Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Intelligent Document Digitization and Classification using Hybrid Deep Learning and NLP Approaches

Fazliya S, Evangline R, Ms. Sumitha T
Department of Computer Science and Engineering College
R. M. K. Engineering College

Abstract: - Handwritten archives are time-consuming, inclined to errors, and prone to bodily harm when stored, retrieved, and categorized manually. An AI-powered device for the digitization and categorization of handwritten police and criminal archives is proposed in this study. After enhancing scanned archives with an picture preprocessing pipeline, the device makes use of optical personality focus (OCR) to retrieve text. A Natural Language Processing (NLP)-based classification model is then used to manner the digitized text, routinely classifying files into preset businesses in a way scientific reports, witness statements, and FIRs. In order to facilitate equipped storage and high-quality retrieval, metadata extraction is used to pick out essential entities consisting of names, dates, places, and kinds of crimes. Unstructured handwriting inputs are transformed into safe, searchable, and analyzable digital statistics thru the integration of digitization and classification modules. Data-driven police work is made feasible with the aid of experimental assessment, which indicates extended accuracy in record categorization and textual content recognition, substantially slicing down on retrieval time. This work gives a scalable, multilingual, and impervious answer for digital policing projects, assisting to modernize the administration of regulation enforcement records.

Keywords: Document Digitization, Handwriting Recognition, OCR, NLP, Document Classification.

INTRODUCTION:

Traditional file administration structures used in regulation enforcement and judicial establishments are dealing with top notch stress due to the speedy enlarge in criminal and crook cases. Despite managing a broad range of vital documents, many of these establishments nonetheless depend closely on handwritten records. Consequently, the routine tasks of accessing, verifying, and reviewing records often prove cumbersome and time-consuming. Artificial Intelligence presents an opportunity to overhaul these workflows by converting paper-based documentation into structured digital formats. However, digital transformation in law enforcement encompasses more than simple archiving. It enables the development of sophisticated systems that can interpret, organize, and retrieve information with remarkable efficiency.

When techniques from computer vision, machine learning, and natural language processing work together, previously manual and tedious processes can be automated. This creates opportunities for predictive analytics, improved transparency, and enhanced operational effectiveness.

2.1 Problem Statement:

For essential files which include cost sheets, witness statements, court docket filings, and FIRs (First Information Reports), the felony and regulation enforcement structures nevertheless basically count on handwritten documentation. There are important operational limitations prompted by way of this traditional method. Investigations and judicial complaints are appreciably slowed down by way of the widespread quantity of bodily labor required to locate, store, and arrange these papers.

Another main fear is human mistake. Inaccuracies delivered by using flawed dealing with or transcription of handwritten files may additionally jeopardize the consequences of a lawsuit. These problems are made worse via the bodily fragility of paper archives; information decay with time and can be broken through fire, water, or ordinary put on and tear.

An extra project is the inconsistent formatting of these publications. Without standardized templates, it turns into needlessly complicated and time-consuming to search for precise facts or cross-reference cases. Delays in decisions, ineffective case management, and a greater danger of dropping necessary proof are the consequences of these interrelated problems.

The most vast difficulty is likely that interdepartmental cooperation is hampered by means of the absence of a centralized digital repository. When working throughout departments or jurisdictions, officers and criminal experts locate it hard to attain the records they require fast. The reply is obvious: present day digitization strategies that can scan, classify, and extract statistics from handwritten archives are integral for the prison and regulation enforcement systems. Automated options have the plausible to substantially extend operational effectivity and produce safe, searchable databases that expedite the prison device and hold essential documents.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

2.2 Scope of the Project

The aim of this challenge is to increase a wise system that can convert handwritten police and crook archives into digital formats. The challenge is divided into three fundamental components: first, creating preprocessing methods to decorate the excellent scanned images; second, incorporating AI fashions into workouts that can become aware of linguistic patterns and come to be conscious of textual content; and third, increasing classification buildings to crew archives into organizations applicable to legislation enforcement operations. Furthermore, the desktop acknowledges and retrieves the names, dates, locations, and criminal classifications that investigators and crook gurus want to manipulate proof and music cases.

Security is important, and the gadget has sturdy get right of entry to controls to make sure that solely licensed people can get entry to non-public information, stopping it from being shared besides consent. The answer is made to be bendy in the future. It can deal with a large vary of file codecs and languages, and its shape lets in for future changes in response to organizational desires or technological advancements. This adaptability ensures that the system will proceed to be advised as lengthy as people's needs change.

3. OVERVIEW OF EXISTING RESEARCH:

Existing lookup on record classification and digitization covers a huge vary of methods, datasets, and software areas. For digitizing handwritten regional language archives [1] and printed and handwritten texts in Tamil [2], CNN-based OCR structures have been developed. These structures have elevated accuracy and adaptability, however they are regularly confined by means of language scope, dataset dependence, and the absence of real-time or multilingual capabilities.

While CNN-CTC architectures have been used to apprehend handwritten textual content with modest accuracy with suggestions for extra complicated preprocessing and decoding [4], hybrid OCR-classification fashions using Naive Bayes have been used to digitize and classify archives [3].

A variety of domain-specific classification strategies have been developed, such as TF-IDF with PAM and Naive Bayes for organizing huge datasets [7], critiques of supervised and unsupervised classification techniques [6], and Naive Bayes-based summarization and classification of lookup publications [5].

Naive Bayes classifiers [8] have been used to categorize IT lookup papers, and bibliometric critiques [9] have examined gaining knowledge of algorithms for handwritten report recognition, pointing up CNN dominance and the necessity of benchmark datasets.

While comparative algorithm research have discovered KNN to be the most high quality classifier throughout a range of datasets [11], NLP-based felony record categorization the usage of KNN, TF-IDF, and Count Vectorizer has been investigated [10]. Despite constraints such as handwriting variability and language limits, font-independent accuracy has been the intention of handwritten text-line focus efforts using IWR and segmentation [12]. Lastly, for scalable, high-accuracy performance, deep studying and CNN-driven structures have been used to digitize clinical prescriptions and different handwritten documents. These structures combine on-line interfaces with cloud deployment [13].

4. PROPOSED WORK:

The advised assignment's intention is to create and enforce an AI-powered gadget that can routinely scan and classify handwritten police and jail records. By combining modern-day photo processing techniques, Natural Language Processing (NLP), and Optical Character Recognition (OCR), the machine creates a clean workflow from report acquisition to storage and retrieval. Handwritten archives, such as witness statements, rate sheets, and FIRs, are scanned and preprocessed the use of methods along with binarization, skew correction, noise reduction, and big distinction augmentation to make certain brilliant OCR accuracy.

The framework makes use of TrOCR (Transformer-based OCR) for handwritten English text, which makes use of a Text Transformer decoder to flip the image patches into specific textual sequences and a Vision Transformer encoder to generate the image patches. A CRNN (Convolutional Recurrent Neural Network with BiLSTM and CTC loss) is used to understand challenging Tamil letters and extract sequential facts from images in handwritten Tamil text. Following textual content extraction, a post-processing stage makes use of language fashions and dictionary-based algorithms to in addition expand accuracy by means of fixing universal OCR problems. An NLP-driven categorization module that employs a BERT + SVM hybrid mannequin receives the digitized textual content after that.

In this case, the SVM classifier classifies the file into types like FIRs, cost sheets, witness statements, court docket orders, or judgment copies, whilst BERT transforms the textual content into contextual embeddings. It is via the implementation of metadata extraction that makes use of a BERT-based Named Entity Recognition (NER) model, which identifies key portions of information-likes names, dates, and places of crime scenes as nicely as the classes of crimes-that most suitable indexing and curated information storage end up possible.

Lastly, a safe, searchable repository homes the labeled and digitalized documents, facilitating speedy access, fantastic analysis, and much less guide labor. In addition to facilitating multilingual support, more suitable analytics, and interplay with policing dashboards, this digitization technique safeguards documents from bodily harm. All matters considered, the framework transforms unstructured handwritten statistics into arranged, trustworthy, and beneficial digital assets, extensively growing the effectiveness of regulation enforcement organizations.

5. METHODOLOGY:

5.1 System Overview:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The advised framework affords an AI-powered answer for the digitization and classification of reports, which is meant to correctly manage handwritten police and jail documents.

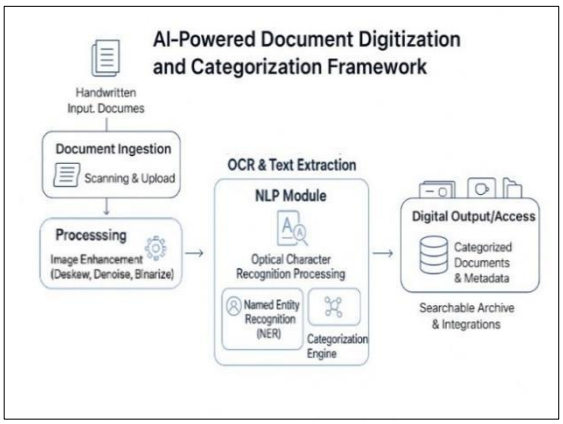

FIG 1. ARCHITECTURE DIAGRAM

By integrating file ingestion, photo preprocessing, OCR-based textual content content material extraction, Natural Language Processing (NLP), and metadata-driven categorization into a easy workflow, the gadget converts uncooked handwritten inputs into searchable, categorized, and structured digital archives.

According to the structure diagram, the device workflow is made up of the following foremost parts:

1. Document Ingestion: Scanning and importing strategies are used to acquire handwritten enter documents. This stage ensures that paper archives are changed into digital picture archives so they can be processed further.

2. Processing Module: Preprocessing techniques which consists of deskewing, denoising, and binarization are used to beautify file readability and accumulate inputs for OCR. Immediate photograph enchancment influences core of interest accuracy, which is crucial for enhancing orientation, reducing noise, and growing the visibility of textual content.

3. Text Mining and OCR Module: Optical Character Recognition (OCR) is a vital NLP technological know-how that converts long-form photographs into machine-readable text.Next, a state-of-the-art NLP strategy recognised as Named Entity Recognition (NER) is used to extract imperative devices such names, dates, case numbers, and jail references. Files are grouped in accordance to predetermined requirements (e.g., a FIR, witness testimony, price sheets, or scientific studies) the use of a categorization engine.

4. Digital Output & Access: A structured, searchable series of documents with metadata and labels is the ultimate product. Authorized customers can easily combine with crook databases and case administration structures by way of retrieving archives thru each metadata-based and free-text searches.

5.2 Digitization Module:

The Digitization Module is responsible for converting handwritten police and prison documents—whether physical or scanned—into structured digital formats.

This ensures that even unstructured, noisy, or partially damaged records are accurately processed and made compatible with downstream analytics.

This module consists of four major stages:

1. Document Acquisition

2. Image Processing

3. Handwriting Recognition (OCR)

4. Structured Document Creation

It integrates deep learning-based OCR models (TrOCR and CRNN), OpenCV for image enhancement, and Python-based tools for generating structured XML or JSON outputs.

5.2.1 Document Acquisition:

Using cell units or laptop computer computer systems with AI-assisted auto-cropping and de-skewing capabilities, the acquisition stage focuses on acquiring bodily documents, such as witness statements, fee sheets, and FIRs. The bought photograph I(x, y) is normalized the use of the following brightness and illumination correction settings:

$$I'(x,y) = \alpha \cdot I(x,y) + \beta$$

The following changes are made to every difficulty in the photo in order to whole standpoint correction the usage of homography transformation:

$$P' = H \cdot P$$

5.2.2 Image Processing:

Preprocessing methods like segmentation to extract textual content material cloth cloth cloth regions, Otsu's approach for binarization, Gaussian or median filtering for noise reduction, CLAHE for distinction enhancement, and Hough Line Transform for skew correction are some of the strategies used to enhance the pictures after they are taken. The most threshold (T*) is observed by way of decreasing intra-class variance:

$$T^* = \arg\min_T [\omega_1(T)\sigma_1^2(T) + \omega_2(T)\sigma_2^2(T)]$$

Text boundaries are sophisticated the usage of morphological strategies such as erosion and dilation, which are mathematically described as:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

$$I_{processed} = (I \ominus B) \oplus B$$

5.2.3 Handwriting Recognition:

This stage transforms the preprocessed handwritten text into machine-readable content using advanced OCR techniques.

(a) English Text – TrOCR Model

The TrOCR model approaches picture patches the use of a Vision Transformer (ViT) encoder and generates the output textual content material sequence the usage of a Transformer decoder. The probability distribution of every estimated token can be expressed as follows:

$$P(y_t|y_{<t}, X) = \text{softmax}(W_o h_t)$$

(b) Tamil Text – CRNN Model

A Convolutional Recurrent Neural Network (CRNN) combines CNN layers for the extraction of spatial features, Bidirectional LSTM for studying sequences, and Connectionist Temporal Classification (CTC) for label alignment in order to system Tamil handwriting. The definition of the CTC loss feature is:

$$L_{CTC} = -\ln \sum_{\pi \in B^{-1}(y)} P(\pi|x)$$

5.2.4 Structured Document Creation:

Following OCR processing, unstructured text is converted into structured formats. Named Entity Recognition (NER) based on BERT is employed to identify entities like Name, Date, Location, and Crime Type. The prediction for each token is computed as:

$$y_i = \arg\max P(y_i|x_i; \theta_{BERT})$$

Lastly, the structured records is stored in a SQL/NoSQL database for rapid indexing, search, and analytical makes use of like crime vogue evaluation and case correlation..

5.3 Classification Module:

After digitization, the recognized and formatted textual content is categorised to mechanically crew files into predefined classes such as witness statements, cost sheets, FIRs. This categorization module is indispensable to take care of giant volumes of police and criminal files extra efficaciously and with much less guide effort.

Hugging Face Transformers (Legal-BERT) and spaCy and PostgreSQL incorporate the unified technological stack of the proposed device that streamlines the pipeline with high-quality accuracy and scalability. Data-driven evaluation, greater organization, and faster retrieval are made possible by the mechanical preprocessing, vectorization, categorization, and tagging of the digital textual content.

The scalability and flexibility of this machine are similarly superior by using its assist for multilingual categorization the use of multilingual transformer fashions to cater police documents in numerous Indian languages.

5.3.1 Text Preprocessing for Classification:

Text content material education is a vital step in getting ready digitized textual content material for environmentally pleasant classification. Fashions find out it tough to analyze the textual material taken from handwritten police and trial files on the grounds that it normally consists of misspellings, noise, and irregularities in structure. This system makes use of spaCy to deal with textual content material teaching operations, tokenization, lowercasing, stopword removal, lemmatization, and pretty a wide variety persona eradication.

SpaCy provides a strong pipeline that correctly handles domain-specific linguistic challenges. It is higher with specialised dictionaries to get higher misspellings and acronyms extraordinary to police and criminal vernacular. These strategies standardize and put collectively the textual fabric whilst getting rid of pointless statistics and holding semantic significance. Instead of the usage of extra traditional vectorization methods like TF-IDF or Word2Vec, the answer makes use of transformer embeddings from a pretrained Legal-BERT model, which provides complete contextual representations of the textual content besides the want for human characteristic engineering. This approach ensures that the classification model's enter is surprisingly informative and precisely represents know-how of crime areas.

5.3.2 AI Classification Model:

Digital police and felony files are classified the use of a Legal-BERT transformer mannequin that has been increased on a labeled dataset that consists of witness testimony, cost sheets, FIRs, and scientific reports. This method eliminates the want for a couple of exceptional laptop gaining knowledge of fashions (such CNNs, SVM, and Naïve Bayes) whilst reaching cutting-edge overall performance and machine simplification. The transformer-based method gives drastically greater accuracy in taking pictures the complicated linguistic shape and domain-specific semantics ordinary of prison archives when in contrast to usual techniques.

The mannequin confirms standard standard overall performance and ensures resilience for the coaching duration the usage of the preferred distinction metrics, which encompass accuracy, precision, recall, F1-score, and confusion matrices. Cross-validation strategies are employed to stop overfitting and make certain that the mannequin features properly when utilized to sparkling data. Using

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Hugging Face's APIs helps seamless mannequin coaching and inference with the resource of frameworks like TensorFlow or PyTorch, which simplify improvement and maintenance. After training, the gadget robotically assigns freshly received files to the suitable category, making sure environment friendly report management.

5.3.3 Metadata Extraction:

In order to arrange and facilitate the retrieval of digital police and criminal records, metadata extraction is essential. Key entities like private data (accused names, sufferer details), case-related attributes (FIR numbers, crime type, felony sections), temporal statistics (dates and times), and geographical statistics (crime scene places and police jurisdictions) are extracted by means of the gadget the usage of the identical spaCy pipeline better via Legal-BERT embeddings after classification.

Dependency parsing aids in the extraction of hyperlinks between entities, such as linking a precise accused character to a crime scene on a positive date, whilst Named Entity Recognition (NER) is used to exactly discover entities from unstructured text. Regular expressions are used for structured fields, such as telephone numbers or case IDs.After being extracted, the metadata is linked to the fabulous digital archives and saved in a PostgreSQL database. This approves for greater state-of-the-art querying features, inclusive of discovering all witness statements citing the equal accused or getting all FIRs filed in a positive 12 months and region. The system's adoption of a single science stack effects in a simplified and maintainable answer that minimizes integration complexity whilst guaranteeing scalability, accuracy, and ease of deployment.

5.4 Integration of Digitization and Classification:

The machine's closing area combines the classification and digitization modules to supply a absolutely computerized gadget for managing handwritten police and detention center documents. The content material of the textual content is without delay despatched to the classification pipeline when scanned archives are transformed into editable textual content the usage of OCR and handwriting recognition. This ensures that the digital material is without problems available and retrievable by means of making sure that it is now not solely organized however additionally classified and more advantageous with metadata. A Flask backend linkages the quite a number modules by using REST APIs, enabling easy verbal conversation between the OCR, preprocessing, and classification services.

Authorized customers can also add documents, habits searches the usage of free-text or metadata filters, and rapidly achieve labeled data the usage of the system's web-based frontend. Real time access, security, and scalability are assured by means of cloud deployment (AWS or GCP). The science enables future-ready purposes like predictive police and crime analytics, improves organization, and speeds up felony operations by using combining digitalization and categorization into a single, environment friendly stack.

6. RESULTS AND FINDINGS:

To investigate the counseled system's potential to digitize and categorize police records, an instance handwritten FIR record used to be used. To enhance image great and get it equipped for recognition, the enter file used to be scanned and preprocessed the use of deskewing, denoising, and binarization techniques. The handwritten FIR was once then transformed into machine-readable textual content by way of the OCR engine. Tokenization and normalization are examples of post-processing strategies that improved textual content readability and ensured higher enter for the classification stage.

After processing the recognized text, the NLP-based categorization module routinely positioned the report in the "Crime Reports" class. To facilitate prepared storage and retrieval, metadata was once additionally retrieved, inclusive of criticism details, location, date, and case type. A structured digital reproduction of the FIR, stored in a searchable archive, was once the cease result. According to experimental evaluation, the system's common accuracy in figuring out and classifying handwritten FIRs was once 83%. Illegible handwriting and overlapping entries in the scanned web page have been related with the majority of attention mistakes. These effects exhibit that the machine can efficaciously convert unstructured handwritten FIRs into digital files that are searchable and structured.

The developed software affords a whole workflow from file add to digital conversion. FIG 2 suggests the most important dashboard, whilst FIG 3 depicts the file add interface. After uploading, the machine tactics the record and generates the transformed output as proven in FIG 4 Users can view all processed archives through the View Files part in FIG 5. FIG 6 suggests the backend running, with FIG 7 representing the uploaded handwritten file and FIG 8 showing the corresponding transformed digital output, confirming correct record digitization and gadget functionality.

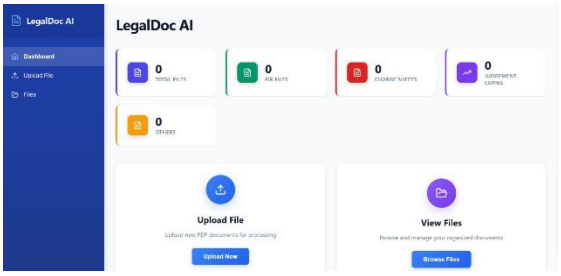

FIG 2. APPLICATION DASHBOARD

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

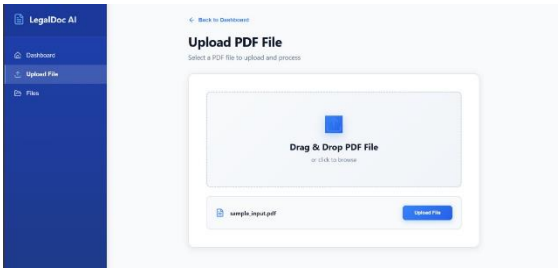

FIG 3. UPLOADING FILES

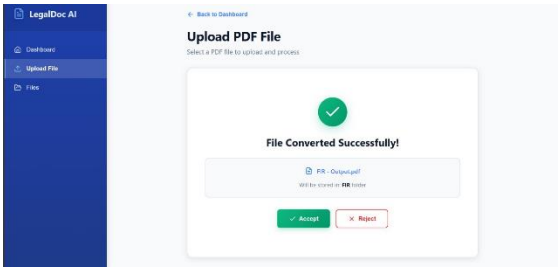

FIG 4. CONVERTED FILE

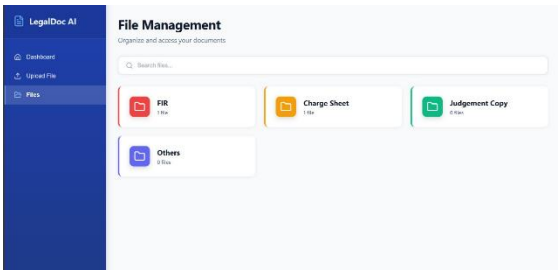

FIG 5. VIEW FILES

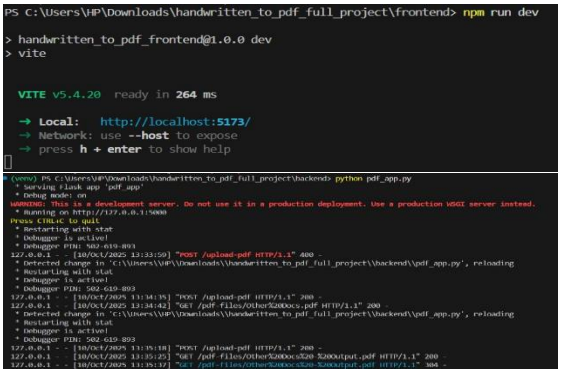

FIG 6. BACKEND RUNNING

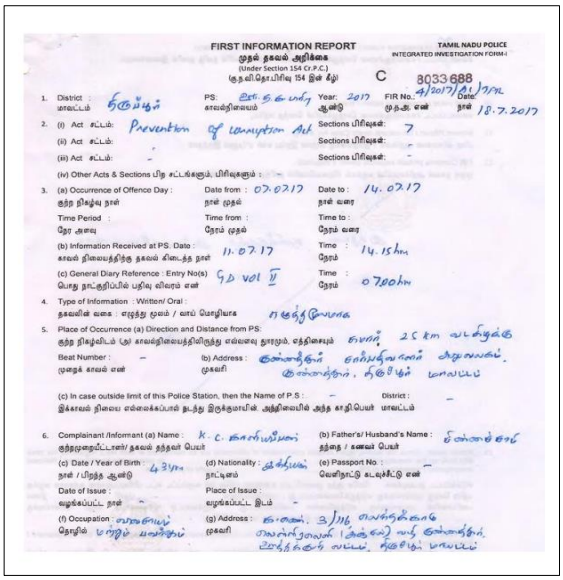

FIG 7. UPLOADED INPUT

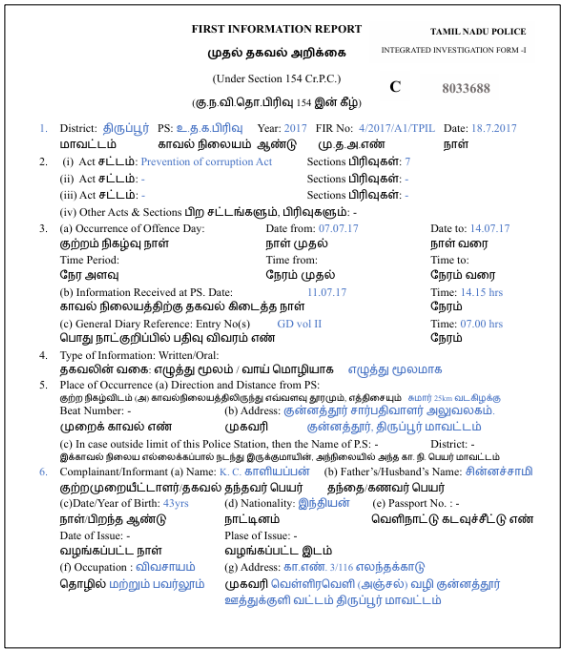

FIG 8. CONVERTED OUTPUT

7. CONCLUSION:

This find out about addressed the drawbacks of guide storage and retrieval by means of supplying an smart technique for scanning and classifying handwritten police and criminal documents. The method transformed unstructured FIRs into geared up and searchable digital documents via combining preprocessing, OCR-based handwriting recognition, and NLP-driven classification. Experiments on a pattern FIR report confirmed that the system's textual content focus and classification accuracy used to be 83%. Usability used to be similarly extended with the aid of the incorporation of metadata extraction, which made it viable to rapidly get fundamental records along with complainant names, dates, and case kinds.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The results show that the recommended strategy can radically extend the effectiveness of keeping police and court docket data. It affords a scalable and reliable approach of updating documentation workflows in judicial and regulation enforcement settings by means of reducing guide labor and minimizing errors. Future tendencies would possibly pay attention on growing the system's cognizance accuracy for extraordinarily difficult-to-read handwriting, including guide for different regional languages, and connecting it to real-time police databases for wider use.

8. REFERENCES:

[1] Ms. S. Shanthi, Kavishri B, Mahalakshmi R, Nivedhitha S, "AI-Based Ocr System For Digitizing Handwritten Historical Documents In Regional Languages",INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY, 2025.

[2] Dr. L. Rasikannan, S. Pratheek, N. Mohammed Riyas, R. Bharathkumar, "AI-POWERED OCR FOR DIGITIZING HANDWRITTEN HISTORICAL DOCUMENTS IN TAMIL", International Research Journal of Engineering and Technology (IRJET), 2025.

[3] Sumita Gupta, Aditya Gupta, Simran Khanna, Shivam Arora, "Digitization of Handwritten text using Deep Learning", IEEE, 2022.

[4] Omer AYDIN, "Classification of Documents Extracted from Images with Optical Character Recognition Methods", Anatolian Journal of Computer Sciences, 2021.

[5] Vinaya Kulkarni, Shruti Rothe, Rasika Deshpande, Shivani Devgirikar, Sneha Sultane, "DOCUMENT CLASSIFICATION AND SUMMARIZER USING PROBABILISTIC CLASSIFIER", Journal of Emerging Technologies and Innovative Research (JETIR), 2020.

[6] Madjid Khalilian, Shiva Hassanzadeh, "Document classification methods", Research Gate, 2019.

[7] Rajnish Virpate, Adityavikram Gurao, Ankit Naik, Mahebub Shaikh, Smita Chaudhari, "Document Classification using Machine Learning Techniques", International Journal Of Engineering Research And Development, 2020.

[8] Prajakta Pawale, Poonam Masal, Ankita Jadhav, Prof. Shivraj B Kone, "DOCUMENT CLASSIFICATION USING MACHINE LEARNING", International Journal For Multidisciplinary Research, 2019.

[9] Vanita Agrawal, Jayant Jagtap, M.V.V. Prasad Kantipudi, "Exploration of advancements in handwritten document recognition techniques", Intelligent Systems with Applications, 2024.

[10] Mr. Nikhil Wani, Ms. Gayatri Mangire, Mr. Aman Kumar, Ms. Nandini Solse, Mrs. P. S. Gaikwad, "Legal Document Classification using TF-IDF and KNN", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2022.

[11] Faizur Rashid, Abdulkadir H. Aden, Suleiman M. A. Gargaare, Afendi Abdi, "Machine Learning Algorithms for Document Classification: Comparative Analysis", International Journal of Advanced Computer Science and Applications (IJACSA), 2022.

[12] Swapnil R. Aghadte , Vansh R. Vyawahare , Talha Ahmad Khan, Pranav R. Zade , Tejas P. Kharkar, Prof P. N. Umekar, "Turning Handwritten Document into Digitized Version", International Journal Of Creative Research Thoughts (IJCRT), 2024.

[13] Koushik S, Mohana Aarthi K, Narmatha R, Niranjan R M, Priyanth V, Asst Prof. Pragadheesh Thirumal, "Turning Handwritten Documents into Digitized Form", International Journal of Research Publication and Reviews, 2024.