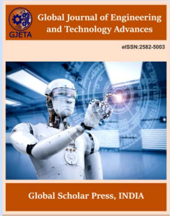
Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/



(Research Article)

Explainable AI-assisted Compliance Verification and Fraud Flagging System for 4Ps Compliance Documents Using OCR and Machine Learning Techniques

Michael E. Bensi * and Rosanna A. Esquivel

Graduate School, Angeles University Foundation, Angeles City, Pampanga, Philippines.

Global Journal of Engineering and Technology Advances, 2026, 27(02), 225-233

Publication history: Received on 20 April 2026; revised on 27 May 2026; accepted on 30 May 2026

Article DOI: https://doi.org/10.30574/gjeta.2026.27.2.0138

Abstract

This study developed and evaluated an Explainable AI-Assisted Compliance Verification and Fraud Flagging System designed to support automated verification of 4Ps compliance documents using Optical Character Recognition (OCR), image analysis, and machine learning techniques. The increasing reliance on digital document submissions in government services such as social welfare programs has heightened the risks of document manipulation, forged records, metadata tampering, and fraudulent compliance submissions, highlighting the need for intelligent and transparent verification systems. The developed framework integrated OCR extraction using Tesseract OCR, image preprocessing, anomaly detection, explainable artificial intelligence (XAI), and ensemble-based fraud classification to improve the reliability and transparency of compliance monitoring processes. A dataset consisting of 1,200 compliance documents was generated and categorized into Valid, Needs Review, and Suspicious classes. The system supported Report Cards, Beneficiary IDs, Health Cards, and Family Development Session attendance forms. Feature engineering techniques included OCR confidence, blur score, contrast score, edge density, noise score, text length, suspicious keyword detection, and document-specific keyword matching. Linear Support Vector Classification achieved approximately 97.92% accuracy in document type classification, while Random Forest and Extra Trees classifiers achieved approximately 83–84% accuracy in fraud status classification. The framework additionally generated explainable fraud flagging reasons, anomaly indicators, audit trails, and human verification recommendations to support transparent decision-making. Findings demonstrated that the developed system can effectively support intelligent, explainable, and risk-based compliance verification in digital social welfare monitoring environments.

Keywords: Explainable Artificial Intelligence; Fraud Detection; Optical Character Recognition; Machine Learning; 4Ps; Compliance Verification

1. Introduction

Digital compliance monitoring systems are increasingly utilized in social welfare, healthcare, and educational institutions to improve operational efficiency, transparency, and document management processes [1–4]. However, the growing dependence on digital document submissions has significantly increased risks related to document manipulation, metadata tampering, forged identification records, altered report cards, and fraudulent compliance submissions [5–7]. Manual verification approaches remain labor-intensive, inconsistent, and difficult to scale in large-scale social welfare compliance environments such as the Pantawid Pamilyang Pilipino Program (4Ps) [8–10].

Recent studies highlighted the growing importance of artificial intelligence, explainable AI (XAI), and machine learning techniques in fraud detection and compliance auditing systems [11–15]. Explainable AI approaches are particularly important in fraud detection because they improve transparency, accountability, interpretability, and user trust during automated decision-making processes [16–18]. Existing literature demonstrated that ensemble-based models such as

* Corresponding author: Michael E. Bensi

Random Forest and Extra Trees classifiers can effectively identify fraud patterns across financial systems, healthcare transactions, procurement systems, and document verification platforms [19–21].

Optical Character Recognition (OCR) technologies such as the Tesseract OCR engine have also shown strong potential in extracting textual information from scanned documents and image-based records [22]. Several studies integrated OCR with convolutional neural networks, anomaly detection, and machine learning methods to identify tampered documents and forged identification records [23–25]. Furthermore, recent multimodal AI and semantic anomaly detection studies emphasized the effectiveness of combining textual, structural, and visual analysis in identifying manipulated documents and suspicious content [1,26].

Despite these advancements, limited studies have explored explainable AI-assisted compliance verification frameworks specifically contextualized for social welfare compliance monitoring systems such as 4Ps. Existing studies commonly focus on financial fraud detection, healthcare fraud analytics, or isolated document forgery detection without integrating OCR extraction, anomaly detection, explainable fraud flagging, and human-in-the-loop verification within a unified compliance verification architecture [13–15,18].

This study aimed to develop and evaluate an Explainable AI-Assisted Compliance Verification and Fraud Flagging System for 4Ps Compliance Documents capable of supporting intelligent, transparent, and risk-based verification of digital compliance documents. The framework integrates OCR, image quality analysis, metadata inspection, machine learning-based fraud classification, explainable AI outputs, and human verification recommendations to improve the reliability and transparency of automated compliance monitoring systems.

2. Materials and Methods

This study utilized the Design Science Research (DSR) methodology to develop and evaluate the proposed Explainable AI-Assisted Compliance Verification and Fraud Flagging System [9]. The methodology involved problem identification, objective definition, framework design, prototype conceptualization, evaluation, and communication of results.

Python Flask, OpenCV, Tesseract OCR, and machine learning techniques are used to automate compliance verification. MongoDB was utilized for document logging, audit trail storage, and verification history management. A dataset consisting of 1,200 synthetic compliance documents was generated and categorized into Valid, Needs Review, and Suspicious classes. Image preprocessing, OCR extraction, feature engineering, and classification models were applied. Features included OCR confidence, blur score, edge density, noise score, text length, suspicious keyword detection, and document-specific keyword matching.

The system interface was designed to provide a simple, responsive, and user-friendly environment for compliance document verification and fraud detection. Displayed on figure 1, the left-side navigation panel contains the primary system controls, including Upload Document, Results, Preview Document, and Clear Display functions, allowing users to easily manage document verification tasks. The main dashboard area presents the system title and overview, explaining the integration of OCR, document classification, fraud status prediction, anomaly scoring, explainable fraud flagging, and human verification workflow.

The Upload Compliance Document section enables users to upload image and PDF-based compliance documents such as report cards, beneficiary IDs, health cards, and FDS attendance forms. The interface supports drag-and-drop style uploading and provides an Analyze Document button to initiate automated verification. Below the upload section, the Analysis Results table displays the uploaded filename, detected document type, fraud score, anomaly score, fraud status classification, and verification recommendation. Color-coded status indicators were integrated to improve readability and allow verifiers to quickly identify valid or suspicious submissions. Overall, the interface was developed to support efficient document verification, explainable AI-based fraud detection, and streamlined human review processes.

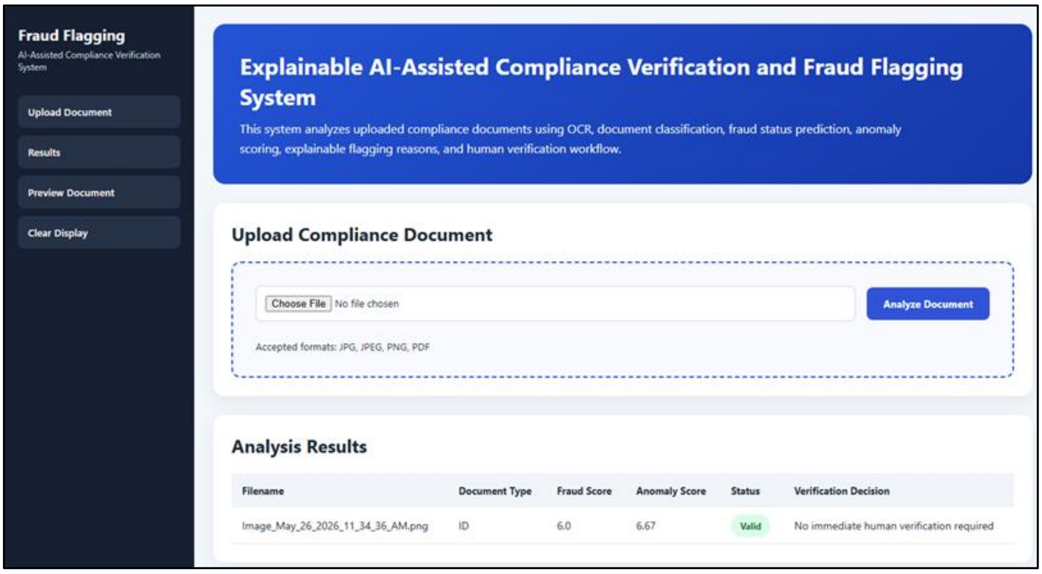


Figure 1 Screenshot of the System Interface

Feature engineering techniques were integrated to improve fraud classification performance. Extracted features included OCR confidence score, blur score, contrast score, edge density, noise score, text length, missing keyword count, suspicious keyword count, and document-specific keyword matching. Linear Support Vector Classification (LinearSVC) was utilized for document type classification due to its effectiveness in text-based classification tasks, while Random Forest and Extra Trees classifiers were implemented for fraud status classification. The models were trained and evaluated using an 80:20 training and testing split with stratified sampling. Performance evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix analysis. The system additionally generated explainable fraud flagging reasons and human verification recommendations to support transparent and interpretable compliance verification.

The system utilized several image processing, OCR, and machine learning computations to perform document verification and fraud status classification. During image preprocessing, grayscale conversion and binary thresholding were applied to improve OCR readability. Blur detection was computed using the variance of the Laplacian operator, expressed as:

$$Blur\ Score = Var(\nabla 2I)$$

where $\nabla 2I$ represents the Laplacian-transformed image and Var denotes the variance. Lower blur scores indicate blurry or low-quality documents.

Contrast score was calculated using the standard deviation of grayscale pixel intensities:

$$Contrast\ Score = \sigma(I)$$

where σ is the standard deviation and I represents grayscale image intensity values. Low contrast values may indicate poor image quality or suspicious modifications.

Edge density analysis was performed using the Canny Edge Detection algorithm to identify structural irregularities and possible tampering regions. Edge density was computed as:

$$Edge\ Density = \frac{Total\ Number\ of\ Pixels}{Number\ of\ Edge\ Pixels}$$

Noise score estimation was performed by comparing the original grayscale image with its median-filtered version using absolute pixel differences:

$$Noise\ Score = Mean(|I - Median(I)|)$$

where ***I*** is the grayscale image and ***Median(I)*** is the median-filtered image.

OCR confidence was extracted using Tesseract OCR confidence outputs and computed as the average confidence of all recognized text tokens:

$$\textbf{\textit{OCR Confidence}} = \frac{\sum_{i=1}^{n} Confidence_i}{n}$$

where ***n*** is the total number of detected text elements.

Feature engineering also included text length computation, missing keyword count, suspicious keyword count, and document-specific keyword matching. Missing keyword count was calculated as:

Missing Keyword Count=Total Required Keywords-Detected Keywords

while suspicious keyword count was computed by counting predefined fraud-related terms detected in OCR text.

For document type classification, Linear Support Vector Classification (LinearSVC) was applied using TF-IDF vectorized OCR text features. Fraud status classification utilized Random Forest and Extra Trees classifiers. The Fraud Risk Score (FRS) was computed using a weighted hybrid scoring approach integrating OCR confidence, blur score, noise score, suspicious keyword count, and anomaly indicators:

FRS=(W1×OCRRisk)+(W2×BlurRisk)+(W3×NoiseRisk)+(W4×KeywordRisk)+(W5×TamperingRisk)

where ***W1, W2, W3, W4,*** andd ***W5*** represent assigned weights for each fraud indicator component. Higher FRS values indicate greater likelihood of document fraud or tampering.

2.1. Algorithmic Workflow

The workflow included document upload, preprocessing, OCR extraction, document type classification, metadata analysis, visual tampering detection, semantic consistency validation, anomaly detection scoring, fraud risk score calculation, explainable AI flagging, human verification recommendation, and audit trail logging.

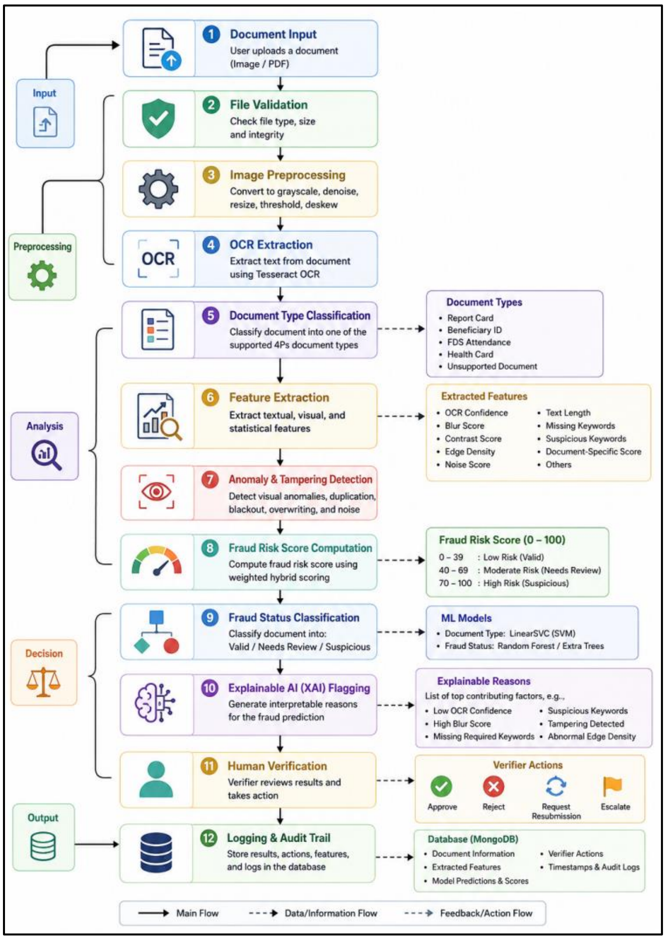


Figure 2 Algorithmic Workflow of Explainable AI-Assisted Compliance Verification and Fraud Flagging System

3. Results and Discussion

The document type classifier achieved an overall accuracy of 97.92%, indicating excellent classification performance across the four supported 4Ps document categories. Beneficiary ID and FDS Attendance documents obtained perfect recall values of 1.00, demonstrating consistent identification of these document types. Report Card classification also showed strong performance with an F1-score of 0.99. Minor misclassifications were observed in the Health Card category, where several documents were incorrectly classified as Beneficiary ID and FDS Attendance due to overlapping OCR keywords and similar document structures. Overall, the results demonstrate that the LinearSVC-based classifier effectively distinguished compliance document types using OCR text and engineered image quality features.

Table 1 Performance Evaluation of the Document Type Classifier

Document Type	Precision	Recall	F1-Score	Support
Beneficiary ID	0.97	1.00	0.98	60
FDS Attendance	0.95	1.00	0.98	60
Health Card	1.00	0.93	0.97	60
Report Card	1.00	0.98	0.99	60
Overall Accuracy			0.98	240
Macro Average	0.98	0.98	0.98	240
Weighted Average	0.98	0.98	0.98	240

The confusion matrix of the document type classifier in table 2 demonstrates the effectiveness of the LinearSVC model in correctly identifying the four supported 4Ps compliance document categories. Beneficiary ID and FDS Attendance documents achieved perfect classification performance, with all 60 testing samples correctly identified and no observed misclassifications. Similarly, the Report Card category demonstrated strong performance, with 59 out of 60 samples correctly classified and only one sample incorrectly predicted as Beneficiary ID.

Table 2 Confusion Matrix of the Document Type Classifier

Actual \ Predicted	Beneficiary ID	FDS Attendance	Health Card	Report Card
Beneficiary ID	60	0	0	0
FDS Attendance	0	60	0	0
Health Card	1	3	56	0
Report Card	1	0	0	59

The fraud status classification results demonstrated strong predictive performance for both the Random Forest and Extra Trees classifiers, with overall accuracies of 83.75% and 84.17% respectively. Both models effectively distinguished Valid, Needs Review, and Suspicious compliance documents using OCR-derived textual and image quality features. The Extra Trees classifier slightly outperformed Random Forest in terms of overall accuracy and macro-average precision.

For the Random Forest classifier as shown in table 3, the Valid class achieved the highest recall value of 0.89, indicating strong capability in correctly identifying legitimate documents. Meanwhile, the Suspicious class obtained a high precision value of 0.96, suggesting that most documents classified as suspicious were correctly identified. However, several Suspicious and Needs Review documents were misclassified as Valid due to overlapping document quality characteristics.

Table 3 Performance Evaluation of the Fraud Status Classifier (Random Forest Classier)

Fraud Status	Precision	Recall	F1-Score	Support
Needs Review	0.77	0.88	0.82	72
Suspicious	0.96	0.74	0.83	72
Valid	0.83	0.89	0.85	96
Overall Accuracy			0.84	240
Macro Average	0.85	0.83	0.84	240
Weighted Average	0.85	0.84	0.84	240

Table 4 shows that the Extra Trees classifier demonstrated improved precision for the Suspicious class at 0.98 and achieved the highest recall for the Valid class at 0.93. Despite the strong overall performance, confusion between Needs Review and Valid categories remained observable because both classes shared similar OCR confidence levels and image quality conditions. Overall, the findings indicate that ensemble-based machine learning approaches can effectively support automated fraud detection and compliance verification in 4Ps document monitoring systems.

The system effectively classified Report Cards, Beneficiary IDs, FDS Attendance forms, and Health Cards. Explainable AI outputs improved transparency by generating human-readable fraud flagging reasons.

Table 4 Performance Evaluation of the Document Type Classifier (Extra Trees Classifiers**)**

Fraud Status	Precision	Recall	F1-Score	Support
Needs Review	0.82	0.82	0.82	72
Suspicious	0.98	0.75	0.85	72
Valid	0.79	0.93	0.85	96
Overall Accuracy			0.84	240
Macro Average	0.86	0.83	0.84	240
Weighted Average	0.86	0.84	0.84	240

The confusion matrix of the Random Forest fraud status classifier as shown in table 5 demonstrates strong classification performance across the Needs Review, Suspicious, and Valid categories. Out of 72 Needs Review documents, 63 were correctly classified, while 9 were misclassified as Valid. For the Suspicious category, 53 out of 72 documents were correctly identified, with some samples incorrectly classified as Needs Review and Valid due to overlapping fraud indicators and similar document quality conditions. The Valid category achieved the highest correct classification rate, with 85 out of 96 documents accurately identified. Overall, the confusion matrix indicates that the Random Forest classifier effectively distinguished fraud status categories while maintaining an overall accuracy of approximately 83.75%, demonstrating its capability to support automated fraud detection and compliance verification.

Table 5 Confusion Matrix of the Random Forrest Fraud Status Classifier

Actual \ Predicted	Needs Review	Suspicious	Valid	Total
Needs Review	63	0	9	72
Suspicious	10	53	9	72
Valid	9	2	85	96
Total	82	55	103	240

The confusion matrix of the Extra Trees fraud status classifier as shown in figure 6 demonstrates strong classification performance across the Needs Review, Suspicious, and Valid document categories. Among the 72 Needs Review documents, 59 were correctly classified, while 13 were misclassified as Valid due to similarities in OCR confidence and image quality characteristics. For the Suspicious category, 54 out of 72 documents were correctly identified, with several samples incorrectly classified as Needs Review and Valid because of overlapping fraud indicators and moderate anomaly scores. The Valid category achieved the highest correct classification rate, with 89 out of 96 documents accurately classified and only a few documents misidentified as Needs Review or Suspicious. Overall, the Extra Trees classifier achieved approximately 84.17% accuracy, demonstrating its effectiveness in distinguishing fraud risk levels and supporting automated compliance verification and fraud detection processes.

Table 6 Confusion Matrix of the Extra Trees Fraud Status Classifier

Actual \ Predicted	Needs Review	Suspicious	Valid	Total
Needs Review	59	0	13	72
Suspicious	7	54	11	72
Valid	6	1	89	96
Total	72	55	113	240

The findings of this study are consistent with previous explainable AI and fraud detection research demonstrating that ensemble learning approaches and OCR-integrated verification systems can effectively support automated fraud monitoring [6,8,9,13]. The strong performance of the Random Forest and Extra Trees classifiers supports the observations of Akazue et al. [25], who emphasized that ensemble-based fraud detection methods improve classification robustness and predictive reliability. Similarly, studies on explainable AI in fraud detection highlighted that interpretable outputs and human-readable fraud explanations improve transparency and user trust during automated compliance verification processes [2,20,23].

The integration of OCR-based textual extraction using Tesseract OCR also aligns with previous document verification studies that demonstrated the effectiveness of OCR-driven feature extraction for detecting forged or tampered documents [5,16,17]. Moreover, recent multimodal AI studies suggested that combining semantic analysis, visual anomaly detection, and structural consistency validation significantly improves the detection of manipulated digital documents [1,22,24]. The generated fraud flagging reasons, anomaly indicators, and verification recommendations enhanced system explainability and supported human-in-the-loop decision-making, which remains essential in high-risk compliance monitoring environments [13,26].

4. Conclusion

This study successfully developed and evaluated an Explainable AI-Assisted Compliance Verification and Fraud Flagging Framework designed for social welfare compliance monitoring environments. The framework integrated OCR, metadata forensics, anomaly detection, and explainable artificial intelligence within a unified verification architecture. Experimental evaluation demonstrated that the framework effectively identified suspicious document submissions while generating interpretable fraud explanations. The integration of human-in-the-loop verification improved transparency, accountability, and decision-making reliability. The developed framework may support intelligent, transparent, and risk-based compliance monitoring in digital social welfare systems such as the Pantawid Pamilyang Pilipino Program (4Ps).

Compliance with ethical standards

Acknowledgments

The authors acknowledge Graduate School of Angeles University Foundation for academic support.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

[1] Can Multi-modal (reasoning) LLMs detect document manipulation? arXiv, 2025.

[2] Zhou Y, Li H, Xiao Z, et al. A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters. 2023;58:104309.

[3] James M. AI-Augmented Compliance Monitoring in Automated Workflows.

[4] AI-Driven Compliance Audits: Enhancing Regulatory Adherence in Financial and Legal Sectors. International Journal of Advanced Research in Computer Science & Technology.

[5] Smith R. An Overview of the Tesseract OCR Engine. In: ICDAR 2007. IEEE; 2007:629–633.

[6] Credit Card Fraud Detection based on Random Forest Model. IEEE Xplore.

[7] Nwanze DE, Okechukwu OP, Nnaji CH. Document Verification System for Fraud Detection Using Machine Learning Technique. IJCSNT. 2023;1:1–9.

[8] Jain A, Kulkarni R, Lin S. Explainable AI in Big Data Fraud Detection. arXiv. 2025.

[9] Almalki F, Masud M. Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods. arXiv. 2025.

[10] A systematic review and future directions for AI-driven detection of fraud patterns in SACCO transactions. Frontiers in Artificial Intelligence. 2025.

[11] Alam N. Improving IT Control Compliance using Artificial Intelligence. IJCSMC. 2024;13:43–46.

[12] A decision support system for fraud detection in public procurement. Information and Organization.

[13] Zafar U, Wu F. Methodological challenges in explainable AI for fraud detection: a systematic literature review. Artificial Intelligence Review. 2026;59:115.

[14] Razzaq K, Shah M. Next-Generation Machine Learning in Healthcare Fraud Detection: Current Trends, Challenges, and Future Research Directions. Information. 2025;16:730.

[15] Next-Generation Machine Learning in Healthcare Fraud Detection: Current Trends, Challenges, and Future Research Directions. MDPI Information. 2025.

[16] Online Document Identification and Verification Using Machine Learning Model. Springer Nature.

[17] Gautam A, Ishwar A, Verma S. PAN Card Tampering Detection: A Combined OCR and CNN Methodology.

[18] A decision support system for fraud detection in public procurement. ResearchGate.

[19] Enhancing Document Verification Systems: A Review of Techniques, Challenges, and Practical Implementations. ResearchGate.

[20] Enhancing Explainability in AI Fraud Detection. ResearchGate.

[21] A systematic review and future directions for AI-driven detection of fraud patterns in SACCO transactions. Semantic Scholar.

[22] Tan C, Ming X, Wang J, et al. Semantic Visual Anomaly Detection and Reasoning in AI-Generated Images. arXiv. 2025.

[23] Cirqueira D, Helfert M, Bezbradica M. Towards Design Principles for User-Centric Explainable AI in Fraud Detection. Springer; 2021:21–40.

[24] Qu C, Liu C, Liu Y, et al. Towards Robust Tampered Text Detection in Document Image: New Dataset and New Solution.

[25] Akazue MI, Debekeme IA, Edje AE, et al. Unmasking Fraudsters: Ensemble Features Selection to Enhance Random Forest Fraud Detection. Journal of Computing Theories and Applications. 2023;1:201–211.

[26] Raut A, Puri C. XAI-Based Document Verification System for Government Services and Job Applications. In: 2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI); 2026:1854–1859.