Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

A Phishing Detection Model Based on Hybrid Natural Language Processing and Graph-Based Learning for Digital Threat Identification

Hermansyah*, Al–Khaidar, Husni

Informatics Study Program, Faculty of Computer Science, Universitas Almuslim, Aceh, Indonesia
Email: [1,*]waxherman@gmail.com, [2]alkhaidar@umuslim.ac.id, [3]husniabda7@gmail.com
Correspondence Author Email: waxherman@gmail.com
Submitted: 07/08/2026; Accepted: 31/08/2026; Published: 31/08/2026

Abstract–Phishing remains one of the most damaging digital threats in Indonesia, yet most detectors rely on a single source of evidence that attackers can easily disguise. This study proposes a hybrid model that combines natural language processing of message text (TF-IDF with logistic regression), graph-based learning on a heterogeneous message–sender–URL-domain graph (label propagation, neighbour label ratios, and graph-aggregated text evidence), and eleven URL and behavioural attributes, fused by XGBoost trained with modality-dropout augmentation (MD-Hybrid). A stratified sample of 7,604 Indonesian-language messages (3,802 phishing, 3,802 legitimate) was split 80:20 and validated with 5-fold cross validation under a realistic adversarial protocol in which the domain reputation score is unavailable and 30% of phishing messages disguise one to three evidence channels. Without camouflage, all models were near-perfect (accuracy 0.998 to 1.000), so clean accuracy could not discriminate between them. Under the adversarial protocol, the NLP, URL, graph, and early-fusion models reached accuracies of 0.896 to 0.901, F1-scores of 0.884 to 0.890, and MCC values of 0.810 to 0.818, whereas MD-Hybrid reached an accuracy of 0.950, an F1-score of 0.947, and an MCC of 0.904 without false positives; its remaining errors were messages whose three channels were all disguised. Modality-dropout fusion prevents shortcut learning and keeps detection reliable while at least one evidence channel remains genuine; validation on additional Indonesian phishing sources is recommended before deployment.

Keywords: Phishing Detection; Natural Language Processing; Graph-Based Learning; Hybrid Model; Cyber Threat

1. INTRODUCTION

Phishing has become one of the most damaging digital threats in Indonesia, and the scale of the problem is now measured in trillions of rupiah. The Indonesia Anti-Scam Centre (IASC) coordinated by the Financial Services Authority (OJK) received 373,129 fraud reports with reported losses of Rp 8.2 trillion between its launch in November 2024 and 30 November 2025, involving 619,394 reported accounts [1]. The Indonesia Anti-Phishing Data Exchange (IDADX) recorded 26,675 phishing attacks targeting .id domains in the first quarter of 2023 alone, an increase of about 220% over the previous quarter, bringing the five-year total to 69,117 attacks [2]. Social media (45%), financial institutions (31%), and retail or e-commerce (20%) were the most targeted sectors, and 99% of the phishing domains already used HTTPS, so the padlock icon no longer signals safety for ordinary users [2], [3]. Globally, the Anti-Phishing Working Group observed 989,123 phishing attacks in the fourth quarter of 2024 alone [4]. These figures show that Indonesia needs automatic detection that is accurate, fast, and difficult for attackers to evade.

The central difficulty is that phishers deliberately disguise the very signal on which a detector relies. A message can be rewritten in a polite, official tone; a malicious link can be hosted on a compromised but reputable domain; and campaign infrastructure can be rotated within hours [5], [6]. Detectors that depend on a single source of evidence therefore fail as soon as that source is camouflaged. Large language models make this worse, because they can generate fluent and personalised phishing text that is hard to distinguish from legitimate correspondence [7]. A robust detector must combine several independent views of a message, namely what it says, where it comes from and links to, and how its URL and delivery behave, and it must keep working when one of these views is disguised.

Natural language processing (NLP) is the most widely studied view. A systematic review of phishing email detection shows that TF-IDF, word embeddings, and transformer encoders dominate the literature and achieve high accuracy on benchmark corpora [8], while deep architectures such as convolutional and recurrent networks have been organised into a detailed taxonomy [5]. Hybrid deep learning has also been applied to SMS phishing (smishing) [9], and BERT-based models have been used for Indonesian-language spam detection [10]. Lexical models are also common in Indonesian sentiment and text classification studies, including Naive Bayes [11], logistic regression [12], and convolutional neural networks [13]. URL-centred detectors form the second view; PhiUSIIL, for example, combines a similarity index with incremental learning on a large URL dataset [14], and visual reference-based systems such as Phishpedia and PhishIntention identify brand impersonation from webpage appearance and dynamic behaviour [15], [16].

Graph-based learning offers a third and complementary view. Phishing campaigns reuse senders, domains, and hosting infrastructure, so the relations between entities carry evidence that cannot be seen in a single message. PhishGNN models the hyperlink structure of webpages with graph neural networks [17], Ariyadasa et al. combine recurrent-convolutional and graph convolutional networks over URL and HTML structure [18], and network embedding has been used to detect phishing accounts in Ethereum transaction graphs [19]. Surveys and

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

benchmarks on graph anomaly detection confirm that relational information improves the detection of rare malicious entities, but also show that heterophily and camouflage remain open problems [20], [21].

Despite this progress, three research gaps remain. First, most studies optimise one modality, and the few hybrid designs are usually evaluated only on clean test data where every modality is already informative, so the benefit of fusion is not measured under realistic evasion. Second, multimodal models are known to rely on the strongest modality and to degrade sharply when that modality is missing or corrupted [22], [23], but this shortcut behaviour has rarely been examined for phishing detection. Third, benchmark results are often inflated by features that leak the label, a problem identified as a major cause of irreproducible machine-learning findings [24]; leakage control is seldom reported in Indonesian phishing studies.

This study addresses these gaps by proposing a phishing detection model based on hybrid NLP and graph-based learning. The objectives are (1) to build an NLP component that converts Indonesian message text into phishing evidence, (2) to build a heterogeneous message–sender–URL-domain graph and derive leakage-safe graph features, including graph-aggregated text evidence, (3) to fuse NLP, graph, and URL/behavioural evidence with XGBoost trained using modality-dropout augmentation, and (4) to evaluate the model under clean conditions and under text, infrastructure, URL, and combined camouflage. The contributions are a reproducible hybrid pipeline, a leakage-controlled evaluation protocol, and quantitative evidence of how modality-dropout fusion prevents shortcut learning and preserves detection when attackers disguise part of a phishing message.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research follows six stages, as shown in Figure 1. The first stage samples the dataset; the second extracts three groups of evidence, namely NLP evidence from the message text, graph evidence from the heterogeneous entity graph, and URL/behavioural attributes; the third builds leakage-safe features in which labels are used only from training messages; the fourth trains single-modality baselines, an early-fusion hybrid, and the proposed modality-dropout hybrid (MD-Hybrid); the fifth evaluates all models on a stratified hold-out set and with 5-fold cross validation; and the sixth stress-tests the models with camouflage scenarios and interprets feature importance.

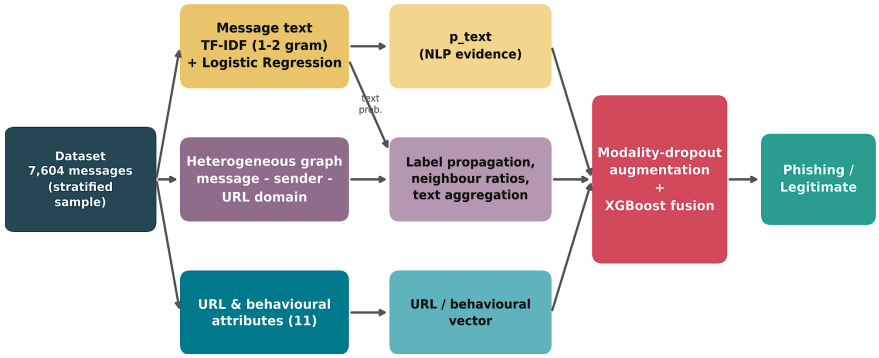

Figure 1. Research framework of the proposed hybrid NLP and graph-based phishing detection model

2.2 Dataset and Sampling

The study uses a phishing dataset of Indonesian-language messages that contains 10,001 Indonesian-language messages (5,001 phishing and 5,000 legitimate), 20 attributes, and 30,003 relational edges linking each message to its sender domain and URL domain. Sender and URL domains are anonymised with placeholder names in the .test and .example namespaces, and no live links are retained. A stratified random sample of 7,604 messages was drawn with random_state = 42, giving 3,802 phishing and 3,802 legitimate messages. The sample contains 20 distinct message texts derived from five base templates per class, 600 sender domains, and 1,396 URL domains. Table 1 summarises the attribute groups.

Table 1. Attribute groups of the sampled dataset and their use in this study

Group	Attributes	Use
Label	label (Phishing / Legitimate)	Target
Text (NLP)	message_text, tokens	TF-IDF + logistic regression
Relational (graph)	message_id, sender_domain, url_domain, Graph_Edges	Heterogeneous graph features
URL attributes	url_length, special_char_count, redirect_count, https_flag, subdomain_count	URL/behavioural block

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

Group Label	Attributes label (Phishing / Legitimate)	Use Target
Behavioural attributes	urgency_keyword_count, credential_request, uppercase_ratio, message_length, has_attachment, send_hour	URL/behavioural block
Excluded	community_id (label prefix), sender_url_shared_domain (constant), node_degree (precomputed), domain_reputation_score (unavailable for new domains)	Leakage and realism control

Two attributes were excluded to prevent label leakage [24]. The community identifier encodes the class directly in its prefix, and the sender–URL shared-domain flag is constant. The precomputed node degree was also excluded because it duplicates the graph structure that is recomputed in this study. The domain reputation score was also excluded from all models, although it is reported descriptively in Table 2, because reputation feeds rarely contain newly registered phishing domains and because this single attribute separates the classes perfectly, which would hide the contribution of the other evidence. The excluded attributes are listed in Table 1 so that the evaluation can be reproduced.

2.3 NLP Component

Message tokens were represented with TF-IDF unigrams and bigrams with sublinear term frequency, as defined in Equation (1), where $\mathrm{tf}_{t,d}$ is the frequency of term t in message d, N is the number of messages, and df_t is the number of messages containing t [25]. A logistic regression classifier converts the TF-IDF vector into a phishing probability p_{text}. For training messages, p_{text} is produced out-of-fold with an inner 5-fold split, so the fusion model never sees a probability computed from the message's own label.

$$w_{t,d} = \left(1 + \log tf_{t,d}\right) \times \log \frac{N}{df_t} \quad (1)$$

2.4 Graph-Based Learning Component

A heterogeneous graph was built with three node types (message, sender domain, and URL domain) and edges from each message to its sender and URL domain. Label propagation spreads the known labels of training messages through entity nodes, as shown in Equation (2), where s_m is the phishing score of message m, $N(m)$ is the set of entities connected to m, and scores of training messages are clamped to their labels; ten iterations were used. For each sender and URL domain, a smoothed phishing ratio was computed with Equation (3), where n_e is the number of labelled neighbouring messages and n_e^ph^ the number of phishing messages among them. Following the neighbourhood-aggregation idea of graph neural networks [21], a label-free graph-aggregated text score was computed with Equation (4), which averages the NLP probability of all messages connected to an entity.

$$s_m^{(k+1)} = \frac{1}{|N(m)|}\sum_{e \in N(m)} s_e^{(k)}, \quad s_e^{(k)} = \frac{1}{|N(e)|}\sum_{j \in N(e)} s_j^{(k)} \setminus \quad (2)$$

$$r_e = \frac{n_e^{ph}+1}{n_e+2} \quad (3)$$

$$a_e = \frac{1}{|N(e)|}\sum_{m \in N(e)} p_{text}(m) \quad (4)$$

The graph features of each message are its label-propagation score, the phishing ratios of its sender and URL domain, the degrees of both entities, the number of distinct URL domains used by the sender, and the aggregated text scores of the sender, the URL domain, and their mean. For training messages, all label-dependent graph features were computed out-of-fold, with the labels of the current fold hidden, so that no message contributes its own label to its features.

2.5 Fusion with Modality-Dropout Augmentation

The early-fusion hybrid concatenates p_{text}, the nine graph features, and the eleven URL/behavioural attributes and trains XGBoost (300 trees, maximum depth 4, learning rate 0.1). When one modality already separates the classes, such a model tends to rely on it and ignore the others [22], [26]. The proposed MD-Hybrid prevents this shortcut by modality-dropout augmentation. For every phishing training message, six additional copies were created in which one or two of the three modality blocks were replaced by the corresponding block of a randomly selected legitimate training message, while the phishing label was kept. Sample weights rebalance the augmented classes. The model is thus forced to recognise phishing from any modality that still carries genuine evidence, which reflects the way attackers disguise only part of a message [23].

2.6 Evaluation and Camouflage Scenarios

The sample was split in a stratified manner into 6,083 training and 1,521 testing messages, and stability was assessed with 5-fold stratified cross validation. Performance was measured with accuracy, precision, recall, F1-score, the area under the ROC curve (AUC), and the Matthews correlation coefficient (MCC), as given in Equations (5) to (8). MCC was included because it remains informative when one class dominates the errors.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 182
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

Four camouflage scenarios were applied to 10% to 50% of the phishing messages in the test set, while all models were trained on clean data. In text camouflage, the message text was replaced with the text of a legitimate message; in infrastructure camouflage, the sender and URL domains were replaced with legitimate entities, simulating compromised accounts and domains; in URL/behavioural camouflage, the eleven attributes were replaced with those of a legitimate message; and in combined camouflage, each attacked message had one, two, or all three randomly chosen modalities disguised. The combined scenario at 30% is used as the main realistic adversarial protocol, because real campaigns differ in how many channels they disguise, and messages with all three channels disguised set an upper limit that no detector can exceed. Graph and NLP features were recomputed after each attack, each scenario was repeated with three random seeds, and the main protocol was additionally evaluated with 5-fold cross validation. All experiments used Python, scikit-learn, XGBoost, and NetworkX.

3. RESULTS AND DISCUSSION

3.1 Characteristics of the Sampled Data

The sample is perfectly balanced, with 3,802 messages in each class (Figure 2a). The chi-square analysis in Figure 2(b) and 2(c) shows that phishing messages are dominated by account and urgency vocabulary, such as *akun* (account), *konfirmasi* (confirm), *segera* (immediately), *verifikasi* (verify), and *pemberitahuan* (notice), whereas legitimate messages use service and scheduling vocabulary such as *portal*, *silakan* (please), *kegiatan* (activity), *jadwal* (schedule), and *layanan* (service). This pattern is consistent with the persuasion cues reported in phishing email studies [7], [8].

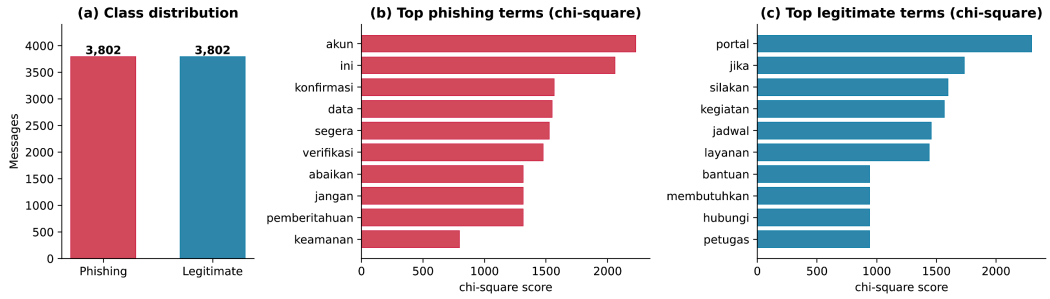

Figure 2. (a) Class distribution of the 7,604 sampled messages; (b) top phishing terms and (c) top legitimate terms ranked by chi-square score

Figure 3 visualises a subgraph of twelve sender domains with their messages and URL domains. Phishing and legitimate entities form separate neighbourhoods, and senders reuse several URL domains, which is the relational signal that label propagation and neighbour ratios exploit. In the full dataset, the sender–URL graph consists of exactly two connected components that coincide with the two classes, so relational evidence alone is sufficient to separate them on clean data.

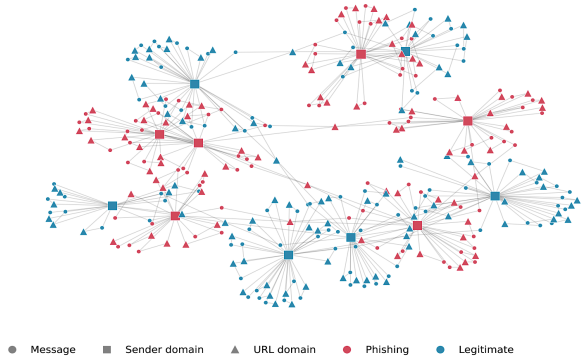

Figure 3. Subgraph of the heterogeneous message–sender–URL-domain graph for twelve sender domains (red: phishing, blue: legitimate)

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

The URL and behavioural attributes also differ sharply between classes, as shown in Figure 4 and Table 2. Phishing URLs are longer (95.5 vs. 51.6 characters), contain more special characters (6.01 vs. 1.98), more redirects (1.59 vs. 0.21), and more subdomains (2.51 vs. 1.01), and come from domains with a much lower reputation score (0.252 vs. 0.758). Phishing messages contain more urgency keywords (2.50 vs. 0.18) and request credentials far more often (88.1% vs. 4.2%). Only 72.0% of phishing URLs use HTTPS, compared with 98.1% of legitimate URLs, but HTTPS alone is a weak indicator, in line with the IDADX finding that most Indonesian phishing domains already use HTTPS [2].

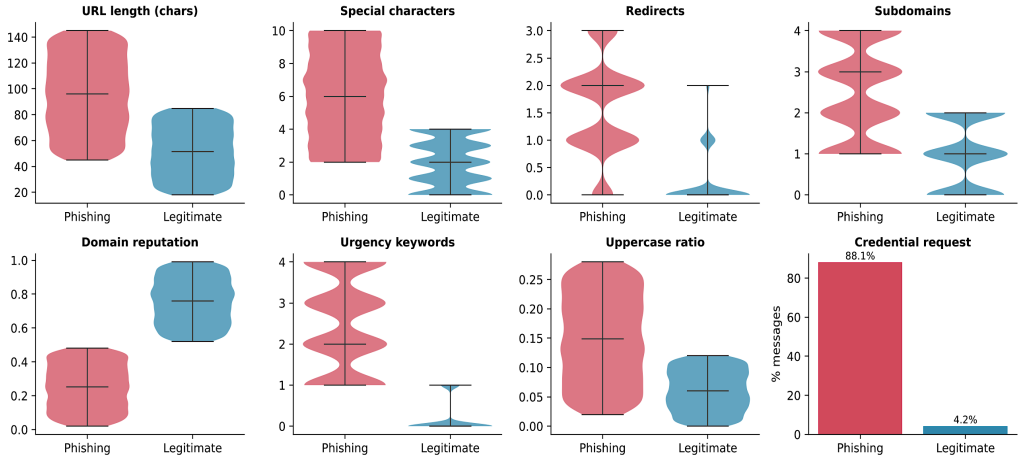

Figure 4. Distribution of URL and behavioural attributes for phishing and legitimate messages

Table 2. Mean values of URL/behavioural attributes per class and single-feature AUC (*descriptive only, excluded from models)

Attribute	Phishing	Legitimate	Single-feature AUC
Domain reputation score*	0.252	0.758	1.000
Urgency keyword count	2.502	0.179	0.978
Credential request (%)	88.1	4.2	0.919
Special character count	6.012	1.983	0.903
Redirect count	1.590	0.207	0.898
URL length (characters)	95.51	51.62	0.880
Uppercase ratio	0.150	0.061	0.838
Subdomain count	2.506	1.008	0.833
HTTPS (%)	72.0	98.1	0.631
Message length (characters)	113.73	116.16	0.567
Has attachment (%)	19.3	27.7	0.542
Send hour	11.45	11.57	0.505

The single-feature AUC values in Table 2 reveal how separable the dataset is. Domain reputation alone reaches an AUC of 1.000, which is why it was excluded from the models, and urgency keyword count reaches 0.978, while message length, attachment, and sending hour are close to random (0.505 to 0.567). Among the graph features, label propagation, the sender and URL phishing ratios, and the aggregated text scores all reach an AUC of 1.000, whereas the structural degree features stay near 0.5. These values show that every modality contains strong evidence, which must be considered when interpreting the clean results.

3.2 Performance on Clean Data

On the clean hold-out set of 1,521 messages, the NLP, graph, early-fusion, and MD-Hybrid models classify every message correctly, and the URL/behavioural model misses three phishing messages (accuracy 0.998, recall 0.996); 5-fold cross validation gives the same picture, with an accuracy of 0.9997 for the URL/behavioural model and 1.000 for the others (Table 3). This ceiling effect means that clean accuracy cannot distinguish a genuinely robust hybrid from a model that relies on a single shortcut feature, and that clean scores overstate the performance that can be expected against attackers who adapt their messages. The realistic adversarial protocol is therefore used as the main basis for comparison.

Table 3. Performance on clean data (hold-out test set and 5-fold cross validation)

Model	Accuracy	Precision	Recall	F1	AUC	MCC
NLP only (TF-IDF + LR)	1.000	1.000	1.000	1.000	1.000	1.000
URL/Behavioural only	0.998	1.000	0.996	0.998	1.000	0.996

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

Model	Accuracy	Precision	Recall	F1	AUC	MCC
Graph only	1.000	1.000	1.000	1.000	1.000	1.000
Early-fusion hybrid	1.000	1.000	1.000	1.000	1.000	1.000
MD-Hybrid (proposed)	1.000	1.000	1.000	1.000	1.000	1.000
5-fold CV, URL/Behavioural	0.9997	0.9997	0.9997	0.9997	1.000	0.9995
5-fold CV, other models	1.000	1.000	1.000	1.000	1.000	1.000

3.3 Robustness under Camouflage

Figure 5 shows the phishing recall of each model as the share of camouflaged phishing messages increases. Under single-modality camouflage, each single-modality model loses recall in proportion to the attack rate: the NLP model falls from 1.000 to 0.500 when 50% of phishing texts are disguised, the graph model falls to 0.500 under infrastructure camouflage, and the URL/behavioural model falls to 0.499 under URL camouflage. Every disguised phishing message is therefore misclassified as legitimate with high confidence, which means that each single view is completely blind once the attacker controls it.

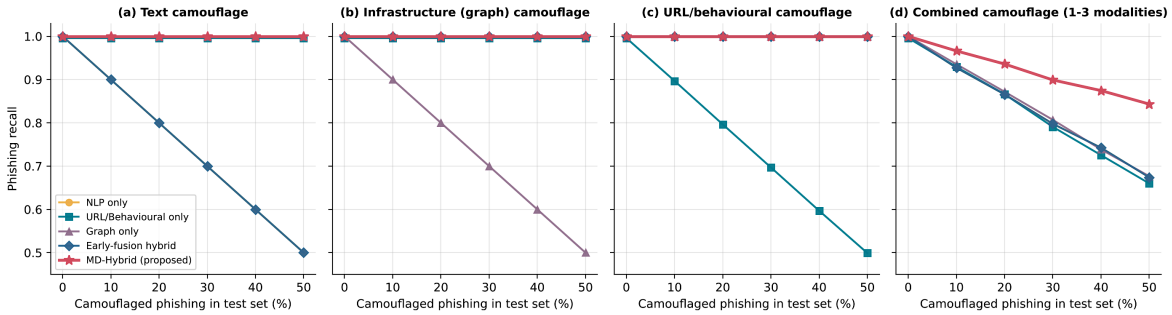

Figure 5. Phishing recall under (a) text, (b) infrastructure, (c) URL/behavioural, and (d) combined camouflage; the NLP-only and early-fusion curves overlap in panels (a) and (d)

The early-fusion hybrid is immune to infrastructure and URL camouflage but follows the NLP curve exactly under text camouflage. This is a clear case of shortcut learning. Because several features already separate the clean training data perfectly, boosting has almost no residual error left to learn from the other features; the model splits its importance between the text probability (52%) and the sender phishing ratio (48%), gives no weight at all to the URL/behavioural attributes, and its predictions follow the text channel, so it collapses when the text is disguised. Such behaviour is consistent with the observation that multimodal models trained on complete data tend to depend on the dominant modality [22].

The proposed MD-Hybrid keeps a recall of 1.000 under every single-modality camouflage and at all attack rates, while its precision stays at 1.000, so robustness is not bought with false alarms. Under combined camouflage, where each attacked message has one, two, or three modalities disguised, all models degrade, but at very different rates. At 30% the recall of the NLP, graph, URL, and early-fusion models falls to 0.791 to 0.807, whereas MD-Hybrid keeps 0.899; at 50% the baselines fall to 0.660 to 0.677 and MD-Hybrid to 0.843. Table 4 reports the 5-fold cross-validation results of the main protocol. The baselines reach accuracies of 0.896 to 0.901, F1-scores of 0.884 to 0.890, and MCC values of 0.810 to 0.818, whereas MD-Hybrid achieves an accuracy of 0.950 ± 0.005, an F1-score of 0.947 ± 0.005, and an MCC of 0.904 ± 0.008, an improvement of about five percentage points in accuracy and nine points in MCC. The early-fusion model has a higher AUC (0.933) than the single-modality models, indicating that its ranking contains some complementary information, but its decision threshold still follows the text channel.

Table 4. 5-fold cross-validation results under the realistic adversarial protocol (mean ± standard deviation)

Model	Accuracy	Recall	F1	AUC	MCC
NLP only (TF-IDF + LR)	0.8993 ± 0.0034	0.7985 ± 0.0068	0.8880 ± 0.0042	0.8974 ± 0.0069	0.8153 ± 0.0058
URL/Behavioural only	0.8964 ± 0.0054	0.7930 ± 0.0103	0.8844 ± 0.0067	0.8998 ± 0.0086	0.8103 ± 0.0092
Graph only	0.9007 ± 0.0051	0.8014 ± 0.0101	0.8897 ± 0.0062	0.9007 ± 0.0050	0.8177 ± 0.0086
Early-fusion hybrid	0.8993 ± 0.0034	0.7985 ± 0.0068	0.8880 ± 0.0042	0.9332 ± 0.0030	0.8153 ± 0.0058
MD-Hybrid (proposed)	0.9498 ± 0.0045	0.8995 ± 0.0090	0.9471 ± 0.0050	0.9495 ± 0.0064	0.9041 ± 0.0082

The confusion matrices in Figure 6 explain these numbers. With 30% combined camouflage, all models classify the 761 legitimate test messages correctly, so the errors are entirely missed phishing messages: 158

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

(20.8%) for NLP only, 151 (19.9%) for URL/behavioural only, 150 (19.7%) for graph only, and 158 (20.8%) for early fusion. MD-Hybrid misses 73 messages (9.6%). Of the 228 attacked phishing messages in this run, 72 had all three channels disguised and therefore carried no phishing evidence at all, so almost all of the errors of MD-Hybrid are attacks that no detector using these three channels could recognise. In a deployment setting, a missed phishing message is far more costly than a false alarm, because it can lead directly to credential theft and the type of financial losses reported by OJK [1].

The behaviour of MD-Hybrid can be traced to the evidence that survives each attack. When the text of a phishing message is replaced with legitimate wording, its URL is still long and contains several redirects and additional subdomains, its behavioural attributes still record urgency cues and a credential request, and its sender still belongs to a neighbourhood whose labelled messages are phishing; the sender-aggregated text score also remains high because the other messages from that sender are not disguised. When the sender and URL domain are moved to legitimate infrastructure, the urgency vocabulary and the credential request remain in the text. When the URL attributes are disguised, both the text and the relational neighbourhood still point to phishing. Because modality-dropout augmentation shows the fusion model exactly these partial patterns during training, it learns that any single genuine channel is sufficient evidence, while legitimate messages, which carry no phishing signal in any channel, are still accepted.

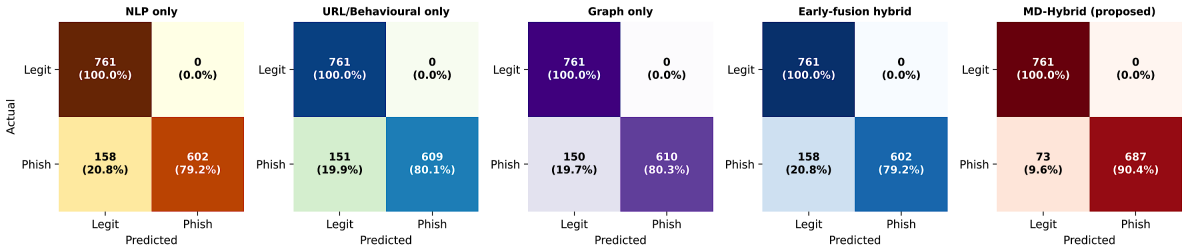

Figure 6. Confusion matrices of the five models on the test set under 30% combined camouflage; percentages are relative to the actual class

3.4 Choice of Fusion Learner and Feature Importance

To check whether the robustness comes from the augmentation strategy rather than from XGBoost itself, the modality-dropout feature set was also used to train logistic regression and random forest. On clean data, random forest and XGBoost achieved an F1-score of 1.000 and logistic regression 0.999. Under the main adversarial protocol, the three learners performed almost identically, with accuracies of 0.951 (logistic regression), 0.952 (XGBoost), and 0.953 (random forest), F1-scores of 0.948 to 0.950, and MCC values of 0.905 to 0.909, with training times of 0.03 s, 0.44 s, and 1.87 s, respectively, on 6,083 original messages and 18,252 augmented phishing samples. The robustness therefore comes from the augmentation strategy, while XGBoost is retained as a fast learner that captures non-linear interactions on tabular evidence [26], [27]. Tree ensembles have also performed well in other Indonesian classification studies that use k-fold validation [28].

Figure 7 compares feature importance. In MD-Hybrid, the most important features are the sender phishing ratio (0.277), the text probability (0.196), urgency keyword count (0.156), the sender-aggregated text score (0.150), credential request (0.094), and special character count (0.066). The importance is spread across the three modalities, with 20% for NLP, 43% for graph, and 38% for URL/behavioural evidence, whereas early fusion uses only the text probability (52%) and the sender phishing ratio (48%) and ignores URL/behavioural evidence entirely. This more balanced use of evidence is the mechanism behind the robustness in Figure 5, and it also makes the decision easier to explain to analysts, in line with Explainable Artificial Intelligence principles [29].

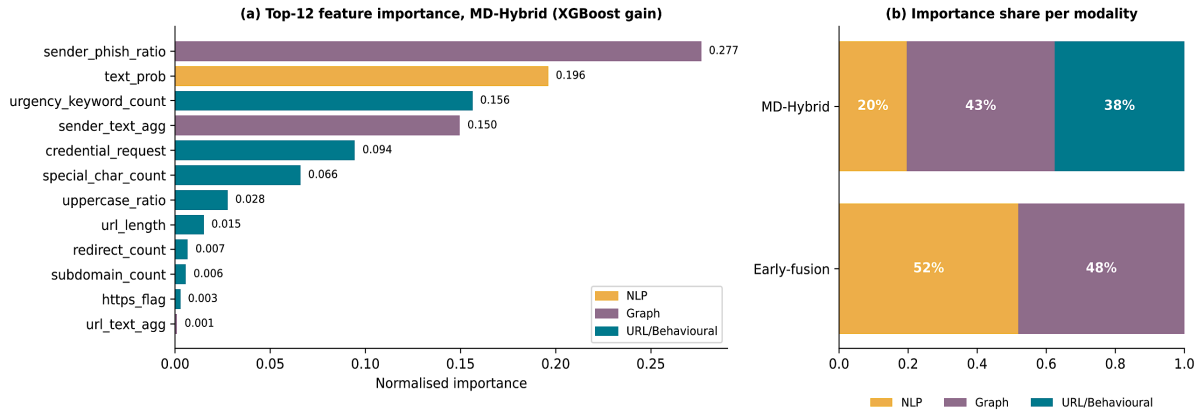

Figure 7. (a) Top-12 feature importance of MD-Hybrid; (b) importance share per modality for early fusion and MD-Hybrid

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

3.5 Discussion

The results answer the research objectives in three ways. First, NLP, graph, and URL/behavioural evidence are each almost sufficient on clean data, which agrees with the high accuracies reported for text-based [8], [10], URL-based [14], and graph-based [17], [18] detectors. Second, each of these views is fragile when it is the only one used, because an attacker who controls that view controls the decision. This supports concerns that LLM-generated text can defeat content-only filters [7] and that compromised infrastructure can defeat reputation-based filters [6]. Third, naive fusion does not automatically solve the problem; the early-fusion hybrid inherits the weakness of the text channel, as also shown for multimodal transformers with missing modalities [22], [23].

Compared with graph neural network detectors such as PhishGNN [17] and anomaly-detection GNNs [20], [21], the proposed graph component is lightweight: it uses label propagation and neighbourhood aggregation of NLP probabilities rather than deep message passing, so it can be updated in seconds when new reports arrive. The graph-aggregated text score is a simple but effective bridge between modalities, because it allows the text of other messages from the same sender to flag a message whose own text has been disguised. This mirrors how Ethereum phishing detection uses the transaction neighbourhood of an account [19]. Compared with Indonesian studies that use a single classifier on text [11]–[13] or tabular data [30], the proposed model shows the value of combining relational and textual evidence.

For Indonesian practitioners, the results suggest a practical design rule. A filter for e-mail, SMS, or social media messages should collect at least three independent kinds of evidence, namely message text, sender and link infrastructure, and URL characteristics, and should be trained so that any one of them can trigger an alert. Because IDADX reports that social media, financial institutions, and e-commerce are the main targets [2], [3], such filters are most relevant for platforms and banks that already hold relational data about senders and domains.

The proposed pipeline is also inexpensive to operate. All components are classical models: TF-IDF with logistic regression for text, sparse matrix operations for label propagation and neighbourhood aggregation, and gradient-boosted trees for fusion. On the 6,083 training messages, the fusion learner trained in well under two seconds, and new phishing reports can be added to the graph simply by updating the labels of the corresponding message nodes and rerunning ten propagation iterations, without retraining the text model. This makes the approach suitable for organisations with limited computing resources, such as universities, local governments, and regional banks, which are frequent targets of credential phishing. At the same time, the augmentation step should be repeated whenever the set of modalities changes, for example when image or attachment features are added, so that the fusion model continues to treat every channel as a possible source of evidence.

This study has important limitations. The dataset comes from a single source, contains only 20 distinct message texts, and is almost perfectly separable by several single features on clean data, so the clean scores should not be generalised to other channels or organisations. When all three modalities are disguised, no evidence remains and any detector using these channels will fail, which explains most of the remaining errors of MD-Hybrid. The camouflage was also simulated by replacing features with values from legitimate messages, which is simpler than adaptive adversarial attacks. Future work should validate the model on curated Indonesian phishing corpora, such as reports collected by IDADX or IASC, evaluate it on unseen campaigns and domains, replace the lexical NLP component with transformer models such as IndoBERT [10], and compare the lightweight graph component with graph neural networks.

4. CONCLUSION

This study proposed a phishing detection model that combines natural language processing of message text, graph-based learning on a heterogeneous message–sender–URL-domain graph, and URL and behavioural attributes, fused by XGBoost trained with modality-dropout augmentation. Using a stratified sample of 7,604 Indonesian-language messages and a leakage-controlled protocol, the study found that every single modality and every fusion model was near-perfect on clean data (accuracy 0.998 to 1.000), which shows that clean accuracy alone cannot reveal whether a hybrid model is robust. The camouflage experiments showed a clear difference. Single-modality models lost phishing recall in direct proportion to the share of disguised messages, falling to about 0.500 when half of the phishing messages were camouflaged in their channel, and the early-fusion hybrid inherited the weakness of the text channel because its decisions were dominated by text evidence and it ignored URL/behavioural attributes entirely. Under the realistic adversarial protocol, in which the domain reputation score is unavailable and 30% of phishing messages disguise one to three channels, the baselines reached accuracies of only 0.896 to 0.901, F1-scores of 0.884 to 0.890, and MCC values of 0.810 to 0.818 in 5-fold cross validation, whereas the proposed MD-Hybrid reached an accuracy of 0.950, an F1-score of 0.947, and an MCC of 0.904 without false positives, and nearly all of its missed messages were attacks with all three channels disguised. The same gain was obtained with logistic regression or random forest as the fusion learner. Feature importance showed that MD-Hybrid spreads its decisions across graph (43%), URL/behavioural (38%), and NLP (20%) evidence, which explains its resilience. The main lesson is that robustness comes from training the fusion model to accept evidence from any genuine channel rather than from simply adding more features. Because the data come from a single source with a limited variety of message texts, future research should validate the model on additional

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 187
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

Indonesian phishing reports, test it on unseen campaigns and domains and against adaptive attackers, use transformer-based Indonesian language models for the text component, and compare the lightweight graph component with graph neural networks before deployment in e-mail, SMS, and social media filters.

REFERENCES

[1] Investortrust, "OJK terima 373.129 laporan penipuan transaksi keuangan, kerugian capai Rp 8,2 triliun," *Investortrust.id*, Dec. 12, 2025. [Online]. Available: https://investortrust.id/financial/88312/ojk-terima-373129-laporan-penipuan-transaksi-keuangan-kerugian-capai-rp-8-2-triliun

[2] IndoTelko, "Domain .id alami 26.675 serangan selama kuartal I 2023," *IndoTelko*, Apr. 26, 2023. [Online]. Available: https://www.indotelko.com/read/1682466431/domain-id-alami-26675-serangan-selama-kuartal-i-2023

[3] N. Naurah, "Serangan phishing di Indonesia terus meningkat, ini datanya," *GoodStats*, May 19, 2023. [Online]. Available: https://goodstats.id/article/serangan-phishing-di-indonesia-terus-meningkat-ini-statistiknya-U8VdY

[4] Anti-Phishing Working Group (APWG), "Phishing activity trends report, 4th quarter 2024," APWG, 2025. [Online]. Available: https://apwg.org/trendreports

[5] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep learning for phishing detection: Taxonomy, current challenges and future directions," *IEEE Access*, vol. 10, pp. 36429–36463, 2022. https://doi.org/10.1109/ACCESS.2022.3151903

[6] W. Li, S. Manickam, Y.-W. Chong, W. Leng, and P. Nanda, "A state-of-the-art review on phishing website detection techniques," *IEEE Access*, vol. 12, pp. 187976–188012, 2024. https://doi.org/10.1109/ACCESS.2024.3514972

[7] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, vol. 12, pp. 42131–42146, 2024. https://doi.org/10.1109/ACCESS.2024.3375882

[8] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022. https://doi.org/10.1109/ACCESS.2022.3183083

[9] T. Mahmud, M. A. H. Prince, M. H. Ali, M. S. Hossain, and K. Andersson, "Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection," *Systems*, vol. 12, no. 11, Art. no. 490, 2024. https://doi.org/10.3390/systems12110490

[10] M. B. M. Amin, G. Hakim, M. T. Maulana, M. F. Alwan, H. S. Anggraheni, M. J. Naufal, and N. Yudistira, "Deteksi spam berbahasa Indonesia berbasis teks menggunakan model BERT," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 6, pp. 1291–1302, 2024. https://doi.org/10.25126/jtiik.2024118121

[11] J. Jessika, A. Khaidar, N. Nurdin, and S. Muliana, "Sentiment analysis of Instagram comments on the BPS Province X account using the naive Bayes algorithm based on machine learning," *Journal of Artificial Intelligence and Software Engineering*, vol. 5, no. 3, pp. 1231–1237, 2025.

[12] A. Khaidar, "Analisis sentimen di Instagram terhadap Menteri Keuangan Purbaya Yudhi Sadewa menggunakan metode logistic regression," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, 2025.

[13] A. P. Erwanda, A. Khaidar, and Nurdin, "Convolutional neural network model for multi-class sentiment analysis on Twitter with deep learning approach," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 1965–1970, 2025.

[14] A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," *Computers & Security*, vol. 136, Art. no. 103545, 2024. https://doi.org/10.1016/j.cose.2023.103545

[15] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 3793–3810.

[16] R. Liu, Y. Lin, X. Yang, S. H. Ng, D. M. Divakaran, and J. S. Dong, "Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1633–1650.

[17] T. Bilot, G. Geis, and B. Hammi, "PhishGNN: A phishing website detection framework using graph neural networks," in *Proceedings of the 19th International Conference on Security and Cryptography (SECRYPT 2022)*, 2022, pp. 428–435. https://doi.org/10.5220/0011328600003283

[18] S. Ariyadasa, S. Fernando, and S. Fernando, "Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML," *IEEE Access*, vol. 10, pp. 82355–82375, 2022. https://doi.org/10.1109/ACCESS.2022.3196018

[19] J. Wu, Q. Yuan, D. Lin, W. You, W. Chen, C. Chen, and Z. Zheng, "Who are the phishers? Phishing scam detection on Ethereum via network embedding," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 1156–1166, 2022. https://doi.org/10.1109/TSMC.2020.3016821

[20] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu, "A comprehensive survey on graph anomaly detection with deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12012–12038, 2023. https://doi.org/10.1109/TKDE.2021.3118815

[21] J. Tang, J. Li, Z. Gao, and J. Li, "Rethinking graph neural networks for anomaly detection," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR vol. 162, 2022, pp. 21076–21089.

[22] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18177–18186.

[23] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023. https://doi.org/10.1109/TPAMI.2023.3275156

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 188
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 180-189
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.10624

[24] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023. https://doi.org/10.1016/j.patter.2023.100804
[25] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Cham, Switzerland: Springer, 2023. https://doi.org/10.1007/978-3-031-38747-0
[26] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*, 2022, pp. 507–520.
[27] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022. https://doi.org/10.1016/j.inffus.2021.11.011
[28] A. Khaidar, N. Nurdin, F. Fajriana, T. Taufiq, and D. Hamdhana, "Classification analysis of single tuition fees using the random forest method with k-fold cross validation," *Journal of Applied Informatics and Computing*, vol. 10, no. 1, pp. 125–133, 2026.
[29] Al-Khaidar, *Explainable Artificial Intelligence (XAI): Teori, Metode, dan Implementasi*. CV. Madza Media, 2026.
[30] A. Khaidar and S. Kurnia, "Classification of health indicators for diabetes mellitus prediction using a TabTransformer model on clinical tabular data," *Journal of Computer Science and Research (JoCoSiR)*, vol. 2, no. 2, pp. 8–15, 2024.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 189
This work is licensed under a Creative Commons Attribution 4.0 International License.