International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

An Ensemble Technique for SQL Injection Attack (SQLiA) Detection in Web Applications

Polireddi Indhumathi[1] *, Tentu Yamuna* [2]*, Mohammad Gaffar*[3]*, Srungavarapu Sai Srinivas*[4]

[1]Polireddi Indhumathi, Information Technology, Razam, Andhra Pradesh
[2]Tentu Yamuna, Information Technology, Razam, Andhra Pradesh
[3]Mohammad Gaffar, Information Technology, Razam, Andhra Pradesh
[4]Srungavarapu Sai Srinivas, Information Technology, Razam, Andhra Pradesh

ABSTRACT -

In SQL Injection (SQLi) remains a critical threat to the security of web applications, enabling attackers to tamper with database queries and potentially access confidential user information. To mitigate such risks, this project proposes a hybrid detection system that leverages both deep learning and traditional machine learning techniques. The system processes SQL queries using natural language processing methods, such as tokenization, normalization, and word embeddings, to capture both semantic and structural patterns.

Deep learning models like 1D-CNN, LSTM, and GRU are utilized to detect complex sequential behaviours in queries, while an ensemble of traditional classifiers—Random Forest, XG-Boost, and Light-GBM—focuses on feature-based detection. These models are integrated through a Flask-based API, allowing real-time analysis of SQL inputs via a web interface. The proposed system demonstrates high accuracy, precision, and recall, making it an effective solution for detecting and preventing SQL injection attacks in dynamic web environments.

Key Words: SQL Injection, Deep Learning, Ensemble Learning, Web Application Security, Natural Language Processing

1. INTRODUCTION

In the current era of digital transformation, web applications have become the backbone of numerous industries, playing an essential role in delivering online services across critical sectors such as banking, healthcare, education, and e-commerce. These applications are designed to provide seamless access to information and services, often requiring constant communication with backend databases to store and retrieve sensitive data like user credentials, medical records, academic information, and financial transactions. As organizations continue to digitize operations and offer services over the internet, the security of these web platforms has become more crucial than ever. One of the most severe and frequently exploited vulnerabilities in web applications is SQL Injection (SQLi).

SQLi attacks occur when an attacker injects malicious SQL code into input fields, thereby manipulating the structure of database queries executed by the application. If not properly detected and prevented, such attacks can lead to unauthorized access to confidential data, corruption or deletion of records, and in extreme cases, complete takeover of the underlying database system. Despite advancements in web application security and the increasing implementation of best practices, SQLi remains a formidable threat. Its persistence is attributed to the simplicity with which it can be executed and the devastating consequences it can produce.

Traditional security mechanisms, such as rule-based detection systems and Web Application Firewalls (WAFs), are commonly deployed to mitigate SQLi threats. However, these systems often rely on predefined patterns and rules, making them ineffective against advanced, obfuscated, or zero-day SQLi attack vectors that deviate from known patterns. To overcome the limitations of conventional methods, researchers and cybersecurity professionals have increasingly turned to intelligent, data-driven approaches. Machine learning (ML) and deep learning (DL) techniques have shown promise in identifying complex attack patterns by analysing large volumes of data and learning from underlying structures in SQL queries. These techniques do not require explicit rules but instead rely on training models to recognize anomalies and detect malicious input with greater precision. This paper proposes a hybrid detection framework that integrates deep learning models with ensemble machine learning algorithms to improve the accuracy and robustness of SQLi detection.

The approach leverages natural language processing (NLP) techniques to preprocess SQL queries, transforming them into a structured format suitable for analysis. Deep learning models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and convolutional neural networks (CNNs), are employed to extract temporal and spatial features from query sequences. In parallel, ensemble methods like Random Forest, XG-Boost, and Light-GBM are used to enhance prediction reliability by combining the strengths of multiple classifiers. The goal of the proposed system is to provide real-time protection for web applications by accurately detecting and flagging SQLi attempts before they can exploit database vulnerabilities.

By combining the learning capacity of deep neural networks with the decision-making efficiency of ensemble models, this hybrid approach aims to strengthen web application security against both known and emerging SQLi threats.

2. LITERATURE SURVEY

The author develops a deep learning-based system for detecting code injection vulnerabilities (SQLi, XSS, and command injection) using a hybrid CNN-GRU model with attention mechanisms for high accuracy. The study employs SXCM1 and DPU-WVD datasets, pre-processed for integrity. CNN extracts features, GRU captures sequential dependencies, and multi-head attention enhances detection. The model achieves 99.99% accuracy on DPU-WVD. However, limitations include dataset dependency, complexity affecting interpretability, real-time challenges, and a narrow focus on specific vulnerabilities, reducing adaptability to emerging threats. The research aims to enhance web security with efficient, automated detection.[1]

The authors explore machine learning and deep learning for SQL injection detection in web applications, evaluating models like Random Forest (99.68% accuracy), LSTM, CNN, CNN-BiLSTM (98% accuracy), and LightGBM (99.34% accuracy). They merge Kaggle and SQL Dataset UmarFarooq into 65,791 balanced records for analysis. The study highlights challenges such as dataset dependency, generalization issues, feature selection complexity, and overfitting risks. Despite these limitations, the research enhances security by improving detection accuracy and enabling integration with automated response systems, significantly contributing to cybersecurity advancements.[2]

3. PROPOSED SYSTEM

This work implements a machine learning pipeline for detecting SQL Injection Attacks (SQLiA) through a combination of deep learning models, machine learning classifiers, and web application security mechanisms. The methodology follows a structured five-stage approach, covering data acquisition, feature extraction, model training, deployment, and visualization.

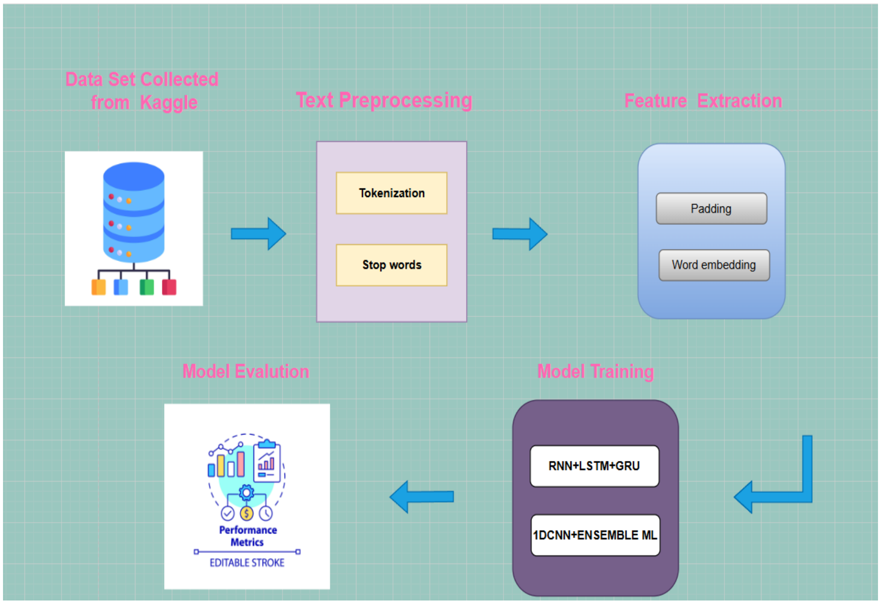


Fig. 1: SQL Architecture

A.Data Collection Methods

The study uses a Kaggle dataset of 30,980 SQL queries, with 15,420 labelled as malicious and 15,560 as benign. It includes various SQL injection types, such as Union-based, Boolean-based, Time-based, and Error-based attacks. Pre-labelled queries ensure reliable supervised learning. The dataset features diverse query structures, including numerical values, logical conditions, and subqueries, enabling machine learning models to detect SQL injection patterns effectively. Its mix of simple and complex queries makes it ideal for training robust classification models for real-world SQL injection detection, supporting the development of intelligent and secure database protection systems.

B. Text Preprocessing Techniques for SQL injection:

Text preprocessing is crucial for preparing SQL queries for deep learning models by structuring and normalizing data while preserving meaningful patterns. Key techniques include tokenization, which breaks queries into components (keywords, operators, identifiers, values), and stop-word removal to eliminate non-informative words while retaining critical terms like DROP. Lowercasing ensures uniformity, while special character removal eliminates unnecessary symbols, enhancing pattern detection. Stemming and lemmatization reduce word variations, and padding standardizes query lengths for deep learning models. These techniques refine SQL queries, improving efficiency, accuracy, and robustness in SQL injection detection systems by enabling effective machine learning analysis.

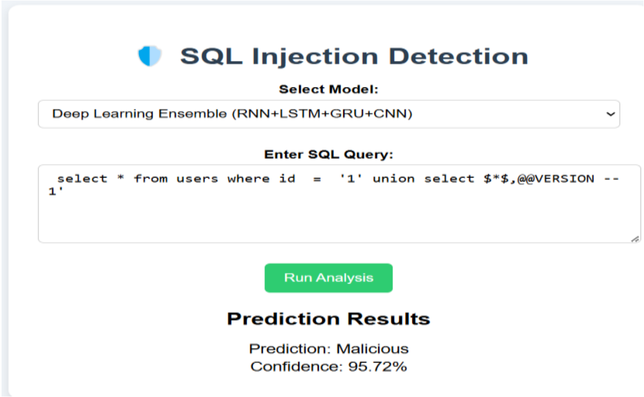
C. Feature Extraction Techniques:

Feature extraction converts SQL queries into numerical representations for machine learning. Key techniques include padding, ensuring fixed-length inputs by adding special tokens to shorter queries for consistency in models like LSTM and CNN. Word embeddings (Word2Vec, Glove) transform SQL terms into dense vectors, capturing contextual similarities (e.g., "SELECT" and "FROM" have similar representations), improving detection accuracy. Sequence based learning (RNN, LSTM, GRU) preserves word order, detecting SQL injection patterns like "OR 1=1." These methods enhance classification accuracy by capturing query structures, relationships, and anomalies, making SQL injection detection more robust and reducing false positives.

D. Model Training

Machine learning and deep learning models were trained to classify SQL queries as benign or malicious. RNN, LSTM, and GRU capture sequential patterns, with LSTM handling long-term dependencies and GRU reducing computational complexity. This combination ensures effective detection of SQL injection indicators. 1D-CNN extracts local query features, detecting attack patterns like "OR 1=1." An ensemble approach combining Random Forest, Gradient Boosting, and SVM enhances accuracy by leveraging multiple classifiers. By integrating CNN-based feature extraction with machine learning, the model improves SQL injection detection, reducing false positives while ensuring robust security analysis.

4. RESULT



SQL Injection Detection

Select Model: Deep Learning Ensemble (RNN+LSTM+GRU+CNN)

Enter SQL Query:

```
 select * from users where id  =  '1' union select $*$,@@VERSION --
1'
```

Prediction Results

Prediction: Malicious
Confidence: 95.72%

Fig. 2. Expected Result

5. CONCLUSION

This project presents a robust and intelligent system for detecting SQL Injection (SQLi) attacks in web applications by combining the strengths of deep learning and traditional machine learning techniques. By leveraging NLP-based preprocessing and word embeddings, the system effectively captures both syntactic and semantic patterns in SQL queries. Deep learning models such as 1D-CNN, LSTM, and GRU enhance the system's ability to identify complex and hidden attack patterns, while ensemble learning methods like Random Forest, XGBoost, and LightGBM improve classification reliability through feature-based analysis. The integration of these models into a web-based application using Flask allows for real-time query analysis and threat detection. Overall, the system demonstrates high accuracy and adaptability, providing a scalable and practical solution for enhancing web application security against SQL injection threats.

REFERENCES -

[1]. Ali, S. H., Mohammed, A. I., Mustafa, S. M., & Salih, S. O. (2025). WEB VULNERABILITIES DETECTION USING A HYBRID MODEL OF CNN, GRU AND ATTENTION MECHANISM. Science Journal of University of Zakho, 13(1), 58-64.

[2]. Seada, Y., Mohamed, A., Hany, M., Mansour, H., & Elsersy, W. (2024, July). A Machine Learning Approach to SQL Injection Detection in Web Applications. In 2024 Intelligent Methods, Systems, and Applications (IMSA) (pp. 26-32). IEEE

[3].Crespo-Martínez, I. S., Campazas-Vega, A., Guerrero-Higueras, Á. M., Riego-DelCastillo, V., Álvarez-Aparicio, C., & Fernández-Llamas, C. (2023). SQL injection attack detection in network flow data. Computers & Security, 127, 103093.

[4].Tram, D. T. N., & Cam, N. T. (2024, August). uitSQLid: SQL Injection Detection Using Multi Deep Learning Models Approach. In 2024 International Conference on Information Management and Technology (ICIMTech) (pp. 765-770). IEEE.

[5]. Leelaprute, P., Kase, Y., Amasaki, S., Aman, H., & Yokogawa, T. (2024, May). A Multi-Aspect Evaluation of DL-based SQLi Attack Detection Models. In 2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 352-355). IEEE.