

AI-Based Cybercrime Attribution Using Probabilistic Clustering and Regularized Multi-Output Learning

John Dien^{1,*}, Les Endahti²

^{1,2}Informatic Management, AMIK YPAT Purwakarta, Indonesia

ABSTRACT

Cybercrime attribution remains a complex challenge due to the dynamic and ambiguous nature of attacker behavior. This study proposes a probabilistic framework for cybercrime attribution that combines Gaussian Mixture Model (GMM) clustering with regularized soft-label learning to identify latent attacker groups and model uncertainty in behavioral patterns. Using a cybersecurity intrusion dataset, clustering performance was optimized through silhouette analysis and stability evaluation, achieving the best configuration at two clusters (silhouette = 0.415, stability ARI = 0.956). The resulting probabilistic cluster memberships were used to train a multi-output Ridge regression model, which demonstrated strong generalization performance (Test Macro F1 = 0.964, CV F1 = 0.972 ± 0.005) with minimal overfitting. Feature analysis and statistical testing revealed that dynamic behavioral indicators, such as session duration and traffic intensity, significantly differentiate attacker groups, while static indicators contribute less. The findings identify two primary attacker archetypes: high-intensity network attackers and stealthy credential-based attackers. From a cyberlaw perspective, the proposed approach supports confidence-based attribution and enhances the transparency and reliability of digital forensic analysis. However, it emphasizes that probabilistic outputs should be interpreted as supporting evidence rather than definitive identification, highlighting the importance of uncertainty-aware AI systems in legal contexts.

Keywords Cybercrime Attribution, Probabilistic Clustering, Soft-Label Learning, Behavioral Analysis, Digital Forensics

INTRODUCTION

The rapid growth of cybercrime has significantly increased the complexity of identifying and attributing malicious activities in digital environments. Cybercrime attribution, defined as the process of linking cyber-attacks to responsible actors or groups, remains a fundamental challenge in both cybersecurity and cyberlaw domains. Traditional attribution techniques often rely on deterministic indicators such as IP tracing, signature-based detection, or known threat intelligence [1]. However, these approaches are increasingly ineffective due to the use of anonymization, proxy infrastructures, and adversarial tactics that obscure attacker identity [2]. As a result, attribution has shifted toward behavioral analysis, where patterns of activity are used to infer attacker characteristics rather than direct identification [3].

Recent advances in machine learning have introduced new approaches to cybercrime attribution, particularly through clustering and classification techniques. State-of-the-art methods include unsupervised clustering algorithms such as K-Means and Gaussian Mixture Models (GMM) to identify latent attacker groups, as well as supervised learning models for classifying attack patterns [4]. More advanced approaches incorporate deep learning and anomaly detection to capture complex behavioral relationships. Despite these developments, most existing studies rely on hard labeling, where each attack is

Submitted 20 May 2026
Accepted 18 July 2026
Published 1 September 2026

Corresponding author
John Dien,
johndien@amikypat.ac.id

Additional Information and
Declarations can be found on
[page 219](#)

© Copyright
2026 Dien and Endahti

Distributed under
Creative Commons CC-BY 4.0

assigned to a single class or actor group. This assumption fails to reflect the inherent uncertainty and overlap in attacker behavior, especially in real-world scenarios where attackers may adopt multiple strategies or mimic other groups.

Furthermore, existing approaches often lack robustness and interpretability. Many models are evaluated solely based on classification accuracy without considering cluster stability or reproducibility. In addition, the absence of probabilistic reasoning limits their applicability in legal contexts, where evidence must account for uncertainty and avoid overconfident conclusions. From a cyberlaw perspective, deterministic outputs can lead to misleading attribution claims, potentially affecting legal judgments and accountability.

Based on these limitations, several research gaps can be identified. First, there is a lack of frameworks that explicitly model uncertainty in cybercrime attribution, particularly through probabilistic representations. Second, few studies integrate unsupervised clustering with supervised learning in a unified pipeline that enhances both interpretability and generalization. Third, there is limited attention to model stability and statistical validation, which are critical for ensuring reliability in forensic and legal applications. Finally, the connection between AI-based attribution methods and legal admissibility or evidentiary standards remains underexplored.

To address these gaps, this study proposes a novel framework for cybercrime attribution that combines probabilistic clustering with regularized soft-label learning. Specifically, Gaussian Mixture Models are used to identify latent attacker groups and generate probabilistic cluster memberships, which are then used as soft labels in a multi-output Ridge regression model. This approach allows the model to capture uncertainty, improve generalization, and provide interpretable behavioral profiles. In addition, clustering quality is evaluated using both silhouette scores and stability analysis, while statistical significance testing is applied to validate feature differences across clusters [5].

Literature Review and Related Works

Cybercrime attribution has been widely studied in the fields of cybersecurity, digital forensics, and artificial intelligence. Traditional approaches to attribution primarily rely on signature-based detection, IP tracing, and rule-based systems to identify malicious actors. These methods, while effective in controlled environments, are often limited in real-world scenarios due to the increasing use of anonymization techniques, proxy networks, and sophisticated evasion strategies [6], [7]. As a result, researchers have shifted toward behavior-based analysis, where patterns of activity are used to infer attacker characteristics rather than direct identification [8].

Unsupervised learning techniques have played a significant role in identifying latent structures within cyber-attack data. Clustering algorithms such as K-Means, hierarchical clustering, and Gaussian Mixture Models (GMM) have been widely applied to group similar attack behaviors [9], [10]. Among these, GMM offers advantages due to its probabilistic foundation, allowing data points to belong to multiple clusters with varying degrees of membership [11]. This property makes GMM particularly suitable for modeling overlapping attacker behaviors, which are common in cyber threat scenarios [12]. However, many studies still rely on hard cluster assignments, limiting their ability to capture uncertainty.

In addition to clustering, supervised machine learning methods have been extensively used for intrusion detection and attack classification. Algorithms such as Support Vector Machines (SVM), Random Forests, and Neural Networks have demonstrated strong performance in identifying known attack patterns [13], [14]. More recent approaches utilize deep learning models, including recurrent and convolutional neural networks, to capture temporal and spatial dependencies in network traffic data [15]. Despite their high accuracy, these models often require large labeled datasets and lack interpretability, which can hinder their application in forensic and legal contexts [16].

To address the limitations of purely supervised or unsupervised approaches, hybrid models have been proposed that combine clustering with classification. These methods aim to leverage the strengths of both paradigms by first identifying latent groups and then learning predictive models based on these structures [17]. However, most existing hybrid approaches still convert probabilistic outputs into hard labels, thereby discarding valuable uncertainty information. This simplification can lead to overconfident predictions and reduced robustness in ambiguous cases [18].

Another important aspect of cybercrime attribution is model evaluation and validation. While many studies focus on accuracy and precision, fewer works consider cluster stability, reproducibility, and statistical significance [19]. These factors are crucial for ensuring that the discovered patterns are not artifacts of specific datasets but represent consistent behavioral structures. Furthermore, the lack of probabilistic reasoning in many models limits their applicability in legal settings, where uncertainty and confidence levels must be explicitly addressed [20].

In summary, the current state of the art demonstrates significant progress in applying machine learning to cybercrime attribution, particularly through clustering and classification techniques. However, key limitations remain, including the reliance on hard labels, insufficient modeling of uncertainty, limited integration between unsupervised and supervised learning, and a lack of focus on legal interpretability. These challenges highlight the need for probabilistic, interpretable, and robust frameworks that can bridge the gap between technical analysis and cyberlaw requirements.

Methodology

This study proposes a probabilistic framework for cybercrime attribution that integrates unsupervised clustering with regularized soft-label learning. The overall research workflow consists of data preprocessing, feature engineering, dimensionality reduction, probabilistic clustering, soft-label modeling, and evaluation. As illustrated in [figure 1](#), the proposed research framework begins with data preparation and feature engineering, followed by preprocessing and dimensionality reduction using PCA. The transformed data is then clustered using Gaussian Mixture Models (GMM) to obtain probabilistic cluster memberships. These probabilities are subsequently used as soft labels in a regularized regression model, and the final outputs are evaluated and interpreted for cybercrime attribution.

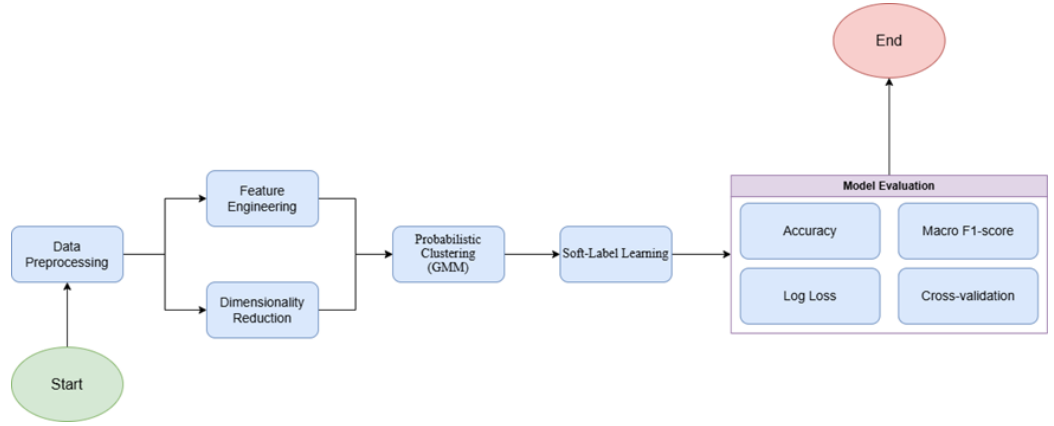


Figure 1 Proposed research workflow

Data Preparation

The dataset consists of network session records that have been preprocessed by selecting only instances labeled as attacks (`attack_detected = 1`), ensuring that the analysis specifically targets malicious behavior patterns. To improve data quality and maintain consistency, duplicate records are identified and removed, preventing redundancy that could bias the model during training. Following preprocessing, the cleaned dataset is divided into training and testing subsets using an 80:20 split, where the training set is used to build and learn the model, and the testing set is used to evaluate its performance and generalization capability on unseen data.

Feature Engineering

To capture attacker behavior, several derived features are constructed:

Login Failure Ratio:

$$\text{Login Failure Ratio} = \frac{\text{Failed Logins}}{\max(\text{Login Attempts}, 1)} \quad (1)$$

Packet per Second:

$$\text{Packet Rate} = \frac{\text{Network Packet Size}}{\max(\text{Session Duration}, 1)} \quad (2)$$

Risk Interaction:

$$\text{Risk Interaction} = \text{IP Reputation Score} \times \text{Unusual Time Access} \quad (3)$$

Attempt Intensity:

$$\text{Attempt Intensity} = \frac{\text{Login Attempts}}{\max(\text{Session Duration}, 1)} \quad (4)$$

These features represent both intensity and behavioral patterns of cyber-attacks.

Data Preprocessing

A preprocessing pipeline is applied using column-wise transformations:

For numerical features:

Median imputation:

$$x = \text{median}(x) \quad (5)$$

Power transformation (Yeo-Johnson):

$$x' = \text{YeoJohnson}(x) \quad (6)$$

Robust scaling:

$$x'' = \frac{x' - \text{median}(x')}{IQR(x')} \quad (7)$$

For categorical features, missing values are handled using mode imputation, where the most frequently occurring value in each feature is used to fill in any gaps, ensuring that the distribution of the data remains representative. After handling missing values, categorical variables are transformed using one-hot encoding, which converts each category into a binary vector representation, allowing the model to process non-numeric data effectively. This preprocessing step ensures that all features are properly structured and compatible with machine learning algorithms, contributing to more reliable and consistent model performance.

Dimensionality Reduction

Principal Component Analysis (PCA) is used to reduce feature dimensionality:

$$Z = XW \quad (8)$$

X is the original feature matrix and W is the projection matrix.

PCA helps remove redundancy and improve clustering performance.

3.5 Probabilistic Clustering

Gaussian Mixture Models (GMM) are used to model latent attacker groups:

$$P(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \quad (9)$$

K = number of clusters

π_k = mixing coefficients

μ_k, Σ_k = mean and covariance

Cluster membership probabilities are computed as:

$$P(z_k \mid x) = \frac{\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x \mid \mu_j, \Sigma_j)} \quad (10)$$

Model Selection for Clustering

Each configuration is evaluated using:

Silhouette score:

$$S = \frac{b - a}{\max(a, b)} \quad (11)$$

Stability (Adjusted Rand Index):

$$ARI = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} \quad (12)$$

Information criteria:

$$BIC = -2\ln(L) + p\ln(n) \quad (13)$$

A composite score is used:

$$\text{Score} = 0.70 \cdot \text{Silhouette} + 0.30 \cdot \text{Stability} - 0.001 \cdot K \quad (14)$$

Soft-Label Learning

Instead of hard labels, probabilistic cluster memberships are used as targets.

The model is defined as:

$$\hat{Y} = f(X) \quad (15)$$

Using Ridge regression:

$$\min_W ||Y - XW||^2 + \alpha ||W||^2 \quad (16)$$

Y = soft labels (cluster probabilities)

α = regularization parameter

Model Evaluation

The model is evaluated using:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{N} \quad (17)$$

Macro F1-score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

Mean Squared Error:

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2 \quad (19)$$

Log Loss:

$$LogLoss = -\sum y \log(\hat{y}) \quad (20)$$

Cross-validation (5-fold) is used to assess generalization.

Statistical Analysis

To validate cluster differences, one-way ANOVA is applied:

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} \quad (21)$$

Low p-values indicate statistically significant differences.

Algorithm 1: Probabilistic Cybercrime Attribution

Input: dataset D , number of clusters K , PCA components m , regularization parameter α

Output: cluster labels Z , soft probabilities P , trained model $f(X)$

Process:

Start

Data Filtering

$$D' = \{x_i \in D \mid \text{attack_detected}(x_i) = 1\}$$

Feature Engineering

for each $x_i \in D'$:

$$f_1 = \frac{\text{failed_logins}}{\max(\text{login_attempts}, 1)}$$

$$f_2 = \frac{\text{network_packet_size}}{\max(\text{session_duration}, 1)}$$

$$f_3 = \frac{\text{ip_reputation} \times \text{unusual_time_access}}{\text{login_attempts}}$$

$$f_4 = \frac{\text{login_attempts}}{\max(\text{session_duration}, 1)}$$

$$X_i = [x_i, f_1, f_2, f_3, f_4]$$

Preprocessing

$$X^{(num)} = \text{RobustScaler}(\text{PowerTransform}(\text{Impute}(X^{(num)})))$$

$$X^{(cat)} = \text{OneHot}(\text{Impute}(X^{(cat)}))$$

$$X = [X^{(num)}, X^{(cat)}]$$

PCA Transformation

$$Z = XW$$

GMM Clustering

$$P(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

$$z_i = \arg \max_k P(z_k \mid x_i)$$

$$P_{ik} = P(z_k \mid x_i)$$

Model Selection

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

$$\text{Score} = 0.70S + 0.30ARI - 0.001K$$

$$(K^*, m^*) = \arg \max \text{Score}$$

Soft-Label Learning

	$\min Y - XW ^2 + \alpha W ^2$
Prediction	$\hat{Y} = f(X_{test})$ $\hat{z}_i = \arg \max_k \hat{Y}_{ik}$
Evaluation	$Accuracy = \frac{TP + TN}{N}$ $F1 = \frac{2PR}{P + R}$ $MSE = \frac{1}{n} \sum (y - \hat{y})^2$
Output	
Return	$(Z, P, f(X))$
End	

Results

Clustering Performance and Model Selection

The clustering phase was conducted to uncover latent behavioral patterns among attack sessions using Gaussian Mixture Models (GMM), with multiple configurations tested across different feature sets, numbers of principal components from PCA, and cluster sizes to identify the most effective setup. The optimal configuration was achieved with two clusters, two PCA components, a silhouette score of 0.415, and a cluster stability score measured by Adjusted Rand Index (ARI) of 0.956. The silhouette score indicates a moderate level of separation between clusters, suggesting that while distinct behavioral groupings exist, there is still some overlap between them. In contrast, the high ARI value demonstrates that the clustering results are highly consistent and reproducible across different runs. As illustrated in figure 2, the distribution of silhouette scores across varying numbers of clusters shows that increasing the number of clusters does not lead to a significant improvement in cluster separation, reinforcing the choice of K = 2 as the most appropriate configuration.

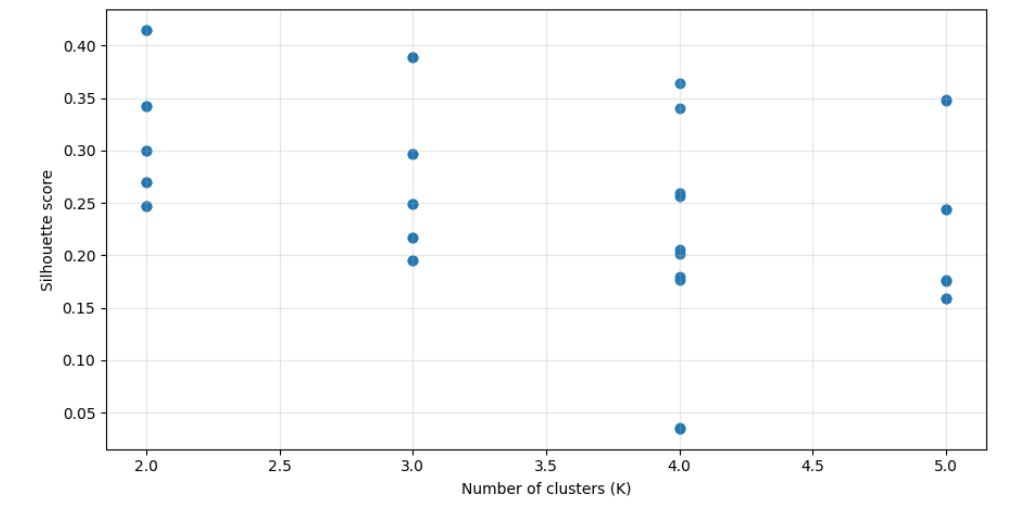


Figure 2 Silhouette scores across candidate clustering configurations

Furthermore, figure 3 illustrates the relationship between silhouette score and cluster stability, showing how different configurations balance separation quality and reproducibility. The selected configuration with K = 2 is positioned in a

region characterized by high stability and relatively strong separation, indicating that it achieves an effective trade-off between these two metrics. This suggests that while higher numbers of clusters do not substantially improve separation, they may introduce unnecessary complexity without enhancing consistency, thereby reinforcing the suitability of the chosen configuration.

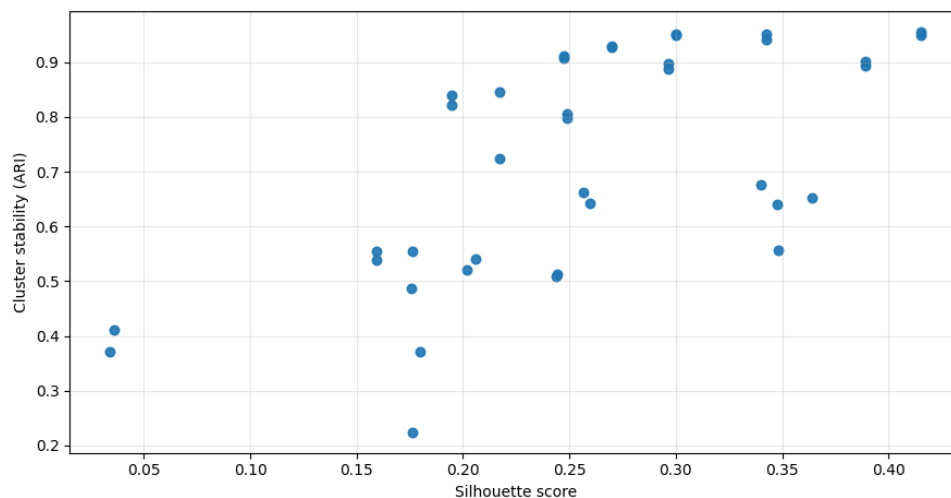


Figure 3 Relationship between silhouette score and cluster stability (ARI)

Visualization of Latent Actor Groups

To interpret the clustering behavior, Principal Component Analysis (PCA) was applied to reduce the data into two dimensions, enabling clearer visualization of the underlying structure. As shown in [figure 4](#), the identified actor groups display a noticeable separation along the first principal component, indicating that this component captures the most significant variation in attack behavior. However, some overlap is observed at the boundary between clusters, suggesting that the behavioral patterns of attackers are not entirely distinct. This overlap highlights the inherent complexity and variability of malicious activities, reinforcing the suitability of probabilistic or soft clustering approaches such as GMM, which allow data points to belong to multiple clusters with varying degrees of membership rather than enforcing rigid boundaries.

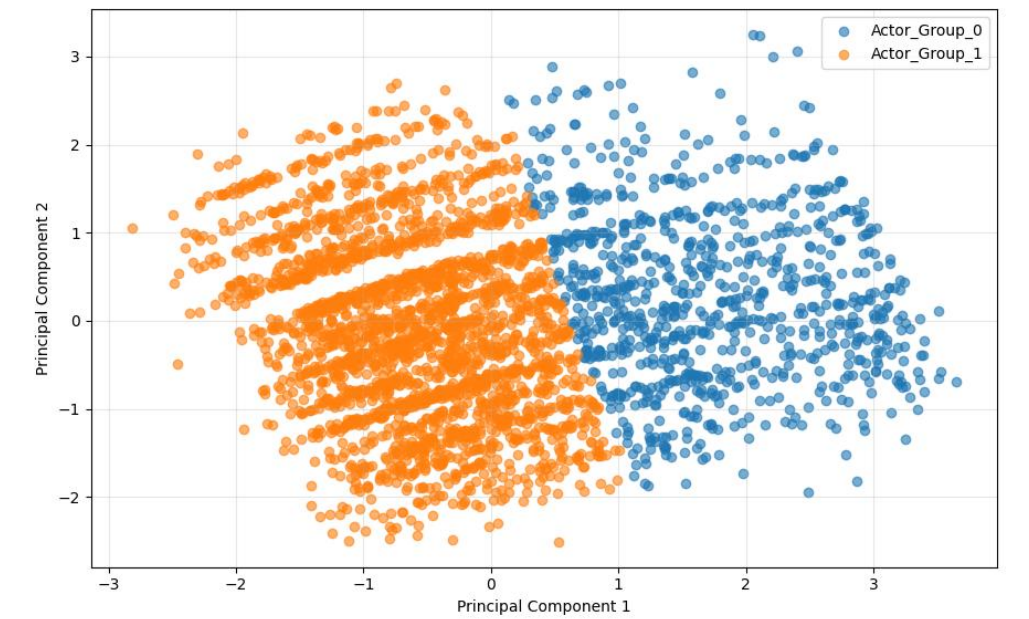


Figure 4 PCA visualization of inferred actor groups

Soft-Label Model Performance

A regularized soft-label learning model was developed using multi-output Ridge regression applied to probabilistic cluster assignments, allowing the model to capture uncertainty in cluster membership rather than relying on hard labels. The regularization parameter was carefully tuned, with the best-performing model obtained at $\alpha = 50$, as this value provided a strong balance between bias and variance while maintaining model stability and generalization capability. This configuration helps prevent overfitting by penalizing large coefficients, ensuring that the model remains robust when applied to unseen data. The overall performance of the model under this optimal setting is summarized in table 1.

Table 1 Performance of the regularized soft-label model		
Metric	Train	Test
Accuracy	0.9807	0.9742
Macro F1-score	0.9744	0.9641
Log Loss	0.2106	0.2237
Mean Squared Error	0.0159	0.0160

The model demonstrates strong generalization performance, as indicated by a cross-validation F1 score of 0.9725 with a low standard deviation of ± 0.0052 , reflecting both high accuracy and consistency across different data splits. Additionally, the small F1 gap of 0.0103 suggests minimal performance degradation between training and validation, indicating that the model is not significantly overfitting. As illustrated in figure 5, increasing the level of Ridge regularization contributes to reducing overfitting by constraining model complexity, while still maintaining stable and high predictive accuracy, further supporting the selection of the optimal regularization parameter.

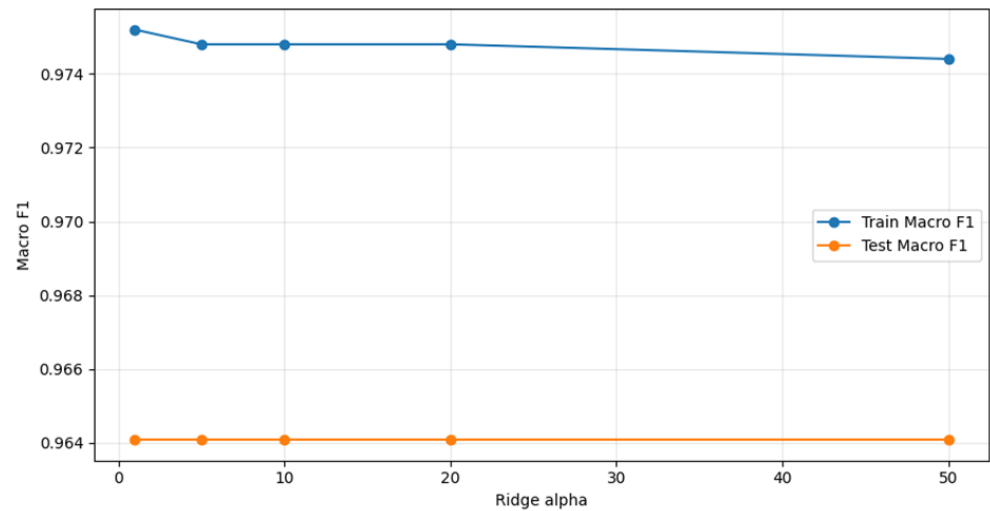


Figure 5 Ridge regularization (alpha) vs model performance

Classification Behavior Analysis

The classification behavior is further evaluated using a confusion matrix to better understand how well the model distinguishes between the identified clusters. As shown in figure 6, the model achieves high overall predictive accuracy, with only minor misclassification observed in one of the clusters. A more detailed analysis reveals that Actor_Group_1 attains near-perfect recall of 1.00, indicating that instances belonging to this group are consistently and correctly identified. In contrast, Actor_Group_0 exhibits a slightly lower recall of 0.90, suggesting that a small portion of its instances are misclassified. This difference implies that Actor_Group_1 represents a more behaviorally consistent pattern, whereas Actor_Group_0 displays greater variability and overlap with other behavioral characteristics, making it slightly more challenging for the model to classify accurately.

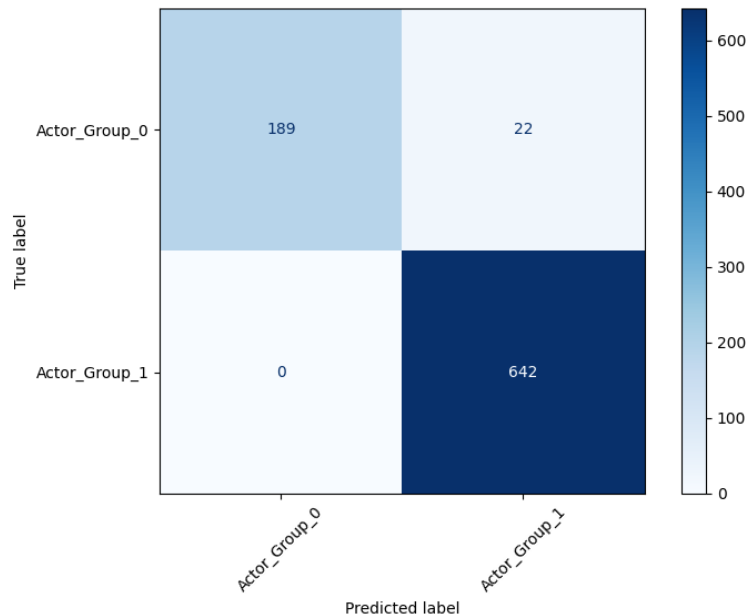


Figure 6 Confusion matrix of the regularized soft-label model

Cluster Behavioral Profiles

To interpret cluster characteristics, feature averages were computed for each group, as summarized in [table 2](#), providing insight into the underlying behavioral patterns captured by the clustering process. The results reveal two distinct profiles, where Actor_Group_0 is characterized by high-intensity and short-duration activity, indicating more aggressive but brief attack behavior, while Actor_Group_1 exhibits low-intensity and long-duration behavior, suggesting more sustained and less conspicuous activity over time.

Table 2 Cluster behavioral profiles		
Feature	Actor_Group_0	Actor_Group_1
Network Packet Size	518.33	492.13
Login Attempts	5.37	4.38
Session Duration	135.02	1072.52
IP Reputation Score	0.367	0.375
Failed Logins	1.82	1.98
Unusual Time Access	0.148	0.155
Login Failure Ratio	0.469	0.673
Packet per Second	14.10	0.78
Risk Interaction	0.052	0.058
Attempt Intensity	0.132	0.006

Statistical Significance Analysis

To validate the differences between clusters, one-way ANOVA tests were conducted on the extracted features, with the results summarized in [table 3](#). The extremely low p-values obtained confirm that there are statistically significant differences between the identified clusters for the majority of features. In contrast, features such as IP reputation and unusual time access were found to be not statistically significant, indicating that these attributes contribute less to distinguishing between groups. This finding suggests that dynamic behavioral features play a more important role in differentiating attack patterns compared to more static or contextual indicators.

Table 3 ANOVA results for feature significance		
Feature	F-statistic	p-value
session_duration	1293.74	1.16e-247
attempt_intensity	281.56	3.05e-61
packet_per_second	248.08	2.27e-54
login_attempts	160.68	3.59e-36
login_failure_ratio	74.72	7.56e-18

Heuristic Interpretation of Threat Actors

ed on the identified cluster characteristics, heuristic labels were assigned to provide meaningful interpretations of each group, where Actor_Group_0 is labeled as a High-Volume Network Attacker and Actor_Group_1 as a Credential-Based Stealth Attacker. These labels reflect the underlying behavioral patterns observed in each cluster, linking quantitative analysis to practical cybersecurity insights. The distribution of the inferred attacker types is

illustrated in figure 7, showing how the dataset is divided between these two categories. Additionally, figure 8 presents the distribution of cluster assignment confidence, highlighting the probability associated with each sample’s cluster membership. The results indicate that most samples exhibit high assignment probabilities, suggesting that the clustering outcomes are reliable and that the model has strong confidence in assigning instances to their respective behavioral groups.

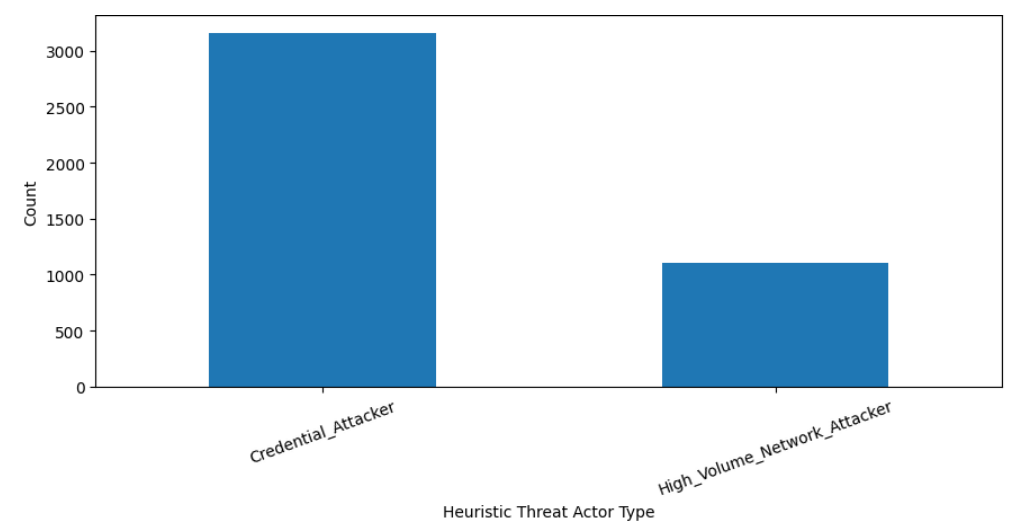


Figure 7 Distribution of inferred threat actor types

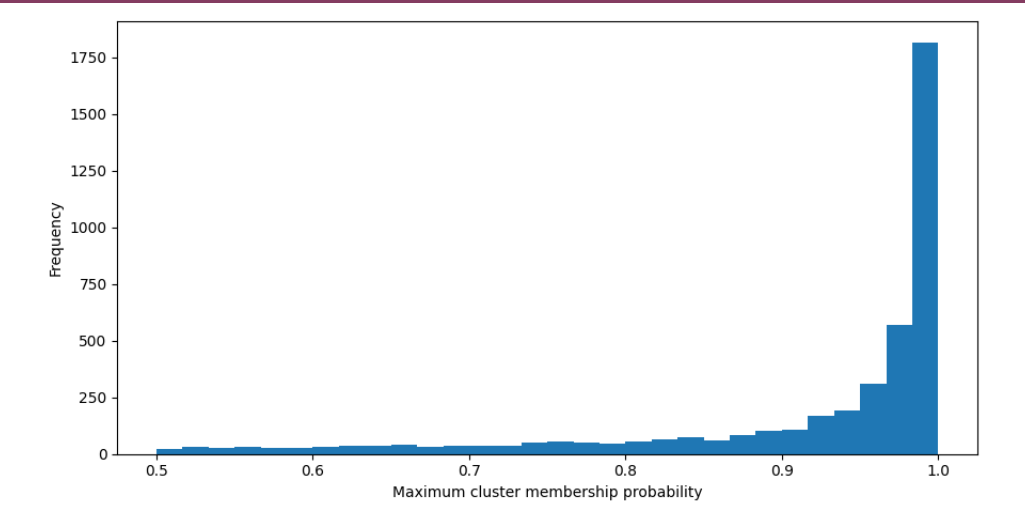


Figure 8 Distribution of cluster membership confidence

Discussion

The results demonstrate that cyber-attack behaviors can be effectively modeled as stable probabilistic clusters, rather than rigid categorical identities. The high cluster stability ($ARI \approx 0.956$) combined with moderate silhouette scores suggests that attacker behaviors exhibit both consistency and overlap, reflecting the complex and adaptive nature of cyber threats. Importantly, the use of soft-label learning allows the model to represent uncertainty in attribution, where a single event may partially belong to multiple behavioral profiles. This probabilistic perspective provides a more realistic representation of cyber

operations compared to traditional deterministic classification approaches, which often overstate certainty and fail to capture ambiguity in attacker behavior.

From a cyberlaw perspective, these findings have significant implications for digital evidence and attribution. The proposed framework supports a shift from definitive attribution claims toward confidence-based, evidence-driven assessments, which are more aligned with legal standards of proof and expert testimony. However, the probabilistic nature of the model also introduces challenges, particularly regarding interpretability, admissibility, and potential misuse in legal proceedings. While the model can strengthen forensic analysis by providing reproducible and statistically validated insights, it should not be interpreted as direct proof of identity but rather as supporting evidence of behavioral similarity. Therefore, its application in legal contexts must be accompanied by transparency, methodological rigor, and careful consideration of uncertainty to avoid over-reliance on automated attribution systems.

Conclusion

This study presents a probabilistic framework for cybercrime attribution that integrates Gaussian Mixture Model-based clustering with regularized soft-label learning to capture latent attacker behaviors and their inherent uncertainty. The results demonstrate that the proposed approach achieves strong clustering stability, high predictive performance, and minimal overfitting, while revealing two distinct behavioral attacker archetypes characterized by intensity and persistence. By modeling attribution as a probabilistic process rather than a deterministic outcome, this work provides a more realistic and legally appropriate representation of cyber-attack analysis. The framework offers valuable support for cybersecurity forensics and cyberlaw applications by enabling transparent, reproducible, and confidence-aware attribution, while emphasizing that such models should be used as complementary evidence rather than definitive proof of attacker identity.

Declarations

Author Contributions

Conceptualization: J.D. and L.E.; Methodology: L.E.; Software: J.D.; Validation: J.D. and L.E.; Formal Analysis: J.D. and L.E.; Investigation: J.D.; Resources: L.E.; Data Curation: L.E.; Writing Original Draft Preparation: J.D. and L.E.; Writing Review and Editing: L.E. and J.D.; Visualization: J.D.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] R. Rid and B. Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies*, vol. 38, no. 1–2, pp. 4–37, Jan. 2015, doi: 10.1080/01402390.2014.977382.
- [2] S. Goel and B. Nussbaum, "Attribution Across Cyber Attack Types: Network Intrusions and Information Operations," *IEEE Open Journal of the Communications Society*, vol. 2, no. April, pp. 1082–1093, Jan. 2021, doi: 10.1109/OJCOMS.2021.3074591.
- [3] J. R. Goodall, W. G. Lutters, and A. Komlodi, "The Work of Intrusion Detection: Rethinking the Role of Security Analysts," in *Proc. 10th Americas Conf. Information Systems (AMCIS)*, New York, NY, USA, Aug. 2004.
- [4] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007, doi: 10.1016/j.comnet.2007.02.001.
- [5] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [6] A. Vikram, Mohana, and R. V. College of Engineering, "Anomaly Detection in Network Traffic Using Unsupervised Machine Learning Approach," in *Proc. 2020 5th Int. Conf. Communication and Electronics Systems (ICCES)*, vol. 2020, no. July, pp. 476–479, Jun. 2020, doi: 10.1109/ICCES48766.2020.9137987.
- [7] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proc. 2010 IEEE Symposium on Security and Privacy (S&P)*, Berkeley/Oakland, CA, USA, vol. 2010, no. July, pp. 305–316, May 2010, doi: 10.1109/SP.2010.25.
- [8] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003, doi: 10.1109/TSP.2003.814797.
- [9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.
- [10] M. Nagode, S. Oman, J. Klemenc, and B. Panić, "Finite Mixture Models: A Key Tool for Reliability Analyses," *Mathematics*, vol. 13, no. 10, p. 1605, pp. 1–24, May 2025, doi: 10.3390/math13101605.
- [11] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [12] S. Jajodia, P. Liu, V. Swarup, and C. Wang, *Cyber Situational Awareness: Issues and Research*. Boston, MA, USA: Springer, 2010, doi: 10.1007/978-1-4419-0140-8.
- [13] W. Lee and S. J. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems," *ACM Transactions on Information and System*

- Security*, vol. 3, no. 4, pp. 227-261, 2001, doi: 10.1145/382912.382914.
- [14] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, no. June, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
- [15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [16] Z. C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36-43, Oct. 2018, doi: 10.1145/3233231.
- [17] X. Xu and X. Wang, "An Adaptive Network Intrusion Detection Method Based on PCA and Support Vector Machines," in *Lecture Notes in Computer Science, Proc. 1st Int. Conf. Advanced Data Mining and Applications (ADMA)*, Wuhan, China, vol. 2005, no. July, pp. 696-703, Jul. 2005, doi: 10.1007/11527503_82.
- [18] T. Fawcett, "Introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
- [19] G. Youness and G. Saporta, "Comparing partitions of two sets of units based on the same variables," *Advances in Data Analysis and Classification*, vol. 4, no. 1, pp. 53–64, Apr. 2010, doi: 10.1007/s11634-009-0057-4.
- [20] H. F. Atlam, E. E.-D. Hemdan, A. Alhomoud, and M. O. Alassafi, "Internet of Things Forensics: A Review," *Internet of Things*, vol. 11, no. September, pp. 1-24, May 2020, doi: 10.1016/j.iot.2020.100220.