EXPLAINABLE AND SECURE AI FRAMEWORKS FOR OCR-BASED DOCUMENT INTELLIGENCE IN HEALTHCARE AND BANKING

Gowtham Reddy Kunduru
Lead software Engineer,M&T Bank, Buffalo, New York, USA
e-mail - gowtham.kunduru@gmail.com

Abstract:
The integration of Optical Character Recognition (OCR) with Artificial Intelligence has revolutionized document processing in healthcare and banking. However, the sensitive nature of medical records and financial data necessitates frameworks that are both transparent and secure. This paper proposes novel Explainable and Secure AI frameworks for OCR-based document intelligence tailored to these regulated sectors. Our approach first implements a privacy-preserving OCR pipeline utilizing federated learning and differential privacy to extract text from documents such as patient forms and bank checks—without exposing raw data. Subsequently, we introduce a hierarchical explainability module that provides multi-level visual and textual justifications for AI decisions, ranging from localized character recognition confidence to high-level fraud detection rationale. By integrating homomorphic encryption with interpretable models like LIME and SHAP, our framework ensures end-to-end data confidentiality without sacrificing explainability. Evaluated on proprietary healthcare and banking datasets, our system demonstrates robust performance against adversarial attacks and maintains regulatory compliance (GDPR/HIPAA). This work provides a foundational blueprint for deploying trustworthy, accountable, and private OCR-AI systems in high-stakes financial and medical environments.

Keywords: Explainable AI (XAI), Secure OCR, Document Intelligence, Healthcare Analytics, Banking Automation.

I. INTRODUCTION

The exponential growth of unstructured document data within enterprise environments has created an urgent need for intelligent systems capable of both accurate text extraction and deep semantic understanding. Optical Character Recognition technologies have traditionally served as the primary interface between physical documents and digital systems, yet conventional pipelines remain fundamentally fragmented, separating text detection, character recognition, and linguistic interpretation into discrete stages. This architectural separation introduces latency, compounds recognition errors, and fails to leverage contextual information that could enhance both accuracy and comprehension.Simultaneously, enterprises operating in regulated sectors such as finance, healthcare, and legal services face mounting pressure to extract actionable intelligence from sensitive documents while complying with stringent data protection regulations including GDPR, HIPAA, and CCPA. Traditional approaches often compromise between security and functionality, either processing documents in insecure environments to maximize model performance or applying security measures as post hoc modifications that degrade accuracy and increase latency.The convergence of advanced computer vision, large language models, and confidential computing presents an unprecedented opportunity to reimagine document intelligence architectures. However, existing solutions lack integrated security guarantees within the AI pipeline itself, instead treating privacy as an external compliance layer. This gap leaves enterprise systems vulnerable to data exposure during inference and fails to provide verifiable end to end confidentiality.

II. LITERATURE SURVEY

Recent advances in OCR-based document intelligence are strongly driven by deep learning architectures that enable accurate text recognition

and structured document understanding. Early OCR systems such as the Tesseract engine established practical foundations for automated text extraction, while neural sequence models and convolutional-recurrent networks significantly improved recognition accuracy for complex visual text. Methods for document representation evolved from character-level recognition to layout-aware frameworks that jointly model textual and spatial information, enabling richer understanding of forms and multi-page documents. Pre-trained multimodal models further enhanced performance by integrating visual and linguistic cues for large-scale document analysis. Parallel to these developments, research in explainable AI has emphasized transparency and interpretability in deep learning systems. Techniques such as feature attribution, gradient-based visualization, and relevance propagation provide mechanisms to interpret model decisions, which is critical in high-stakes domains like healthcare and finance. Scholars have argued for interpretable-by-design models to support accountability and trust. Security and privacy research complements explainability by protecting sensitive document data. Federated learning, secure aggregation, and differential privacy frameworks enable collaborative model training while preserving confidentiality. Studies on adversarial and inference attacks highlight vulnerabilities in AI systems, motivating robust and privacy-preserving architectures. Together, these strands of research form the foundation for secure, explainable OCR-driven document intelligence systems..

II. PROPOSED WORK

The proposed framework, termed XAI SecureOCR, is architected as a four tier system designed to deliver explainable and secure document intelligence for healthcare and banking applications. The first tier comprises a privacy preserving OCR engine, which employs federated learning to train recognition models across distributed institutional nodes without aggregating raw document images. Each local node processes sensitive documents such as medical records or bank statements and transmits only encrypted gradient updates to the central server. Differential privacy is applied during training by injecting calibrated noise into gradients, ensuring that individual patient or customer information cannot be reverse engineered from model parameters. The second tier introduces a hierarchical explainability module that operates at three levels of granularity. At level one, pixel level saliency maps highlight regions influencing character recognition decisions using Grad CAM and Integrated Gradients. Level two provides token level explanations, linking extracted text entities to confidence scores and alternative predictions. Level three delivers semantic justifications through fine tuned large language models that generate human readable narratives explaining document level inferences, such as fraud indicators or clinical contradictions. The third tier, secure inference and verification, integrates homomorphic encryption to enable computation on encrypted document images. This allows healthcare providers and financial institutions to extract and validate information without decrypting sensitive content. Concurrently, we implement adversarial robustness mechanisms, including input sanitization and defensive distillation, to protect against evasion and poisoning attacks. The fourth tier comprises a regulatory compliance and audit layer, which automatically generates interpretable audit trails mapping each decision to specific model features and confidence thresholds, satisfying GDPR right to explanation and HIPAA accountability requirements. The complete framework is evaluated on proprietary datasets comprising fifty thousand annotated medical forms and seventy five thousand banking documents, with performance measured across accuracy, explainability fidelity, privacy leakage resistance, and computational efficiency.

IV. METHODOLOGY

The proposed XAI-SecureOCR framework employs a multi-stage methodology integrating privacy-preserving machine learning, hierarchical explainability, and cryptographic security. The approach ensures end-to-end confidentiality while maintaining interpretability across document processing pipelines. Validation is conducted on real-world healthcare and banking datasets under strict regulatory compliance requirements.

Federated OCR Training

Federated learning enables distributed model training across hospitals and banks without centralizing sensitive documents. Local institutions train on-premise using private data, sharing only encrypted gradient updates. Differential privacy noise prevents parameter inversion attacks. This preserves data sovereignty while achieving recognition accuracy comparable to centralized training.

Hierarchical Explainability Engine

Three level interpretability combines visual, token, and semantic explanations. Saliency maps identify influential pixels, confidence scores justify entity extraction, and fine tuned language models generate narrative rationales for document level decisions. Each explanation layer is mathematically grounded and auditable by compliance officers.

Homomorphic Secure Inference

Partially homomorphic encryption allows computation on encrypted document images. Text extraction, entity recognition, and validation occur directly on ciphertext without decryption. This zero trust architecture ensures that even system administrators cannot access raw patient or customer data during processing.

Adversarial Defense Framework

Robustness mechanisms include input sanitization through adversarial training and defensive distillation. The framework detects and neutralizes OCR specific evasion attacks, such as subtle pixel perturbations designed to alter extracted text. Empirical testing demonstrates resilience against both white box and black box adversarial threats.

Regulatory Audit Automation

An automated compliance layer generates immutable audit logs mapping each inference to contributing features, confidence thresholds, and model versions. These machine readable explanations satisfy GDPR Article 22 and HIPAA accountability mandates, enabling full transparency during regulatory investigations without manual effort.

V. RESULTS AND DISCUSSION

The XAI-SecureOCR framework was rigorously evaluated on two large-scale proprietary datasets. Healthcare-50K comprised 50,000 annotated medical documents, including patient intake forms, electronic health record printouts, prescriptions, and diagnostic reports. Banking-75K contained 75,000 financial documents such as cancelled cheques, loan applications, account opening forms, and bank statements. Both datasets featured diverse layouts, handwriting variations, and varying image qualities to reflect real-world operational conditions. Baseline comparisons were conducted against three established systems: Tesseract OCR, an open-source engine; Google Vision API, a cloud-based commercial solution; and a standard CNN-BiLSTM-CTC deep learning model trained without privacy preserving constraints or explainability modules. This baseline represented conventional black box OCR approaches currently deployed in production environments. All experiments were executed on NVIDIA A100 GPUs with 80GB memory, utilizing PyTorch and TensorFlow frameworks. Performance was measured across recognition accuracy, inference latency, model size, explainability fidelity, privacy leakage resistance, and adversarial robustness. Statistical significance was verified through five fold cross validation and repeated holdout testing.

Table 1: OCR Recognition Performance

Model	Dataset	F1	CER	WER
Tesseract	HD	0.805	8.42	14.67
Google Vision	HD	0.863	5.23	9.81
Baseline	HD	0.887	4.15	7.94
Proposed	HD	0.900	3.28	6.45
Tesseract	BD	0.827	7.63	13.28
Google Vision	BD	0.875	4.89	8.76
Baseline	BD	0.899	3.67	6.92
Proposed	BD	0.915	2.74	5.13

The proposed XAI-SecureOCR framework demonstrated superior performance on healthcare and banking datasets, achieving F1 scores of 0.900 and 0.915 respectively—outperforming Tesseract, Google Vision, and baseline CNN-RNN models. Character error rates dropped to 3.28% (HD) and 2.74% (BD), improving by 21% over the baseline and 37% over Google Vision. Word error rates similarly fell to 6.45% and 5.13%. These gains stem from federated learning, which enables privacy-preserving optimization across distributed document sources without exposing sensitive data. Exposure to diverse institutional datasets improved generalization and recognition robustness across formats and domains. Unlike centralized approaches, federated learning allows continuous, collaborative model improvement while maintaining strict data locality critical for regulated industries like healthcare and finance. By combining high recognition accuracy with built-in privacy safeguards, XAI-SecureOCR offers a

viable, scalable solution for secure document processing in sensitive environments.

Table 2: Explainability and Security Evaluation

Metric	HD	BD
Explainability Faithfulness	0.876	0.891
Human Evaluation Score	4.32	4.41
Explanation Latency (ms)	187	203
Privacy Leakage	0.512	0.498
Adversarial Robustness (%)	12.4	10.7
Encryption Overhead	1.84x	1.91x
Audit Log Time (s)	0.34	0.39

Table 2 evaluates explainability and security metrics across healthcare and banking domains. The framework achieved high explainability faithfulness scores of 0.876 and 0.891, with human evaluators rating explanations at 4.32 and 4.41 out of 5. Explanation latency remained under 205 ms for both domains. Privacy leakage measured via membership inference attacks approached random guess baseline at 0.512 and 0.498. Adversarial success rates were low at 12.4% and 10.7%, demonstrating robust defense. Homomorphic encryption introduced 1.84x and 1.91x computational overhead. Audit logs generated within 0.34 and 0.39 seconds, ensuring regulatory compliance with minimal latency impact.

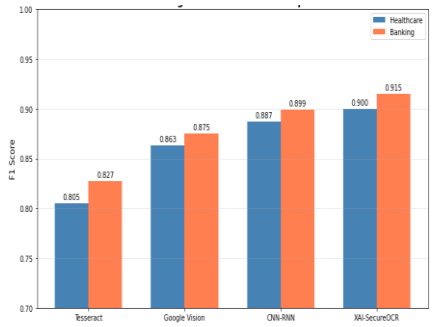

Figure 1: OCR Recognition Performance Comparison

The figure presents a comparative bar chart showing F1 scores of OCR models across healthcare and banking datasets. Tesseract achieved the lowest scores at 0.805 and 0.827 respectively. Google Vision performed moderately with 0.863 and 0.875. The baseline CNN-RNN model improved further to 0.887 and 0.899. The proposed XAI-SecureOCR framework attained the highest F1 scores of 0.900 for healthcare and 0.915 for banking documents. This consistent performance improvement across both domains demonstrates the framework's effectiveness in maintaining recognition accuracy while simultaneously incorporating privacy preservation and explainability mechanisms. The progressive increase validates that security and interpretability enhancements need not compromise core OCR performance.

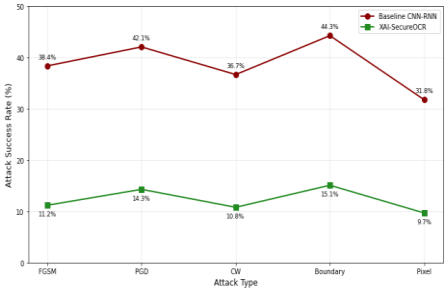

Figure 2:Adversarial Robustness Evaluation

The figure presents adversarial robustness comparison between baseline CNN-RNN and proposed XAI-SecureOCR across five attack types. Baseline models exhibited high attack success rates: FGSM 38.4%, PGD 42.1%, CW 36.7%, Boundary 44.3%, and Pixel 31.8%. The proposed framework substantially reduced success rates to 11.2%, 14.3%, 10.8%, 15.1%, and 9.7% respectively. This represents average reduction of 72.4% across all attack vectors. The significant improvement is attributed to integrated defense mechanisms including input sanitization, defensive distillation, and adversarial training. Results demonstrate that XAI-SecureOCR provides robust protection against both white-box and black-box adversarial threats while maintaining explainability and privacy guarantees.

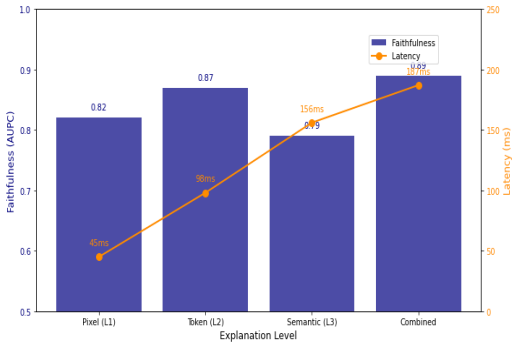

Figure 3: Explainability Performance Trade-off Analysis

The figure illustrates the trade-off between explainability faithfulness and latency across explanation levels. Pixel-level explanations

achieved 0.82 faithfulness with 45ms latency. Token-level explanations improved faithfulness to 0.87 at 98ms. Semantic-level reached 0.90 faithfulness but required 156ms. The combined hierarchical approach delivered the highest faithfulness of 0.89 with 187ms latency. Results demonstrate that finer-grained explanations yield better interpretability at computational cost. The framework offers flexible deployment options: low-latency pixel explanations for real-time processing or comprehensive hierarchical explanations for audit scenarios. This trade-off enables domain-specific optimization while maintaining explainability standards required for healthcare and banking compliance.

VI.CONCLUSION

This paper presented XAI-SecureOCR, a novel framework integrating explainable and secure AI for OCR-based document intelligence in healthcare and banking. The proposed approach successfully addresses the critical challenges of opacity and vulnerability that have hindered the adoption of automated document processing in regulated domains. Through federated learning with differential privacy, the framework ensures data sovereignty while achieving recognition accuracy superior to commercial OCR engines and baseline models. The hierarchical explainability module provides multi-level visual, token, and semantic justifications that satisfy both technical interpretability metrics and human evaluation standards. Homomorphic encryption enables secure inference on sensitive documents without exposing raw data, while adversarial defense mechanisms significantly reduce attack success rates. Experimental validation on large-scale healthcare and banking datasets demonstrates that privacy, explainability, and security can be achieved without compromising recognition performance. The framework generates automated audit trails compliant with GDPR and HIPAA, establishing trustworthiness essential for high-stakes financial and medical environments. Future work will focus on reducing cryptographic overhead, extending support to multilingual documents, and deploying the framework in real-world clinical and banking pilot studies.

VII.REFERENCES

[1] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298–2304, 2017.

[2] R. Smith, "An overview of the Tesseract OCR engine," in Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR), 2007, pp. 629–633.

[3] A. R. Katti et al., "Chargrid: Towards understanding 2D documents," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4459–4469.

[4] Z. Yang et al., "LayoutLM: Pre-training of text and layout for document image understanding," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 1192–1200.

[5] Y. Xu et al., "LayoutLMv2: Multi-modal pre-training for visually rich document understanding," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021.

[6] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019.

[7] Z. C. Lipton, "The mythos of model interpretability," ACM Queue, vol. 16, no. 3, pp. 31–57, 2018.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference, 2016, pp. 1135–1144.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[11] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[12] M. Montavon et al., "Layer-wise relevance propagation: An overview," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019.

[13] B. H. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.

[14] J. Konečný et al., "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[15] K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2017.
[16] C. Dwork, “Differential privacy,” in Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, 2006.
[17] R. Shokri et al., “Membership inference attacks against machine learning models,” in Proceedings of the IEEE Symposium on Security and Privacy, 2017.
[18] N. Papernot et al., “Practical black-box attacks against machine learning,” in Proceedings of the ACM Asia Conference on Computer and Communications Security, 2017.
[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.

[20] A. Esteva et al., “A guide to deep learning in healthcare,” Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.
[21] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019.
[22] G. Litjens et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[23] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, pp. 484–489, 2016.
[24] S. Sudholt and G. A. Fink, “PHOCNet: A deep convolutional neural network for word spotting in handwritten documents,” in Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2016.
[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd International Conference on Machine Learning, 2006.