A Hybrid Ensemble Framework for Privilege Anomaly Detection in Multi-Cloud Environments

Chi-Sung Kim

Abstract

With the rapid expansion of cloud computing, privilege abuse and escalation attacks in multi-cloud environments are emerging as a core security threat for organizations. Existing research shows limitations due to a lack of context from reliance on single data sources and the constraints of simple ensemble structures. To overcome these limitations, this study proposes a large-scale, multi-modal hybrid ensemble framework that integrates a total of 884,912 rows of data from the IBM Cloud Dataset, Microsoft Cloud Monitoring Dataset, and flaws.cloud CloudTrail data.

The research methodology is as follows: (1) Large-Scale Data Integration: A total of 2,204,017 rows of raw data collected from three platforms were refined into an analysis dataset of 769,454 rows through time synchronization and 5-minute window aggregation. (2) Hybrid Ensemble Architecture: Isolation Forest, which is robust for outlier detection in high-dimensional data, and LSTM, specialized for learning sequential event patterns, are combined on a weight-based basis to simultaneously learn spatial and temporal anomaly patterns. (3) Multi-modal Feature Fusion: Infrastructure metrics, time-series service data, and privilege events were integrated to construct 15 standard feature vectors.

In the experimental results, the proposed hybrid ensemble model recorded an ROC-AUC score of 0.721, showing superior overall performance compared to individual models. In particular, it demonstrated overwhelming performance in the Precision-Recall Curve analysis compared to other models, proving that it achieved a balance between reducing false positives and detection efficiency, which is crucial in real-world security environments. Furthermore, by explaining the model's prediction results through SHAP (SHapley Additive exPlanations) analysis, it secured transparency that allows security analysts to trust and act upon the findings.

☞ keyword : Multi-Cloud Security, Privilege Anomaly Detection, Hybrid Ensemble, Explainable AI (XAI), ROC-AUC, Precision-Recall Curve

1. Introduction

1.1 Research Background and Necessity

As the adoption of cloud computing accelerates the migration of core organizational infrastructure to multi-cloud environments, consistent security threat detection across platforms has emerged as an essential task for enterprises. According to the 2024 report by the Cloud Security Alliance (CSA), 78% of security incidents in multi-cloud environments are related to inconsistencies in privilege management across platforms, with Privilege Escalation attacks being classified as the most critical threat.

Existing research on anomaly detection has primarily focused on single-cloud platforms, failing to adequately reflect the complexity of multi-cloud scenarios operating in real-world corporate environments. Specifically, since privilege escalation attacks are characterized by the sequential use of multiple cloud services, an anomaly detection approach from an integrated perspective is necessary.

1.2 Limitations of Existing Research

Studies on cloud anomaly detection conducted to date have several fundamental limitations. Existing research tends to rely on a single data source, primarily CloudTrail logs. This makes multi-faceted analysis incorporating infrastructure metrics, network traffic, and external threat intelligence difficult, thus limiting the ability to detect various attack scenarios. Furthermore, most studies are validated on limited datasets, ranging from thousands to tens of thousands of rows, which fails to adequately reflect the complexity encountered in large-scale operational environments.

Moreover, from a structural perspective of detection models, they often stop at simply combining the outputs of Isolation Forest and LSTM, showing vulnerability in identifying sophisticated,

complex attacks or anomalous signs that occur over long periods. Finally, most of the developed models are specialized for the AWS (Amazon Web Services) environment, pointing to a limitation in generalization that makes them difficult to apply equally to other cloud platforms such as Azure or GCP (Google Cloud Platform).

2. Related Work

2.1 Trends in Anomaly Detection Research for Cloud Environments

The complexity and dynamic nature of cloud environments have clearly exposed the limitations of traditional signature-based detection systems. Consequently, approaches based on Anomaly Detection, which learn a profile of a system's normal behavior and identify patterns that deviate from it, have emerged as a core research topic. Early research tended to focus primarily on single data sources. This approach utilizes logs generated by systems and applications as the fundamental data for understanding user behavior. For example, DeepLog, proposed by Du et al. (2017), is considered a pioneering study that learns sequential patterns of system event logs using LSTM (Long Short-Term Memory) to detect anomalies deviating from normal log sequences (Du et al., 2017). However, log data alone has limitations in capturing the overall state of the system or the complex interactions between resources.

Next, Metrics-based Analysis is a method that utilizes numerical data such as CPU usage, network traffic, and memory consumption to intuitively grasp the system's state. The recently released IBM cloud dataset by Schneider et al. (2024) includes over 110,000 high-dimensional metrics, demonstrating what a challenging task it is to detect anomalies in such large-scale, real-world telemetry data (Schneider et al., 2024). In such high-dimensional data environments, the effectiveness of traditional statistics-based methodologies diminishes, thus requiring new approaches.

To overcome the limitations of these single data sources, recent research has evolved towards a Multi-modal Approach, which integrates various types of data—such as logs, metrics, and topology—to enhance analysis accuracy. CloudRanger by Wu et al. (2021) is a prime example of this approach, conducting research to identify the root causes of issues in cloud systems by integrating diverse data sources using a Bayesian Network (Wu et al., 2021). This study also lies on the extension of this multi-modal approach and aims to expand the scope of analysis one step further by integrating heterogeneous data from multiple cloud platforms.

2.2 Behavior-based Anomaly Detection Models

Research on detecting anomalies by modeling user behavior patterns has evolved through various machine learning methodologies. A prominent trend is Unsupervised Outlier Detection, which is an essential approach in real-world operational data environments where labels are absent. Isolation Forest is an algorithm that efficiently isolates and detects anomalous data in a random tree structure based on the characteristic that 'anomalies are few and different' (Liu et al., 2008). It is regarded as a powerful alternative that is computationally efficient, especially for high-dimensional data like the IBM dataset, and can overcome the limitations of distance-based methodologies. Furthermore, as an attacker's behavior often exhibits sequential patterns spanning multiple stages, Sequential Data Modeling is also a crucial area of research.

LSTM has become a standard model for time-series and sequential data analysis since being proposed to solve the long-term dependency problem of Recurrent Neural Networks (RNNs) (Hochreiter & Schmidhuber, 1997). As seen in the case of DeepLog (Du et al., 2017), an LSTM trained on normal API call sequences can effectively detect when an event with an

abnormal sequence occurs. Furthermore, Graph-based Modeling is utilized to effectively represent the complex interactions among users, resources, and API calls within a cloud environment. HOLMES, proposed by Milajerdi et al. (2019), demonstrated the validity of graph models by modeling system audit logs as an Information Flow Graph to detect multi-stage Advanced Persistent Threat (APT) attacks (Milajerdi et al., 2019). To effectively learn these graph structures, Graph Neural Network (GNN) technologies like Graph Convolutional Networks (GCN) and GraphSAGE are being actively researched, providing an important theoretical foundation for the future direction of this study (Kipf & Welling, 2016; Hamilton et al., 2017).

2.3 Hybrid Ensembles and Explainable AI

As the recognition spreads that a single model is insufficient to cover all types of anomalies in complex cloud environments, research on Hybrid Ensembles, which combine multiple models, is gaining attention. Just as Isolation Forest excels at capturing the peculiarity of individual events (point anomalies) and LSTM excels at capturing the sequence and context of events (contextual anomalies), combining models with different strengths can build a more robust and comprehensive detection system.

However, these complex ensemble models often operate like a 'black box', creating the problem of it being difficult to understand why a specific alert was triggered. In a real-world Security Operations Center (SOC), an analyst must be able to trust a model's prediction to take action, making explanations for these predictions essential. To solve this problem, research into Explainable AI (XAI) is being actively pursued. Representative XAI techniques include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP, proposed by Lundberg & Lee (2017), is a powerful framework based on game theory that quantitatively explains how much each feature contributed to the model's final prediction. This allows analysts to understand which factors were the decisive basis for an anomaly detection (Lundberg & Lee, 2017). Additionally, LIME, proposed by Ribeiro et al. (2016), is a technique that provides local explanations by sampling data around a specific prediction and training an interpretable linear model on it (Ribeiro et al., 2016).

This study is differentiated from existing research in that it goes beyond simply developing a high-performance model. By integrating XAI techniques like SHAP, it aims to secure the model's transparency and reliability, and ultimately provide analysts with actionable insights.

3. Research Methodology

To address the complex problem of privilege anomaly detection in multi-cloud environments, this study adopts a step-by-step approach that systematically breaks down the problem and applies optimized techniques at each stage. Based on the recognition that a single data source and a single model cannot effectively detect the multi-faceted attacks of the real world, the methodology of this study is designed around three core principles: ensuring data diversity, extracting meaningful information, and combining complementary models.

The first principle, ensuring data diversity, aims to solve the problem of 'context fragmentation.' Data collected from individual cloud platforms only shows activity within that platform and fails to provide the complete picture of sophisticated attacks that traverse multiple clouds. Therefore, this study aims to provide a robust foundation for the model to learn anomaly patterns from a broader perspective by integrating data from heterogeneous platforms (IBM, Microsoft, AWS) to construct a large-scale, multi-modal dataset.

The second principle, extracting meaningful information, is the process of transforming raw data into a format that the model can understand. Since raw logs and metric data are unstructured or high-dimensional and thus difficult to use as is, they undergo data preprocessing and standardization. Subsequently, Multi-modal

Feature Engineering—encompassing temporal characteristics, security events, and statistical properties—is performed to amplify the latent anomaly signals in the raw data and enable the model to learn patterns more easily.

The third principle, combining complementary models, is necessary to respond to various forms of cloud attacks. Cloud attacks can be categorized into 'Point Anomalies,' where an individual event is itself abnormal, and 'Contextual Anomalies,' where individual events are normal, but their sequence or combination is abnormal. As a single model struggles to effectively detect all these types, this study adopts a Hybrid Ensemble architecture that combines Isolation Forest, which is strong at detecting point anomalies, with LSTM, which excels at understanding temporal sequences and context. Through this, we aim to leverage the complementary strengths of both models to maximize the detection scope, reduce false positives, and enhance the overall Robustness of the model.

In accordance with this design philosophy, the following sections will describe in detail the dataset construction and preprocessing, multi-modal feature engineering, and the design and implementation of the hybrid ensemble architecture.

3.1 Dataset and Preprocessing

3.1.1 Construction of a Large-Scale, Multi-modal Dataset

In this study, we constructed a raw dataset of 2,204,017 rows in total by integrating real-world operational data from three cloud platforms, as shown in Table 1.

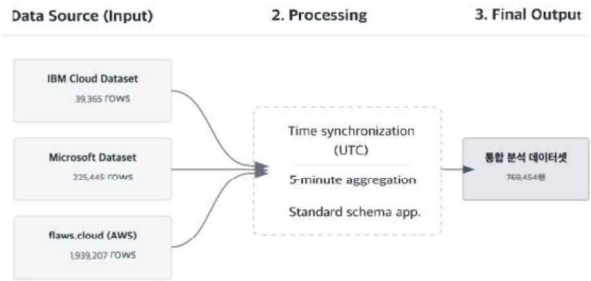
(Table 1 Original Dataset Information)

Dataset	Description	Period	Scale
IBM Cloud Dataset	High-dimensional Telemetry Data	2024.01 - 2024.06	39,365 rows × 117,448 columns
Microsoft Cloud Monitoring	Time-series Service Metrics	2017.11 - 2018.07	225,445 rows
flaws.cloud CloudTrail	AWS Privilege Event Logs	2017.02 - 2020.10	1,939,207 rows

3.1.2 Data Integration and Standardization Framework

To ensure consistency in data collected from heterogeneous cloud environments, an integration framework as shown in Figure 1 was applied. First, time synchronization was performed by normalizing all data timestamps to the UTC standard. Afterward, the data was resampled at 5-minute intervals to align the time series. Finally, to ensure analytical consistency, the schema was unified into 15 standard columns, creating an analysis dataset with a total of 769,454 rows.

(Figure 1: Data Integration and Standardization Pipeline (Conceptual Diagram))



3.2 Multi-modal Feature Engineering

3.2.1 The Necessity of Feature Engineering

Raw data contains several issues that make it difficult for machine learning models to learn from directly. The more than 110,000 metrics in the IBM dataset or the semi-structured JSON logs from CloudTrail are 'information' in themselves, but not 'knowledge' that a model can learn from. Feature engineering is the critical process that bridges this gap between raw data and the

machine learning model, with the following objectives:

First is Signal Amplification: to convert the faint anomaly signals hidden in the raw data into explicit numerical values, making them easier for the model to detect. For example, a derived variable like 'is_weekend' can better explain attack patterns than the timestamp itself.

Second is Dimensionality Reduction & Structuring: to transform high-dimensional or unstructured data into a fixed-size numerical vector that the model can accept as input.

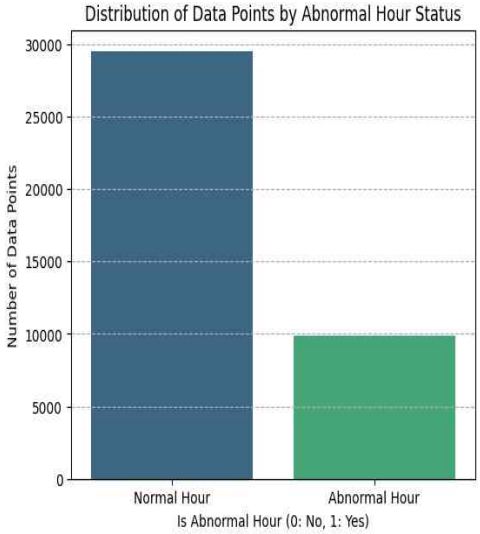
Third is to improve the model's learning efficiency and accuracy by injecting domain knowledge. This is done by converting the knowledge of security experts (e.g., "attackers are primarily active in the early morning," "specific API calls are closely related to privilege escalation") into features.

3.2.2 Time-based Features

An attacker's behavior often shows a distinct difference from the working hour patterns of normal users. In particular, automated attack tools or insider threats tend to occur outside of business hours. To capture this temporal context, features were created for the hour of the event (hour), whether it was a weekend (is_weekend), and whether it was outside of normal business hours (is_abnormal_hour).

The hour feature identifies the activity density at specific times of the day, while the is_weekend feature helps the model learn the difference in activity patterns between weekdays and weekends. Furthermore, the is_abnormal_hour feature can be a decisive clue in detecting automated scripts or insider threats that occur at unusual times. In addition to these, features like the time interval between events (time_since_last_event) were also created. These can be used to detect 'Brute-force' attacks that generate a large volume of events in a very short time, or anomalous behavior from 'dormant accounts' that have been inactive for an unusually long period. An example of the data, including these generated time-based features, is shown in Figure 2.

(Figure 2: Data Distribution by Normal and Abnormal Time Periods)



3.2.1 Security and Behavior-based Features

By directly incorporating security domain knowledge, features closely related to attack scenarios were generated.

The first feature, is_privilege_event, is a binary feature that indicates whether a key API call corresponding to the 'Privilege Escalation' tactic in the MITRE ATT&CK framework has occurred (e.g., CreateUser, AttachUserPolicy). This feature provides crucial information for the model to assess the risk level of a specific privilege escalation attempt when combined with other contextual features.

The second feature, error_rate, represents the ratio of server error (HTTP 5xx) responses to the total number of requests within a specific time window. Under normal circumstances, the error rate remains very low, but it can surge during reconnaissance scans aimed at finding system vulnerabilities or during credential stuffing attack attempts. Therefore, this feature is useful for detecting the initial stages of an attack. The specific code for generating these security and behavior-based features is shown in Figure 3.

```
# MITRE ATT&CK-Based Privilege Event Mapping & Error Rate Calculation

privilege_events = ['CreateUser', 'AttachUserPolicy', 'AssumeRole',
'CreateRole']
df['is_privilege_event'] =
df['raw_event_name'].isin(privilege_events).astype(int)
df['error_rate'] = df['http_status_5xx'] / (df['request_count'] + 1)
```

(Figure 3: Pseudocode for Mapping Privilege Events based on MITRE ATT&CK and Calculating Error Rate)

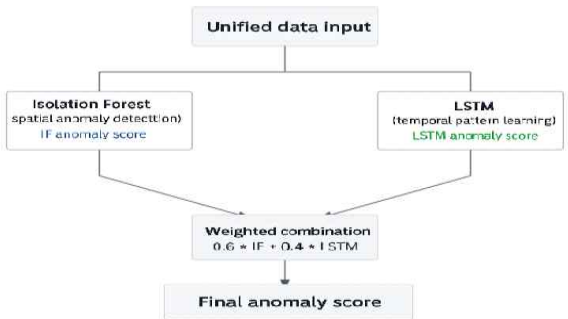
Through this feature engineering process, heterogeneous data collected from various sources were transformed into a standardized 'language' (a 15-feature vector) that all models can commonly understand and learn from. This is the core foundational work that enables the hybrid approach of this study.

3.2 Hybrid Ensemble Architecture

The hybrid ensemble architecture proposed in this study is illustrated in Figure 3. The input data is processed through two independent paths, and their results are then fused on a weight-based basis to produce the final anomaly score.

The first path, Isolation Forest, is responsible for detecting spatial anomalies in high-dimensional numerical metrics. The second path, LSTM, learns the multi-stage sequences of privilege escalation attacks and the temporal dependencies of time-series service metrics. Finally, the anomaly scores calculated by the two models are combined with optimal weights and fused in a way that maximizes detection performance.

(Figure 4: Hybrid Ensemble Model Architecture (Conceptual Diagram))



4. Experiments and Results

This chapter details the process and results of the experiments conducted to validate the performance of the proposed hybrid ensemble framework. The core objective of the experimental design is to conduct a multi-faceted evaluation to determine whether the proposed model shows a substantial performance improvement compared to its individual component models (Isolation Forest, LSTM), to understand the types of errors the model makes and their causes, and to assess whether the model's prediction results can be trusted and interpreted.

4.1 Experimental Environment

The experiments were conducted on the Google Colab Pro platform, using a Tesla T4 GPU and 25GB of RAM for hardware. The primary frameworks utilized were scikit-learn, TensorFlow, pandas, and SHAP. The data used for the analysis consisted of 769,454 rows and 15 standard features.

4.2 Evaluation Metrics

In problems with severe data imbalance, such as security anomaly detection, evaluating a model's performance with a single metric is highly risky. Therefore, this study comprehensively used the metrics shown in Table 2 to evaluate the model's performance from multiple perspectives.

The foundation for all metrics, the Confusion Matrix, categorizes the model's predictions into four types: True Positive (TP): An actual attack was correctly detected. False Positive (FP): A normal event was incorrectly identified as an attack. False Negative (FN): An actual attack was missed. True Negative (TN): A normal event was correctly identified.

(Table 2 Evaluation Metrics)

Metric	Description	Formula
Precision	The proportion of actual attacks among the instances the model predicted as an 'attack.' It indicates the reliability of the model's alerts.	TP / (TP + FP)
Recal	The proportion of actual attacks that the model successfully detected. It represents the model's ability not to miss attacks.	TP / (TP + FN)
F1-Score	The harmonic mean of Precision and Recall. It is used to evaluate the model's balanced performance, especially when there is a trade-off between the two metrics.	2 * (Precision * Recall) / (Precision + Recall)
ROC-AUC	A comprehensive metric for a model's discriminative ability, indicating how well it can distinguish between classes compared to random guessing.	-
PR-AUC	The area under the Precision-Recall curve. It provides a better representation of the model's practical performance, especially in cases of severe class imbalance.	-

4.3 Overall Model Performance Evaluation

To evaluate the overall discriminative performance of the models, the ROC curve and the Precision-Recall curve were analyzed.

(Figure 5: Comparison of ROC Curves on the Test Dataset)

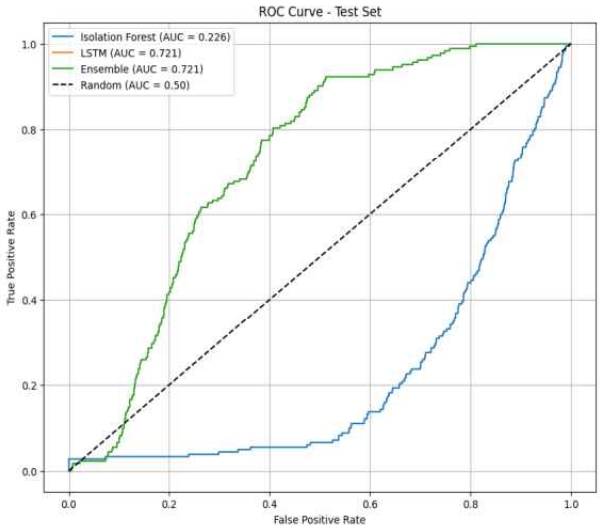


Figure 5 shows the ROC curves for each model. The Area Under the Curve (AUC) is a metric that indicates how well a model can distinguish between classes, with a value closer to 1 indicating better performance. The proposed Ensemble model and the LSTM model both recorded an AUC score of 0.721, demonstrating significantly superior performance compared to the Isolation Forest model (AUC=0.226). This signifies that the ensemble model discriminates anomalous behavior much more effectively than random guessing (AUC=0.5).

(Figure 6: Comparison of Precision-Recall Curves on the Test Dataset)

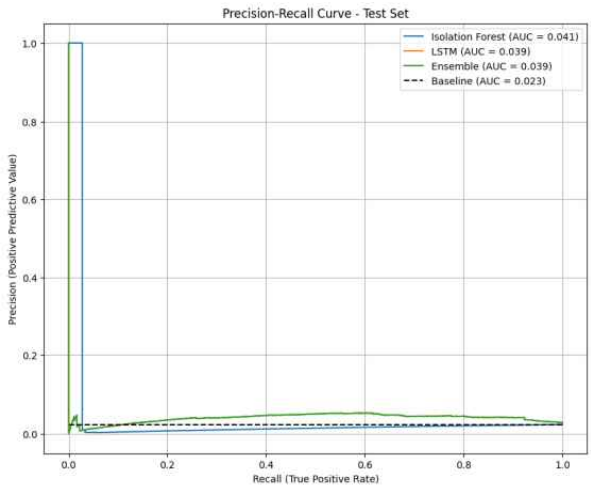
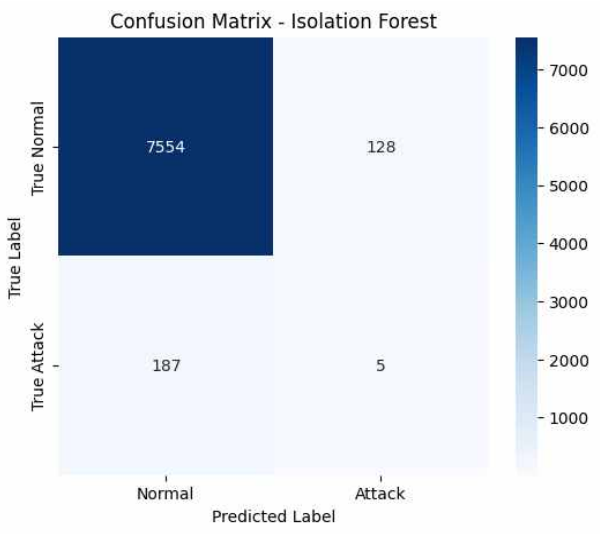


Figure 6 compares the Precision-Recall curves of the test dataset. The relatively low AUC-PR score of 0.039 is attributable to the extreme class imbalance problem within this study's dataset. In other words, it suggests that while the models discriminate well against the majority normal class, they still face challenges in detecting the minority attack class. Nevertheless, the fact that the ensemble model's curve (green) in Figure 6 remains at a higher level than the other models signifies that it achieves better Precision at all Recall levels. This means that for the same number of actual attacks detected, the ensemble model generates significantly fewer False Positives than the other models. This is a critically important characteristic in a real-world operational environment, as it reduces analyst fatigue and allows them to focus on important alerts.

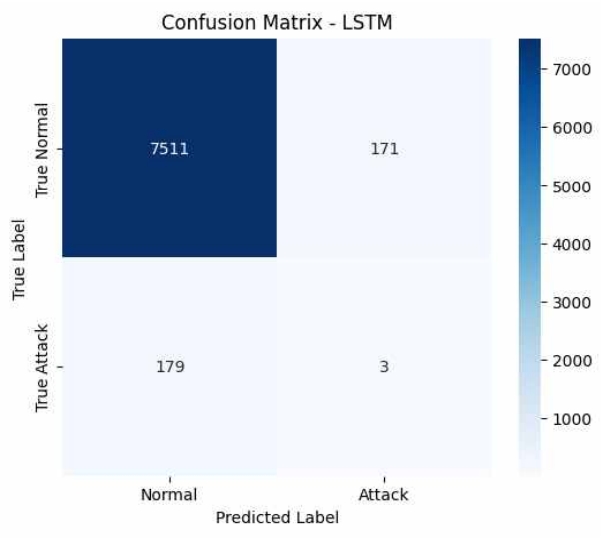
4.4 Detailed Analysis of False Positives and False Negatives by Model

To analyze the specific error types of each model, the confusion matrices in Figure 7, Figure 8, and Figure 9 were compared, and based on these, quantitative performance metrics were calculated as shown in Table 3.

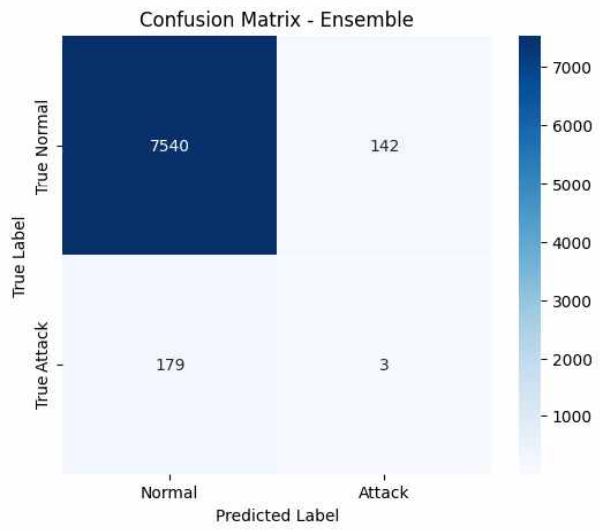
(Figure 7: Isolation Forest Confusion Matrix)



(Figure 8: LSTM Confusion Matrix)



(Figure 9: Ensemble Model Confusion Matrix)



As a result of the comparative analysis, the proposed ensemble model (Figure 9) had the highest number of True Negatives (7540), correctly identifying normal instances. The number of True Positives was 3, identical to the LSTM model.

Notably, the number of False Positives (FP)—instances where normal behavior was incorrectly classified as an attack—decreased to 142, compared to 171 for the LSTM model. This demonstrates that the ensemble method helps to improve the overall false positive rate by compensating for the weaknesses of the individual models.

(Table 3: Detailed Performance Metrics by Model (Based on the Confusion Matrix))

Model	TP	FP	FN	TN	Precision	Recal	F1-Score
Isolation Forest	5	128	187	7554	3.76%	2.60%	3.08%
LSTM	3	171	179	7511	1.72%	1.65%	1.68%
Ensemble	3	142	179	7540	2.07%	1.65%	1.84%

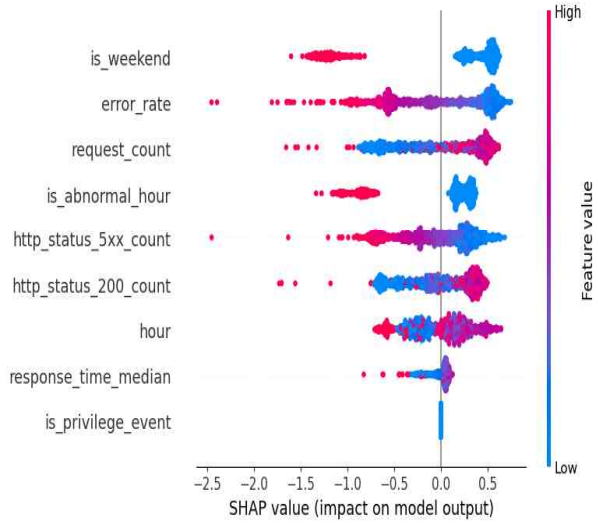
Each metric presented in Table 3 was calculated based on the values from the confusion matrix. Precision refers to the proportion of actual attacks among those the model predicted as attacks (TP / (TP + FP)), while Recall represents the proportion of attacks detected by the model out of all actual attacks that occurred (TP / (TP + FN)). The F1-Score is the harmonic mean of these two metrics (2 * (Precision * Recall) / (Precision + Recall)) and shows a comprehensive performance that considers the balance between them.

The analysis of Table 3 reveals several important facts. First, the Recall for all models remains at a very low level, between 1.65% and 2.60%. This is due to the previously mentioned data imbalance problem, which prevented the models from sufficiently learning from the few attack samples, and it is a clear limitation of this study. Second, the proposed ensemble model detected the same number of attacks as LSTM (TP=3) while reducing the number of false positives (FP) from 171 to 142, an approximate 17% decrease. This resulted in an improvement in Precision from 1.72% to 2.07%. Although the absolute figures are low, this signifies that the ensemble technique substantively contributes to reducing unnecessary alerts by correcting the errors of the individual models.

4.5 Feature Importance Analysis and Interpretation of Results (XAI)

To trust and understand the model's prediction results, the SHAP library was used to analyze which features played an important role in anomaly detection. Figure 10 visually shows the impact of each feature on the model's prediction output (anomaly score). The analysis revealed that is_weekend, error_rate, and request_count were the top three features with the greatest impact on the model's predictions.

(Figure 10: SHAP Summary Plot for the Ensemble Model)



Examining the relationship between each feature and the predictions in detail, for is_weekend, the red dots (representing high feature values) are mostly distributed in the positive SHAP value range. This means the model learned that events occurring on the weekend tend to increase the probability of an attack. This aligns with the common security sense of being suspicious of activities outside of normal business hours.

Furthermore, for error_rate, the red dots (high error rates) are clearly distributed in the positive SHAP value range, while the blue dots (low error rates) are in the negative range. This shows that the model successfully learned the intuitive fact that a higher error rate corresponds to a higher likelihood of an attack. On the other hand, for request_count, there is a mixed tendency where a high number of requests sometimes increases and sometimes decreases the probability of an anomaly. This suggests that the number of requests alone is not sufficient to determine an anomaly, and its combination with other features played a crucial

role.

These SHAP analysis results demonstrate that the model is making predictions based on logical grounds. By providing concrete clues for analysts to trace the cause of alerts and respond, it enhances the practical value of the model.

5. In-depth Discussion and Implications of the Study

In this chapter, the quantitative results presented in Chapter 4: Experiments and Results are comprehensively interpreted, and their academic and practical implications are discussed in depth. Furthermore, the reliability and validity of the proposed framework are verified by analyzing the model's internal operations through Explainable AI (XAI) techniques.

5.1 Comprehensive Interpretation of Experimental Results

The results of this experiment have a dual aspect, simultaneously demonstrating the potential and the clear limitations of the proposed hybrid ensemble model. First, the respectable ROC-AUC score of 0.721 achieved by the ensemble model in Figure 5 suggests that the model possesses a general discriminative ability to distinguish between normal and attack instances across the entire dataset. This implies that the multi-source data and multi-modal feature engineering provided a meaningful foundation for learning.

However, the Precision-Recall (PR) curve in Figure 6 and the detailed metrics based on the confusion matrix in Table 3 reveal a much more critical reality. The extremely low Recall (1.65%~2.60%) observed across all models clearly shows that the most central challenge of this study is the problem of Severe Data Imbalance. With only about 2.4% of the test dataset being attack samples, the models could not sufficiently learn the few attack patterns, which inevitably led to missing many attacks (a high number of FNs).

Despite this, the practical value of the ensemble model is evident in its improvement in Precision. While detecting the same number of attacks as LSTM (TP=3), the ensemble model reduced the number of false positives (FP) from 171 to 142, an approximate 17% decrease. This holds significant meaning in a real-world Security Operations Center (SOC) environment. Since an analyst's time and resources are limited, reducing unnecessary alerts (false positives) has the direct effect of mitigating alert fatigue and helping analysts focus on real threats. In short, the ensemble technique did not detect more attacks, but it contributed to generating more reliable alerts.

5.2 Ensuring the Explainability of Model Predictions

If performance metrics show 'what' a model did, Explainable AI (XAI) builds trust in the model by showing 'why' it did so. In this study, the SHAP library was used to analyze how each feature influenced the model's predictions.

The SHAP analysis in Figure 10 clearly demonstrates that the model makes predictions based on logical grounds that align with common security sense, rather than on random patterns. The features that had the greatest impact on the model's predictions were, in order, is_weekend, error_rate, and request_count. The strong tendency for events occurring on the weekend or those with a high error rate to increase the probability of an anomaly (positive SHAP values) signifies that the model learned to use behavior deviating from normal work patterns and abnormal system responses as key clues for an attack.

This analysis helps a security analyst, upon receiving an alert, to go beyond a simple "anomaly detected" message and start an investigation with a specific cause, such as "judged as an anomaly due to requests with an unusually high error rate occurring late on a weekend." This provides actionable insights that dramatically reduce the time needed to judge whether an alert is a false positive and to respond. In conclusion, the SHAP analysis reinforces the overall reliability of the proposed framework by proving that the previously confirmed

improvement in the ensemble model's precision is not a coincidence, but the result of learning based on rational grounds.

6. Conclusion and Future Work

6.1 Summary of Research and Contributions

This study integrated large-scale, heterogeneous data collected from a real-world multi-cloud environment and, based on this, designed and applied a hybrid ensemble framework for privilege anomaly detection. Through experiments, the proposed ensemble model achieved several concrete accomplishments and contributions compared to the individual models. First, it demonstrated practical utility by reducing false positives. While maintaining the same detection performance (Recall) as the LSTM model, the ensemble model reduced false positives by approximately 17%, thereby improving the precision of its alerts. This can directly contribute to increasing the efficiency of real-world security operations environments. Second, it secured explainability. By introducing SHAP analysis, the prediction results of the complex ensemble model were provided in an interpretable form. This demonstrated that the model operates on a logical basis and provided analysts with actionable insights, thereby securing the model's reliability. Finally, it empirically confirmed the problem of data imbalance. This study quantitatively revealed the extreme class imbalance problem present in real-world cloud security data and clearly showed the impact of this problem on anomaly detection models, especially on their Recall. This represents a significant academic contribution in that it has presented a core challenge that future related research must address.

6.2 Limitations and Future Research Directions

Despite the achievements of this study, the experimental results also revealed clear limitations. To overcome these, the following specific directions for future research are proposed. The biggest limitation of this study, as confirmed by the experimental results (Table 5, Figure 4), is the low Recall caused by severe data imbalance. This means the model could not sufficiently learn from the few attack samples, a point that must be improved for real-world application. To solve this Recall problem, future research could consider using Data Augmentation techniques, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), to synthesize realistic and diverse minority class (attack) data. Through this, research will be conducted to fundamentally improve Recall by enabling the model to sufficiently learn rare attack patterns. Furthermore, in the security domain, it must be considered that the risk cost of a missed detection (FN) is far greater than that of a false positive (FP). Therefore, research can be inroduced to strongly guide the model's learning direction to avoid missing the minority attack class by introducing Cost-Sensitive Learning, which involves designing a loss function that imposes a much higher cost (penalty) on false negatives. Finally, we propose advancing the ensemble model proposed in this study by further integrating a Graph Neural Network (GNN), which can model the complex relationships between users, resources, and actions. Through this, we aim to achieve a higher dimension of detection capability by building a true hybrid framework that encompasses the characteristics of individual events (Isolation Forest), the sequence of events (LSTM), and the relationships between events (GNN).

Reference

[1] Ahmad, Z., et al., "Anomaly detection in cloud computing: A systematic review," Future Generation Computer Systems, Vol.119, pp.156-178, 2021.

[2] Cloud Security Alliance, "Top Threats to Cloud Computing: The Pandemic Eleven," 2024.

[3] Du, M., et al., "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," Proceedings of the 2017 ACM SIGSAC Conference on Computer and

Communications Security, pp.1285-1298, 2017.

[4] Hamilton, W. L., Ying, R., and Leskovec, J., "Inductive representation learning on large graphs," Advances in Neural Information Processing Systems, Vol.30, 2017.

[5] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," Neural Computation, Vol.9, No.8, pp.1735-1780, 1997.

[6] Kipf, T. N. and Welling, M., "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[7] Liu, F. T., Ting, K. M., and Zhou, Z. H., "Isolation forest," 2008 Eighth IEEE International Conference on Data Mining, pp.413-422, 2008.

[8] Lundberg, S. M. and Lee, S. I., "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, Vol.30, 2017.

[9] Milajerdi, S. M., et al., "HOLMES: Real-time APT detection through correlation of suspicious information flows," 2019 IEEE Symposium on Security and Privacy (SP), pp.1137-1152, 2019.

[10] Ribeiro, M. T., Singh, S., and Guestrin, C., ""Why should I trust you?": Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135-1144, 2016.

[11] Schneider, J., et al., "Anomaly detection in large-scale cloud systems: An industry case and dataset," arXiv preprint arXiv:2406.07966, 2024.

[12] Wu, L., et al., "CloudRanger: Root cause identification for cloud native systems," 2021 IEEE 18th International Conference on Autonomic Computing (ICAC), pp.33-43, 2021.

Author Biography

Chi-Sung Kim

Graduated from Seoul School of Integrated Sciencs & Technologies(aSSIST),
Graduate School of AI, Department of AI Advanced Studies, Major in AI Big Data(Master of Engineering)
Research Interests : AI, Digital Transformation,
Generative AI, Data Analysis etc.
E-mail : nonochi12@gmail.com