SentinelAI: A Deployable, Explainable, and Auditable AI-Driven Fraud-Detection System for Philippine E-Wallet Transactions

Jabes P. Pauya*
Mapua University, Philippines

Ma. Abegail F. Mabalot
Mapua University, Philippines

Abstract:

Philippine e-wallet platforms screen billions of pesos of retail payments each month, yet most operators still run a static, analyst-authored rule engine that is backward-looking, costly to maintain, and unable to emit the per-decision traceability expected under Bangko Sentral ng Pilipinas Circular 1140. Published machine-learning work reports strong detection figures but stops at the notebook, supplying no streaming integration, serving topology, observability, or audit trail. This paper presents the design, development, deployment, and evaluation of an AI-Driven Fraud Detection System engineered as an end-to-end, deployable, observable, and auditable IT system rather than an isolated model. We give a formal detector model — a causal, leakage-free point-in-time feature map, a precision-weighted Random-Forest/XGBoost/Deep-Neural-Network ensemble, and a weight-consistent SHAP decomposition — and prove four structural properties (non-anticipation, monotone escalation, bounded single-signal influence, and fail-safe availability). We state the security goals and a STRIDE threat model explicitly and evaluate them on a live, integrated deployment, including an adversarial penetration test of the running backends. On a held-out labelled set the ensemble reached precision, recall, and F1 of 0.696 with a 0.52% false-positive rate and 0.989 AUC-ROC, against a rule baseline that raised 388 false alarms to the ensemble's 7; the improvement is statistically significant (McNemar, $\chi^2 = 356.7$, $p < 0.001$). A survey of one hundred Filipino e-wallet users grounds the design. The contribution is a reference implementation showing how AI-augmented fraud screening can be formalized, secured, deployed, and operated under standard fintech IT discipline.

CCS CONCEPTS • Security and privacy → Intrusion/anomaly detection and malware mitigation • Computing methodologies → Machine learning, Ensemble methods • Applied computing → Electronic commerce.

Keywords and Phrases:

AI-driven fraud detection, e-wallet transaction screening, explainable AI, SHAP, threat model, ensemble learning, MLOps and observability, reference architecture.