iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: https://doi.org/10.22214/ijraset.2026.84812

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IX Sep 2026- Available at www.ijraset.com

An LDA-Enhanced Machine Learning Approach for Efficient SQL Injection Detection from Web Traffic Logs

Dr. Amit Hariyani
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, CHARUSAT, Changa, Anand

Abstract: ***SQL injection (SQLI) remains a critical web application security threat, while conventional signature-based detection methods face limitations in identifying evolving attack patterns and handling high-dimensional web request data. This study proposes an improved machine learning framework for SQLI detection based on TF-IDF N-grams and Latent Dirichlet Allocation (LDA). HTTP requests from the CSIC-2010 and ECML/PKDD-2007 datasets are processed to extract SQLI-related traffic using SQL keyword-based filtering. TF-IDF trigram features are generated to represent request characteristics, followed by LDA-based dimensionality reduction to obtain a compact latent semantic representation. Support Vector Machine (SVM) and Random Forest (RF) classifiers are employed for classifying malicious and genuine requests. Experimental results demonstrate the effectiveness of the proposed LDA-based approach. On the CSIC-2010 dataset, the SVM and RF models achieve 99.90% and 98.13% accuracy, respectively. On the ECML/PKDD-2007 dataset, the improved framework achieves 98.70% accuracy using SVM and 99.85% using RF, with approximately 99% true-positive rate and mean ROC performance. The findings indicate that LDA-based feature reduction can improve the effectiveness and efficiency of machine-learning-based SQLI detection.***
Keywords: SQL Injection (SQLI), Machine Learning, Web Application Security, TF-IDF, N-gram; Latent Dirichlet Allocation (LDA), Support Vector Machine (SVM), Random Forest; HTTP Logs, Dimensionality Reduction

I. INTRODUCTION

Web applications have become an essential component of modern information systems, supporting information exchange, communication, financial transactions, administrative services, and data processing across public and private organizations. The increasing dependence on database-driven web applications has consequently expanded the potential attack surface for cyber threats. Among the vulnerabilities affecting web applications, SQL Injection (SQLI) remains particularly serious because malicious input can manipulate database queries, potentially compromising an application's confidentiality, integrity, authentication, and authorization mechanisms [1, 2]. The ongoing presence of SQLI vulnerabilities underscores the need for effective detection mechanisms that can identify malicious requests before they lead to significant security consequences. The source study identifies SQLI as a major web application security concern and emphasizes the continuing challenges associated with protecting database-driven applications [3]. Traditional SQLI detection approaches commonly rely on predefined signatures, patterns, lexical rules, regular expressions, or application-specific analysis. Although these approaches can identify known attack forms, their effectiveness depends substantially on the availability and maintenance of attack signatures. Consequently, previously unseen or modified SQLI patterns may not be detected effectively. Dynamic approaches, including behavior- and taint-based techniques, can provide additional detection capabilities but may introduce application dependencies, false alarms, and difficulties associated with under- and over-tainting and threshold selection. These limitations necessitate detection mechanisms that can learn discriminative characteristics from observed HTTP requests rather than relying exclusively on manually defined attack patterns. The reviewed research specifically identifies new-signature detection, false alarms, performance impact, and application-specific behavior as important limitations of conventional approaches.

Machine learning (ML) offers a promising alternative, as classification models can learn patterns from labeled data and subsequently identify previously unseen patterns in test data [4, 5]. However, applying ML directly to web traffic introduces another challenge: HTTP logs are generally available in raw form and may contain different categories of web requests and attack traffic. Therefore, effective SQLI detection requires appropriate preprocessing and extraction of SQLI-related requests before classification. Furthermore, textual HTTP requests can generate a very large and sparse feature space when represented using conventional text-vectorization techniques.

Such high-dimensional representations can increase computational complexity and adversely affect the efficiency of subsequent classification. The source research consequently emphasizes three closely related requirements: filtering SQLI traffic from available web logs, constructing balanced datasets, and extracting informative features through dimensionality reduction.

Feature representation and dimensionality reduction therefore constitute important components of an ML-based SQLI detection system [6]. In the initial combined approach, Term Frequency–Inverse Document Frequency (TF-IDF) with N-gram representation was employed to transform filtered HTTP requests into numerical feature vectors. Because the resulting feature space can be large, Singular Value Decomposition (SVD) was subsequently applied to obtain a reduced representation before supervised classification [7]. Experiments using CSIC-2010 and ECML/PKDD-2007 SQLI logs demonstrated that the combined TF-IDF-N-gram and SVD approach could reduce the feature space and support effective classification using Support Vector Machine (SVM) and Random Forest (RF) algorithms. However, the observed performance difference, particularly on the ECML dataset, indicated that further improvements in feature representation and reduction were necessary [8].

A key limitation of SVD in this context is that it primarily performs statistical dimensionality reduction and does not explicitly model the latent semantic relationships among terms in the textual request corpus. To address this limitation, this study introduces an improved combined framework based on TF-IDF N-grams and Latent Dirichlet Allocation (LDA). LDA is employed as a topic-modeling-based dimensionality-reduction technique to identify latent topics associated with word co-occurrence in SQLI-related HTTP requests [9, 10]. The resulting latent topics provide a compact feature representation for supervised classification. The improved framework is evaluated using SVM and RF, enabling a direct assessment of whether semantic-oriented dimensionality reduction can improve both detection accuracy and computational performance.

The experimental investigation focuses on the CSIC-2010 and ECML/PKDD-2007 datasets, with SQLI requests extracted from web-log data and organized into balanced datasets for classification. The study evaluates the proposed approach using classification performance and execution-time measures. In particular, the improved approach replaces SVD with LDA and investigates whether the resulting latent representation can provide more informative features for SQLI detection. The experimental findings reported in the study demonstrate that the LDA-based approach substantially improves detection performance on the ECML dataset while reducing training and testing requirements compared with the earlier SVD-based approach.

The principal contributions of this study are summarized as follows:

1) An improved ML-based SQLI detection framework is developed by integrating TF-IDF trigram feature representation with LDA-based dimensionality reduction.
2) SQLI-specific HTTP traffic is extracted and processed from publicly available web-log datasets, addressing the challenge of using raw mixed web-attack logs for SQLI classification.
3) LDA is investigated as an alternative to SVD for reducing the high-dimensional TF-IDF feature space while retaining latent information from the request corpus.
4) SVM and RF classifiers are comparatively evaluated to determine their effectiveness for SQLI classification using the reduced feature representation.
5) The proposed approach is evaluated in terms of detection accuracy and computational performance, demonstrating the practical value of semantic-oriented feature reduction for ML-based SQLI detection.

Overall, the study addresses the challenges of feature representation and dimensionality reduction in ML-based SQLI detection and demonstrates that an LDA-based representation can provide a more informative and computationally efficient basis for classifying malicious HTTP requests. The research therefore contributes an experimentally evaluated approach for improving the reliability and efficiency of SQLI detection from web-log data. The underlying research identifies dimensionality reduction as a critical factor in improving ML classification and reports that LDA is more effective than SVD for the SQLI detection task investigated [11, 12].

II. RELATED WORK AND RESEARCH GAP

SQL injection detection has been investigated through a range of static, dynamic, signature-based, anomaly-based, and machine-learning approaches. Traditional detection mechanisms generally depend on predefined attack signatures, lexical patterns, regular expressions, or manually constructed rules [13]. Such mechanisms can be effective against recognized attack patterns, but their reliance on predefined knowledge limits their ability to detect modified or previously unseen SQLI requests. The literature reviewed in this study similarly identifies the inability to detect new attack signatures, maintenance complexity, false alarms, performance overhead, and application-specific behavior as persistent challenges in conventional SQLI detection [14].

Dynamic approaches seek to overcome some limitations of static pattern matching by analyzing application behavior at runtime. However, these approaches may require access to the application's source code and strongly depend on application-specific usage models. Taint-based techniques introduce additional challenges, including under-tainting, over-tainting, threshold selection, and false-positive generation. Consequently, although dynamic analysis can provide valuable runtime information, its practical applicability may be constrained in heterogeneous web environments.

The emergence of machine learning has provided an alternative direction for SQLI detection. Instead of depending entirely on manually specified signatures, ML classifiers can learn distinguishing characteristics from labeled malicious and legitimate requests and subsequently classify unseen instances. Recent studies demonstrate the continued interest in learning-based SQLI detection, including approaches based on neural networks, NLP-derived features, and network or web logs. For example, neural-network-based detection has been investigated using authentic URL access logs and engineered statistical features, while more recent work has explored NLP-based TF-IDF representations combined with classifiers such as SVM, logistic regression, Naïve Bayes, and convolutional neural networks [15].

Despite these developments, feature representation remains a critical challenge. HTTP requests and SQLI payloads are both textual and structurally diverse, and directly converting these requests into machine-readable vectors can yield a large, sparse feature space. Recent research on network-log-based SQLI detection also confirms the importance of selecting appropriate input features and evaluating multiple ML classifiers; one such study reported Random Forest as the strongest among the evaluated tree-based methods, achieving 97.58% accuracy and a 0.976 ROC-AUC. This indicates that classifier selection is important and reinforces the need to investigate how the underlying feature representation influences detection performance.

In the present research, TF-IDF with N-gram representation is employed to capture the importance of terms and local sequential patterns in HTTP requests. However, the resulting representation can still contain a very large number of dimensions. The initial combined approach therefore incorporated Truncated SVD to reduce dimensionality before classification. Experiments on the CSIC-2010 and ECML/PKDD-2007 datasets demonstrated that the combined TF-IDF-N-gram and SVD representation could support effective SQLI classification using SVM and RF [16]. However, the ECML results were substantially lower than those obtained on CSIC-2010, motivating improvements to the feature-reduction stage.

A. Research Gap

The analysis of the existing approaches and the experimental progression of this study reveal four important gaps. First, although ML-based methods have demonstrated strong SQLI detection capability, SQLI-specific labeled data derived from realistic web logs remain challenging to obtain and prepare. Available web logs can contain multiple types of attack traffic, requiring appropriate filtering and preprocessing before they can be used for SQLI classification.

Second, many ML-based approaches concentrate primarily on classifier selection while giving comparatively less attention to the quality of the reduced feature representation. The source analysis identifies dimensionality reduction as an important component of ML classification and argues that more informative features can improve detection accuracy.

Third, the initial TF-IDF-N-gram + SVD approach reduces dimensionality but does not explicitly exploit latent semantic relationships among terms. SVD provides an algebraic/statistical representation, whereas LDA can model latent topics based on the co-occurrence of terms. The improved approach consequently investigates whether topic-based representation can produce a more informative low-dimensional feature space for SQLI classification.

Fourth, there is a need to evaluate feature-reduction strategies not only by classification accuracy but also by computational performance and overall detection capability. The research therefore compares the SVD-based combined approach with an improved LDA-based approach using SVM and RF and evaluates the resulting models using accuracy, detection-related measures, ROC analysis, and execution time.

Accordingly, this study addresses the following research gap: existing ML-based SQLI detection approaches lack an effective method to transform high-dimensional HTTP request representations into compact, informative features without losing useful semantic information. To address this gap, the proposed framework replaces SVD with LDA after constructing TF-IDF N-gram features and investigates whether latent topic-based dimensionality reduction improves SQLI classification performance. The improved framework is evaluated on filtered and balanced SQLI data using SVM and RF classifiers. The experimental results reported in the study indicate that the LDA-based approach achieves higher accuracy and lower computational requirements than the earlier SVD-based configuration.

III.PROPOSED METHODOLOGY

The proposed methodology is designed to detect SQL injection attacks in HTTP web log data by combining text-based feature representation, semantic dimensionality reduction, and supervised machine learning. The central premise is that SQLI requests contain textual and sequential characteristics that can be represented computationally and learned by classification models. The methodology, therefore, focuses on transforming raw web requests into a compact, informative feature space before performing binary classification.

The complete detection process consists of four principal stages: (i) web-log preprocessing and SQLI traffic extraction, (ii) TF-IDF N-gram feature construction, (iii) LDA-based dimensionality reduction, and (iv) supervised classification using SVM and RF. This structure follows the experimental framework developed in the source study, while the proposed improvement specifically replaces the earlier SVD-based reduction stage with LDA.

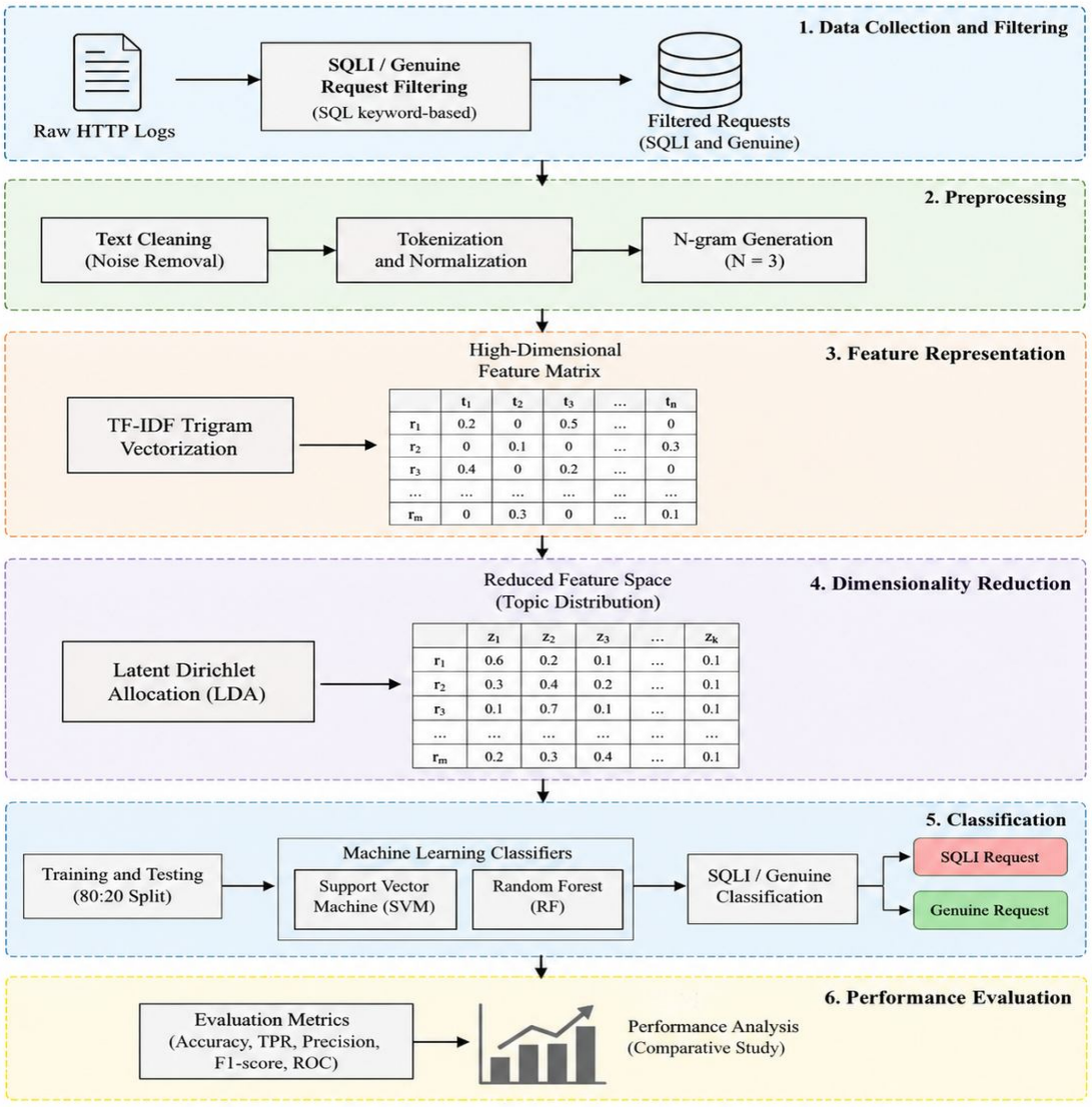


Fig.1 Proposed TF-IDF-N-gram and LDA-based SQL injection detection framework.

A. Web-Log Preprocessing and SQLI Extraction

The first stage converts raw HTTP web logs into a dataset suitable for machine learning. Web-log datasets may contain both legitimate requests and different categories of malicious traffic; therefore, directly applying a classifier to the complete log can introduce irrelevant patterns and increase the complexity of the learning problem. The study addresses this issue by identifying and extracting SQLI-related HTTP requests from the available logs.

A SQL keyword-based filtering mechanism is used to identify requests that contain SQL-related characteristics. The resulting SQLI traffic is then separated from genuine requests and prepared for binary classification. This preprocessing step is important because the objective is not general web-attack classification, but specifically the distinction between SQLI and legitimate HTTP requests. The research framework emphasizes that raw logs contain a mix of web-attack requests and consequently require filtering and processing before they can be effectively used for SQLI detection.

The processed data are balanced between malicious and genuine classes to reduce the potential influence of class imbalance on the learning process. For the improved ECML/PKDD-2007 experiment, the final dataset contains 2,274 SQLI requests and 2,274 genuine requests

B. TF-IDF N-Gram Feature Construction

HTTP requests are textual sequences and cannot be directly supplied to conventional supervised classifiers. Therefore, the second stage transforms each request into a numerical representation using Term Frequency–Inverse Document Frequency (TF-IDF) combined with an N-gram representation.

TF-IDF assigns greater importance to terms that are frequent within a particular request but relatively uncommon across the complete corpus. In addition to individual terms, N-grams capture local sequential relationships between neighboring tokens. In this study, trigram representation (N = 3) is employed to capture short sequential patterns within HTTP requests.

The resulting TF-IDF-N-gram representation provides a richer description of SQLI requests than a simple term-frequency representation. However, it can generate a very large sparse feature matrix. A large number of dimensions may increase computational requirements and introduce less informative or redundant characteristics into the classification process. The source study consequently identifies feature extraction and dimensionality reduction as critical components of the ML-based SQLI detection pipeline.

C. LDA-Based Dimensionality Reduction

The principal methodological improvement of the proposed framework is the use of Latent Dirichlet Allocation (LDA) instead of SVD for dimensionality reduction. The earlier combined model used TF-IDF-N-gram followed by SVD to obtain a lower-dimensional representation. Although SVD effectively reduces the size of the feature space, it primarily provides a statistical decomposition and does not explicitly model latent semantic relationships among terms.

LDA addresses this limitation through probabilistic topic modeling. It assumes that documents are associated with latent topics and that topics are represented through distributions over words. Applied to SQLI HTTP requests, LDA can identify latent patterns arising from the co-occurrence of terms and represent each request through a smaller number of topic-based dimensions. Thus, instead of retaining a large sparse TF-IDF feature space, the proposed approach converts the requests into a compact latent representation.

The source experimentation specifically motivates this replacement, as LDA can identify semantic co-occurrence between observed words and latent topics, thereby generating informative, reduced features for subsequent classification.

The number of latent components is selected experimentally rather than arbitrarily. The source methodology uses dimensionality-reduction evaluation and model mean and standard deviation measures to identify an appropriate number of components.

D. Supervised Classification

After dimensionality reduction, the resulting feature vectors are supplied to two supervised classifiers: Support Vector Machine (SVM) and Random Forest (RF).

SVM is employed because it is well-suited to classification problems involving high-dimensional representations. It identifies a decision boundary that separates the malicious and legitimate classes. RF provides a complementary tree-based classification approach based on an ensemble of decision trees. Using both classifiers enables the proposed feature representation to be evaluated across different learning paradigms rather than attributing performance improvements to a single classifier.

The classification task is formulated as a binary decision:

$$y \in \{0,1\}$$

where y=1 represents an SQLI request and y=0 represents a genuine request.

The source experiments use an 80:20 training-to-testing split and evaluate SVM and RF on the resulting reduced datasets.

E. Performance Evaluation

The proposed framework is evaluated using classification and computational-performance measures. The principal classification measure is accuracy, while detection capability is assessed using true positive rate (TPR)/recall, precision, F1-score, and ROC analysis. Training and testing time are additionally considered to determine whether dimensionality reduction improves computational efficiency.

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

The true positive rate is calculated as:

$$TPR = \frac{TP}{TP + FN} \times 100$$

Precision and F1-score are calculated as:

$$Precision = \frac{TP}{TP + FP} \times 100$$

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

These measures provide complementary information: accuracy indicates overall classification correctness, TPR reflects the ability to identify SQLI requests, precision indicates the reliability of positive predictions, and F1 provides a combined measure of precision and recall. The source methodology explicitly uses confusion matrices, accuracy, detection rate (TPR), precision, F1 Score, ROC curves, and execution time for model evaluation.

F. Overall Proposed Pipeline

The proposed TF-IDF-N-gram + LDA framework can therefore be summarized as:

Raw HTTP Logs → SQLI/Genuine Filtering → Preprocessing → TF-IDF Trigram Representation → LDA Dimensionality Reduction → SVM/RF Classification → Performance Evaluation

The key distinction from the earlier combined model is the dimensionality-reduction stage. The earlier configuration followed TF-IDF-N-gram → SVD → classifier, whereas the proposed improved configuration follows TF-IDF-N-gram → LDA → classifier. This modification aims to obtain a more informative low-dimensional representation by incorporating latent topic information while reducing the computational burden of the original high-dimensional feature space. The experimental evidence in the source study reports that the LDA-based configuration improved both classification performance and training/testing efficiency on the investigated ECML dataset.

IV. EXPERIMENTAL SETUP AND DATASET DESCRIPTION

The proposed framework was evaluated using two publicly available web-traffic datasets: CSIC-2010 and ECML/PKDD-2007. These datasets were selected to evaluate the proposed approach on HTTP requests that contain both malicious and legitimate traffic. Since the original logs contain various web requests and attack traffic, SQLI-related requests were extracted and processed first before model training.

A. Dataset Preparation

SQLI requests were identified in the raw web logs using a SQL keyword-based filtering mechanism. The extracted requests were then processed using the TF-IDF N-gram representation. To obtain a suitable classification dataset, SQLI and genuine requests were balanced. This preprocessing step was necessary because the availability of appropriately prepared and labeled SQLI data is a major challenge to applying ML techniques to web logs.

For the CSIC-2010 dataset, the resulting balanced dataset contained 4,549 SQLI requests and 4,549 genuine requests. For ECML/PKDD-2007, the balanced dataset contained 2,274 SQLI requests and 2,274 genuine requests. The improved LDA-based experiment was conducted on the ECML dataset using these 4,548 requests.

B. Experimental Configuration

The experiments followed an 80:20 training-to-testing ratio. TF-IDF with a trigram (N=3) representation was used for feature construction. The resulting high-dimensional feature space was reduced using LDA in the proposed framework. The reduced feature vectors were subsequently provided to Linear SVM and Random Forest classifiers for binary SQLI classification.

To demonstrate the effectiveness of the proposed improvement, the LDA-based framework was compared with the earlier TF-IDF-N-gram + SVD configuration. Model performance was assessed using accuracy, TPR/recall, precision, F1 Score, ROC analysis, and training/testing times.

The experiments were implemented using Python and the scikit-learn machine learning library.

This experimental design enables a direct assessment of whether LDA-based semantic dimensionality reduction provides a measurable advantage over SVD for SQLI detection from HTTP web logs.

V. RESULTS AND DISCUSSION

The proposed TF-IDF-N-gram + LDA framework was evaluated using SVM and Random Forest classifiers and compared with the earlier TF-IDF-N-gram + SVD configuration. The comparison primarily examines whether replacing SVD with LDA improves SQLI detection accuracy while maintaining computational efficiency.

For the CSIC-2010 dataset, the earlier combined approach achieved 99.90% accuracy with SVM and 98.13% with RF. On the ECML/PKDD-2007 dataset, the corresponding SVD-based approach achieved 84.61% with SVM and 82.97% with RF. The comparatively low ECML performance motivated improvements to the dimensionality-reduction stage.

TABLE I DATASET CONFIGURATION

Dataset	SQLI Requests	Genuine Requests	Total Samples
CSIC-2010	4,549	4,549	9,098
ECML/PKDD-2007	2,274	2,274	4,548

The improved approach replaced SVD with LDA to obtain a more informative latent representation of the HTTP-request features. On the ECML/PKDD-2007 dataset, the improved model achieved 98.70% accuracy using SVM and 99.85% using RF. The study also reports approximately 99% TPR and mean ROC performance for the improved scheme. Both classifiers demonstrated improved performance compared with the earlier SVD-based configuration, while training and testing time were also reduced.

TABLE II CLASSIFICATION PERFORMANCE

Dataset	Dimensionality Reduction	Classifier	Accuracy
CSIC-2010	SVD	SVM	99.90%
CSIC-2010	SVD	RF	98.13%
ECML/PKDD-2007	SVD	SVM	84.61%
ECML/PKDD-2007	SVD	RF	82.97%
ECML/PKDD-2007	LDA (Proposed)	SVM	98.70%
ECML/PKDD-2007	LDA (Proposed)	RF	99.85%

The results indicate that the principal improvement is attributable to the feature-reduction strategy rather than to simply changing the classifier. The LDA representation captures latent relationships among terms in the HTTP request corpus and yields a compact feature space for classification. This provides a plausible explanation for the substantial improvement observed on ECML/PKDD-2007. The source study concludes that LDA generated more informative features, reduced model complexity, and provided better detection performance than SVD in the investigated setting.

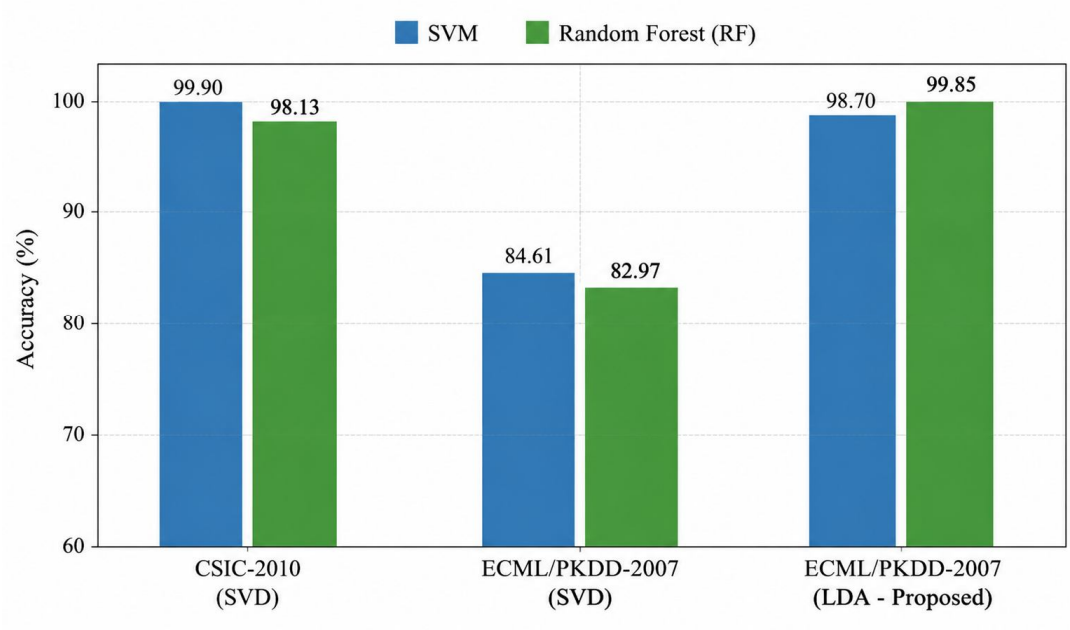


Fig. 2. Accuracy comparison between SVD- and LDA-based SQLI detection approaches.

Overall, the experimental results support the effectiveness of combining TF-IDF N-gram representation with LDA-based dimensionality reduction for SQLI detection. The proposed framework achieved high classification accuracy while addressing the high-dimensional nature of HTTP-log features, demonstrating its potential as an efficient ML-based approach for SQLI detection.

VI. CONCLUSION AND FUTURE SCOPE

This study presented an improved machine-learning framework for detecting SQL injection attacks in HTTP web log data using TF-IDF N-gram feature representation and Latent Dirichlet Allocation (LDA). The framework addresses the high dimensionality of textual web requests by applying LDA to obtain a compact latent representation prior to classification. SVM and Random Forest were employed to evaluate the effectiveness of the resulting feature space.

Experimental results demonstrate that replacing SVD with LDA substantially improved SQLI detection performance on the ECML/PKDD-2007 dataset. The proposed approach achieved 98.70% accuracy with SVM and 99.85% with Random Forest, compared with 84.61% and 82.97%, respectively, for the earlier SVD-based configuration. The improved framework also demonstrated approximately 99% TPR and mean ROC performance, while requiring lower training and testing time.

These findings indicate that semantic-oriented dimensionality reduction can play an important role in improving ML-based SQLI detection. In particular, LDA provides a compact representation based on latent topic relationships and, for the datasets investigated, proved more effective than SVD in reducing dimensionality.

Future research can extend this framework to multiclass SQLI detection, including categories such as tautology, logically incorrect queries, union-based queries, and alternate-encoding attacks. Further investigation can also consider other web vulnerabilities, deep-learning architectures, larger labeled SQLI corpora, and integrated detection-and-prevention mechanisms.

VII. ACKNOWLEDGMENT

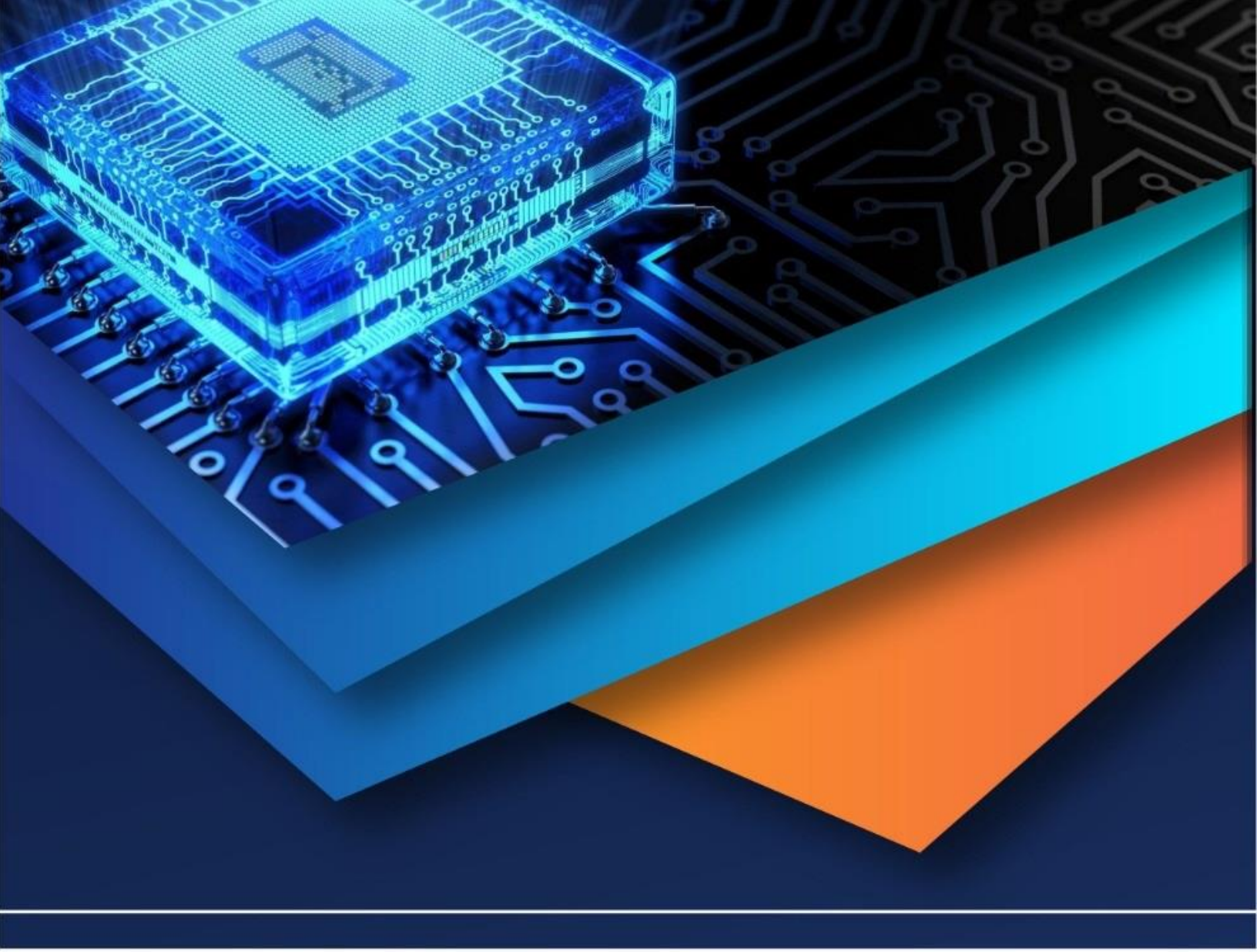
The authors would like to express their sincere gratitude to their institution and colleagues for their support and encouragement throughout this research. The authors also acknowledge the availability of the CSIC-2010 and ECML/PKDD-2007 datasets, which enabled the experimental evaluation of the proposed SQL injection detection framework.

The authors further appreciate the researchers and organizations whose publicly available resources and prior studies contributed to the development and evaluation of this work.

REFERENCES

[1] A. Hariyani and P. Dolia, “Comprehensive review of advanced techniques for mitigating SQL injection vulnerabilities in modern applications,” International Journal of Innovative Science and Research Technology, vol. 10, no. 3, pp. 3063–3070, Mar. 2025, doi: 10.38124/ijisrt/25mar1982.

[2] K. S. Fathi, S. I. Barakat, and A. Rezk, "An effective SQL injection detection model using LSTM for imbalanced datasets," Computers & Security, vol. 153, Art. no. 104391, Jun. 2025, doi: 10.1016/j.cose.2025.104391.
[3] A. Hariyani and P. Dolia, "An innovative method for detecting SQLi attacks by altering SQL query attribute values," International Journal of Advanced Computer Research, vol. 14, no. 68, pp. 89–96, Sep. 2024, doi: 10.19101/IJACR.2024.1466005.
[4] E. Casmiry, N. Mduma, and R. Sinde, "Enhanced SQL injection detection using chi-square feature selection and machine learning classifiers," Frontiers in Big Data, vol. 8, Art. no. 1686479, Nov. 2025, doi: 10.3389/fdata.2025.1686479.
[5] R. Bakır, "UniEmbed: A novel approach to detect XSS and SQL injection attacks leveraging multiple feature fusion with machine learning techniques," Arabian Journal for Science and Engineering, vol. 50, no. 19, pp. 15591–15604, 2025, doi: 10.1007/s13369-024-09916-4.
[6] A. Hariyani and P. Dolia, "CryptoSQLShield: A comprehensive study on cryptography-assisted methods for SQL injection defense," International Journal of Engineering Research & Technology (IJERT), vol. 15, no. 1, Jan. 2026, Art. no. IJERTV15IS010090, doi: 10.17577/IJERTV15IS010090.
[7] X. Wang, Y. Zheng, Z. Wan, and M. Zhang, "SVD-LLM: Truncation-aware singular value decomposition for large language model compression," in Proc. 13th Int. Conf. Learn. Representations (ICLR), 2025.
[8] A. Damayanti and A. Baita, "Comparison of support vector machine (SVM) and random forest (RF) algorithm performance with random undersampling technique to predict gestational diabetes mellitus risk," Journal of Applied Informatics and Computing, vol. 9, no. 2, pp. 328–337, Mar. 2025, doi: 10.30871/jaic.v9i2.9009.
[9] A. Hariyani and P. Dolia, "A cryptography-enforced SQL query integrity framework for complete SQL injection prevention," International Journal of Scientific Research in Engineering and Management, vol. 10, no. 3, pp. 1–9, Mar. 2026, doi: 10.55041/IJSREM58457.
[10] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1627078, Sep. 2025, doi: 10.3389/frai.2025.1627078.
[11] A. Hariyani, "CDiCENet: A structure-aware lightweight deep learning model for SQL injection attack detection on resource-constrained devices," International Journal of Engineering Research & Technology (IJERT), vol. 15, no. 8, Aug. 2026, Art. no. IJERTV15IS080682, doi: 10.17577/IJERTV15IS080682.
[12] M. E. D. Rafi, M. H. M. Fajar, M. S. Purwanto, A. Hilyah, A. S. Bahri, and H. K. Rahayu, "Analysis of formation Ronggojalu spring and Probolinggo active fault continuity with satellite data gravity method," Jurnal Penelitian Pendidikan IPA, vol. 9, no. 10, pp. 8456–8461, Oct. 2023, doi: 10.29303/jppipa.v9i10.3399.
[13] A. Hariyani, "A structured framework for detection and mitigation of SQL injection vulnerabilities in web applications," International Journal of Creative Research Thoughts (IJCRT), vol. 14, no. 8, pp. e332–e343, Aug. 2026, Paper ID: IJCRT2608464.
[14] A. Hariyani and P. Dolia, "SecureSQL: Preventing SQL injection attacks using cryptographic query protection and intelligent detection," in Proc. 4th Int. Conf. Cybersecurity and Generative Artificial Intelligence (CyberGenAI'2026), Mar. 2026, doi: 10.5281/zenodo.18957529.
[15] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," Array, vol. 27, Art. no. 100467, Sep. 2025, doi: 10.1016/j.array.2025.100467.
[16] S. He, Y. Zhang, D. Liang, and P. K. Sharma, "An unsupervised malicious web request detection based on transformer and contrastive learning," IEEE Transactions on Network and Service Management, vol. 22, no. 4, pp. 3281–3294, Aug. 2025, doi: 10.1109/TNSM.2025.3563089.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY