Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(ARTICLE TYPE)

ML TO MEASURE EFFECTIVENESS OF DATABASE SECURITY

Dr. Upakar Bhatta

Assistant Professor;
Department of Information Technology Management (ITM)
Central Washington University

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2026, 3(2), 1–12

Article DOI: https://doi.org/10.70715/jitcai.2026.v3.i1.049

Abstract

Enforcing database security to protect data is crucial in today's digital world. As modern organizations increasingly rely on cloud infrastructure to run their business operations, database security becomes essential to protect the data of critical assets against malicious actors, while also maintaining compliance and regulatory requirements. This paper explores a machine learning approach to measuring the effectiveness of Database security. It utilizes logs from the AWS RDS service to construct a representative dataset, with key features including query latency ms, user role, access hour, query type, security alert, and top threat type. These features are used to train a model to predict the effectiveness of database security using AWS RDS log patterns. This enables organizations to prevent data breaches, protect sensitive information, and ensure ongoing compliance with relevant regulations.

Keywords: Artificial Intelligence, Machine Learning

1. INTRODUCTION

Databases, which provide a systematic way of managing data, are a cornerstone of today's digital landscape, where data are continuously produced and consumed. The database is a structured collection of data designed to be easily utilized in different systems within companies or organizations (Cui et al., 2015). As databases are widely used across modern organizations, including the government, private, academic, and e-commerce sectors, their security has become increasingly important. This growing concern is impacting industries, governments, and individuals (Ige, Kupa, Ilori, 2024). Data protection solutions are vital in the digital world today to protect the privacy and confidentiality of data, but current implementations and controls remain inadequate and vulnerable (Sun, Yu, Zhang, 2021). In recent years, the rise of encryption techniques capable of securing data both in transit and at rest has gained significant momentum. Protecting data requires a combination of data encryption, access control, monitoring, classification, and other security measures (Sun, Yu, Zhang, 2021). The traditional approach of using deterministic models and rules based system to protect data is no longer sufficient to address newly evolving threats (Butt et al., 2023). Therefore, this study aims to fill these gaps by proposing an enhanced machine learning approach integrated with best security practices to measure the effectiveness of Database security. The research is guided by three research questions:

[1] How can machine learning (ML) techniques leveraging log features enhance database security effectiveness?
[2] What are the benefits of integrating encryption into the ML pipeline?
[3] Which machine learning (ML) classification algorithms provide the robust predictive performance?

1.1. Research Purpose

The purpose of this research is to explore machine learning (ML) approaches to evaluate the effectiveness of Database security. This study aims to develop a predictive model to identify the pattern of threats, allowing organizations to protect their sensitive data while maintaining compliance with regulatory requirements in cloud environments.

1.2. Research Questions

The following are the key research questions that this study addresses:

RQ1: How can machine learning (ML) techniques, supported by log-based features analysis, be integrated to enhance database security effectiveness in cloud environments?

RQ2: What benefits does incorporating encryption within machine learning (ML) pipeline provide in mitigating modern cybersecurity challenges?

RQ3: Which machine learning (ML) classification techniques provide robust performance for predicting database security effectiveness?

2. PRE-REQUISITE KNOWLEDGE

2.1. Cloud computing

Cloud computing, on-demand delivery of virtual IT resources over the internet, serves as a foundational infrastructure for deploying scalable AI systems. It enables organizations to align technical architectures with their business goals through service models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), as well as deployment models including public, private, community, and hybrid clouds (SIIA, 2017).

2.2. Big data analytics

Big data analytics is vital component of modern businesses that drives innovation. It is used to analyze large datasets and visualize user query behavior and data access patterns. It is important that modern organizations leverage machine learning, statistics, predictive analysis, behavioral analysis to enhance big data analytics to combat evolving threats and attacks.

2.3. Machine learning

Machine learning, a field of Artificial Intelligence (AI) closely related to computational statistics, uses a set of rules to predict behavior. There are three main categories of machine learning: supervised learning, unsupervised learning, and reinforcement learning that can be used to solve machine learning tasks. Supervised learning can be beneficial for predicting risk indicators using labeled datasets. The most popular supervised learning techniques are known as classification and regression methods, which are used to classify or predict the future for a particular security problem (Sarker, Kayes, Watters, 2019). In unsupervised learning, the algorithm is trained on unlabeled data to predict the hidden pattern to identify indicator of attacks. In reinforcement learning, ML model is trained through trial and error learning, allowing ML agents to optimize its behavior overtime.

3. Database Security Background in the Age of Artificial Intelligence

Modern enterprises are increasingly adopting data-driven technologies to make effective decisions and enhance operational efficiency. As organization expands and scales their enterprise infrastructure, the complexity of database grows, bringing new challenges in maintaining confidentiality, integrity, and availability of information. Traditional database security mechanisms, which rely on reactive, rule based security controls, struggle to address these challenges. Innovations such as quantum-resistant encryption, homomorphic encryption, AI-driven threat detection, and zero-trust architectures are reshaping how organizations secure cloud-based databases (Kora, 2024). A survey conducted by the Cloud Security Alliance found that organizations implementing zero-trust architectures for their cloud databases reported a 60% reduction in successful breach attempts (Chen & Patel, 2023). Machine learning is gaining momentum as a transformative tool to enhance database security. ML offers a framework to address these challenges by analyzing the logs related to database access patterns, breach attempts, user behavioral trend to measure the effectiveness of database security controls. By applying a data-centric threat model, ML can identify anomalous database interactions, enabling organization enforce fine-grained access policies and implement security control based on evolving threats. Data-centric threat model focuses on protecting data within the systems by identifying data as sets, mapping data flows and trust boundaries, evaluating threats, and applying security controls (Souppaya Scarfone, 2015). This paper references data-centric threat model framework to identify sensitive data, map data flows, collect data logs, and perform feature engineering before selecting and training machine learning models to measure effectiveness of Database security.

4. RELATED WORK

Several recent studies have explored the database security using machine learning algorithms. The detection of anomalies and attack response system were designed to monitor data traffic (Sallam et al., 2017). Deep learning applied for database anomaly detection using log-sequence modeling (Zhang et al. 2022). LLMs have been trained on vast amounts of cross-platform data, which gives them a strong degree of generality (Zhang, Jia et al., 2024). LLMs have shown strong capabilities in interpreting unstructured data, making them highly effective for log analysis and system diagnosis (Vitui & Chen, 2025). Hybrid Supervised Machine Learning Framework was purposed for predicting Insider Threat Detection (Eguavoen & Nwelih, 2025). ML based secure database system, including access control and intrusion detection were explored (Deepa Dhili pan, 2024). ML techniques were leveraged for cloud security, outlining both the strengths and weaknesses of various ML algorithms in cloud environments (Babaei et al., 2023). Precision health systems were empha sized and discussed data security and privacy strategies (Thapa Camtepe, 2021). Their study focused on maintaining compliance and ethical requirements to foster innovation in the healthcare field. A systematic evaluation of emerging cybersecurity risks and mitigation strategies were purposed to address the growing threat landscape (Aslan et al., 2023). They reviewed recent technological trends, including machine learning approaches for detecting and mitigating cyber threats, outlined the potential of ML in identifying malware and intrusions. This work is relevant to RQ1, as it explores the potential machine learning techniques to protect the system against emerging cybersecurity risks. Explainable artificial intelligence (XAI) investi gated in the context of cybersecurity, addressing the challenges posed by the black-box nature of ML and DL algorithms, which reduce user trust and hinder understanding of how these models detect or respond to threats (Zhang et al., 2022). While previous studies have emphasized database security using machine learn ing, they haven't explored experiment methodologies to address the risks associated with database systems. This paper fills that gap by focusing on RQ1: How can machine learning techniques be integrated to en hance database security? It utilize a supervised learning techniques for predictive analysis and incorporates encryption into the ML pipeline to address modern cybersecurity challenges, support effecting deployment of encryption technique (RQ2), and providing comprehensive performance overview of different classification technique(RQ3).

5. RESEARCH METHOD

This paper leverages machine learning approach to evaluate effectiveness of database security. The experimental methodology employed in this paper incorporates five key components: data construction, exploratory data analysis, data preprocessing, feature selection, and model training and evaluation. This study utilizes AWS RDS logs, which captures real operational indicators of database activity, to construct a synthetic sample dataset that includes features such as query latency ms, user role, access hour, query type, security alert, top threat type. These features are selected as indicators for potential anomalous behavior and insider thereat. Features including security alerts, threat types, and user roles directly relate to potential malicious activity, whereas features like query latency and access hour capture deviations from normal operational patterns and serve as early indicators of anomalous behavior. Exploratory data analysis (EDA) was employed to identify patterns and correlations among features and to ensure that the selected features provide non-redundant signals that contribute meaningfully to model performance. EDA also helped identify distribution patterns, detects skewness, and examines inter-feature relationships. Visualizations such as heatmaps, boxplots, and distribution plots were used to assess the feature relationships. The initial preprocessing steps included cleaning to remove missing values and filter incomplete log entries, encoding to convert categorical variables into numerical value using one-hot encoding, and normalization to scale numerical data to ensure uniformity across features. These steps help prevent bias, improve model convergence, and ensure that the machine learning algorithms are properly trained on the constructed dataset. The preprocessing steps allow machine learning algorithm to perform accurately. The full dataset is then partitioned into training (80) and testing (20) segments, with Synthetic Minority Oversampling Technique (SMOTE) applied to the training set to address class imbalances. Multiple advanced machine learning models were employed, trained, and tuned using grid search and cross-validation to optimize performance. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. A modular ML pipeline was designed with the following components:

Feature Extraction: Using PCA to reduce dimensionality Classification Methods

Model Architecture: RF, SVM, KNN, Gradient Boosting, Logistic Regression.

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.049

Model Evaluation: Assessing model effectiveness using accuracy and other metrics such as precision, recall, and F1-score to evaluate their performance.

Hyperparameter tuning was performed using GridSearchCV to determine the optimal configuration for each model. The tuning process focused on correctly identifying abnormal data activities by leveraging performance metrics such as precision, recall, and F1-score.

The parameter grids used for machine learning models in this experimental research are as follows:

Random Forest: n estimators = [50, 100], max depth = [10, 20]

Support Vector Machine (SVM): C = [0.1, 1], kernel = [’linear’, ’rbf’]

K-Nearest Neighbors (KNN): n neighbors = [3, 5]

Gradient Boosting: n estimators = [50, 100], learning rate = [0.01, 0.1], max depth = [3, 5] Logistic Regression: C = [0.1, 1], penalty = [’l1’, ’l2’], solver = [’liblinear’]

Each model was trained on the labeled dataset to evaluate database security risks. The best-performing models were selected based on their ability to handle mixed dataset feature, accuracy, robustness, and effectiveness in handling imbalanced classification tasks.

This experimental study evaluates both the validity and reliability of the machine learning models used to predict database security. Construct validity was addressed by selecting the features that represent database abnormal activity, ensuring that the machine learning model accurately measures the security-related behaviors. To support internal validity, the study performed preprocessing steps and used GridSearchCV to tune hyperparameters and reduce model bias.

This experiment research also examines the reliability and robustness of the models by evaluating their performance across multiple validation folds. Cross-validation procedures were used to ensure that machine learning model demonstrated genuine generalization rather than overfitting to a single training–testing split, making the approach suitable for real-world database security applications.

6. DATA ANALYSIS

Proposed experimental study utilizes Using Azure agentic AI servcie logs to create a sample dataset and leverage exploratory data analysis (EDA) techniques to understand the dataset and detect any anomalies prior to training a machine learning model.

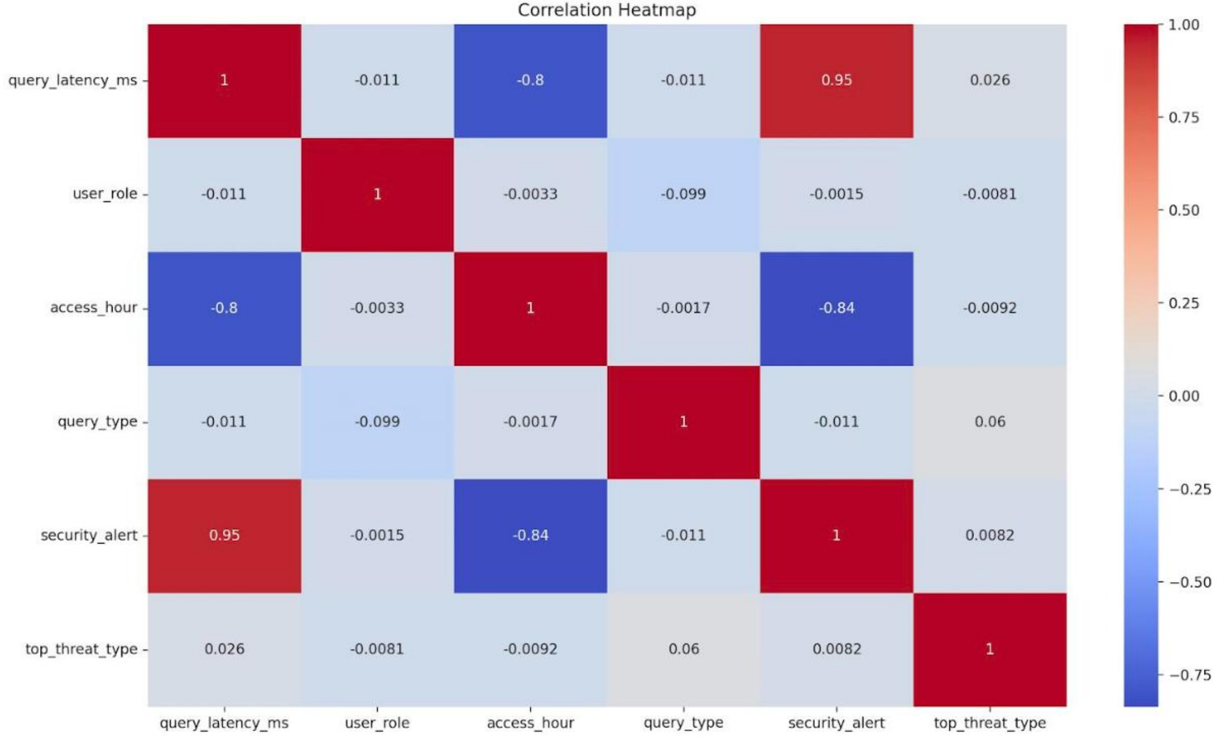


Fig. 1. Correlation heatmap.

Figure 1 shows a correlation heatmap that visualizes correlation coefficient between 6 variables related to system performance, user behavior, and database risk indicators. Features such as query latency ms and security alert exhibit strong positive correlations, indicating that increased latency is associated with the presence of security alert. In contrast, query latency ms and access hour, as well as access hour and security alert, show strong negative correlations. Variables that demonstrate unique relationships with others include query latency ms, access hour, and security alert. These features which provides non redundant signal and are well-suited to enhance machine learning model performance for database security risk prediction, addressing RQ1 by demonstrating how log-based feature engineering improves prediction accuracy.

Figure 2 shows box plots that highlight the statistical distribution of five features: query latency ms, user role, access hour, query type, and top threat type. Query latency (ms), which indicates occasional high latency queries, remains within acceptable thresholds for near-real-time security monitoring, thus supporting RQ2. User role and query type suggest stable categorical behavior, while access hours reflects varied access times. Top threat type indicates irregular occurrences of threat categories. This visualization is essential for preprocessing and model selection as it identifies feature skewness, outlier presence, and distribution characteristics.

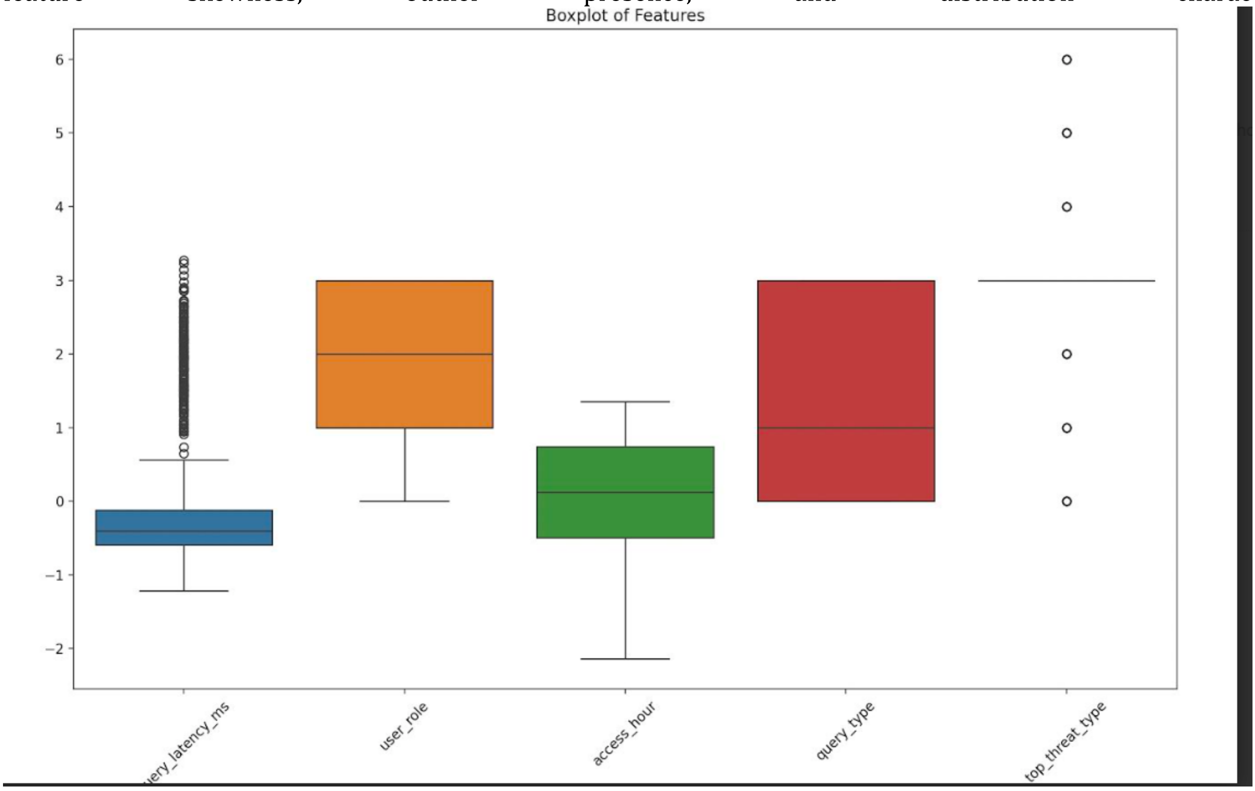


Fig. 2. Boxplot of features.

Figure 3 shows the distribution patterns of 6 features related to query behavior, access patterns, and security indicators. These visualizations are key to ensuring appropriate fea ture engineering techniques are applied before selecting and training the machine learning model. Features such as query latency ms and access hour exhibit skewed distribution, indicating the need of normalization. Categorical features such as user role, query type, and top threat type show discrete value ranges, with top threat type dominated by a single category, which may led to model bias. Feature like security alert require resampling techniques to address class imbalance.

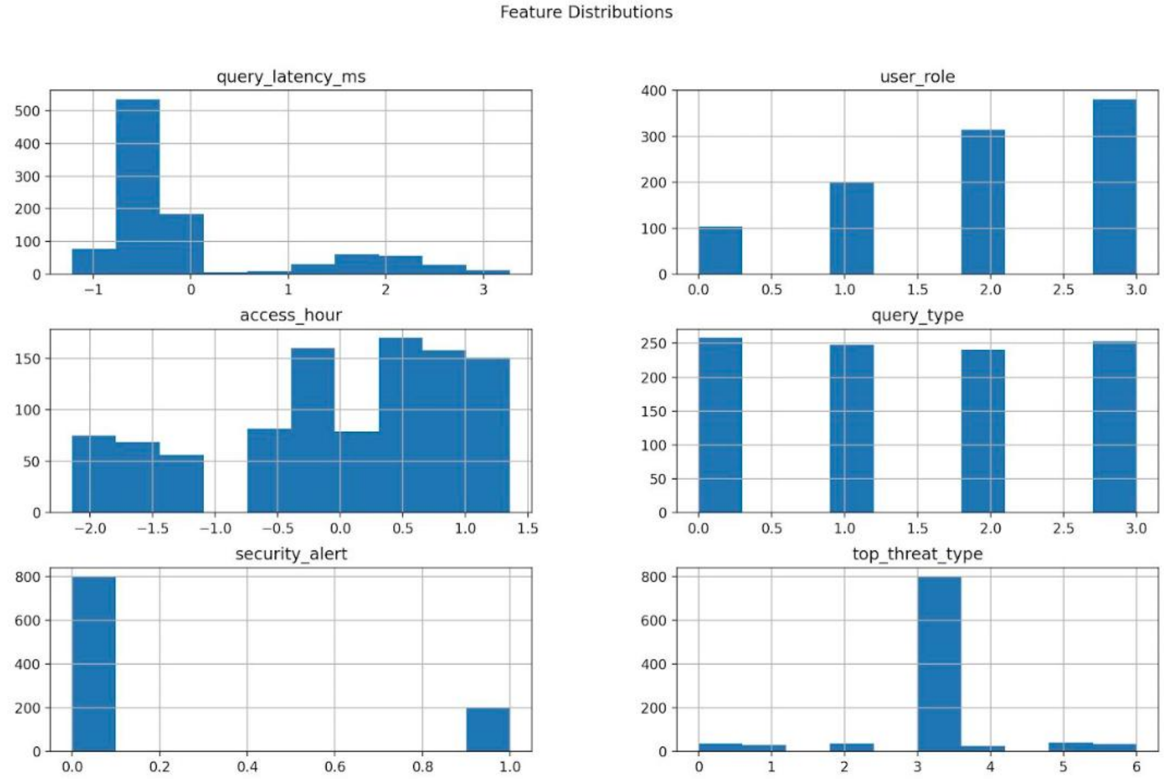


Fig. 3. Distribution of system and user interaction features.

Figure 4 demonstrate the SMOTE (Synthetic Minority Over-Sampling Technique) to mitigate class imbalance enabling machine learning models to avoid bias toward the majority class and achieve more balanced and accurate results. This supports RQ3 by identifying which machine learning classifiers are most robust for evaluating database security effectiveness in cloud environments.

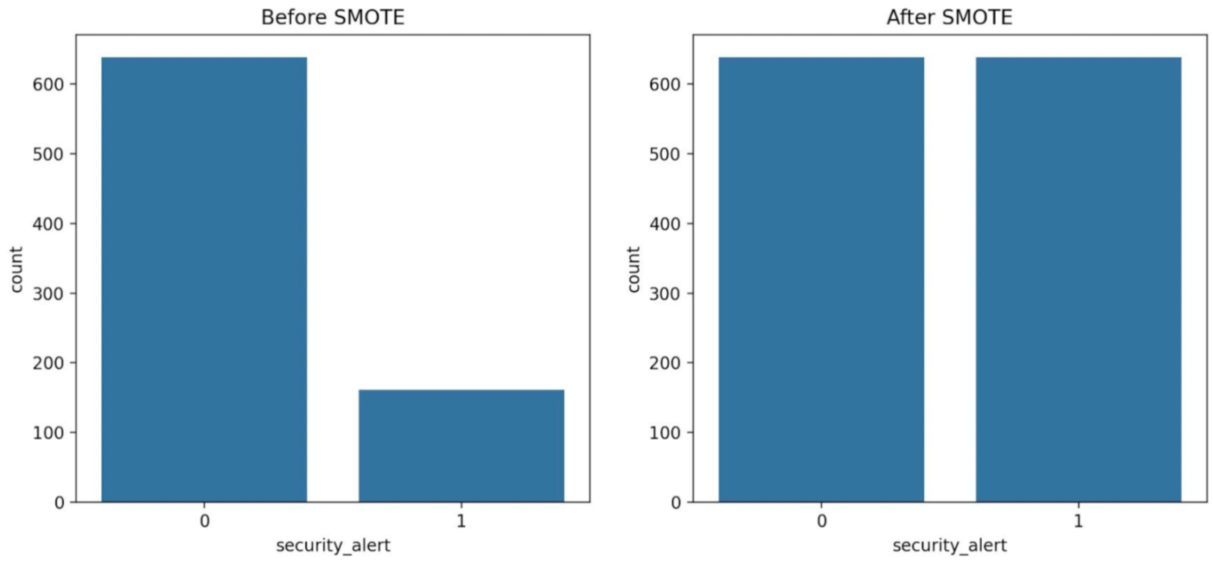
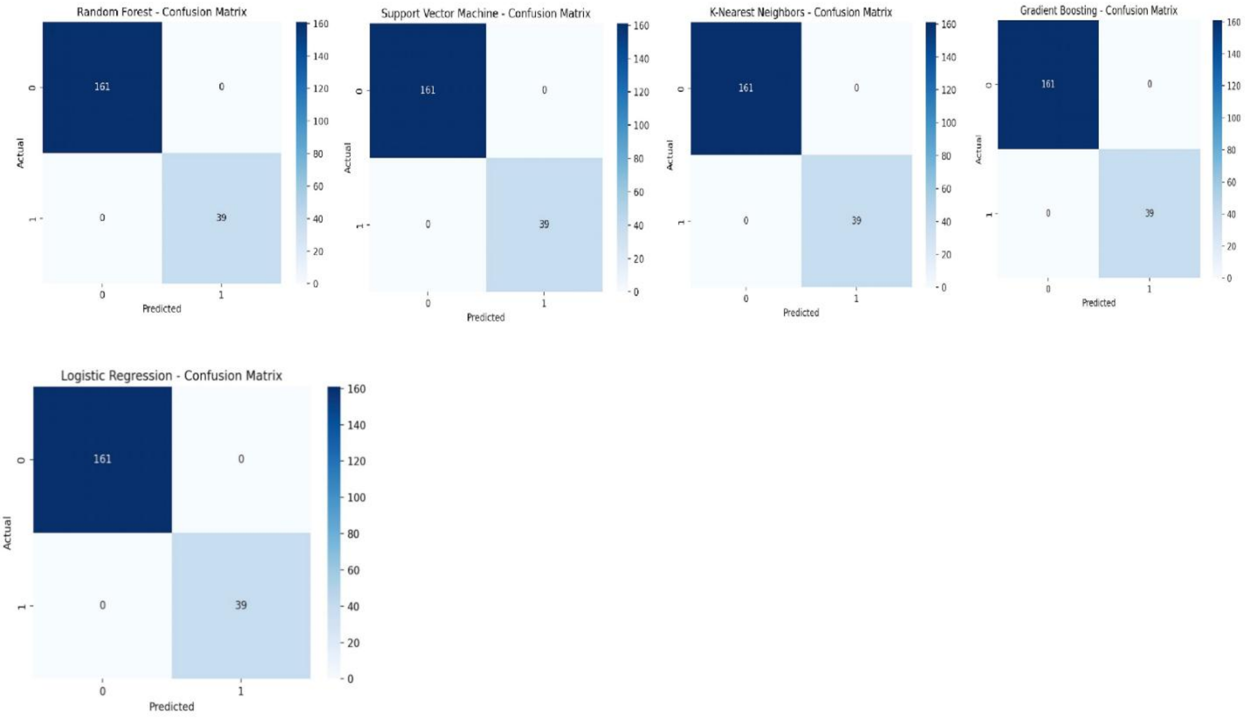


Fig. 4. Class distribution before and after SMOTE



Random Forest - Confusion Matrix

Support Vector Machine - Confusion Matrix

K-Nearest Neighbors - Confusion Matrix

Gradient Boosting - Confusion Matrix

Logistic Regression - Confusion Matrix

Fig. 5. Consolidated confusion matrix.

Figure 5 shows consolidated confusion matrix for five models. Based on the sample outlined in the confusion matrix, machine learning models produced the following results:

Table 1: Model metrics

Random forest	SVM	KNN	Gradient boosting	Logistic regression
TP= 39 **TN= 161** **FP= 0** **FN= 0** **A= 1** **P= 1** **R= 1**	**TP= 39** **TN= 161** **FP= 0** **FN= 0** **A= 1**	**TP= 39** **TN= 161** **FP= 0** **FN= 0** **A= 1** **P= 1**	**TP= 39** **TN= 161** **FP= 0** **FN= 0** **A= 1** **P= 1** **R= 1**	**TP= 39** **TN= 161** **FP= 0** **FN= 0** **A= 1** **P= 1** **R= 1**

F1-Score= 1	P= 1
	R= 1
	F1-Score= 1
		R= 1
		F1-Score= 1
			F1-Score= 1	F1-Score= 1

$$\boldsymbol{Accuracy\ (A)} = \frac{\mathbf{tp+tn}}{\mathbf{tp+tn+fp+fn}} \qquad \mathbf{(1)}$$

$$\boldsymbol{Precision\ (P)} = \frac{\mathbf{tp}}{\mathbf{tp+fp}} \qquad \mathbf{(2)}$$

$$\boldsymbol{Recall\ (R)} = \frac{\mathbf{tp}}{\mathbf{tp+fn}} \qquad \mathbf{(3)}$$

$$\boldsymbol{F1 - Score} = \frac{\mathbf{2*(p*r)}}{\mathbf{p+r}} \qquad \mathbf{(4)}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

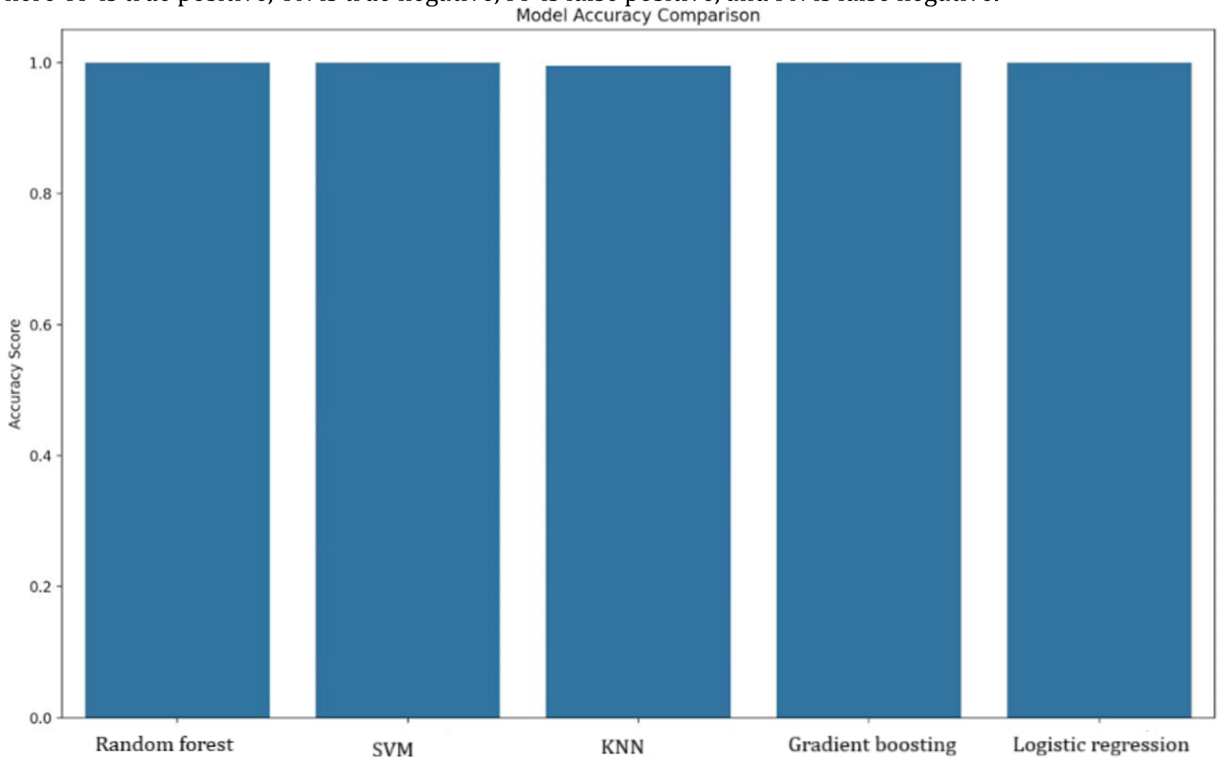


Fig. 6. Model accuracy comparison.

This experimental research considered five models: Random Forest, Support Vector Machine, K-Nearest Neighbors, Gradient Boosting, and Logistic Regression, to perform classification tasks aimed at evaluating the effectiveness of database security. Based on the confusion matrix metrics, the machine learning model perfectly predicted all the samples. The model accuracy 100% because AWS RDS service log were generated using an algorithm to mimic the pattern of real logs for machine learning testing. The synthetic dataset was created from AWS RDS service log with features such as query latency ms, user role, access hour, query type, security alert, and top threat type. A single train-test split and SMOTE (Synthetic Minority Over-Sampling Technique) were used to address class imbalance issues. Each model was evaluated using key parameters such as Recall, Precision and F-Measure, error rate and overall model accuracy. The results showed that each model accurately distinguished between security alerts and non-alerts in the test set. Despite the uniform performance, Random Forest was chosen for this research paper due to its high accuracy, robustness, and effectiveness in handling imbalanced classification tasks. It handles mixed dataset features used in this research without extensive preprocessing, and its ensemble learning approach reduced the risk of overfitting, making it ideal for real-world database security applications.

7. CONCLUSION

This paper demonstrates the integration of machine learning to evaluate and enhance database security by analyzing operational logs from AWS RDS. These logs are used to construct a dataset, containing 5,000 labeled entries. The experimental methodology leverages an ML pipeline to enable predictive measurement by extracting key dataset features such as query latency, user roles, access times, query types, security alerts, and threat classifications. This dataset is used to train supervised learning models capable of identifying potential data risks. With a predictive accuracy of 100 percent, the study underscores the effectiveness of ML-driven database security techniques in proactively identify vulnerabilities and mitigating risks. This experiment study contributes to the advancement of development of database security monitoring by demonstration how ML pipeline can empower organizations to safeguard sensitive digital assets.

8. LIMITATION AND FUTURE WORK

While the proposed ml pipeline shows promise, the research can be enhanced by applying larger dataset to improve robustness. The current study limited on AWS RDS operational logs; future validation across platform such as OpenAI API and Google Gemini would help ensure cross-platform applicability. Future work should focus on incorporating additional data sources, such as behavioral telemetry and network traffic logs, and applying data anonymization techniques to those datasets, which would enable the capture emergent risks and improve threat detection granularity in dynamic database environments. Furthermore, this research study can be enhanced by employing recurrent neural networks (RNNs) that could help identify evolving patterns of risk over time. Finally, integrating explainable AI (XAI) tools such as SHAP to transparency in regulatory audits and deploying the ML pipeline in live enterprise environments under operational constraints will further contribute to the advancement of intelligent database security monitoring.

9. REFERENCES

[1] A. Babaei, P. M. Kebria, M. Moradi Dalvand, and S. Nahavandi. A review of machine learning-based security in cloud computing. arXiv preprint arXiv:2309.04911, 2023. Available at https://arxiv.org/pdf/2309.04911.

[2] A. B. Ige, E. Kupa, and O. Ilori. Best practices in cybersecurity for green building management systems: Protecting sustainable infrastructure from cyber threats. International Journal of Science and Research Archive, 12(1):2960–2977, 2024.

[3] A. Fatani, A. Dahou, M. A. Al-Qaness, S. Lu, and M. A. Elaziz. Advanced feature extraction and selection approach using deep learning and aquila optimizer for IoT intrusion detection system. Sensors, 22(1):140, 2021.

[4] A. U. R. Butt, T. Mahmood, T. Saba, S. O. Bahaj, F. S. Alamri, M. W. Iqbal, and A. R. Khan. An optimized role-based access control using trust mechanism in eHealth cloud environment. IEEE Access, 2023.

[5] A. Sallam, E. Bertino, S. R. Hussain, D. Landers, R. M. Lefler, and D. Steiner. DBSAFE—An anomaly detection system to protect databases from exfiltration attempts. IEEE Systems Journal, 11(2):483–493, June 2017. doi:10.1109/JSYST.2015.2487221.

[6] Chen, S., & Patel, R. (2023). AI in cybersecurity: A double-edged sword*.* IEEE Security & Privacy, 21*(3),* 45–52. https://fptsoftware.com/resource-center/blogs/ai-in-cybersecurity-a-double-edged-sword

[7] C. Thapa and S. Camtepe. Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in Biology and Medicine, 129:104130, 2021.

[8] I. H. Sarker, A. Kayes, and P. Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1):1–28, 2019.

[9] Kora, P. R. (2024). Innovations in cloud database security: Addressing emerging threats. International Journal for Multidisciplinary Research (IJFMR), 6(6), 1–10. https://www.ijfmr.com/papers/2024/6/30209.pdf

[10] M. Deepa and J. Dhilipan. Security in database management system using machine learning. International Journal of Electronic Security and Digital Forensics (IJESDF), Inderscience, 2024. doi:10.1504/IJESDF.2024.10057609. Available at https://www.inderscienceonline.com/doi/pdf/10.1504/IJESDF.2024.136024.

[11] M. Souppaya and K. Scarfone. Guide to data-centric system threat modeling (Draft NIST Special Publication 800-154). National Institute of Standards and Technology, 2015. Available at https://csrc.nist.gov/files/pubs/sp/800/154/ipd/docs/sp800_154_draft.pdf.

[12] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin. A comprehensive review of cybersecurity vulnerabilities, threats, attacks, and solutions. Electronics, 12(6):1333, 2023.

[13] Software and Information Industry Association (SIIA). Ethical principles for artificial intelligence and data analytics. Pages 1–25, 2017.

[14] V. H. Le and H. Zhang. Log-based anomaly detection with deep learning: How far are we? *arXiv preprint*, February 2022. doi:10.48550/arXiv.2202.04301

[15] Vitui, A., & Chen, T.-H. (2025). Empowering AIOps: Leveraging large language models for IT operations management. Available at http://arxiv.org/abs/2501.12461.

[16] V. O. Eguavoen and E. Nwelih. HSML-ITD: Hybrid supervised machine learning framework for insider threat detection. *Quantum Journal of Engineering, Science and Technology*, 6(1):100–110, 2025. doi:10.55197/qjoest.v6i1.202.

[17] X. Sun, F. R. Yu, and P. Zhang. A survey on cyber-security of connected and autonomous vehicles (CAVs). IEEE Transactions on Intelligent Transportation Systems, 23(7):6240–6259, 2021.

[18] Z. Cui, J. Zeng, C. Wu, and S. Zhang. Design and implementation of a new database security model based on hopping mechanism. In Proceedings of the IEEE 9th International Conference on Anti-Counterfeiting, Security, and Identification (ASID), pages 1–5, September 2015. IEEE. doi:10.1109/ICASID.2015.7405649.

[19] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10:93104–93139, 2022.

[20] Zhang, L., Jia, T., Jia, M., Wu, Y., Liu, A., Yang, Y., Wu, Z., Hu, X., Yu, P. S., & Li, Y. (2024). A survey of AIOps for failure management in the Era of large language models. Available at http://arxiv.org/abs/2406.11213.