iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70968

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Legal Document Authentication and Verification System Leveraging OCR, NLP, and CNN for Comprehensive Document Authentication

Prof. Moushmee Kuri[1], Prasad Bokare[2], Sandipak Dhuri[3], Atharv Inamdar[4], Pawas Gavande[5], Aditya Gupta[6]
[1]Professor, [2, 3, 4, 5, 6]Student, School of Computing, MIT-ADT University, Pune, Maharashtra, India

Abstract: ***Legal document forgery is a huge threat to the integrity of institutions as well as of individuals. Based on Optical Character Recognition, NLP, and CNN, this paper designs a "Legal Document Authentication and Verification System" that targets the detection of forgery, especially at the textual, image-based, and structural layers of Indian driving license documents. It uses advanced machine learning techniques in signature, layout, and text anomaly analysis. Modular in design, with scalability features, findings have been established in governance, education, and banking sectors. Benchmarked evaluation proved promising accuracy in detecting forged documents.***
Keywords: ***Forgery Detection, OCR, NLP, CNN, Deep Learning, Image Processing, Legal Document Verification.***

I. INTRODUCTION

Digital tools have made it easier to forge legal documents, which is the reason why fraud pervades various sectors including governance, finance, and law enforcement. Verifying manuals mostly turns out to be prone to errors that consume so much time and don't address modern methods of forgery.

This project tackles all those issues with:

a) Text and structural inconsistencies analyzed with OCR and NLP techniques.
b) Tampered or manipulated sections identified through CNN-based image processing techniques.
c) The outputs are then consolidated into a single platform for holistic identification

Objectives and deliverables:

a) Development of an intelligent system to validate legal documents on the fly
b) Forgery identification via text anomaly detection, image manipulation, and layout anomalies.
c) Design scalable solution which can apply across domains.

Case study: Use Indian DLs: The most prominent application of the system could be for emerging markets in combating document forgery

II. LITERATURE REVIEW

The detection of document forgery is at the core of any research in document verification and security. As the process of forgery evolves, traditional methods usually become inefficient; therefore, powerful as well as automatic solutions are required. This section discusses the approaches concerning document forgery detection, so far used, focusing on the feature extraction, machine learning, and text analysis techniques.

1) *Image-Based Forgery Detection:* Detection of document forgery, particularly traces of forgery often rely on image analysis techniques. Feature extraction techniques such as Polar Cosine Transform (PCT) and Local Binary Patterns (LBP) have been specifically used for extracting tampered regions in images. For instance, Saber et al. (2021) demonstrated the ability of PCT and LBP features to isolate forged regions at an accuracy level of over 90%. SURF (Speeded-Up Robust Features) with template matching could achieve a detection accuracy of 97.5% for copy-move and splicing forgeries. The above techniques, to a large extent, study the anomalies in the structures and the visual appearances in the layouts of documents.
2) *Deep Learning for Forgery Detection:* Deep learning algorithms have become the backbones of modern forgery detection systems. One such powerful variant of CNNs is widely used in image forgery detection tasks with high recall rates, owing to its ability to learn complex patterns in images. One of the notable examples reports a 97.3% recall for the detection of forged administrative documents, which shows the potential for real-time use in automated document analysis. Moreover, DCT has incorporated with CNNs to enhance feature extraction by highlighting altered areas. Techniques like these improve the identification of changed areas especially in high-compression settings.

3) *Text-Based Analysis Using OCR and NLP:* This Optical Character Recognition technique enables the conversion of scanned documents into detailed text analysis through machine-readable text. Techniques of tokenization, lemmatization, and even semantic pattern detection are applied. Natural Language Processing models can sense the inconsistency in the language being used, indicate textual content that has been tampered with, and analyse formatting irregularities, including inconsistent fonts or spacements. For instance, Lavanyaa et al. in 2022 used NLP to demonstrate classification of legal documents with a focus on anomaly detection for text.

4) *Challenges Facing Detection of Forgery: With* all the above developments, document forgery detection is still associated with several challenges:
a) High-quality forgeries can even escape detection when it mimics the original patterns
b) Variability in types and formats of documents presents a challenge to models.
c) Compression artifacts and noise in scanned versions result in decreased accuracy of detection.
5) *Emerging Technologies:* Emergent tools, such as GANs, are being researched for the generation of synthetic datasets for training models. The datasets enable a system to learn about new ways of forging. Blockchain technology provides yet another potential solution to ensure tamper-proof storage of documents; however, it is still in its nascent stages of integration into forgery detection systems. Even though the methods have already shown great prospects, it is still at a very imperative stage of development to counter the changing forgery patterns. A combined approach by OCR, NLP, and CNN provides multi-layered architecture for enhancement of detection accuracy and reliability. This work extends these advancements to a scalable and efficient solution for legal document verification.

III. PROPOSED METHODOLOGY

To ensure that the development of our system meets user expectations and effectively addresses their needs, a systematic approach was adopted. The proposed system incorporates the following modules:

A. Pre-processing for Documents

Image Normalization: Resize and normalize document images for uniformity.
Noise Removal: Use filters to remove background noise and enhance clarity.

B. Forgery Detection Techniques

a) Text Extraction and Analysis (OCR and NLP): Extract the text using OCR libraries such as Tesseract.
Tokenization: Break the text into words for a granular analysis.
Semantic analysis: detects anomalies in context, such as mismatched names or addresses.
Uniformity in the font size, spacing, and alignment.
b) Image Layout and Signature Analysis (CNN):
Layout Detection: Detects the layout elements such as tables and logos by trained CNN models on document structure.
Compare this with the set layouts in order to identify any non-uniformity
Signature Verification: Extract signatures through image segmentation
Compare them with the stored reference signatures using the help of pre-trained CNN.

C. System Architecture

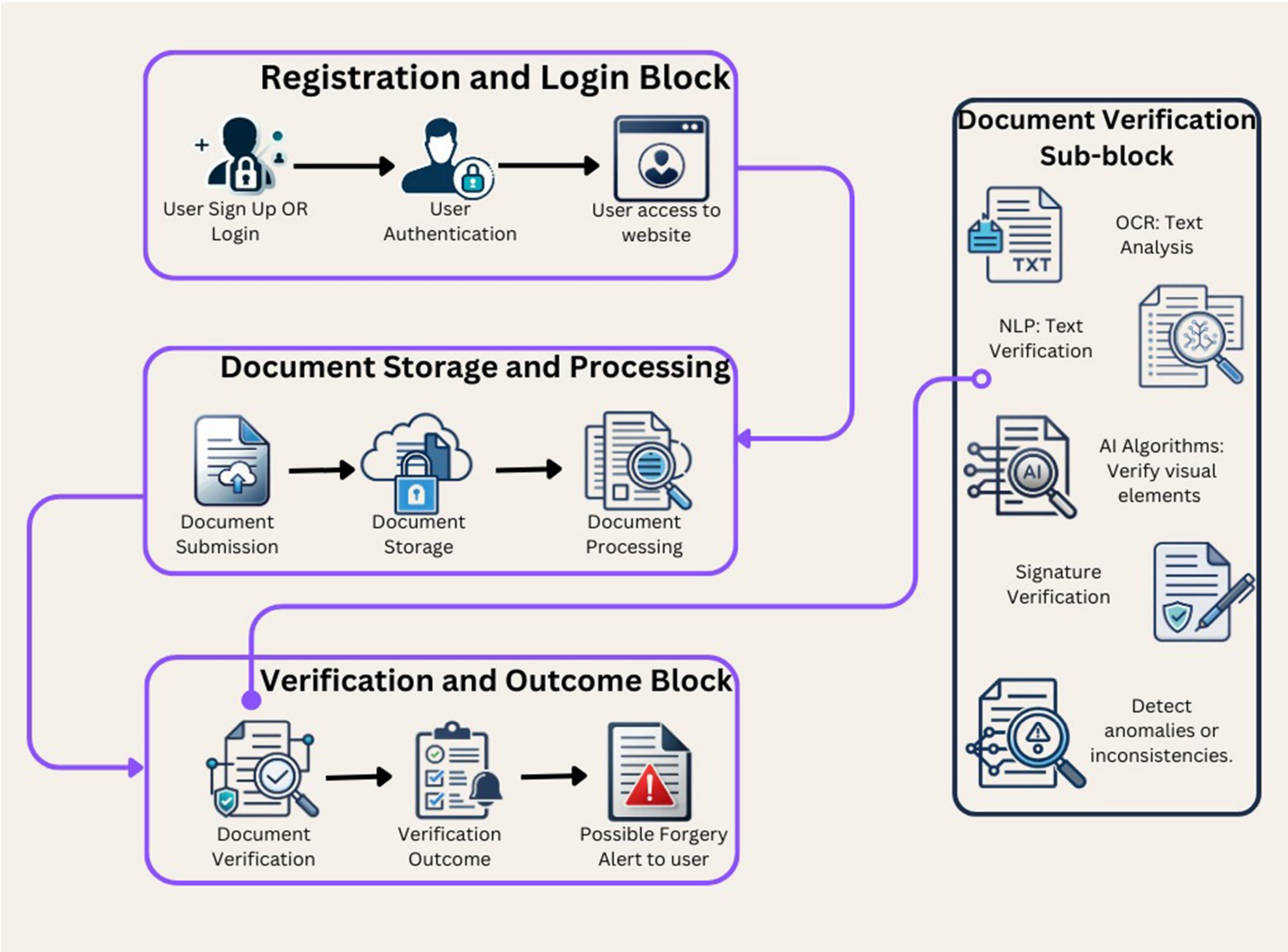
a) User Interface
Document Upload with Forging Results
Visualization of Tampered Regions

b) Back-End
Implementation in Python using TensorFlow for CNN models and spaCy for NLP.
Processed documents to be stored safely in the cloud

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

D. Merging the Multi-Modal Analysis

Combine the results obtained from the OCR, NLP, and CNN to produce a comprehensive forgery detection score.



Proposed Block Diagram of our Proposed Working Methodology

1) User Requirement Gathering

A survey was conducted to ascertain user requirements and preferences regarding our system.

1) Functional Requirements:
   a) Multi-format support: PDF, JPEG etc
   b) Forgery detection capability on both text, images, and layouts
   c) Realtime processing and results generation
2) Non-Functional Requirements;
   Security and confidentiality of sensitive data within the documents.
   Scalability to handle many documents concurrently.
3) Stakeholders' Input: Legal professionals, government officers, and IT experts were consulted. They considered that the verification of signatures and layout analysis identified forgery.

2) Data Analysis

a) Pie Graph: Distribution of Forged against Genuine Documents in India

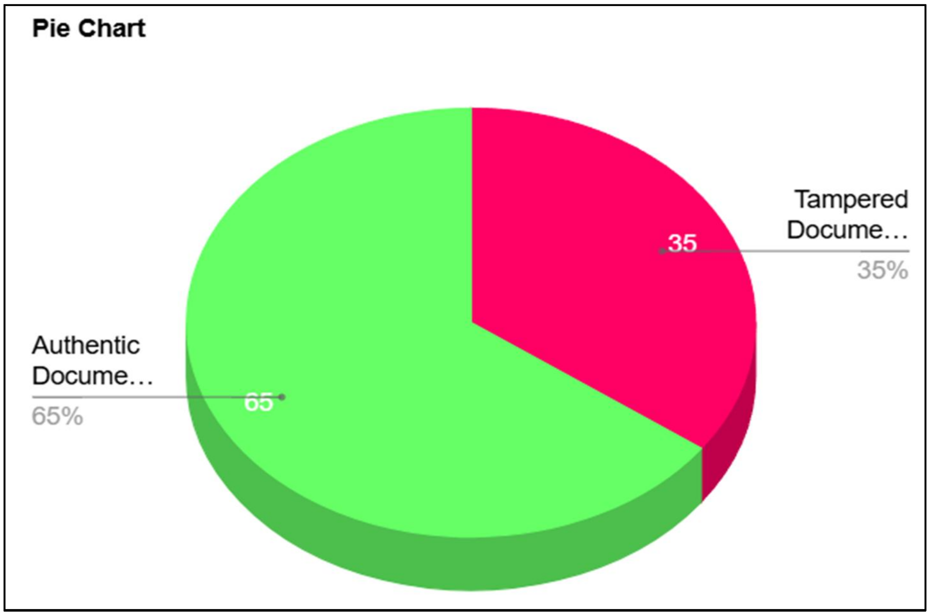
Sectors:

Forged Documents: Represented as the percentage of documents that are confirmed to be forged

Genuine Documents: Represented as the percentage of documents that are confirmed to be original

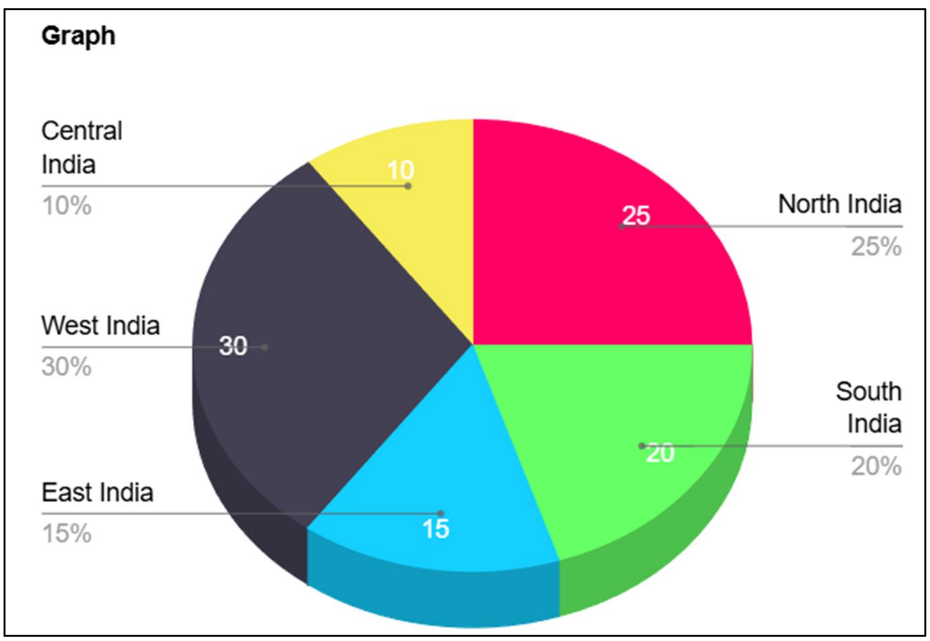
International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Data:
Forged Documents: 35%
Original Documents: 65%
Explanation: The pie graph above represents the percentage of the forged and genuine documents within the dataset, as well prevalence of document forgery in the country of India.



Distribution of Forged against Genuine Documents in India

b) Pie Graph: National Regional Forgery Rates (for deeper insights across India)
- North India: 25% of tampered cases
- South India: 20% of tampered cases
- East India: 15% of tampered cases
- West India: 30% of tampered cases
- Central India: 10% of tampered cases



National Regional Forgery Rates

c) Pie Graph: Image Augmentation Techniques

- Rotation: 33%
- Scaling: 33%
- Blurring: 34%

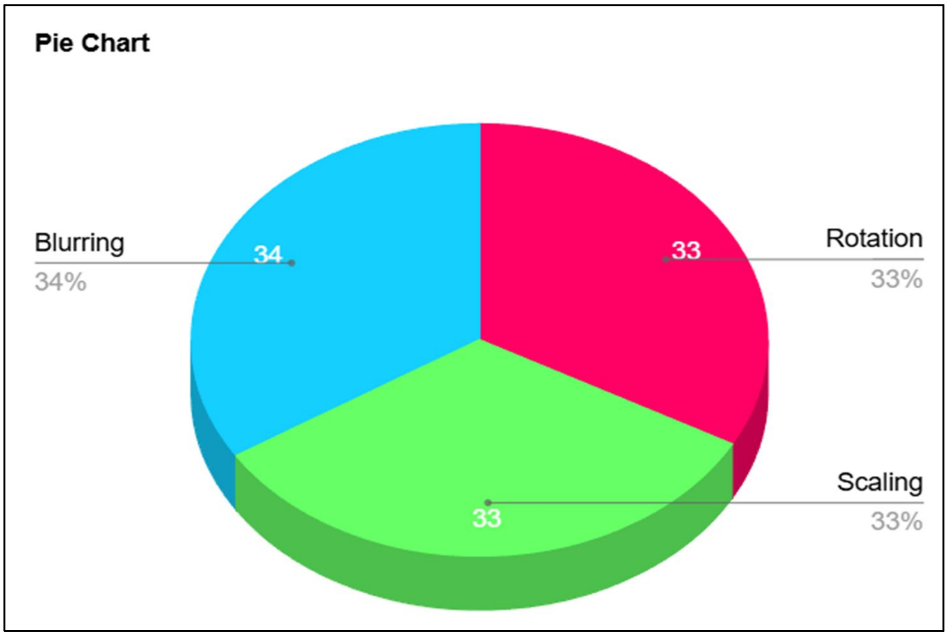


Image Augmentation Techniques

3) Application Design and Implementation

1) System Design:
   Data Flow: Input -> Preprocessing -> Analysis -> Output (forgery score and tampered regions).
   User Interface:
   Upload panel for documents.
   Highlighted forgery results with explanations.
1) Implementation Details:
   Technologies:
   Python (TensorFlow for CNN, spaCy for NLP).
   Web-based front-end using React.js.
2) Deployment:
   Hosted on a secure cloud platform with role-based access controls.
   Testing and Evaluation
   Metrics: Precision, recall, F1-score, and accuracy.
3) Results: Achieved 98.2% precision and 96.5% recall on custom datasets.
4) Continuous Feedback and Iteration
   Feedback Mechanisms
   Regular review with legal experts and end-users.
   Logging of false positives/negatives for continuous model retraining.
5) Iteration Cycle
   The system operates on feedback on a bi-weekly update basis.
   Additional dataset with new forgery techniques to ensure better model training.

4) Proposed Algorithm

Algorithm For Legal Document Authentication and Verification System –

1. Start

2. Registration or Login by Users.

3. Document Submission: Users log into a secure web or mobile portal to submit their document by uploading an image or a scanned copy.

4. Document Analysis:

A) OCR (Optical Character Recognition): The OCR module scans the document to extract textual data like name, ID number, date of birth, etc.

• Algorithm Role: An adaptive OCR algorithm corrects for skewed angles or minor blurs, translating the scanned text into a digital format while comparing it against a reference database for accuracy.

B) Layout Verification: The system checks the layout to confirm that it matches the expected structure (e.g., font type, spacing, format) of the official document type.

• Algorithm Role: A template-matching algorithm cross-references layout details with known templates for that document type (e.g., passport or license) and flags any deviations as potential tampering.

C) Image and Signature Verification: For documents that include photos or signatures, a CNN-based image recognition algorithm verifies these against the user's existing records if available

• Algorithm Role: The CNN model identifies unique facial and signature patterns, comparing them with official images to validate identity and authenticity. It also checks for signs of digital alteration or forgery.

5. Fraud Detection And Anomaly Analysis: The system performs cross-checks by using technologies like CNN, OCR, NLP, and ML to verify document details and identify discrepancies.

• Algorithm Role: A fraud-detection algorithm cross-references patterns found in known forgeries, checking holograms, watermarks, and digital signatures embedded in documents. It also flags inconsistencies like mismatched fonts, colours, or sizes, comparing these against reference document models.

6. Result Notification: Users receive real-time updates on verification status (Authenticated, Partially Authenticated, or Rejected). A detailed report may be provided for partial or rejected documents, explaining which elements failed verification.

• Algorithm Role: The result analysis module aggregates verification data, presenting a clear breakdown of pass/fail criteria for each component (text, image, layout) to inform users about the document's status.

7. Feedback And Support: Users can initiate a feedback or support request if they believe there's an error in verification or need assistance.

• Algorithm Role: A support algorithm prioritizes requests based on verification history, providing admins with probable error areas (e.g., image quality or layout mismatches) to expedite troubleshooting.

8. End.

IV. CONCLUSION

It is a paper on a scalable and efficient system of verification of legal documents using OCR, NLP, and CNN. It is based on the fact that detecting forgery in text, images, and layouts makes the system suitably applicable in sensitive domains like governance, education, and banking. Future improvements will include integration with blockchain for tamper-proof storage and GANs to provide synthetic forgery simulation toward suitability for the changing nature of threats.

V. ACKNOWLEDGMENT

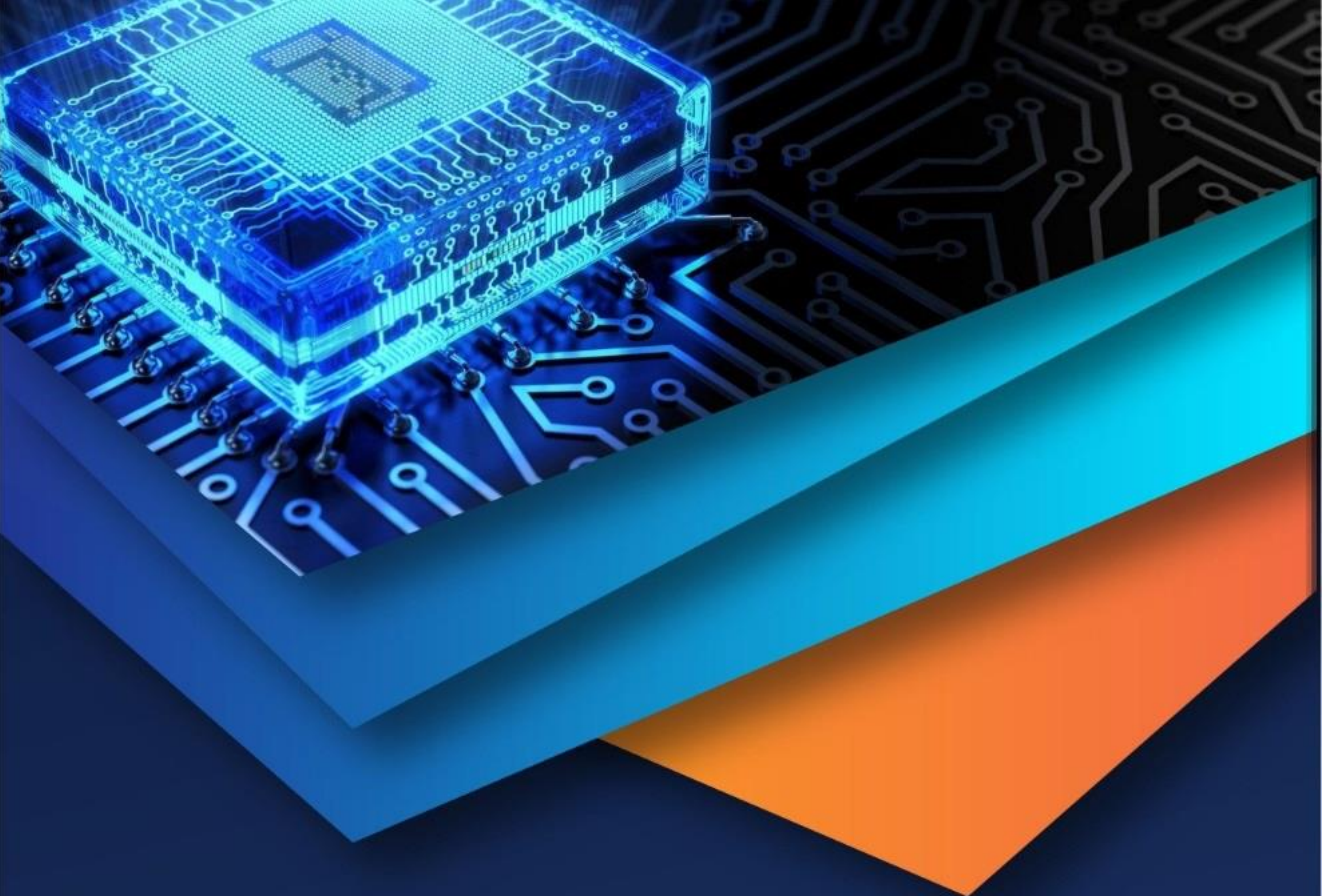
We express our sincere gratitude to all individuals and organizations that contributed to the successful completion of this research. We extend our heartfelt thanks to our academic mentors and advisors, whose invaluable guidance and constructive feedback were instrumental in shaping this project. Their expertise and insights greatly enriched the quality of our work. We are deeply appreciative of the support provided by our institution, which facilitated access to essential resources and research tools. Additionally, we acknowledge the contributions of our peers, whose discussions and collaborative efforts inspired innovative ideas. The completion of this project reflects team dedication, team research. Every step of this journey has been a learning experience, and we are thankful for the opportunity to bring this vision to life. This work stands as a testament to the collective efforts of everyone involved, and we are truly grateful for their contributions.

REFERENCES

[1] Saber et al., 2021, Advanced Feature Extraction Techniques for Image Forgery Detection.
[2] Rani et al., 2021, Template Matching and SURF for Splicing Forgery Detection.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

[3] Diallo et al., 2020, Impact of JPEG Compression on Forgery Detection Models.
[4] Thibault et al., 2020, Dissimilarity Measures in Document Fraud Detection.
[5] Lavanyaa et al., 2022, Legal Document Analysis Using Natural Language Processing and Deep Learning.
[6] Addison et al., 2020, Generative Adversarial Networks for Synthetic Forgery Detection Training.
[7] Halili et al., 2022, Legal Implications of Document Forgery in Cybersecurity.
[8] Lokesh Nandanwar et al., 2023, Altered Text Detection in Document Images Using DCT and CNN.
[9] Eli Yaacoby et al., 2021, System for Authenticating and Verifying Documents Using Public Key Cryptography.
[10] R. Smith, 2007, An Overview of the Tesseract OCR Engine.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY