International Journal of Novel Trends and Innovation
IJNTI.ORG | ISSN : 2984-908X
An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY ON SQL INJECTION DETECTION

[1]Deeksha M B, [2]Dr. Kavyasri M N, [3]Anusha K R, [4]Harshil K, [5]Aniketh Patil

[1]Student, [2]Associate Professor, [3]Student, [4]Student, [5]Student

[1]Department of Computer Science and Engineering,

[1]Malnad College of Engineering, Hassan, India

Abstract : The increasing reliance on web-based applications and databases has increased the risk of SQL Injection Attacks (SQLIAs), which is one of the most widespread threats to web platforms. SQLIAs take advantage of weaknesses in user input validation to run malicious SQL queries, leading to unauthorized access, data theft, and extreme economic loss. In response to these issues, many research studies have looked into machine learning (ML) and deep learning (DL) solutions to improve SQLIA detection. Methods like 2D-Convolutional Neural Networks (2D-CNNs), Elastic-Pooling CNNs, Multi-Layer Perceptrons (MLPs), and hybrid approaches like CNN-BiLSTM have shown higher accuracy and adaptability compared to conventional rule-based systems. These models utilize features like word embeddings, syntactic parsing, and vector representations to detect SQLIA patterns efficiently. Using Kaggle and other data sets, experiments show the resilience of DL models in real-time detection and capacity to classify intricate SQLIAs, guaranteeing a substantial improvement in protecting database-driven systems.

IndexTerms - SQL Injection, Anomaly Detection, Convolutional Neural Networks (CNN), Multi-Layer Perceptron (MLP), Word Embedding, Malicious Query.

INTRODUCTION

The exponential growth of the web and heavy use of web-based applications have served to make sure that information is readily available, radically transforming the way human beings and organizations deal with information. But further dependence on the internet-based technology has introduced more issues towards security loopholes leaving users vulnerable to possible attacks by scam artists. One of the most common and dangerous forms of online attacks among all such attacks is SQL Injection, also referred to as SQLi. SQLi is when the attackers take advantage of weaknesses in the database layer of a web application by injecting malicious SQL commands into input fields. These queries, when they are run by the database, tend to undermine authentication and authorization checks, hence providing attackers with unauthorized access to confidential data, changing or deleting data, and even take over whole systems. The effect of SQLi attacks is far-reaching, with attackers having the potential to retrieve sensitive data, manipulate data integrity, and take control of the administrative level. Conventional defenses like firewalls and web application firewalls do provide some level of protection, but they tend to fall short when handling the dynamic and changing nature of SQL Injection Attacks (SQLIAs). In reaction to this, sophisticated detection techniques like machine learning (ML) and deep learning (DL) have emerged as dominant solutions. These approaches utilize tools such as word embeddings to convert text-based SQL queries into numerical representation and use neural architectures, such as Convolutional Neural Networks (CNNs) [2][4], Long Short-Term Memory networks (LSTMs), and hybrid models, for automatic feature extraction, pattern learning, and anomaly detection. Through the greater dramatic effectiveness of their execution as compared with conventional methods, these

newer techniques increase the need for robust, scalable systems to defend against SQLIAs with maximum effectiveness, providing complete protection of sensitive web content in the fast-paced environment of the age of the internet.

LITERATURE REVIEW

In this section a concise detail about the works of various authors about the Detection of SQL Injection Attack , in order to secure a web application has been discussed.

Raheem and Rana suggested a 2D-Convolutional Neural Network (2D-CNN) to detect SQL Injection Attacks (SQLIAs) with a novel approach where one-dimensional textual SQL data is converted to a two-dimensional matrix by applying word embedding techniques such as the Skip-Gram model [1]. This process maintains the context and syntax of the queries to give a comprehensive feature extraction. The dataset, obtained from Kaggle, had varied types of SQLIAs and normal queries such that binary classification into normal and attack queries could be achieved. Preprocessing involved data preparation by structuring the dataset and expressing SQL queries in vectorized forms such that a fixed dimension was maintained to ensure uniformity. The 2D-CNN was implemented in Python and TensorFlow with convolutional, pooling, and fully connected layers and a sigmoid activation function for classification. The model achieved 99.66% accuracy, establishing the superiority of the proposed technique over traditional techniques.

Andri Setiyaji and others designed CNN-based approach to identify SQLIAs, with focus on log analysis and real-time performance [2]. Preprocessing of the dataset, obtained from a big public repository hosted on Kaggle, included tokenization of input queries, stripping of extra characters, and conversion of text data to numerical values. Convolutional and pooling layers were used in the CNN architecture to derive meaningful features from the SQL queries, and then a softmax function for classification was applied. The authors also performed a comparison between CNN model and other machine learning methods, such as Naive Bayes, Decision Trees, and Logistic Regression, on the basis of precision, recall, and F1 score. The CNN model performed better than others with an accuracy level of 91%, indicating its ability to generalize and learn from unseen data patterns. The research showed the merits of integrating machine learning and deep learning to effectively identify SQLIAs.

Jothi K. R and team introduced a deep learning system with an MLP architecture for SQLIA detection [3]. The approach was centered around automatic feature extraction from raw SQL queries through embedding layers, which converted text data into a format that can be processed by a neural network. The model was trained on a carefully curated dataset with different kinds of SQL injection techniques, including error-based, union-based, and blind attacks. Preprocessing involved tokenization, stemming, and word index generation to provide uniformity of input. The MLP model was assessed on cross-validation accuracy at 98%, with high recall and precision scores. The research highlighted the scalability and capability of the model to detect all known forms of SQLIA techniques, overcoming drawbacks in conventional systems that are based on predetermined feature sets.

Misquitta and Dr. Asha S utilized CNNs to develop a real-time SQLIA detection model by converting SQL query data into structured inputs with the help of word embeddings and vectorization methods [4]. Preprocessing included data cleaning, tokenization, and padding to standardize inputs. The authors utilized a hybrid dataset with benign SQL queries and injection patterns. The architecture of the CNN model included convolutional, pooling, and fully connected layers, specifically optimized to detect high-dimensional injection activity. Moreover, the research compared the performance of CNN with traditional machine learning algorithms like Support Vector Machines (SVM) and K- Nearest Neighbors (KNN). The CNN surpassed the traditional approaches in recall and precision, proving its adaptability and robustness to mutating attack strategies. The approach highlighted the significance of feature engineering and the capability of CNNs to identify sophisticated SQLIAs.

A. Sivasangari et al. suggested a detection system utilizing the AdaBoost algorithm to detect SQLIAs with improved accuracy [5]. The approach began with the collection of samples of SQL, such as different attack patterns like union-based, error-based, and blind SQL injections. Preprocessing included dividing data into stumps and giving weights to the stumps based on their accuracy in classification. The AdaBoost algorithm iteratively learned weak classifiers and combined their outputs to create a strong predictive model. The iterative weighting scheme corrected misclassifications, and high overall accuracy was ensured. The architecture showed better results than neural network-based approaches, especially in handling imbalanced datasets and accurately classifying SQL queries. The research identified the benefits of adaptive boosting in noise reduction and enhancing model robustness.

Table 1: Comparative Analysis of Machine Learning-Based Approaches for SQL Injection Attack Detection

Title	Proposed Approach	Advantages	Limitations	Research Result
SQL Injection Detection Using 2D-Convolutional Neural Networks.	2D-CNN model employing word embedding (Skip-Gram) to convert 1D SQL queries to 2D matrices for extracting features.	- Feature extraction automated - High accuracy (99.66%) - Resilient to diverse SQLIA.	- Dependent on quality datasets - High training cost computationally.	Accomplished 99.66% accuracy.
A Technique Utilizing Machine Learning and CNN for SQL Injection Detection.	CNN-based model for log analysis and real-time SQLIA detection with preprocessing for tokenizing and normalizing the input data.	- Real-time detection - High adaptability - CNN performed better than conventional ML methods such as Naive Bayes and Decision Trees.	- Needs regular updates to manage new SQLIAs - Moderate accuracy (91%) with respect to deep DL models.	Accomplished 91% accuracy, proving the efficiency of CNN in detecting SQLIA.
An Efficient SQL Injection Detection System Using Deep Learning.	MLP model with embedded feature extraction through embedding layers for vectorization of SQL queries.	- Scalable and adaptive - Senses all known SQLIA types - High precision (98%) and recall (97%).	- Low ability to find new patterns not included in training data.	Reached 98% cross-validation accuracy with high precision and recall scores.
SQL Injection Detection Using Machine Learning and CNN.	Preprocess CNN model with cleaning and vectorization of SQL queries to make high-dimensional analysis of features feasible.	- Resilient to high-dimensional data - Performed better than conventional ML methods such as SVM and KNN.	- Heavy reliant on heavy preprocessing - Needs large, diverse training data.	CNN exhibited better recall and accuracy than classical ML techniques.
SQL Injection Attack Detection Using Machine Learning Algorithm.	AdaBoost algorithm with weighted stumps to classify the SQL queries and resolve class imbalance problems.	- Performed well in imbalanced datasets - High precision due to iterative weight update - Less susceptible to overfitting.	- Vulnerable to noisy data - Needs considerable computing resources to train.	AdaBoost demonstrated better detection accuracy than existing classifiers, especially on noisy datasets.

CONCLUSION

The review of the five SQL Injection Detection papers showcases great progress in the use of machine learning and deep learning models to improve cybersecurity. The approaches, from 2D-CNN models to AdaBoost algorithms have shown great strength in detecting SQL injection attacks with high precision and reliability. These models perform well in automatic feature extraction, dealing with different attack patterns, and offering scalability for different applications. Yet, significant limitations remain. Most models are heavily dependent on large, high-quality datasets, which constrain their capability in identifying new or emerging attack patterns. In addition, the computational requirements of training complex neural networks like CNNs and MLPs can complicate deployment in low-resource environments. The use of static datasets, which do not account for emerging SQLIA variants, is also a limitation, as this constrains their capability to generalize effectively in real-world scenarios. In order to overcome these weaknesses, the research in the future should aim to design lightweight and adaptive models that can perform efficiently without demanding much computational power. Attempts should also be made to develop dynamic datasets that are periodically updated

to feature new attack patterns, so that models can keep themselves updated with changing threats. Also, hybrid methods that combine various learning paradigms, for example, supervised learning with unsupervised learning can make detection more robust and decrease false positives. All these improvements can work together to construct more robust and adaptive systems for SQL injection detection, which will provide greater security for database-driven applications in the more complicated cyber environment.

REFERENCES

[1]Fouad Raheem Abdulhamza and Rana Jumaa Surayh Al-Janabi, "SQL Injection Detection Using 2D-Convolutional Neural Networks (2D-CNN)," International Conference on Data Science and Intelligent Computing (ICDSIC 2022), IEEE, 2022.

[2]Andri Setiyaji, Zulkifli Yasin Hidayatulloh, Kalamullah Ramli, and G.S. Budhi Dharmawan, "A Technique Utilizing Machine Learning and Convolutional Neural Networks (CNN) for the Identification of SQL Injection Attacks," 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA 2024), IEEE, 2024.

[3]Jothi K.R., Pradyumn Beriwal, Saravana Balaji B., Abhinandan Amarajan, and Nishant Pandey, "An Efficient SQL Injection Detection System Using Deep Learning," International Conference on Computational Intelligence and Knowledge Economy (ICCIKE 2021), IEEE, 2021.

[4]Jason Misquitta and Dr. Asha S., "SQL Injection Detection Using Machine Learning and Convolutional Neural Networks," 5th International Conference on Smart Systems and Inventive Technology (ICSSIT 2023), IEEE, 2023.

[5]A. Sivasangari, J. Jyotsna, and K. Pravalika, “SQL Injection Attack Detection Using Machine Learning Algorithm,” Fifth International Conference on Trends in Electronics and Informatics (ICOEI 2021), IEEE,2021.