ABNUS JOURNAL
Your Research Partner

Jurnal Teknika dan Informatika (JTI)
E-ISSN: 2797-0000 | Vol. 6, No. 2 (2026) | pp. 177-191

OSIRA, an AI Early Warning System Using Open Source Intelligence for Digital Fraud Prevention

Nafila Kurnia Sari[1*], Alya Zaskya[2], Muhammad Zakiyyuddin Abdul Adhiim[3]

Universitas Gadjah Mada, Yogyakarta, Indonesia[1,2,3]

***Corresponding author:** nafilakurniasari@mail.ugm.ac.id |

Received: 21 September 2026 | Revised: 10 October 2026 | Revised: 28 October 2026 | Published: 6 November 2026

Abstract

Purpose: This study designs and evaluates OSIRA (Open Source Intelligence Risk Analyzer), an early warning system based on artificial intelligence that integrates Open Source Intelligence (OSINT) and Natural Language Processing (NLP) to identify digital fraud indications before they enter formal reporting mechanisms in a cashless society.

Research Methodology: The study uses a design science approach consisting of problem analysis, system design, risk score formulation, and a limited scale proof of concept simulation. The simulation used 540 synthetic texts from 50 template groups, divided into 394 training and 146 testing texts through a group based split, and classified them with Term Frequency–Inverse Document Frequency (TF IDF) features and logistic regression.

Results: OSIRA integrates web crawling, text preprocessing, fraud classification, entity extraction, a risk engine, and an early warning module. At a threshold of 0.50, the model achieved 82.88% accuracy, 79.66% precision, 98.95% recall, an F1 score of 88.26%, and a ROC AUC of 0.8528, but its false positive rate reached 47.06%, mainly for normal texts discussing fraud related topics.

Conclusions: OSIRA shows initial feasibility as an additional intelligence layer that can shift fraud mitigation from reactive response toward proactive fraud intelligence.

Limitations: The evaluation used synthetic rather than real OSINT data, a baseline classifier with limited contextual understanding, and risk score weights and thresholds that have not been empirically calibrated.

Contributions: The study offers a conceptual and technical framework for using pre reporting public signals in fraud risk management and provides guidance for banks, digital wallet providers, and regulators.

Keywords: *Early Warning System, Fraud Prevention, Natural Language Processing, Open Source Intelligence, Risk Scoring*

How to Cite: Sari, N. K., Zaskya, A., & Adhiim, M. Z. A. (2026). OSIRA, an AI Early Warning System Using Open Source Intelligence for Digital Fraud Prevention. *Jurnal Teknika dan Informatika (JTI)*, 6(2), 177-191.

1. Introduction

The development of information technology has driven the transformation of Indonesia's financial sector toward a cashless society. Transactions that previously relied on cash have shifted toward electronic money, digital wallets, mobile banking, the Quick Response Code Indonesian Standard (QRIS), and other digital payment services. This transformation offers convenience, speed, and efficiency while expanding public access to digital financial services. Globally, the fintech revolution has transformed payments and other financial services through new business models (Gomber, Kauffman, Parker, & Weber, 2018), and digital finance has been shown to support financial inclusion

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470

and stability in emerging economies (Ozili, 2018). The use of mobile payments, however, depends heavily on users' trust and their perceptions of security (Patil, Tamilmani, Rana, & Raghavan, 2020).

This development is reflected in the growth of electronic money transactions in Indonesia. According to Bank Indonesia, the value of electronic money transactions reached IDR 344.12 trillion in May 2026 (Bank, 2026). The high transaction value indicates growing public dependence on digital payment infrastructure. Yet the growth of the digital economy also presents a new paradox. The wider the use of digital financial services, the larger the space that criminals can exploit, targeting both technological weaknesses and human vulnerabilities.

One of the growing threats is social engineering combined with account takeover. Through social engineering, perpetrators manipulate victims into revealing sensitive information or credentials, or into making particular transactions. This information can then be used to take over accounts, misuse identities, and move funds through mule accounts. Digital fraud schemes are also becoming more adaptive through phishing, impersonation, and even the use of artificial intelligence by perpetrators (Otoritas, 2026). Research shows that phishing attacks on digital banking exploit both technical and behavioral weaknesses (Yuspin, Putri, Fauzie, & Pitaksantayothin, 2024), that susceptibility to cyber fraud is shaped by individual characteristics (Whitty, 2019), and that awareness alone is often insufficient to prevent victimization (Chen, Li, Chua, & Chen, 2026).

The urgency of the threat is reflected in data from the Indonesia Anti Scam Centre (IASC). From November 22, 2024, to January 2026, IASC received 432,637 fraud complaints with public losses reaching IDR 9.1 trillion. By July 31, 2026, the number of complaints had increased to 637,055 reports involving more than 1.17 million accounts reported as linked to fraud (Otoritas, 2026). The growth in complaints is shown in Figure 1.

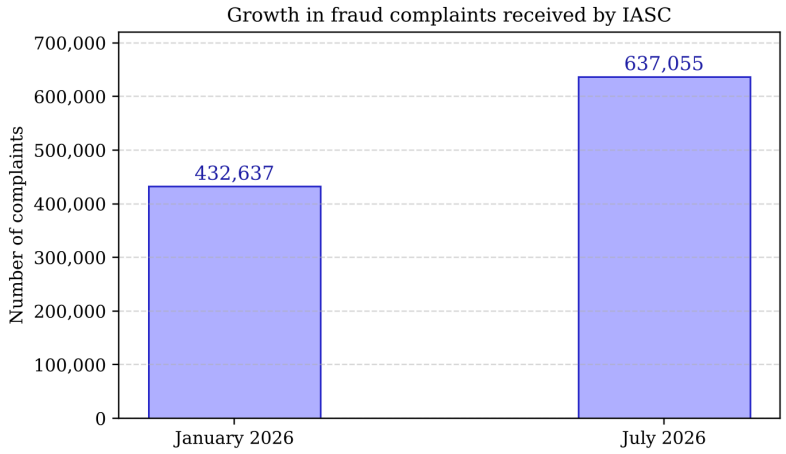


Figure 1. Growth in the number of fraud complaints received by the Indonesia Anti Scam Centre (IASC), processed from Otoritas Jasa Keuangan (2026)

Figure 1 shows that the number of fraud complaints received by IASC rose from 432.637 in January 2026 to 637.055 in July 2026, an increase of more than 200.000 complaints, or about 47%, in roughly six months. This sharp rise illustrates that the challenge of handling fraud lies not only in the large number of cases but also in the speed of response. IASC has established mechanisms for coordinating reporting, tracing, and blocking accounts. However, the effectiveness of these mechanisms is still affected by delays in victim reporting, the rapid movement of funds between accounts, and the use of temporary accounts and identities. These conditions create a time gap between the emergence of fraud indications and their recording in formal reporting mechanisms, which is the gap that this study seeks to narrow.

This gap is all the more important because early information about potential fraud often appears first in open digital spaces. Victims and members of the public share experiences about suspicious

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470

accounts, phone numbers, links, or schemes through social media, discussion forums, complaint sites, and other public platforms. This information, however, remains scattered and unstructured and has not been used systematically as a source of risk warnings. Blacklist based systems are effective once an entity has been identified, but they are limited in dealing with new accounts, mule accounts, or temporary identities that have not yet been recorded. This situation creates an information lag, namely a delay during which threat signals have already appeared in public spaces but have not yet been converted into risk information that can support transaction prevention.

One approach that can narrow this gap is Open Source Intelligence (OSINT), the process of collecting and analyzing information from open sources to produce valuable intelligence for decision making (Pastor-Galindo, Nespoli, & Martínez, 2020; Yadav, Kumar, & Singh, 2023). In the context of fraud, public posts about accounts, phone numbers, communication patterns, and fraud experiences can serve as early fraud signals. The large volume and diversity of such data, however, make manual analysis difficult. Artificial Intelligence (AI), particularly Natural Language Processing (NLP), is therefore needed to extract entities, understand context, and classify fraud indications automatically (Salloum, Gaber, Vadera, & Shaalan, 2022).

Fraud handling therefore needs to shift from reactive fraud response toward proactive fraud intelligence. Systems should not rely only on information after an account has entered formal databases, but should also use public signals that appear earlier as supporting information for estimating transaction risk. This study proposes OSIRA (Open Source Intelligence Risk Analyzer), an AI based early warning system that integrates OSINT and NLP to identify fraud indications from public digital traces. OSIRA is designed to collect information through web crawling, extract entities such as account or phone numbers, classify fraud indications, and generate dynamic risk scores. This information can then be integrated into mobile banking or digital wallet interfaces to provide warnings before transactions are completed.

The study addresses three research questions. The first asks how the OSIRA fraud mitigation system can be designed by integrating OSINT and NLP to identify digital fraud indications earlier in a cashless society. The second asks how NLP can automate the extraction, processing, and classification of OSINT data into risk information relevant for detecting fraud indications. The third asks how accurate OSIRA is, what its false positive rate is, and how effective its detection timing is compared with conventional reactive reporting mechanisms. Accordingly, the study aims to design OSIRA, to apply NLP to automate the conversion of OSINT data into risk information, and to evaluate the accuracy and false positive rate of OSIRA in a proof of concept simulation.

The study is expected to benefit several parties. For the public, OSIRA can raise awareness of risky transactions by providing early warnings before transactions are completed, thereby reducing financial losses from fraud. For financial service providers such as banks and digital wallet operators, OSIRA can serve as an additional intelligence layer that supports fraud mitigation by using risk signals from public information. For regulators and stakeholders, the study can serve as a reference for developing more proactive fraud prevention systems through the integration of public data, AI, and early warning mechanisms.

2. Literature Review and Hypothesis/es Development

2.1 Fraud and Fraud Prevention Strategies

Fraud is a deliberate act of deception carried out to obtain a benefit by misleading, manipulating, or harming others. In risk management, fraud must be managed through continuous identification, assessment, prevention, detection, investigation, and monitoring. The fraud risk management guide issued by the Committee of Sponsoring Organizations of the Treadway Commission and the Association of Certified Fraud Examiners emphasizes that fraud risk management focuses not only on handling incidents after they occur but also on governance, risk assessment, and the implementation of preventive and detective controls (Committee & Association, 2023). In the digital context, fraud

develops through various schemes such as phishing, impersonation, social engineering, account takeover, and the use of mule accounts (Abdallah, Maarof, & Zainal, 2016).

The fast and adaptive nature of digital fraud makes purely repressive approaches inadequate. Prevention aims to reduce opportunities for fraud, whereas detection aims to recognize fraud indications as early as possible so that responses can occur before losses grow (Sudarmanto, 2020). Reviews of financial fraud detection show that the volume of digital transactions and the rarity of fraud among legitimate transactions make manual and rule based detection increasingly ineffective (Hilal, Gadsden, & Yawney, 2022; West, & Bhattacharya, 2016). Fraud patterns also change as perpetrators adapt to new controls, so detection models trained on historical data gradually lose accuracy unless they are updated, a problem known as concept drift (Lu, Liu, Dong, Gu, Gama, & Zhang, 2019). Mitigation strategies therefore need to move from predominantly reactive mechanisms toward more preventive and proactive approaches, including the use of risk indicators and early warning systems.

2.2 Open Source Intelligence and Early Fraud Signals

OSINT is the process of collecting and analyzing publicly available information to produce intelligence that supports decision making. OSINT sources include websites, social media, discussion forums, public databases, and other open digital platforms. OSINT has been described as a largely unexploited source of value for cybersecurity, crime investigation, and social monitoring, although it raises challenges of data volume, reliability, and ethics (Pastor-Galindo et al., 2020). Yadav et al. (2023) explain that OSINT is increasingly used in cybersecurity and digital crime investigation because it can turn open data into information about threat patterns. In digital financial services, monitoring public information can also help identify consumer risks and detect early warning signals before problems spread (Duflos et al., 2023), and OSINT has been studied as an early warning instrument in other high stakes settings (Safira et al., 2026).

The use of early fraud signals allows fraud mitigation to move from dependence on historical data and blacklists toward more dynamic risk monitoring. Social media monitoring can provide relatively fast information about consumer problems and emerging risks, but it is limited because not every incident is reflected in public data (Duflos et al., 2023). OSINT data may also contain incomplete information, duplicates, or unverified context, so they require analysis and validation before use. In this study, OSINT is treated as a source of risk signals rather than as sole evidence of fraud, which allows it to support a more proactive early warning system.

2.3 Natural Language Processing for Fraud Detection

NLP is a branch of AI that enables computers to process, understand, and analyze human language in text or speech. AI is increasingly implemented in business processes, including customer service, risk analysis, and fraud prevention (Marsella, Wijaya, Wijaya, Shidqi, & Novita, 2023). Deep learning has substantially advanced text classification, although classical approaches such as term weighting combined with linear classifiers remain strong and interpretable baselines (Minaee, Kalchbrenner, Cambria, Nikzad, Chenaghlu, & Gao, 2021). In fraud detection, NLP can process unstructured text through preprocessing, classification, and entity extraction. NLP based approaches have been applied to detect online fraud by recognizing patterns that distinguish normal from suspicious activity, and their effectiveness has been evaluated with various performance measures (Boulieris, Pavlopoulos, Xenos, & Vassalos, 2024). A systematic review of phishing email detection shows that NLP techniques such as feature extraction, text representation, and machine learning classification are widely used to identify deceptive messages (Salloum, Gaber, Vadera, & Shaalan, 2022).

NLP enables large scale analysis to be performed automatically rather than through manual monitoring. In OSIRA, texts obtained from OSINT sources are classified according to fraud indications and combined with the extracted entities as the basis for risk information. Model effectiveness can be assessed through accuracy, precision, recall, F1 score, and false positive rate. Research on fraud detection emphasizes that these metrics must be interpreted in light of class

imbalance and the practical costs of errors, because a model that misses fraud and a model that raises too many false alarms impose different burdens on users and investigators (Dal, Boracchi, Caelen, Alippi, & Bontempi, 2018). AI based fraud detection is also moving toward hybrid and explainable systems that combine several techniques (Ali, Othman, Eisa, Al-Dhaqm, Nasser, Elhassan, Elshafie, & Saif, 2022; Yang, Shukur, & Sahran, 2026). NLP thus acts as the analytical component that converts unstructured OSINT text into structured information to support early fraud detection.

2.4 Early Warning Systems and Risk Scoring in Fraud Mitigation

An Early Warning System (EWS) is designed to recognize risk indicators at an early stage and issue warnings before a threat develops into larger losses. In finance, an EWS can use risk indicators and predictive analysis to detect changes in conditions early and support preventive decisions. Pehlivanlı et al. (2024) show that an overall risk scoring approach combining financial and nonfinancial indicators has high predictive power and can serve as an early warning system. A similar approach has been used to develop AI based financial risk warning models that identify risky conditions before their impact grows (Zhang, 2024).

To avoid purely binary warnings, an EWS can be combined with risk scoring, which assigns a risk value based on a combination of indicators. Combining several detection signals has been shown to improve fraud detection, as hybrid models that integrate different techniques outperform single models (Carcillo, Caelen, Kessaci, Oblé, & Bontempi, 2021), and ensemble methods reduce misclassification (Randhawa et al., 2018). The quality of the input features is also decisive, as aggregated behavioral features substantially improve detection performance (Correa Bahnsen et al., 2016). In OSIRA, the risk score is formed from report intensity, entity occurrence patterns, classification confidence, and links to other fraud indicators. Setting warning thresholds must be done carefully, because changes in thresholds affect system sensitivity and the timing of alerts (Pehlivanlı et al., 2024).

2.5 Explainability, Trust, and Human Oversight

Warnings generated by AI systems are useful only when users and institutions understand and trust them. Many machine learning models operate as black boxes whose reasoning is difficult to explain (Adadi & Berrada, 2018), and explainable AI seeks to make model decisions understandable through techniques that reveal which inputs contributed to a prediction (Barredo Arrieta et al., 2020). For high stakes decisions, interpretable models or traceable decision rules are preferred (Rudin, 2019). Explainability affects user trust and acceptance (Shin, 2021), and research on human trust in AI shows that both overreliance and underreliance on automated outputs can lead to poor decisions (Glikson & Woolley, 2020). Recent work on explainable fraud detection in online payments shows that transparent models can link fraud risk to observable user behavior (Erbuğa & Ünal, 2026). These findings support a human in the loop design in which AI provides risk information while the final decision remains with the user or the institution.

2.6 Research Gap

Previous studies have examined fraud detection using transaction data (Abdallah, Maarof, & Zainal, 2016; Hilal, Gadsden, & Yawney, 2022), NLP for phishing and online fraud detection (Salloum, Gaber, Vadera, & Shaalan, 2022), OSINT for cybersecurity (Yadav, Kumar, & Singh, 2023), and risk scoring for early warning (Pehlivanlı & Katanalp, 2024). However, these streams have rarely been integrated into a single system that uses public, pre reporting signals to warn users about risky transaction counterparties before a transaction is completed. Most fraud detection systems rely on transaction histories or blacklists that become available only after fraud has been reported, which leaves the information lag unaddressed. In Indonesia, where fraud complaints are rising rapidly and digital payments are expanding through QRIS and digital wallets (Nurqamarani, Fadilla, & Juliana, 2024), evidence on such an integrated approach remains limited. This study addresses the gap by designing OSIRA and testing its classification component in a proof of concept simulation.

3. Research Methodology

3.1 Research Design

This study uses a design science approach, which aims to create and evaluate an artifact that solves a practical problem. The research proceeded in four stages. The first stage was problem analysis, which examined the gap between the emergence of fraud signals in public spaces and their entry into formal reporting and mitigation mechanisms. The second stage was system design, which developed the OSIRA architecture by integrating OSINT data acquisition, NLP based classification and entity extraction, a risk engine, and an early warning module. The third stage was risk score formulation, which defined the indicators and the scoring rule used to convert NLP outputs and entity patterns into risk levels. The fourth stage was a limited scale proof of concept simulation to test the feasibility of the classification component, which is the core analytical element of OSIRA.

3.2 Simulation Dataset

To test the initial feasibility of OSIRA, a limited scale simulation was carried out using a synthetic dataset representing fraud and nonfraud texts. Synthetic data were used because real OSINT data on fraud reports raise privacy and verification issues that are beyond the scope of an initial proof of concept. The dataset was built from 50 template groups covering explicit fraud, implicit fraud, normal nonfraud texts, hard negatives, and ambiguous texts. Hard negatives are normal texts that use vocabulary similar to fraud reports, such as security education about one time passwords or phishing, while ambiguous texts contain incomplete or unclear context. After deduplication, 540 texts were obtained, consisting of 394 training texts and 146 testing texts.

The data were divided using a group based split, so variations of the same template did not appear in both the training and testing sets. This approach was used to reduce template leakage and to test the model's ability to generalize to text patterns it had not learned before. Group based splitting provides a more conservative estimate of performance than random splitting, because random splitting would allow near duplicate texts from the same template to appear in both sets and inflate accuracy.

3.3 Classification Model and Evaluation Metrics

The baseline model used Term Frequency-Inverse Document Frequency (TF IDF) features with logistic regression to classify texts as fraud or nonfraud. This combination was chosen because it is transparent, computationally light, and widely used as a strong baseline in text classification (Minaee et al., 2021), which makes it suitable for an initial proof of concept. Model performance was evaluated at a classification threshold of 0.50 using accuracy, precision, recall, F1 score, false positive rate, and the area under the Area Under the Curve (AUC) of the Receiver Operating Characteristic Curve (ROC). Recall was treated as the priority metric because, in an early warning system, failing to detect fraud is more costly than raising a warning that the user can verify. The false positive rate was monitored because frequent false alarms can erode user trust and cause warning fatigue. Performance was also analyzed by text category, and the effect of the classification threshold on precision, recall, F1 score, and false positive rate was examined across thresholds from 0.25 to 0.80.

3.4 Risk Score Formulation

The risk engine forms a risk score from four main indicators, namely NLP classification confidence (C), entity occurrence frequency (F), source diversity (S), and report recency (T). Each indicator is normalized to a range of 0 to 100 so that the indicators can be compared proportionally. Conceptually, the OSIRA risk score is formulated in Formula 1.

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470

$$R = \sum_{k=1}^{4} w_k x_k \quad (1)$$

Here R is the final risk score and w_1 to w_4 are the weights of the indicators. At the simulation stage, the weights are not treated as fixed values but as parameters that can be tested under several scenarios to examine their effect on system performance. This approach is consistent with the concept of risk scoring that combines several indicators into an integrated measure of risk (Pehlivanlı et al., 2024). The final score is classified into three risk levels as initial simulation parameters, with scores of 0 to 39 categorized as low risk, 40 to 69 as medium risk, and 70 to 100 as high risk. Entities with medium or high risk trigger the early warning system before a transaction is completed. These thresholds are experimental and can be adjusted based on test results, score distributions, and the balance between false positives and false negatives.

4. Results and Discussions

4.1 Gap Analysis of Current Fraud Mitigation

Current fraud mitigation systems generally rely on a combination of user reporting, verification, transaction monitoring, and account blocking after a fraud indication has been identified. These mechanisms are important for reducing losses and stopping fund flows, but they depend heavily on information that has already entered formal systems. In practice, the effectiveness of the response can be affected by delays in victim reporting, the rapid movement of funds between accounts, and the use of new or mule accounts by perpetrators. These conditions create an information lag between the emergence of early fraud indications and the availability of information that can be used in formal mitigation.

At the same time, signals of fraud activity often appear earlier in open digital spaces through victim posts, discussion forums, complaint sites, and social media. This information can serve as early fraud signals, but it is not yet systematically integrated into transaction risk assessment. The main weakness of current systems is therefore not the absence of response mechanisms but the limited ability to use pre reporting information as a basis for earlier risk identification. An approach is needed that complements conventional mechanisms by using open data dynamically. Integrating OSINT, NLP, and risk scoring allows previously scattered and unstructured public information to be converted into risk indicators that can be processed automatically. This approach is consistent with the principles of fraud risk management, which emphasize preventive and detective controls (Committee of Sponsoring Organizations of the Treadway Commission & Association of Certified Fraud Examiners, 2023), and with the early warning concept, which focuses on identifying risk before its impact grows (Pehlivanlı et al., 2024).

4.2 OSIRA Design and Architecture

OSIRA is designed as an early warning system that uses public information to identify fraud indications before an account or digital identity is formally recorded as a risky entity. The system integrates four main components, namely OSINT based data collection, text analysis using NLP, risk score formation, and early warnings to users. Architecturally, the process begins with collecting public data through web crawling of open sources such as social media, discussion forums, and complaint sites. The data then undergo preprocessing to clean and normalize the text before being analyzed by the NLP module, which classifies fraud indications and extracts important entities such as account numbers, phone numbers, and links. The results are passed to the risk engine, which forms a risk score based on model confidence, entity frequency, source diversity, and report recency. The architecture is presented in Figure 2.

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470

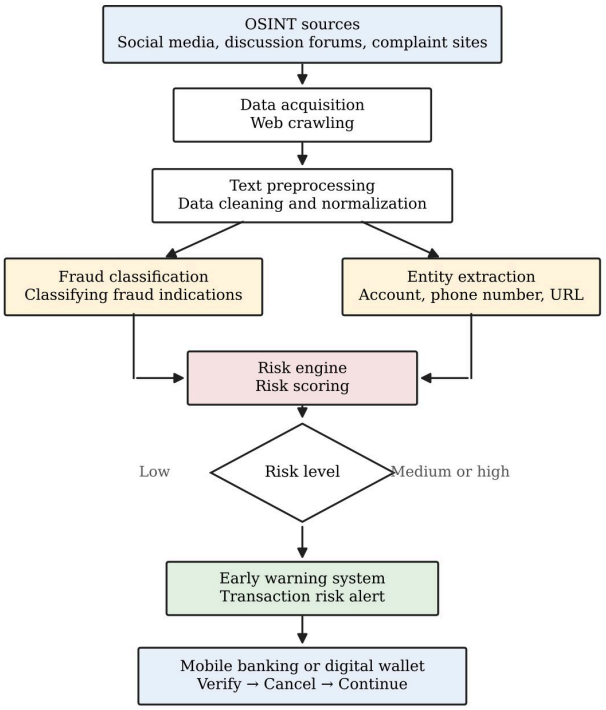


Figure 2. Architecture of the OSIRA system

Figure 2 shows that the OSIRA workflow moves from OSINT sources through data acquisition and text preprocessing, after which the NLP module splits into two parallel functions, fraud classification and entity extraction, whose outputs are combined in the risk engine. The risk engine assigns a risk level, and entities with low risk do not trigger warnings, whereas entities with medium or high risk are passed to the early warning module. The warning is then delivered through the mobile banking or digital wallet interface before the transaction is completed, and the user can choose to verify the counterparty, cancel the transaction, or continue. OSIRA is not designed to label an account as a perpetrator automatically but to provide supporting information so that users can carry out additional verification. The architecture thus functions as an additional intelligence layer that connects early fraud signals in public spaces with transaction mitigation in a more proactive way.

4.3 Detection Mechanism and Risk Scoring

The detection mechanism of OSIRA begins with collecting public data through web crawling of OSINT sources. The collected text is preprocessed, and the NLP module then performs fraud classification, which assesses the likelihood that a text contains a fraud indication, and entity extraction, which recognizes account numbers, phone numbers, and links. The detected entities are then aggregated according to occurrence patterns, report frequency, source diversity, and report recency. The four indicators that form the OSIRA risk score are summarized in Table 1.

Table 1. Indicators forming the OSIRA risk score

Indicator	Symbol	Description
NLP confidence level	C	Measures the model's confidence that a text contains a fraud indication
Entity occurrence frequency	F	Measures how often the same account, phone number, or entity appears in public reports
Source diversity	S	Measures the number of independent sources that report fraud indications for the same entity
Report recency	T	Measures how recently reports have appeared in order to capture increasing risk in the latest period

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470

Table 1 shows that the OSIRA risk score combines one indicator derived from the NLP classifier with three indicators derived from the aggregation of extracted entities. The NLP confidence level captures the textual evidence of fraud in individual posts, while entity occurrence frequency and source diversity capture how widely and independently the same entity is reported, which helps prevent a single unverified post from producing a high risk score. Report recency gives greater weight to recent reports, reflecting the fact that fraud accounts are often used for a short period before being abandoned. Combining these four indicators means that the NLP output is not used as the sole basis for a warning but as one signal among several, which is important given the false positive behavior of the classifier reported below.

4.4 Classification Performance

The evaluation results of the baseline model at a classification threshold of 0.50 are presented in Table 2.

Table 2. Evaluation results of the OSIRA classification model

Metric	Result
Accuracy	82.88%
Precision	79.66%
Recall	98.95%
F1 score	88.26%
False positive rate	47.06%
ROC AUC	0.8528

Table 2 shows that the model achieved a recall of 98.95%, meaning that it recognized almost all texts that actually contained fraud indications. This result is relevant to the function of OSIRA as an early warning system, because failing to detect fraud would leave a threat without a warning. The F1 score of 88.26% indicates a reasonably good balance between precision and recall, and the ROC AUC of 0.8528 shows that the model separates fraud from nonfraud texts considerably better than chance across thresholds. However, the false positive rate of 47.06% shows that the model still frequently classifies nonfraud texts as fraud, especially texts that use vocabulary similar to fraud reports. The distribution of correct and incorrect predictions is shown in Figure 3.

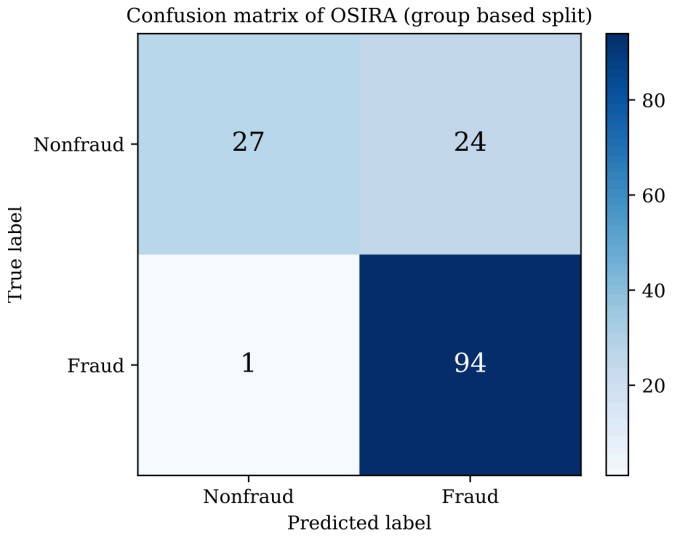


Figure 3. Confusion matrix of the OSIRA classification model

Figure 3 shows that, of the 95 fraud texts in the test set, 94 were correctly classified as fraud and only 1 was missed, which explains the very high recall. Of the 51 nonfraud texts, 27 were correctly

classified as nonfraud, while 24 were incorrectly flagged as fraud, which produces the false positive rate of 47.06% and a specificity of only about 53%. The matrix therefore makes clear that the model's errors are concentrated almost entirely on the nonfraud side. For an early warning system this pattern is preferable to missing fraud, but in practice it means that nearly half of legitimate entities discussed in fraud related language could receive unnecessary warnings if the classifier were used alone.

Analysis by data category shows that the model achieved 100% accuracy for explicit fraud, 96.67% for implicit fraud, and 100% for ambiguous fraud. In contrast, performance dropped to 41.67% for ambiguous normal texts and 32.00% for hard normal texts. These findings indicate that the model is relatively good at recognizing fraud patterns but still struggles to distinguish fraud reports from normal texts that discuss similar topics, such as education about one time passwords, phishing, or security warnings. This limitation is typical of bag of words representations, which capture vocabulary but not context (Minaee et al., 2021).

4.5 Effect of the Classification Threshold

The threshold analysis shows a trade off between sensitivity and the rate of false alarms, as presented in Figure 4.

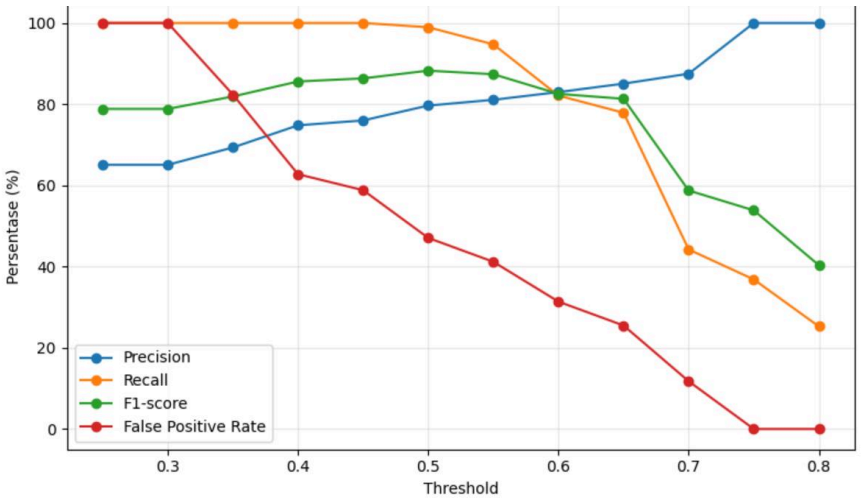


Figure 4. Effect of the classification threshold on the performance of the OSIRA model

Figure 4 shows that at low thresholds of 0.25 to 0.30 the model reaches a recall of 100% but also a false positive rate of 100%, meaning that every text is flagged as fraud. As the threshold rises, the false positive rate falls steadily, to about 47% at 0.50, about 31% at 0.60, and to zero at 0.75 and above, while precision increases to 100% at the highest thresholds. Recall remains close to 100% up to a threshold of about 0.55, but it then declines sharply, falling below 50% at 0.70 and to about 25% at 0.80. The F1 score peaks at around 88% near a threshold of 0.50 and declines at higher thresholds, which confirms that 0.50 offers the best balance for the classifier alone. These results show that lowering the threshold increases recall at the cost of more false alarms, while raising it suppresses false alarms at the cost of missing more fraud.

For this reason, the NLP classification result in OSIRA is not used as a single decision but as one indicator in the risk engine, together with report frequency, source diversity, and report recency. This integration is expected to reduce dependence on a single classifier output and produce more proportional risk warnings. The trade off also shows why threshold selection should be treated as a policy decision that weighs the cost of missed fraud against the cost of user friction, rather than as a purely technical parameter (Dal Pozzolo et al., 2018).

4.6 Effectiveness and Implementation

The simulation results indicate that OSIRA has potential as a supporting system for fraud mitigation oriented toward early detection. The high recall shows strong sensitivity in capturing fraud indications, although the high false positive rate confirms that the NLP output cannot yet serve as the sole basis for decisions. The effectiveness of OSIRA therefore lies in integrating NLP classification with risk scoring that considers entity frequency, source diversity, and report recency, which allows an indication to be treated first as a risk signal requiring further validation rather than immediately as fraud.

Unlike blacklist based approaches that depend on entities that have already been identified, OSIRA uses pre reporting information to form early fraud signals. This approach can narrow the information lag between the appearance of fraud indications in public spaces and the entry of information into formal mechanisms. OSIRA is not intended to replace IASC, the internal detection systems of banks, or regulatory verification processes, but to function as an additional intelligence layer that strengthens the preventive approach. In implementation, OSIRA can be integrated through an Application Programming Interface (API) into mobile banking or digital wallet applications. When a user enters the destination account or identity, the system queries the OSIRA risk database, and if the score exceeds a certain threshold, the user receives a warning together with risk information before the transaction is completed. The user can then continue, carry out additional verification, or cancel the transaction. This intervention model preserves the human in the loop principle, under which the final decision remains with the user rather than being determined solely by an automated system (Glikson & Woolley, 2020).

4.7 Discussion

The findings show that integrating OSINT and NLP can provide an initial basis for proactive fraud intelligence, but they also reveal the limits of a simple text classifier. The very high recall is consistent with research showing that NLP can recognize fraud related language effectively (Salloum, Gaber, Vadera, & Shaalan, 2022), while the high false positive rate reflects a well known challenge in fraud detection, namely that the rarity and diversity of legitimate behavior make it easy for models to raise false alarms (Dal, Boracchi, Caelen, Alippi, & Bontempi, 2018; West, & Bhattacharya, 2016). The difficulty with hard negatives in particular shows that vocabulary alone cannot distinguish a fraud report from a security warning, which points to the need for models with stronger contextual understanding (Minaee, Kalchbrenner, Cambria, Nikzad, Chenaghlu, & Gao, 2021).

The design choice to treat NLP outputs as one input to a risk engine rather than as a final verdict is consistent with research showing that combining several detection signals improves fraud detection (Kim, Lee, Shin, Yang, Cho, Nam, Song, Yoon, & Kim, 2019) and with early warning research that relies on integrated risk scores (Pehlivanlı & Katanalp, 2024). It also responds to concerns about explainability and trust. Users are more likely to act on warnings when they understand why a counterparty is considered risky (Shin, 2021), so OSIRA warnings should display the reasons behind the score, such as the number of independent reports and how recent they are, in line with recommendations for explainable fraud detection (Erbuğa & Ünal, 2026).

Finally, the use of public data raises governance questions. OSINT can contain false accusations, coordinated misinformation, or personal data, so a system that assigns risk scores to accounts and phone numbers must include safeguards against defamation, misuse, and privacy violations (Pastor-Galindo, Nespoli, & Martínez, 2020). Multisource validation, transparent scoring rules, appeal mechanisms for wrongly flagged entities, and compliance with data protection rules are therefore necessary conditions for operational use. Because fraud schemes evolve, the models must also be monitored and retrained to cope with concept drift (Lu, Liu, Dong, Gu, Gama, & Zhang, 2019).

5. Conclusions

This study designed OSIRA (Open Source Intelligence Risk Analyzer), a fraud mitigation system that integrates OSINT, NLP, risk scoring, and an early warning system to identify fraud indications from public digital traces before they enter formal reporting mechanisms. NLP processes unstructured OSINT data through text classification and entity extraction, and its output is combined in the risk engine with entity frequency, source diversity, and report recency, so classification serves as one risk signal rather than a single decision. The proof of concept simulation using TF IDF and logistic regression produced an accuracy of 82.88%, precision of 79.66%, recall of 98.95%, an F1 score of 88.26%, and a ROC AUC of 0.8528. The high recall shows strong initial sensitivity, but the false positive rate of 47.06% shows difficulty in distinguishing fraud reports from nonfraud texts with similar language. OSIRA therefore shows potential as an additional intelligence layer that supports a shift from reactive to proactive fraud mitigation.

This study is limited by its use of synthetic data, a baseline classifier with limited contextual understanding, uncalibrated risk score weights and thresholds, and the absence of tests on detection time, user responses, and legal and privacy implications. Future development should use real OSINT data, adopt models with stronger contextual understanding such as transformer based language models, calibrate risk scoring with empirical data, and apply multisource validation. Implementation should involve collaboration among financial service providers, regulators, fraud reporting operators, and digital platforms, beginning with limited trials or regulatory sandboxes before integration with mobile banking and digital wallets.

Acknowledgement

The authors thank the Accounting Student Association, Faculty of Economics and Business, Universitas Bangka Belitung, as the organizer of the National Accounting Fair 2026, and all parties who supported the preparation of this study.

Author Contributions

NKS contributed to the conceptualization of the study, the design of the OSIRA system, the simulation and analysis, and the writing and revision of the manuscript. AZ contributed to the conceptualization of the study, the design of the OSIRA system, the simulation and analysis, and the writing and revision of the manuscript. MZAA contributed to the conceptualization of the study, the design of the OSIRA system, the simulation and analysis, and the writing and revision of the manuscript. All authors read and approved the final manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68, 90–113*. https://doi.org/10.1016/j.jnca.2016.04.007

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6, 52138–52160*. https://doi.org/10.1109/ACCESS.2018.2870052

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, *12*(19). https://doi.org/10.3390/app12199637

Bank Indonesia. (2026). Statistik sistem pembayaran dan infrastruktur pasar keuangan (SPIP). Retrieved from https://www.bi.go.id/id/statistik/ekonomi-keuangan/spip/Default1.aspx

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58, 82–115*. https://doi.org/10.1016/j.inffus.2019.12.012

Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2024). Fraud detection with natural language processing. *Machine Learning, 113, 5087–5108*. https://doi.org/10.1007/s10994-023-06354-5

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences, 557, 317–331*. https://doi.org/10.1016/j.ins.2019.05.042

Chen, L., Li, H., Chua, Y. P., & Chen, J. (2026). When awareness is not enough: Online fraud susceptibility, threat awareness, anti-fraud self-efficacy and online scam prevention behavior among Chinese university students. *Frontiers in Psychology, 17, Article 1857855*. https://doi.org/10.3389/fpsyg.2026.1857855

Committee of Sponsoring Organizations of the Treadway Commission & Association of Certified Fraud Examiners. (2023). *Fraud risk management guide (2nd ed.)*.

Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51, 134–142*. https://doi.org/10.1016/j.eswa.2015.12.030

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3784-3797. https://doi.org/10.1109/TNNLS.2017.2736643

Duflos, E., Izaguirre, J. C., Datwani, L., Mishra, A., Collins, D., Bathula, P. R., Mathias, L., & Sawant, D. (2023). Social media monitoring to assess consumer risks in digital credit apps: Guidance for supervisors from an India pilot. *Consultative Group to Assist the Poor*.

Erbuğa, G. S., & Ünal, C. (2026). Explainable artificial intelligence-driven fraud detection: Behavioral proxies of digital financial literacy in online payment models. *Borsa Istanbul Review, 26, Article 100840*. https://doi.org/10.1016/j.bir.2026.100840

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of Management Information Systems*, *35*(1), 220-265. https://doi.org/10.1080/07421222.2018.1440766

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications, 193, Article 116429*. https://doi.org/10.1016/j.eswa.2021.116429

Kim, E., Lee, J., Shin, H., Yang, H., Cho, S., Nam, S., Song, Y., Yoon, J., & Kim, J. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications, 128, 214–224*. https://doi.org/10.1016/j.eswa.2019.03.042

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346-2363. https://doi.org/10.1109/TKDE.2018.2876857

Marsella, M., Wijaya, C. S., Wijaya, I., Shidqi, M. T., & Novita, D. (2023). Analisis implementasi artificial intelligence untuk bisnis: Systematic literature review. *DEVICE: Journal of Information System, Computer Science and Information Technology*, *4*(2). https://doi.org/10.46576/device.v4i2.4037

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3). https://doi.org/10.1145/3439726

Nurqamarani, A. S., Fadilla, S., & Juliana, A. (2024). Revolutionizing payment systems: The integration of TRAM and trust in QRIS adoption for micro, small, and medium enterprises in Indonesia. *Journal of Information Systems Engineering and Business Intelligence*, *10*(3), 314-327. https://doi.org/10.20473/jisebi.10.3.314-327

Otoritas Jasa Keuangan. (2026). Indonesia Anti-Scam Centre berhasil kembalikan Rp161 miliar dana masyarakat korban scam. Retrieved from https://ojk.go.id/id/berita-dan-kegiatan/siaran-pers/Pages/IASC-Berhasil-Kembalikan-Rp161-Miliar-Dana-Masyarakat-Korban-Scam.aspx

Ozili, P. K. (2018). Impact of digital finance on financial inclusion and stability. *Borsa Istanbul Review*, *18*(4), 329-340. https://doi.org/10.1016/j.bir.2017.12.003

Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., & Martínez Pérez, G. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access, 8, 10282–10304*. https://doi.org/10.1109/ACCESS.2020.2965257

Patil, P., Tamilmani, K., Rana, N. P., & Raghavan, V. (2020). Understanding consumer adoption of mobile payment in India: Extending Meta-UTAUT model with personal innovativeness, anxiety, trust, and grievance redressal. *International Journal of Information Management, 54, Article 102144*. https://doi.org/10.1016/j.ijinfomgt.2020.102144

Pehlivanlı, D., Aykaç Alp, E., & Katanalp, B. (2024). Introducing the overall risk scoring as an early warning system. *Expert Systems with Applications, 246, Article 123232*. https://doi.org/10.1016/j.eswa.2024.123232

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. *IEEE Access, 6, 14277–14284*. https://doi.org/10.1109/ACCESS.2018.2806420

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Safira, S., Sembiring, A. M. C., Firdaus, M. Z., Tsabitah, N. D., Longdong, G. A. C., & Solihin, I. (2026). Peran open-source intelligence (OSINT) sebagai early warning Ukraina menjelang invasi Rusia Februari tahun 2022 [The role of open source intelligence (OSINT) as an early warning for Ukraine before the Russian invasion of February 2022]. *Jurnal Wacana Sosial Nusantara*. Retrieved from https://ejournal.lembagarisetmandirinusantara.or.id/index.php/jwsn/article/view/122/153

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access, 10, 65703–65727*. https://doi.org/10.1109/ACCESS.2022.3183083

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146, Article 102551*. https://doi.org/10.1016/j.ijhcs.2020.102551

Sudarmanto, E. (2020). Manajemen risiko: Deteksi dini upaya pencegahan fraud. *Jurnal Ilmu Manajemen*, *9*(2), 107-121. https://doi.org/10.32502/jimn.v9i2.2506

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57, 47–66*. https://doi.org/10.1016/j.cose.2015.09.005

Whitty, M. T. (2019). Predicting susceptibility to cyber-fraud victimhood. *Journal of Financial Crime*, *26*(1), 277-292. https://doi.org/10.1108/JFC-10-2017-0095

Yadav, A., Kumar, A., & Singh, V. (2023). Open-source intelligence: A comprehensive review of the current state, applications and future perspectives in cyber security. *Artificial Intelligence Review, 56, 12407–12438*. https://doi.org/10.1007/s10462-023-10454-y

Yang, H., Shukur, Z., & Sahran, S. (2026). A review of artificial intelligence for financial fraud detection. *Applied Sciences*, *16*(4). https://doi.org/10.3390/app16041931

Yuspin, W., Putri, A. O., Fauzie, A., & Pitaksantayothin, J. (2024). Digital banking security: Internet phishing attacks, analysis and prevention of fraudulent activities. *International Journal of Safety and Security Engineering*, *14*(6), 1699-1706. https://doi.org/10.18280/ijsse.140605

Zhang, W. (2024). Dynamic monitoring of financial security risks: A novel China financial risk index and an early warning system. *Economics Letters, 234, Article 111445*. https://doi.org/10.1016/j.econlet.2023.111445

© 2026 Abnus | Open Access — CC BY 4.0 | abnus.co | DOI:10.52909/jti.v6i2.470