جمهورية العرا ق

وزارة التعليم العالي والبحث العلمي

جامعة ديالى

كلية العلوم

قسم علوم الحاسبات

تصميم نموذج لاكتشاف هجوم حقن SQL اعتمادا على التعلم العميق

رسالة مقدمة

الى كلية العلوم في جامعة ديالى وهي جزء من متطلبات نيل

شهادة الماجستير في علوم الحاسبات

تقدم بها الطالب

باسم حسين علي

بإشرا ف

أ.م.د عبد الباسط كاظم شكر

1444هـ

2023م

Chapter One
General Introduction

Chapter One

Introduction

1.1 Introduction

One of the most crucial areas of study in network security is web attacks. Network information has multiplied due to the rapid development of Internet technology [1]. Due to their vulnerability and network accessibility, web applications are often a simple target for cyberattacks [2]. Although there are numerous online attacks, "Structured Query Language" (SQL) injection represents one of the most public and will likely rank among the top 10 web dangers in 2021 [1].

SQL Injection attacks enable attackers to change or retrieve sensitive data. to take advantage of the operating system of the database server and change their focus to other targets on the victim's network [3].

Most of the available solutions can only find a small number of SQL injection attempts and can't adapt to new ways of attacking. So, there needs to be research and development on a deep learning-based detection solution. By using a deep learning classifier, SQL injection threats can be discovered [4].

This chapter will give background of SQL injection Attack, in addition to the description of the problem, the Aim, and an outline of the thesis.

1.2 Background of SQL Injection

Data in databases using Structured Query Language may be "added," "deleted," "changed," and "queried". SQL only works with a single kind of database, called a (relational database), and it also enables a user to choose the organization and structure of the data that is kept as well as the relationships between the data items [5].

SQL is a non-procedural programming language. It is a specialized language that differs greatly from other programming languages, such as Java or C.

SQL is the language of control and dealing with coherent and interconnected databases by working through data entry operations, data structures, filtering, searching, deletion, sorting, data modification, and other tasks.

SQL injection vulnerabilities occur when programmers utilize strings as part of SQL commands sent to a database. In this way, attackers can alter SQL statements by introducing keywords or special symbols. The model gets attacked after execution.

The following figure (1.1) presents the procedure of the SQL injection attack[6].

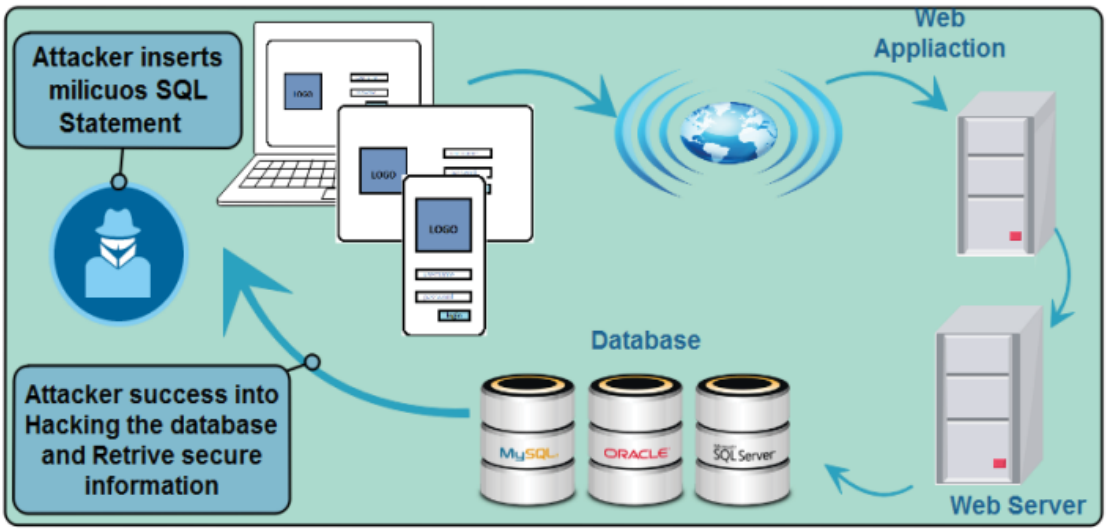

Figure (1.1) Procedure of the SQL Injection Attack [6]

SQL Injection attacks categorize according to the attacker's goals (such as data extraction or database schema discovery include (finding database structure, add records, avoid detection, cause a denial of service, issue commands remotely) and tautology, egal wrong queries, union queries, stored

procedures, other encodings, blind injections, and timing assaults are all examples of technical methods used in SQL injection attacks [7][8]. Figure (1.2) depicts the recent rise in the amount of SQL injection attacks [9].

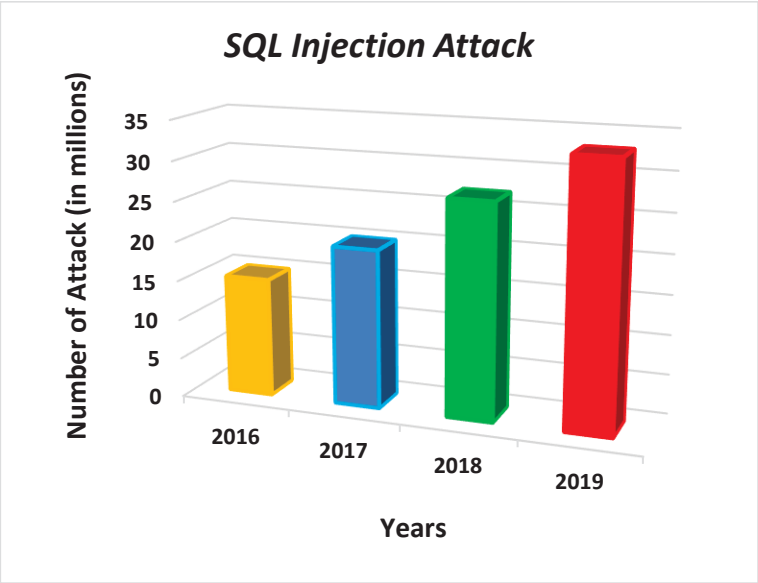

Figure (1.2) The histogram of the Increase in the Number of SQLI Attacks Globally [9]

1.3 Problem Statement

Due to the many uses of web applications in our daily lives, especially in commercial and financial transactions, dealing with such a thing is a double-edged sword.

Due to the large number of attacks on web applications, one of the most famous and most harmful is the SQL attack.

The main problem is how to protect these applications from those attacks.

1.4 The Aim of the Thesis

develop and build a model to detect SQL injection attacks.

1.5 The Objective of the Thesis

1. study and analyze machine learning (MLP model) and deep learning (1D-CNN model)

2. The comparison between the result MLP model and the 1D-CNN model.

3. Get the best model able the detection from SQL injection attack.

4. Evaluate the performance of the proposed model.

5. To run the proposed DSQLIAM model in a real-time environment based on the H5 model. The H5 form is used to save parameters of a CNN, Such as weights, kernels, and layers, we implement the online model on a website.

1.6 The contributions

1. propose a model to compare machine learning and deep learning to detect SQL injection.

2. detect SQL injection by 1D-CNN model.

3. propose DSQLIAM to classify normal attacks and SQL attacks.

1.7 Related Works

This section reviews related works in SQL injection attack detection and compares it with them, as shown below:

- **P. Tanget et al. in 2020** [10]: They suggested a neural network-based method for detecting SQL injection that is both efficient and robust. To begin, They sifted through a mountain of SQL injection data for clues on what to focus on. In the following years, various neural network models were created, including the multi-layer perceptron (MLP) and the long short-term memory (LSTM). The proposed method was put to the test using an open-source dataset consisting of 6080 samples, 3040 of which were malicious and 3040 of which were benign. The suggested model had an accuracy level greater than **99.5%.**
- **D. Chen et al. in 2021** [1]: devised a lightweight approach to protecting against SQL injection threats by using word embedding, CNN (convolutional neural network), and MLP (multi-layer perceptron). The authors begin by de-noising and decoding HTTP queries to identify the fraudulent request. The decrypted letters are then used to train a CNN and MLP classifier, which is subsequently used to detect the fake request. Both proposed models were tested using a total of 4,000 normal and 4,000 SQLI data samples, respectively. The models' stated accuracy is above **98%**, however as different neural networks have distinct application settings, performance may vary significantly.
- **A. Krishnan et al. in 2021**[8]: conducted research about SQL injection detection-based machine learning and deep learning, evaluating the effectiveness of some selected algorithms in the quest to establish the most robust learning model, including Naive Bayes, Logistic Regression, CNN, SVM, and passive-aggressive algorithms. The

proposed model achieved the best accuracy value of **97%** with CNN's deep learning algorithm.

- **N. Gandhi et al . in 2021**[9]: suggest a hybrid CNN-BiLSTM(A Convolutional Neural Network-Bidirectional Long short-term memory-based) strategy for detecting SQLI attacks. The authors provided a thorough comparative examination of There are many machine-learning strategies used to identify SQL injection attacks. By lowering the frequency of SQL injection attacks by foreseeing them using a suggested hybrid CNN-BiLSTM-based machine learning model, the study contributes to the area of machine learning. 3072 SQL injection requests and 1128 regular data searches were included in the data collection, which was collected from multiple websites. Comparing the CNN-BiLSTM method to other machine learning algorithms, it offered an accuracy of around **98%**.

- **K.R. Jothi et al. in 2021**[11]: They developed an artificial neural network (ANN)-based detection model for SQL injection attacks. The suggested model the has benefit of being able to identify all varieties of injection procedures. The model itself will extract and choose all of the features. The suggested model uses data from Lib-injection, a Python library with SQL injection queries. The authors used 3,692 simple-text phrases and 5,928 SQLI queries to construct the system. MLP modeling is used in the proposed model, which achieves 98% cross-validated accuracy, 98% precision, and 97% recall.

- **M. Arock in 2021** [12]: gives a framework for obtaining the "WHERE" clause of a SQL query, including the usage of a SQL parser, a word tokenizer, and a tagger to get the "WHERE" clause's tagged pattern. Separating legal and injected queries based on their WHERE clause patterns is the key objective. The uniqueness of the labeled patterns is

found when they are retrieved, and this uniqueness helps with model training and query classification using an MLP based on a network of deep neural networks. The dataset development process is incorporated by obtaining queries from a current large-scale human-labeled dataset and constructing on the sql-injection-payload-list challenges. Each of the 500 legitimate searches and 500 injection queries may be found in the dataset. The accuracy of the proposed technique is **94.4**%.

- **M.A Azman el al. in 2021**[13]: They Using data from server access logs, ML was applied to the problem of spotting SQL injection attacks. Researchers propose a three-part system architecture. To separate the recovered log file into training and testing datasets, the recovered log file must first be separated into attribute values that are extracted from log files by searching for distinctive phrases. Following training, to spot injection, the classifier builds a KB (Knowledge Base) that includes both good and bad requests. To identify injections, the proposed system compared log strings for harmful characteristics using Boyer's Moore string matching technique. To get training datasets and run tests, investigators used the Damn Vulnerability Web Application (DVWA). Five minor groups were created from the testing sets. The accuracy score for the proposed model is **93%**.
- **W. Zhang et al. in 2022** [4]: Create a model for an SQLNN deep neural network. Using the "ReLU function" as the basis for a deep neural network architecture with multiple hidden layers, which optimizes the conventional loss function and develops the "Dropout" strategy to expand the applicability of this model, one of the most essential techniques is to use word pauses to transform the data into word vectors, then create a sparse matrix and feed it into the framework to be trained. This research makes use of a publicly available SQL

injection dataset containing a total of 30,919 data items, 11,330 of which are examples of SQL injection phrases and "19589" of which are examples of non-SQL injection statements. Above **96%** accuracy was achieved in the final model.

- **P. Roy et al. in 2022** [14]: They offer a machine learning model for detecting SQL injection attacks utilizing the 3951 different data sets in the Kaggle dataset, together with the "logistic regression, AdaBoost (adaptive boosting), random forest, naive Bayes, and XGBoost (extreme gradient boosting) " classifier algorithms. The authors found that "Naive Bayes", which has a precision of **98.33%,** represents the most effective method for identifying SQL injection payloads. It can distinguish the payload and defend against SQL injection.

- **W.B. Demilie and F.G.Deriba in 2022** [15]: They developed a hybrid system based on "Navies Bayes" (NB), "Decision Trees" (DT), "Support Vector Machines" (SVM), "Random Forests "(RF), "Logistic Regression" (LR), and "Multilayer Perceptron" (MLP) Neural Networks for identifying and blocking SQLI attacks. To train and evaluate the suggested system, Weblogs, cookie files, session logs, and "HTTP" request files provided 54,306 data points for this study. Of the whole dataset, 16,292 were utilized for model testing and 38,014 were used to train the suggested system. 47,343 legitimate inquiries and 6,963 malicious queries were made using the datasets. With better accuracy (**98.87%** and **99.20%**) than previous ML techniques, a hybrid strategy (ANN plus SVM) yields the greatest result.

- **A. Falor et al. 2022** [16] They investigated the many strategies for detecting and avoiding SQL injection threats. One of the main goals of our study was to create a comprehensive dataset that included all potential payload and query types that may be utilized in SQL injection

attacks. The suggested system made use of the "Convolutional Neural Network" (CNN) deep learning method in addition to the decision tree, "support vector machine" (SVM), and "K-nearest neighbor" (KNN) machine learning algorithms. The researchers discovered that CNN exhibits the greatest performance metrics results because of its steady, well-balanced performance with high recall (**96.56%**), precision (**85.67%**), and accuracy **(94.84%).**

- **M. A. Oudah et al. in 2023** [3]: They showed how several NLP approaches may be utilized to extract text characteristics to prepare data for SQL injection detection. Six phases make up the suggested model: To acquire clean data, it is necessary to first extract and decode SQL queries from the user access log file. Next, use the "character level", "word level", and "n-gram level "levels of the "TF-IDF" feature extraction approach to discover the level that is best for detecting SQL injection. The dataset for the ML classifier was then updated with the retrieved features. The dataset, which may be available on Kaggle.com, consists of 37,093 records of online requests gathered from various domains and classified as either benign or malicious. Naive Bayes, a linear classifier, is used as a first step in building the classifier. "a support vector machine", and "extreme gradient boosting" after training and testing. A precision of **99.7%** was achieved by using the SVM model in conjunction with character-level TF-IDF feature extraction .

According to the findings. The following **Table (1.1)** highlights the ML (machine learning) and DL (deep learning) avoidance methods for SQL injection attacks that are cited in the background literature.

Table (1.1) Summarizes the Related Works

Year and Reference	Methodology and Disadvantage	Accuracy	Dataset
2020[10]	• employed deep learning methods and extraction features to identify the payload of the SQL injection attack. • This paper has a flaw in that the suggested model has an overfitting issue.	99.5%	Collected dataset from open-source website Github , and the normal data is from the ISP.
2021[1]	• creates a system for detecting SQL injections using lexical analysis and deep learning algorithms. To compare the trials, a CNN and MLP were used. • This paper has a flaw since it does not concentrate on advanced SQL injection attack techniques like hybrid and second-order injection assaults.	98%	Collect dataset from HTTP request
2021[8]	• Employed feature extraction-based classification methods such as (Naive Bayes, Logistic Regression, Passive Aggressive, SVM, and CNN),	97 % with CNN	dataset from GitHub

	word-level TF-IDF vectors, and NLP (natural language processing). • The proposed model is not real-time capable, and the attack types covered by the database need to be increased.		
2021[9]	• SQL injection attack detection using a hybrid CNN-BiLSTM technique. • The weakness of the presented approach is the need to enhance model performance by feature extraction, tokenization, and stemming from the provided data set.	98%	Collected datasets from various websites
2021[11]	• Employing a deep learning method based on MLP to identify SQL injection attacks. • It does not use feature extraction to investigate how they affect performance metrics.	98%	Public dataset named Lib-injection
2021[12]	• Classified SQL injection attacks using MLP built on a DNN (Deep Neural Network) model. • Other forms of SQL injection attacks are not included in the dataset.	94%	Created a dataset by extracting queries from an existing dataset and SQL-injection-payload list.

2021[13]	• Used String Matching and Boyer's Moore to predication SQL injection attacks. • The model should be evaluated against bigger datasets rather than smaller ones in the tested data sets, and additionally, to reading user access log files, it should also consider real-time internet requests.	93%	Damn Vulnerability Web Application (DVWA) and wrap
2022[4]	• SQLNN injection detection model used IF-TDF feature extraction and deep neural networks algorithms. The performance of SQLNN was compared with that of KNN, decision trees, and LSTM algorithms. • The model used 1D word factors to seed the NN method which reduce the proposed model's performance.	SQLNN 96%	Public available SQL injection attack dataset
2022[14]	• For the identification of SQL injection attacks, machine learning methods were used. Five distinct classification models' performances were compared for effectiveness. • There is a need to expand this research by using deep learning techniques and exploring other SQL injection threats.	98.33% with Naive Bayes	Kaggle SQL Injection Dataset

2022[15]	• Employ machine learning, deep learning, and a hybrid algorithm to detect SQL injection attacks. • The data set needs to be expanded to obtain all SQL injection types.	Hybrid ANN+SVM 98.87% and 99.20%	Dataset amassed from HTTP(S) request files, cookies, and weblogs
2022[16]	• Used KNN, DT, SVM, NB machine learning, and CNN deep learning to detect SQL injection attacks. • It doesn't employ the feature selection approach to investigate how it affects performance metrics. Only payloads and queries that may be exploited in SQL injection attacks were discovered in this investigation.	CNN 94.84%	Kaggle dataset
2023[3]	• Word-level, character level , and N-gram level feature extraction levels have all been implemented. Using Naive Bayes, Linear classified, SVM, and Extreme gradient boosting (EGB) machine-learning classification algorithms. • This work is limited to the detection of traditional SQL injection attacks.	SVM with IF-TDF 99.7%	Kaggle dataset

1.8 Layout of Thesis

The other chapters in this thesis are as follows:

Chapter Two: **Theoretical Background**, describes the methods and techniques used in this thesis.

Chapter Three: **The Proposed Model**, offerings in detail the proposed (DSQLIAM) algorithm used for the detection of SQL injection attacks.

Chapter Four: Experimental **Results and Evaluation** covers the implementation, analysis, and testing outcomes of the suggested model, assesses them, and contrasts the outcomes of the suggested algorithm with outcomes from comparable works.

Chapter Five: findings, Challenges, and Proposals for **Future Work** highlight the thesis's findings and future growth proposals.

الخلاصة

تعد SQL ("لغة الاستعلام الهيكلية") نوعًا من تهديد الأمن السيبراني الذي يمكن أن يسمح هجمات حقن للمهاجمين بالوصول غير المصرح به إلى قواعد البيانات وسرقة المعلومات الحساسة.

تعد هذه الهجمات في الوقت الفعلي ضرورية لمنع انتهاكات البيانات وحماية أمان بيانات . المؤسسات.

واحدة من أكثر الطرق الواعدة للكشف عن هجمات حقن SQL هي من خلال استخدام التعلم العميق. تقدم هذه الأطروحة نظامًا محسنًا يسمى 'DSQLIAM (Deep SQL Injection Attack Model) والذي يكتشف بدقة ويصنف ما إذا كان طلب الإدخال طبيعيًا أو هجوم حقنSQL.

يشتمل النظام على مراحل ، بما في ذلك المعالجة المسبقة لبيانات حقنSQL المدخلات باستخدام الرموز المميزة ، واستخراج الميزات المهمة باستخدام متجه العد وتقنيات معالجة اللغة الطبيعية TF-IDF، وتطبيع البيانات باستخدام مقياس الحد الأدنى ، وأخيراً تصنيف هجمات حقن SQL باستخدام النموذج المقترح (1D-CNN، MLP) تساعد هذه التقنيات في زيادة فعالية ودقة النموذج للتعرف على هجمات حقنSQL وتجنبها.

تم اختبار DSQLIAM المقترح باستخدام عينات من مجموعة بيانات هجوم حقن SQL، والتي تتكون من 30919 عينة مقسمة إلى فئات حقن عادية وSQL تمت مقارنة أداء النظام باستخدام طريقتين مختلفتين لاستخراج الميزات ، وهما Count Vectorization و TF-IDF. أظهرت النتائج أن طريقة استخراج ميزة TF-IDF أنتجت دقة أفضل من Count Vectorization.

بالإضافة إلى ذلك ، تمت مقارنة أداء خوارزميات التصنيف المختلفة ، مع أساليب التعلم العميق التي تفوقت على خوارزميات التعلم الآلي من حيث الدقة. تم تحقيق أفضل أداء باستخدام TF-IDF مع 1D-CNN حيث بلغ معدل الدقة 0.98214 ، ومعدل الحساسية هو 1 ، ومعدل الخصوصية 0.97449 ومعدل الدقة 0.98315 ، ومعدل درجةF1.

علاوة على ذلك ، تم اختبار DSQLIAM المقترحة على قاعدة بيانات ثانية تشتمل على بيانات حقيقية تسمى "مجموعة بيانات حمولة تطبيقات الويب" . تحتوي مجموعة البيانات هذه على 4201 عينة

مصنفة في فئتين: عادي وهجوم SQL . تم تقييم أداء النظام المقترح بناءً على مدى فعالية تصنيف البيانات في الوقت الفعلي. أظهرت النتائج التي تم الحصول عليها أن DSQLIAS المقترح كان قادرً على تحديد هجمات حقن SQL بنجاح وبدقة أكبر مع تجنب مشكلة فرط التجهيز.

أخيرًا ، تمت مقارنة DSQLIAM المقترح مع النتائج ذات الصلة ، وأظهرت النتائج أنه حصل على دقة ممتازة. أنتجت خوارزمية 1D-CNN جنبًا إلى جنب مع تقنية استخراج ميزة TF-IDF أفضل نتائج دقة 98.214 لمجموعة البيانات الأولى و 98.09 لمجموعة البيانات الثانية.