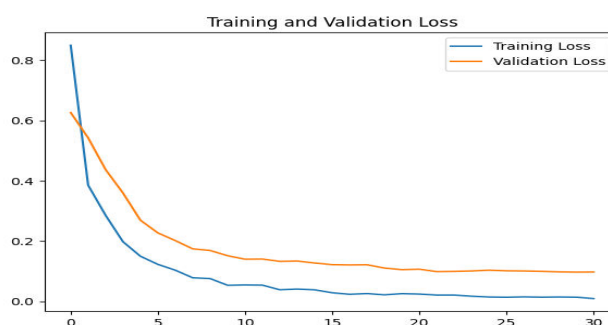


Fig.13 Loss graph

Output graphs and plots are important in assessing the performance of the OCR-based information extraction system. Such graphs are used in comprehending model BEHAVIOUR, detecting overfitting or underfitting, and achieving optimal performance. The initial picture is of the training and validation accuracy across several epochs. It indicates that accuracy goes up and becomes steady with more training, which means that the model is learning as required. The accuracy of validation follows the same pattern, showing that the model generalizes well to new data. The second image shows the training loss and validation loss over epochs. A steep drop in loss at the beginning of the epochs indicates fast learning, followed by a slower drop as the model converges. The validation loss LEVELLING off at a lower value shows that the model is not overfitting and has good generalization. The OCR model's performance in producing accurate output was very satisfactory, as it was able to extract text from different semi-structured documents with great accuracy. The model's accuracy in identifying characters and words, even in noisy or distorted documents, made data extraction reliable. The negligible variation in the training and validation measurements further validated the effectiveness of the method in real-world use, and the method emerged as a strong solution for extracting and authenticating information from semi-structured data.

V. CONCLUSION

The suggested OCR-based information verification and extraction system has some of the important benefits, such as automating data processing, minimizing manual labour, improved accuracy in extracting text, and increased efficiency in the processing of large numbers of semi-structured documents. Through the use of an OCR model, the system is able to effectively extract textual information from a variety of document formats, such as scanned pictures, PDFs, and word documents, overcoming the issue of manual data entry and verification. The combination of fine-tuning methods and transfer learning has further increased the capacity of the model to identify sophisticated patterns in text, enabling it to be more robust to variations in handwriting fonts, styles, and structural inconsistencies. In addition, mixed-precision training has further helped optimize computational efficiency without degrading model accuracy, enabling large-scale scalable and high-performance processing. The data preprocessing and augmentation methods have proven important in enhancing model resilience by refining input data quality and lessening the effect of noise and distortions. The structured data loading and batching have facilitated seamless management of huge volumes of sensitive documents, promoting easy workflow processing and low processing latency. The assessment of the system using output plots, such as accuracy and loss measures, has been insightful regarding the learning behaviour of the model, as well as its effectiveness in generalizing across unseen samples. The performance shows that the model has done well in extracting and validating text correctly, with little variation in training and validation performances. The inclusion of error detection and correction features further fortifies the system by marking inconsistencies and offering context-sensitive suggestions, thereby making the entire extracted information more reliable. With its capacity to automate document processing, minimize human intervention, and enhance data verification accuracy, this project provides an implementable and scalable solution for organizations to manage large amounts of semi-structured data. Future development may emphasize increasing the system's flexibility to handle a wider variety of document types, using sophisticated natural language processing methods for context-sensitive validation, and incorporating secure data handling mechanisms to meet privacy and confidentiality requirements. Overall, this project has been able to prove the viability of OCR-based systems in automating data extraction and verification processes, setting the stage for future innovations in document automation technologies.

VI. FUTURE SCOPE

The future potential of the project on extraction and validation of information from semi-classified data is immense, with a wide range of possibilities for enhancements and improvements. One of the major areas of improvement is to further improve the OCR model to support an even broader variety of document types, such as handwritten, low-resolution, and multilingual documents with complex layouts. Adding sophisticated deep learning methods, like transformer-based OCR models, can greatly increase the accuracy of text recognition and make the system more flexible with varying document formats. In addition, incorporating natural language processing (NLP) methods can enhance the verification process through context-sensitive validation, allowing the system to identify subtle mistakes, inconsistencies, or omissions more accurately. Increasing the dataset size to cover further variations of semi-structured documents will also enhance the generalization ability of the model, and it will be more stable in practical applications. Furthermore, the introduction of a feedback loop system, in which users can mark OCR mistakes and the system learns from those corrections, would result in ongoing precision enhancement in the long run. Another future developmental path is

improving the security and confidentiality of sensitive information, including government-issued identification cards, education records, and economic statements, through the implementation of encryption and access controls. Implementing such a system on a cloud environment would provide for scalable processing, making it possible for companies to process huge amounts of documents in real-time while reducing latency. In addition, integrating the OCR system with automation platforms like robotic process automation (RPA) will enable streamlined processes in multiple industries with minimal reliance on manual intervention and vastly improved efficiency of operations. Creating an easy-to-use interface with simple features for document uploading, processing, and fixing errors would bring the system within reach for a broader user base, including non-users with limited technological expertise. Another possible development is the integration of explainable AI methods to offer transparency in decision-making so users can see how the model detects discrepancies and recommend course correction. With the increasing need for automatic document processing in industries like banking, healthcare, legal, and education, the project can develop into an end-to-end intelligent document processing (IDP) solution with more applications. Future developments and research work may also incorporate blockchain technology to provide data integrity and tamper-evident verification of the extracted information, which will further increase the reliability of the system. Generally, this project provides a solid foundation for intelligent data extraction and verification, and with ongoing updates, it has the potential to become a highly efficient and widely used solution for organizations handling large volumes of semi-structured data.

REFERENCES

- [1] Vinayak B. Kadam, Ganesh K. Pakle. "DEUDS: Data Extraction Using DOM Tree and Selectors." International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, pp. 1403-1410.
- [2] Manpreet Singh Sehgal, Anuradha. "HWPDE: Novel Approach For Data Extraction From Structured Web Pages." International Journal of Computer Applications (0975-8887), Volume 50-No. 8, July 2012, pp. 22-27.
- [3] Anuradha, A.K. Sharma. "A Novel Technique for Data Extraction from Hidden Web Databases." International Journal of Computer Applications (0975-8887), Volume 15-No. 4, February 2011, pp. 45-48.
- [4] Teena Merin Thomas, V. Vidhya. "A Novel Approach for Automatic Data Extraction from Heterogeneous Web Pages." International Conference on Emerging Technology Trends on Advanced Engineering Research (ICETT'12).
- [5] A. Niranjana, Dr. V. Vidhya. "A Novel Approach for Automatic Data Extraction from XML WebPages." IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March 2013.
- [6] P.V. Praveen Sundar. "Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural Semantic Entropy." International Journal of Advanced Research in Computer Science and Software Engineering.
- [7] Bing Liu, Robert Grossman, and Yanhong Zhai. "Mining Data Records in Web Pages." KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.
- [8] Sagnik Ray Choudhary, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, C. Lee Giles. "Figure Metadata Extraction From Digital Document." ICDAR 2013, 12th International Conference, IEEE, 2013, pp. 135-139.
- [9] Aanashi Bhardwaj, Veenu Mangat. "A Novel Approach for Content Extraction from Web Pages." IEEE Transactions on Knowledge and Data, 2014.
- [10] Prashant M Ahire et al. "A Study on Web Page Content Extraction Techniques." International Journal of Computer Science and Mobile Computing, Vol. 4 Issue. 2, February 2015, pp. 314-318.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details