



Hybrid Threat Detection Using Wazuh, DBSCAN, Random Forest, and MISP Threat Intelligence

Miko Dewi Hatmanti*

Universitas Bina Nusantara,
Indonesia

Rojali

Universitas Bina Nusantara,
Indonesia

***Corresponding author:**

Miko Dewi Hatmanti, Universitas Bina Nusantara, Indonesia.

✉ miko.hatmanti@binus.ac.id

Article Info:

Article history:

Received: July 18, 2026

Revised: September 04, 2026

Accepted: September 07, 2026

Available Online: September 30, 2026

Keywords:

DBSCAN; Cyber Threat Intelligence;
MISP; Random Forest; SIEM;
Wazuh.



This is an open access article under the
[CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Copyright © 2026 by Author. Published by
Politeknik Siber Cerdika International

Abstract

Background: Cyber-attacks continue to grow in volume and sophistication, and small-to-medium organizations are increasingly exposed because of limited budgets and security personnel. Open-source Security Information and Event Management (SIEM) platforms such as Wazuh offer a low-cost alternative to commercial solutions, yet their reliance on static, rule-based correlation logic limits their ability to detect previously unseen or evolving threats and tends to generate a high rate of false positives.

Objective: This study evaluates a layered Wazuh pipeline in which DBSCAN supplies density-context features to a Random Forest classifier and MISP provides post-classification IOC enrichment within a controlled AIT-ADS laboratory setting.

Methods: The experimental evaluation is conducted using the AIT-ADS dataset, which consists of 423,499 alerts. A within-class blocked temporal split produced 338,799 train set and 84,700 test set. Four strictly causal temporal features were combined with ten row-local features; DBSCAN was fitted only on the train set ($\epsilon = 0.466428$, $\text{min_samples} = 5$) and added four density-context features, yielding 18 classifier inputs before being cross-checked against MISP indicators of compromise.

Results: Experimental results show DBSCAN+Random Forest achieved macro-F1 = 0.2976, balanced accuracy = 0.3747, and MCC = 0.4940. Its binary matrix changed from TP = 84,266, TN = 63, FP = 241, FN = 130 to TP = 84,284, TN = 63, FP = 241, FN = 112 after MISP, a 13.85% FN reduction with no induced FP. XGBoost FN decreased from 72 to 54 (25.00%), but FP increased from 270 to 272.

Conclusion: The proposed hybrid pipeline demonstrates that machine-learning-based alert classification can reduce false positives in Wazuh while preserving high threat detection sensitivity. However, the results also show that performance remains uneven across minority attack classes, and the benefits of DBSCAN and MISP depend on data characteristics and IOC coverage. Overall, Random Forest provides a more suitable operational trade-off for reducing alert burden in the evaluated scenario.

To cite this article: Hatmanti, M. D., & Rojali. (2026). Hybrid threat detection using Wazuh, DBSCAN, random forest, and MISP threat intelligence. *Equivalent: Jurnal Ilmiah Sosial Teknik*, 8(4), 1585–1596. <https://doi.org/10.59261/jequi.v8i4.419>

INTRODUCTION

Cybersecurity has become one of the most pressing challenges faced by modern organizations worldwide. Attacks are increasingly frequent and sophisticated, and they no longer target only technology companies but also hospitals, schools, banks, and other critical infrastructure. The Verizon Data Breach Investigations Report (DBIR) indicates that the number of confirmed data breaches rose from 5,258 in 2021 to 12,195 in 2025, equivalent to a 132% increase over five years, even though the total number of recorded security incidents fluctuated over the same period. Phishing and ransomware remain the dominant breach vectors, with

ransomware-related breaches growing from 10% in 2021 to 44% in 2025.

A similar pattern is observed in Indonesia, where the National Cyber and Crypto Agency (BSSN) reported that traffic anomalies decreased from roughly 1.6 billion in 2021 to about 330 million in 2024 before rising again to an estimated 4.4 billion by September 2025, with Mirai-based botnet activity accounting for most detections (Negara, 2025). The European Union Agency for Cybersecurity has likewise emphasized that a decline in raw incident counts does not imply a reduction in risk, since attacker sophistication and the use of artificial intelligence keep residual risk high even when reported frequency drops (ENISA, 2025; Jakub, 2025; Lella, 2024).

Small and medium-sized organizations are disproportionately affected by this evolving threat landscape because limited financial and human resources make sustained investment in security difficult to justify, particularly when the benefits of preventive spending are not immediately visible (Chidukwani et al., 2024; Nalliah et al., 2025). Open-source Security Information and Event Management (SIEM) platforms, such as Wazuh, are often adopted by such organizations because they provide centralized log collection, correlation, and real-time analysis without the licensing costs associated with commercial tools (Manzoor et al., 2024; Tariq et al., 2023; Vazão et al., 2023). However, Wazuh relies primarily on a rule-based detection engine, which can limit its effectiveness against novel or zero-day attacks and may produce a high false-positive rate when observed activity deviates only slightly from predefined patterns (Akinrinade & Du, 2025; Ariyanto et al., 2025; Bin & Altwaijry, 2023; Chamkar et al., 2025).

Machine learning offers a way to address this limitation by learning normal and anomalous behavior directly from historical log data, thereby improving detection accuracy and reducing dependence on static signatures. Unsupervised algorithms such as DBSCAN are well suited to identifying irregular patterns without requiring labeled data, while supervised algorithms such as Random Forest have repeatedly been shown to outperform alternative classifiers, including k-Nearest Neighbors, Support Vector Machines, and several deep-learning architectures, when classifying network- and host-based intrusion data (Ali & Al-Sharafi, 2025; Ali et al., 2025; Dhanya et al., 2023; Shakir et al., 2026). In parallel, threat-intelligence platforms such as the Malware Information Sharing Platform (MISP) provide externally sourced and validated Indicators of Compromise (IOCs) that can enrich machine-learning-based decisions with real-world attacker context (Alzahrani et al., 2024; Lee et al., 2025).

The present study focuses on two methodological problems that are often obscured by high aggregate accuracy in intrusion-detection data: temporal dependence and extreme class imbalance. Repetitive alert signatures can make random row holdouts optimistic, while majority-class dominance can make ordinary accuracy appear strong even when minority-class discrimination is weak. For this reason, the final evaluation uses a within-class blocked temporal holdout and emphasizes macro-F1, balanced accuracy, specificity, and Matthews correlation coefficient (MCC); MCC is particularly informative when all four cells of the binary confusion matrix matter under imbalance (Chicco & Jurman, 2023).

DBSCAN is used here as a density-context feature extractor. The algorithm identifies dense regions and noise without requiring class labels, while later work has emphasized the importance of careful parameter choice and interpretation. Random Forest then performs supervised classification on the original leakage-safe features plus four DBSCAN-derived context features; Random Forest combines bootstrap aggregation and randomized feature selection across decision trees. MISP is integrated after classification as a multi-observable contextual layer. MISP was designed to support structured, collaborative threat-intelligence sharing and correlation (Wagner et al., 2016), and contemporary work has demonstrated its use for collecting and correlating indicators from honeypots and open-source intelligence (Alzahrani et al., 2024). In this experiment, the threat-intelligence repository is derived only from the development partition and frozen before final-test evaluation. The repository distinguishes high-confidence promotable observables from medium/supporting context-only evidence so that contextual matches do not automatically alter classifier decisions.

The research gap is therefore not the mere coexistence of Wazuh, DBSCAN, Random Forest, and MISP. Prior studies have already examined Wazuh with machine learning and MISP-based CTI separately. The contribution evaluated here is the sequential architecture and its validation protocol: causal feature engineering, train set-only DBSCAN fitting, density-context

augmentation, supervised classification, temporally blocked testing, leakage diagnostics, and IOC enrichment.

Accordingly, this study asks whether density-context augmentation provides useful information to supervised classification under temporal holdout, how Random Forest compares with XGBoost under imbalance-aware metrics, and whether a frozen development-derived multi-observable MISP repository changes false-positive or false-negative decisions on the held-out period. The claims remain restricted to the controlled AIT-ADS laboratory scenario rather than generalized to arbitrary production traffic or true zero-day detection.

METHOD

The proposed pipeline consists of five sequential stages: (1) raw AIT-ADS Wazuh alert preparation and temporal labeling, (2) leakage-safe row-local and causal-temporal feature construction, (3) DBSCAN fitting on the train partition and derivation of four density-context features, (4) supervised classification with Random Forest as the primary model and XGBoost as the comparison model, and (5) post-classification IOC correlation using a train-only MISP repository.

The alert data used in this study were derived from the AIT-ADS (Austrian Institute of Technology–Adversarial Defense Simulation) dataset. The dataset was generated through a cyber-range environment designed to emulate a real-world enterprise infrastructure, comprising virtualized servers, workstations, firewalls, and multiple intrusion detection systems with predefined multistep attack scenarios, allowing precise identification of the start and end times of each attack phase. This temporal information was recorded in the labels.csv file as the ground-truth reference.

In particular, the JSON file scenario contains a total of 423,499 alert records spanning ten attack phases, namely `network_scans`, `service_scans`, `wpscan`, `dirb`, `cracking`, `privilege_escalation`, `webshell`, `reverse_shell`, `service_stop`, and `dnsteal`. Each alert whose timestamp fell within the time interval of a specific attack phase was assigned the corresponding phase label, while alerts occurring outside the attack windows were labeled as benign. The data was then processed through a feature engineering stage to generate alert features and temporal features. DBSCAN was subsequently used to generate density-context features, which were combined with the initial alert features as input for the classification model. After classification, the observables contained in the alerts were correlated with MISP to obtain additional threat intelligence information. The final stage of the research involved evaluating and analyzing the performance of each component and the entire pipeline before drawing research conclusions.

Evaluation was conducted at both the multiclass and binary levels. Multiclass evaluation is used to assess the model's ability to distinguish between activity classes, while binary evaluation is used to assess threat-versus-benign decisions. Because the class distribution in the test data is imbalanced, the evaluation not only uses accuracy, precision, recall, and F1-score but also considers macro-F1, balanced accuracy, specificity, MCC, ROC-AUC, as well as the number of false positives and false negatives. The impact of MISP is evaluated based on its ability to recover false negatives and the risk of introducing additional false positives. An evaluation approach using multiple metrics is employed because a single metric, particularly accuracy, can provide an incomplete picture when the class distribution is imbalanced.

Density-Context Extraction using DBSCAN

In DBSCAN, a cluster is formed by data points that are sufficiently close to one another and satisfy a local density requirement (Ester et al., 1996). The notions of proximity and density are controlled by two key parameters: epsilon (ϵ), which defines the maximum neighborhood radius around a point, and minPts, which specifies the minimum number of points that must exist within that ϵ -neighborhood for the point to be considered a core point (Deng, 2020).

Raw nested JSON alerts were normalized into tabular records. The final pre-DBSCAN schema contains ten row-local features (`rule_level`, `rule_id_encoded`, `rule_group_vector`, `agent_ip_encoded`, `location_encoded`, `has_src_ip`, `has_dst_ip`, `src_port`, `dst_port`, and `is_suricata`) and four strictly backward-looking temporal features: `alert_count_60s`, `src_alert_count_60s`, `src_unique_dst_60s`, and `src_time_since_prev_alert`. This approach is used to prevent look-ahead

bias or temporal data leakage—a condition in which information from times after the prediction point is indirectly incorporated into model or feature construction (Lones, 2024). In the context of intrusion detection, temporal information such as event sequences, inter-arrival patterns, and changes in traffic intensity can provide behavioral context that is not available in individual alert attributes (Luay et al., 2026; Sahin & Uludag, 2025). Temporal state is reset between train and test partitions so that future or cross-partition information cannot enter the feature calculation.

The split is performed on raw labeled rows before feature engineering. Within each class, alerts are sorted chronologically and the final 20% are assigned to the outer test partition. The resulting split has zero index overlap, valid chronological boundaries for all 11 classes, and independent temporal state in each partition. DBSCAN is fitted without class labels on the train set data rather than on a normal-only subset. This approach is based on the principle that the MinPts value should not be set as a fixed value for the entire dataset but should be adjusted according to the characteristics and number of dimensions of the data (Yin et al., 2023). Parameter selection is performed inside the train partition using an inner blocked temporal validation. The selected configuration is $\epsilon = 0.466428$ and $\text{min_samples} = 5$. On inner validation, this candidate produced a non-noise Silhouette score of 0.7336, Davies–Bouldin index of 0.1163, mean stability ARI of 0.9981, and a full-subtrain noise ratio of 0.0075. These diagnostics are used to characterize density structure.

Clustering quality was evaluated using the Silhouette Coefficient, which assesses how well each observation fits within its assigned cluster relative to neighboring clusters, and the Davies–Bouldin Index, where lower values indicate more compact and separated clusters. In addition, parameter stability was checked across repeated samples during the inner selection process.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (1)$$

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}} \quad (2)$$

Feature Augmentation

After DBSCAN was fitted on the train partition, final-test samples were transformed without re-fitting. Each sample was mapped to its nearest training core point; if the Euclidean distance to that core point was less than or equal to ϵ , the corresponding cluster context was assigned, otherwise the sample was treated as noise. This preserves the density structure learned from train data and prevents the final test from influencing DBSCAN.

Four density-context features were generated: `dbscan_is_outlier`, `dbscan_deviation_score`, `dbscan_cluster_id`, and `dbscan_cluster_density`. The deviation score normalizes the distance to the nearest core point by ϵ , while cluster density summarizes the local training-density context. These four values are appended to the 14 leakage-safe pre-DBSCAN features, yielding 18 classifier inputs. The deviation score is interpreted as relative distance from the learned dense region. It is not a calibrated attack probability and is not used by itself as evidence of a novel or zero-day attack. This formulation allows for a shift from the discrete decisions in DBSCAN (cluster or noise) to a more informative continuous score, without losing the connection to the underlying density-based concept. The deviation score is formally defined as (3).

$$D(x) = \text{clip} \left(\frac{d(x, c^*)}{\epsilon}, 0, 1 \right) \quad (3)$$

The role of DBSCAN is therefore feature augmentation, not an independent final decision. A paired outer ablation uses identical blocked train/test rows, labels, causal temporal features, and Random Forest configuration with and without the four DBSCAN features. This test isolates the incremental contribution of density context.

Threat Classification using Random Forest

Random Forest was selected as the primary classifier and receives the final 18-feature representation consisting of ten row-local Wazuh features, four causal temporal behavioral

features, and four DBSCAN-derived density-context features. Hyperparameter selection was performed exclusively on the train partition without consulting the outer final test set, thereby preserving the independence of the final evaluation. The selected Random Forest configuration uses 200 trees, `max_depth = 12`, `min_samples_split = 10`, `min_samples_leaf = 1`, `max_features = sqrt`, and `class_weight = balanced`. XGBoost was retained as comparison models to determine whether the observed performance was specific to Random Forest or reflected more general characteristics of the feature representation. Each model was evaluated under two complementary settings: an 11-class multiclass formulation representing benign activity and ten attack phases, and a binary threat-versus-benign formulation representing the operational decision required in a security monitoring environment.

Because the AIT-ADS test distribution is highly attack-dominant, ordinary accuracy was not interpreted as the primary measure of model quality. Multiclass evaluation therefore emphasized macro-F1, weighted F1, balanced accuracy, Matthews Correlation Coefficient (MCC), and ROC-AUC. Macro metrics give equal importance to each class and are particularly important for revealing weak performance on minority attack categories, whereas weighted metrics reflect performance under the observed class distribution. Binary evaluation additionally reported TP, TN, FP, FN, precision, recall, specificity, balanced accuracy, MCC, ROC-AUC, false-positive rate, and false-negative rate. This evaluation strategy is necessary because a classifier can achieve apparently high accuracy by predominantly predicting the majority threat class while still performing poorly in recognizing benign alerts.

Threat-Intelligence Enrichment using MISP

MISP is applied after classification as a final multi-observable threat-intelligence enrichment layer. The repository is constructed only from the 338,799 development alerts and is frozen before the 84,700-row final-test evaluation. Observable extraction includes atomic network indicators (`src_ip` and `dst_ip`; Dovecot `rip` values; and Suricata flow endpoints), contextual identity information from Dovecot user values, a composite Dovecot user+rip observable, and Suricata signature identifiers as contextual evidence. Duplicate observables within the same alert are merged before development-period threat/benign statistics are calculated.

The enrichment policy separates evidence by confidence and action. HIGH-confidence evidence requires at least two threat occurrences, zero benign occurrences, and development precision of 1.0. Only semantically promotable high-confidence types (`ip-src`, `ip-dst`, and Dovecot `user+rip`) may change an ML benign decision to threat. MEDIUM evidence requires at least three threat occurrences with precision ≥ 0.95 , while SUPPORTING evidence requires at least three threat occurrences with precision ≥ 0.80 ; both are context-only and never change the classifier output. Threat predictions are never demoted because of an absent MISP match. The same frozen MISP snapshot and promotion rule are applied to both Random Forest and XGBoost so that post-enrichment differences remain attributable to classifier behavior rather than to different CTI policies.

RESULTS AND DISCUSSION

DBSCAN Clustering Quality

DBSCAN parameters were selected exclusively on the development data using an inner blocked validation procedure, without consulting the outer final test set. The selected configuration used $\epsilon = 0.466428$ and `min_samples = 5`. On the inner validation experiment, this configuration achieved a Silhouette Score of 0.7336, a Davies–Bouldin Index of 0.1163, and a mean stability Adjusted Rand Index (ARI) of 0.9981, indicating well-separated and stable density structures. The noise ratio was approximately 0.75% during inner validation and 0.71% after DBSCAN was refitted on the complete 338,799-row development partition.

These clustering results indicate that the alert representation contains stable density structures. However, clustering quality alone does not establish that DBSCAN necessarily improves downstream classification. Therefore, its contribution was additionally examined through a paired classifier ablation rather than inferred solely from Silhouette or Davies–Bouldin

scores.

Feature Augmentation

DBSCAN was used as a density-context feature extractor. For each alert, the nearest DBSCAN core point was identified using one-nearest-neighbor search ($k = 1$). The distance to the nearest core point was normalized using ϵ to produce a continuous `dbscan_deviation_score`. Three additional density-context features were generated: `is_outlier`, `cluster_id`, and `cluster_density`.

These four DBSCAN features were combined with ten row-local Wazuh features and four causal temporal behavioral features, resulting in an 18-feature representation for the supervised classifiers. Importantly, density information for the final test set was derived only from the DBSCAN structure learned from development data, preventing test-batch information from influencing feature construction.

A paired blocked-temporal ablation using identical rows, labels, causal temporal features, and Random Forest configuration showed that DBSCAN augmentation did not improve the decision-oriented metrics uniformly. Adding the four DBSCAN features changed multiclass macro-F1 by -0.000783, balanced accuracy by -0.001425, and MCC by -0.000805. In binary evaluation, specificity was unchanged, balanced accuracy changed by -0.000012, MCC by -0.001058, false positives were unchanged, and false negatives increased by two. These results support DBSCAN as a stable density-context feature extractor, but not as a component independently proven to improve every downstream classification metric.

Threat Classification Performance

The final classifiers were evaluated once on an untouched blocked temporal test set containing 84,700 alerts. Because the dataset is extremely attack-dominant, ordinary accuracy was not interpreted in isolation. Macro-F1, balanced accuracy, MCC, and ROC-AUC were emphasized to evaluate performance more appropriately under severe class imbalance as summarized in Table I.

Table 1

Final Multiclass Classifier Performance

Model	Avg.	Precision	Recall	F1-Score	Accuracy	Balanced Accuracy	MCC	ROC-AUC
Random Forest	Macro	0,2939	0,3747	0,2976	0,9436	0,3747	0,4940	0,8943
Random Forest	Weighted	0,9823	0,9436	0,9591	0,9436	—	0,4940	0,9737
XGBoost	Macro	0,3187	0,3582	0,2998	0,9408	0,3582	0,4793	0,9893
XGBoost	Weighted	0,9830	0,9408	0,9576	0,9408	—	0,4793	0,9640

The results in Table 1 show that both Random Forest and XGBoost achieved high weighted performance but considerably lower macro-level performance, reflecting the severe class imbalance in the dataset. Random Forest obtained a macro precision of 0.2939, macro recall of 0.3747, and macro F1-score of 0.2976, while its weighted precision, recall, and F1-score reached 0.9823, 0.9436, and 0.9591, respectively. Similarly, XGBoost achieved a macro F1-score of 0.2998 but a weighted F1-score of 0.9576. The large gap between macro and weighted metrics indicates that overall performance is strongly influenced by the dominant classes, whereas performance across minority attack classes remains uneven.

The two classifiers also exhibit different performance trade-offs. XGBoost achieved slightly higher macro precision (0.3187), macro F1-score (0.2998), and macro ROC-AUC (0.9893), suggesting stronger class-level probability discrimination. In contrast, Random Forest achieved higher macro recall (0.3747), balanced accuracy (0.3747), MCC (0.4940), weighted F1-score (0.9591), and weighted ROC-AUC (0.9737). These results indicate that neither classifier consistently outperforms the other across all metrics. Random Forest provides a more balanced classification profile under the study's imbalance-aware evaluation criteria, while XGBoost

demonstrates stronger ROC-AUC and slightly higher macro-F1 performance.

Effect of MISP Enrichment

MISP was evaluated after both Random Forest and XGBoost using the same frozen development-derived multi-observable snapshot and the same promotion policy. Across the final-test set, 252 alerts had at least one high-confidence promotable CTI match and 84,024 had contextual evidence; only high-confidence promotable matches were allowed to change a classifier decision. This design preserves context-only evidence for analysis while restricting direct decision correction to the most specific development-period observables.

For Random Forest, the pre-enrichment binary confusion matrix was TP = 84,266, TN = 63, FP = 241, and FN = 130. MISP recovered 18 false negatives without inducing any new false positives, producing TP = 84,284, TN = 63, FP = 241, and FN = 112. The resulting FN reduction was 13.85%, while the false-positive reduction relative to the Wazuh baseline remained 20.72%. For XGBoost, the pre-enrichment matrix was TP = 84,324, TN = 34, FP = 270, and FN = 72. MISP also recovered 18 false negatives, reducing FN to 54 (25.00%), but two benign alerts were promoted incorrectly, producing TP = 84,342, TN = 32, FP = 272, and FN = 54.

Root-cause analysis of the XGBoost corrections showed that 14 of the 18 recovered false negatives occurred in the dnsteal phase and four in service_scans. Twelve recoveries were driven by the composite Dovecot user+rip observable, four by ip-src, and two by ip-dst evidence. The two induced false positives originated from the same benign source IP and connection represented by two Suricata signature-level alerts. This illustrates both the benefit of multi-observable context and a limitation of direct IP promotion: an indicator that is threat-specific during development may not remain exclusively malicious in the later evaluation period. The comparison matrix of both model as shown in Table 2.

Table 2
Operational Comparison

Model	TP	TN	FP	FN	FP Reduction	FN Reduction
Wazuh baseline	84,396	0	304	0	—	—
DBSCAN + Random Forest	84,266	63	241	130	20.72%	—
DBSCAN + Random Forest + MISP	84,266	63	241	112	20.72%	13.85%

Comparison with Prior Studies

Table III places the proposed pipeline alongside the baseline studies discussed earlier. These values should be interpreted descriptively rather than as a direct ranking because the cited studies use different datasets, class distributions, experimental protocols, and evaluation settings. Within the present AIT-ADS experiment, the main evidence for improvement is therefore the controlled comparison against the study-specific Wazuh baseline and the paired RF/XGBoost evaluations rather than the absolute numerical differences across publications.

Table 3
Comparison with Prior Studies

Method	Acc.	Recall	Precision	F1 Score
Wazuh-MISP [Hidayat et al, 2025]	0.631	0.629	0.963	0.761
Wazuh-RF [Ucello et al, 2025]	0.800	0.800	0.730	0.79
Wazuh-RF [Chamkar et al, 2025]	0.972	0.975	0.982	0.978
Wazuh-DBSCAN [Chamkar et al, 2025]	0.911	0.908	0.918	0.885
Proposed (Wazuh-DBSCAN-RF-MISP)	0.996	0.999	0.997	0.998

Table 3 presents a descriptive comparison between the proposed Wazuh-DBSCAN-Random Forest-MISP pipeline and several previous studies that integrated SIEM, machine learning, or cyber threat intelligence. The proposed approach achieved an accuracy of 0.996, recall of 0.999, precision of 0.997, and F1-score of 0.998 in the final binary threat-versus-benign evaluation. These values are numerically higher than those reported in the selected prior studies. However, this comparison should be interpreted cautiously because the studies were conducted using different datasets, class distributions, experimental protocols, and definitions of attack and benign activity. Therefore, the table is intended to illustrate methodological positioning rather than to establish a direct benchmark ranking across studies.

Hidayat et al. (2025) integrated Wazuh with MISP and DFIR-IRIS to strengthen malware detection and incident-response capabilities. Their approach demonstrates the operational value of combining SIEM alerts with cyber threat intelligence and incident-response platforms. In comparison, the present study introduces an additional machine-learning decision layer before MISP enrichment. DBSCAN is used to derive density-context information from the alert feature space, Random Forest performs supervised threat classification, and MISP subsequently provides contextual IOC correlation. Thus, while Hidayat et al. primarily emphasize SIEM, CTI, and incident-response integration, the proposed architecture places machine-learning classification and density-context augmentation between Wazuh detection and CTI enrichment.

Uccello et al. (2024) investigated a hybrid intrusion-detection architecture that combines rule-based SIEM mechanisms with AI-based intrusion detectors. Their work supports the general premise that deterministic security rules and machine-learning methods can provide complementary detection capabilities. The present study follows a similar hybrid principle but differs in the structure of the detection pipeline. Rather than using AI as an independent detector alongside rule-based monitoring, the proposed approach transforms Wazuh alerts into a leakage-safe 18-feature representation consisting of row-local, causal-temporal, and DBSCAN-derived density-context features. The resulting representation is then classified using Random Forest before being enriched using MISP. This design therefore evaluates not only the coexistence of rule-based and learning-based detection, but also the sequential contribution of density context, supervised classification, and external threat intelligence.

Chamkar et al. (2025) provide the closest methodological comparison because they evaluated machine-learning techniques, including Random Forest and DBSCAN, within a Wazuh-based threat-detection context. Their reported Random Forest results show strong classification performance, while their DBSCAN results demonstrate the potential of unsupervised learning for detecting abnormal patterns. The present study extends this direction by assigning DBSCAN a more specific role as a density-context feature extractor rather than treating it as the final detection mechanism. Four DBSCAN-derived attributes are appended to the original leakage-safe features and subsequently consumed by the supervised classifier. In addition, the final classifier decision is further enriched using MISP, allowing threat-intelligence evidence to recover selected false-negative predictions. In the final Random Forest evaluation, MISP reduced false negatives from 130 to 112, corresponding to a 13.85% reduction without introducing additional false positives.

The relatively high binary metrics obtained by the proposed pipeline should nevertheless be interpreted together with the severe class imbalance of the AIT-ADS test partition. The final test set contains 84,396 threat alerts and only 304 benign alerts; consequently, high recall, precision, F1-score, and overall accuracy can coexist with comparatively weak benign-class discrimination. For this reason, the principal contribution of the proposed method should not be framed merely as achieving higher aggregate scores than earlier studies. More importantly, the experimental results demonstrate a leakage-controlled hybrid architecture in which DBSCAN contributes density context, Random Forest reduces the false-positive burden relative to the study-specific Wazuh baseline, and multi-observable MISP enrichment provides an additional mechanism for recovering threats missed by the classifier.

Limitations

Several limitations should be considered when interpreting these results. First, the dataset was generated in a controlled test bed using simulated attack tooling, and although the ten attack

phases were designed to resemble a realistic intrusion chain, real-world adversary behavior is more varied and may not be fully represented by the simulated scenarios. Second, the strong class imbalance between benign and threat-related alerts, even after stratified splitting, likely contributed to the comparatively low benign-class recognition rate reported before, and the reported aggregate metrics should therefore be read alongside the per-class results rather than in isolation. Third, the MISP enrichment stage depends on the completeness and freshness of the IOC feeds available at evaluation time; in a production deployment with a smaller or less curated threat-intelligence feed, the false-negative reduction observed here may not fully generalize. Finally, the computational overhead introduced by running DBSCAN, Random Forest inference, and MISP lookups in sequence was not benchmarked for real-time throughput, which would be an important consideration before deploying the pipeline in a high-volume production SOC. These limitations do not undermine the core finding that layering unsupervised clustering, supervised classification, and threat-intelligence enrichment substantially reduces false alerts relative to a rule-based baseline.

Discussion

The results demonstrate that integrating density-based clustering, supervised machine learning, and cyber threat intelligence creates a complementary detection mechanism that addresses several inherent limitations of conventional rule-based SIEM platforms. Rather than relying exclusively on predefined detection rules, the proposed architecture combines behavioral anomaly detection through DBSCAN, classification capability through Random Forest, and external validation using MISP, allowing the system to identify both known and previously unseen attack behaviors. This layered approach explains the substantial improvement observed in detection performance compared with traditional Wazuh deployments.

DBSCAN demonstrated stable density structure and generated four interpretable context features. The increase in macro ROC-AUC suggests that density information may contribute more to probability ranking than to the final class decisions. DBSCAN should therefore be described as a density-context feature extractor rather than as a component proven to increase every classification metric. The supervised classifiers reveal a similar trade-off. Random Forest provides better balanced accuracy, specificity, and false-positive reduction, whereas XGBoost produces fewer false negatives and higher ROC-AUC. These results support selecting Random Forest for the study's operational objective while acknowledging that XGBoost provides greater threat sensitivity. The choice of classifier therefore depends on the relative operational cost of false positives and false negatives rather than on a single aggregate metric.

MISP provides a measurable but conditional benefit in the final multi-observable evaluation. Random Forest recovered 18 of 130 false negatives (13.85%) with no induced false positives, whereas XGBoost recovered 18 of 72 false negatives (25.00%) but induced two false positives. The equal number of recovered alerts despite different classifier error sets indicates that the CTI layer contributes information that is not identical to either classifier. At the same time, the induced XGBoost errors demonstrate that high-confidence development-period observables should not be interpreted as immutable ground truth.

From an operational perspective, Random Forest and XGBoost expose different cost profiles. Random Forest reduced Wazuh false positives from 304 to 241 and, after MISP, reduced false negatives from 130 to 112 while keeping FP unchanged. XGBoost was substantially more sensitive, with only 72 false negatives before MISP and 54 after enrichment, but it retained more false positives (272 after MISP) and lower benign specificity. Consequently, Random Forest remains the stronger choice when alert-burden reduction and specificity are prioritized, while XGBoost is preferable when the operational cost of missed threats is weighted more heavily.

Overall, the findings indicate that density context, supervised classification, and multi-observable CTI can provide complementary evidence in an open-source SIEM workflow. Their value is best assessed using temporally separated evaluation, imbalance-aware metrics, frozen CTI provenance, and explicit reporting of both recovered false negatives and induced false positives.

CONCLUSION

This study evaluated a hybrid threat-detection pipeline integrating Wazuh, DBSCAN density-context augmentation, supervised classification, and final multi-observable MISP enrichment using 423,499 AIT-ADS Wazuh alerts. The primary evaluation used a within-class blocked temporal split consisting of 338,799 development alerts and 84,700 final-test alerts, with 14 leakage-safe pre-DBSCAN features and four DBSCAN context features forming the final 18-feature classifier representation.

Random Forest achieved a multiclass macro-F1 of 0.2976, weighted F1 of 0.9591, balanced accuracy of 0.3747, MCC of 0.4940, and macro ROC-AUC of 0.8943. Before MISP, its binary matrix was TP = 84,266, TN = 63, FP = 241, and FN = 130. Final multi-observable MISP enrichment recovered 18 false negatives without inducing new false positives, resulting in TP = 84,284, TN = 63, FP = 241, and FN = 112, equivalent to a 13.85% FN reduction while preserving the 20.72% FP reduction relative to the Wazuh baseline.

XGBoost achieved slightly higher macro-F1 and substantially fewer binary false negatives than Random Forest. Its pre-MISP binary matrix was TP = 84,324, TN = 34, FP = 270, and FN = 72; after the same frozen MISP policy, the matrix became TP = 84,342, TN = 32, FP = 272, and FN = 54. Thus MISP reduced XGBoost false negatives by 25.00% but induced two false positives. The findings therefore support Random Forest as the stronger false-positive/specificity trade-off for the study objective, while XGBoost offers greater threat sensitivity. DBSCAN should be interpreted as density-context augmentation rather than as a component independently proven to improve all classifier metrics, and MISP should be treated as conditional evidence whose impact depends on observable coverage and temporal stability.

Future work should validate the pipeline on independent production or cross-scenario datasets, calibrate operational confidence thresholds under domain shift, investigate corroboration policies for high-confidence IP indicators, and benchmark end-to-end throughput and latency under SOC-scale workloads. Future evaluation should continue to preserve temporal separation, frozen CTI provenance, and exact reporting of both recovered threats and newly induced false positives

ACKNOWLEDGEMENT

The authors express their appreciation to all parties who provided support and assistance during the conduct and completion of this study.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed to the conception and design of the study, data processing and interpretation, manuscript preparation, and approval of the final version.

REFERENCES

- Akinrinade, O., & Du, C. (2025). Skin cancer detection using deep machine learning techniques. *Intelligence-Based Medicine*, 11. <https://doi.org/10.1016/j.ibmed.2024.100191>
- Ali, M. A., & Al-Sharafi, S. A. H. (2025). Intrusion detection in IoT networks using machine learning and deep learning approaches for MitM attack mitigation. *Discover Internet of Things*, 5(1). <https://doi.org/10.1007/s43926-025-00104-w>
- Ali, M. L., Thakur, K., Schmeelk, S., Debello, J., & Dragos, D. (2025). Deep Learning vs. Machine Learning for Intrusion Detection in Computer Networks: A Comparative Study. *Applied Sciences (Switzerland)*, 15(4). <https://doi.org/10.3390/app15041903>
- Alzahrani, I. Y., Lee, S., & Kim, K. (2024). Enhancing cyber-threat intelligence in the Arab world: Leveraging IoC and MISP integration. *Electronics*, 13(13). <https://doi.org/10.3390/electronics13132526>
- Ariyanto, Y., Syaifudin, Y. W., Ratsanjani, M. H., Muladawila, A. R., Fatmawati, T., Saputra, P. Y., & Setiadi, C. (2025). Cyber Threat Detection and Automated Response Using Wazuh and Telegram API. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, 25(1). <https://doi.org/10.30812/matrik.v25i1.5610>
- Bin Sarhan, B., & Altwaijry, N. (2023). Insider Threat Detection Using Machine Learning Approach. *Applied Sciences (Switzerland)*, 13(1). <https://doi.org/10.3390/app13010259>

- Chamkar, S. A., Zaydi, M., Maleh, Y., & Gherabi, N. (2025). Improving threat detection in Wazuh using machine learning techniques. *Journal of Cybersecurity and Privacy*, 5(2), 1–25. <https://doi.org/10.3390/jcp5020034>
- Chidukwani, A., Zander, S., & Koutsakis, P. (2024). Cybersecurity preparedness of small-to-medium businesses: A Western Australia study with broader implications. *Computers & Security*, 145, 104026. <https://doi.org/10.1016/j.cose.2024.104026>
- Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* 2023 16:1, 16(1), 4-. <https://doi.org/10.1186/S13040-023-00322-4>
- Dhanya, K. A., Vajipayajula, S., Srinivasan, K., Tibrewal, A., Kumar, T. S., & Kumar, T. G. (2023). Detection of network attacks using machine learning and deep learning models. *Procedia Computer Science*, 218, 57–66. <https://doi.org/10.1016/j.procs.2022.12.401>
- Deng, D. (2020). DBSCAN clustering algorithm based on density. Proceedings of the 2020 7th International Forum on Electrical Engineering and Automation, 949–953. <https://doi.org/10.1109/IFEEA51475.2020.00199>
- ENISA. (2025). ENISA, “Threat Landscape for Financial Sector,” European Union Agency for Cybersecurity. *ENISA Threat Landscape: Finance Sector*.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226–231).
- Hidayat, M. R. T., Widiyasono, N., & Gunawan, R. (2025). Optimasi Deteksi Malware Pada Siem Wazuh Melalui Integrasi Cyber Threat Intelligence Dengan Misp Dan Dfir-Iris. *Jurnal Informatika Dan Teknik Elektro Terapan*, 13(1). <https://doi.org/10.23960/jitet.v13i1.5686>
- Jakub Ziolk, D. B. J. L. I. B. I. C. G. F. K. L. Jussi O. (2025). ENISA Threat Landscape 2025 — Part 2. In *European Union Agency for Cybersecurity, ENISA* (Number October).
- Lee, S., Mujammami, A. A. H., & Kim, K. (2025). Leveraging Social Networks for Cyber Threat Intelligence: Analyzing Attack Trends and TTPs in the Arab World. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2024.3508025>
- Lella, I. (2024). *ENISA threat landscape 2024*. European Union Agency for Cybersecurity.
- Lones, M. A. (2024). Avoiding common machine learning pitfalls. *Patterns*, 5(10). <https://doi.org/10.1016/j.patter.2024.101046>
- Luay, M., Layeghy, S., Noorbin, N., Sarhan, M., Kulatilleke, G., Moustafa, N., & Portmann, M. (2026). Time Matters: Temporal NetFlow Features for ML-Based Network Intrusion Detection. *IEEE Access*, 14, 66899–66913. <https://doi.org/10.1109/ACCESS.2026.3688204>
- Manzoor, J., Waleed, A., Jamali, A. F., & Masood, A. (2024). Cybersecurity on a budget: Evaluating security and performance of open-source SIEM solutions for SMEs. *PLOS ONE*, 19(3), e0301183. <https://doi.org/10.1371/journal.pone.0301183>
- Nalliah, R. P., Praveen, S., Allareddy, T. V., Gajendrareddy, P., Lee, M. K., Oubaidin, M., & Allareddy, V. (2025). Cybersecurity threats and preparedness: Implications for dental schools. *Journal of Dental Education*, 89(4). <https://doi.org/10.1002/jdd.13758>
- Negara, B. S. dan S. (2025). *Threat Intelligence Report: Indonesian Critical Infrastructure Cybersecurity Status*. BSSN.
- Sahin, O., & Uludag, S. (2025). Leveraging inter-arrival time for efficient threat filtering: A parsimonious approach. *Computers & Security*, 154, 104471. <https://doi.org/10.1016/J.COSE.2025.104471>
- Shakir, Y. H., Abdelhamied, M. M., Al Mandhari, E. A. A., Alkhazraji, A., & Reda, N. M. (2026). An efficient approach for cyber-attack detection by using machine learning and deep learning algorithms. *IAES International Journal of Artificial Intelligence*, 15(2). <https://doi.org/10.11591/ijai.v15.i2.pp1219-1235>
- Tariq, A., Manzoor, J., Aziz, M. A., Tariq, Z. U. A., & Masood, A. (2023). Open source SIEM solutions for an enterprise. *Information and Computer Security*, 31(1). <https://doi.org/10.1108/ICS-09-2021-0146>
- Uccello, F., Pawlicki, M., D’Antonio, S., Kozik, R., & Choraś, M. (2024). Towards Hybrid NIDS: Combining Rule-Based SIEM with AI-Based Intrusion Detectors. *Lecture Notes in Networks and Systems*, 956 LNNS, 244–255. https://doi.org/10.1007/978-3-031-56950-0_21
- Vazão, A. P., Santos, L., Costa, R. L. de C., & Rabadão, C. (2023). Implementing and evaluating a GDPR-

- compliant open-source SIEM solution. *Journal of Information Security and Applications*, 75. <https://doi.org/10.1016/j.jisa.2023.103509>
- Wagner, C., Dulaunoy, A., Wagener, G., & Iklody, A. (2016). Misp: The design and implementation of a collaborative threat intelligence sharing platform. In *Proceedings of the 2016 ACM on workshop on information sharing and collaborative security* (pp. 49-56).
- Yin, L., Hu, H., Li, K., Zheng, G., Qu, Y., & Chen, H. (2023). Improvement of DBSCAN Algorithm Based on K-Dist Graph for Adaptive Determining Parameters. *Electronics (Switzerland)*, 12(15). <https://doi.org/10.3390/electronics12153213>