**Chapter 3**

**Conception and Implementation**

**3.1Introduction**

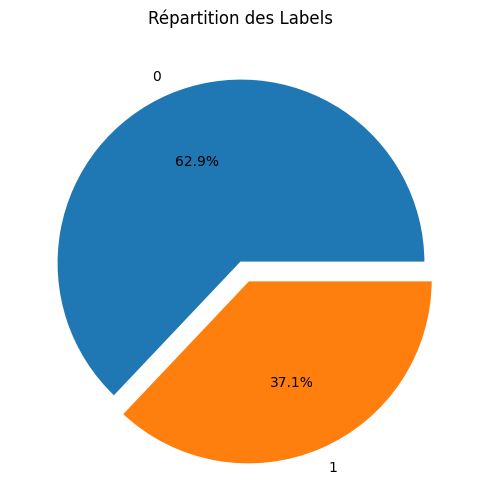
The chapter describes the models' design and implementation for SQL injection detection. Different architectures evaluated in our work include the traditional machine learning, deep learning, and transformer-based approaches. The rationale behind this approach is to assess and compare the models' performances on the same dataset in order to find the one most suitable for accurately detecting SQL injection attempts. The next sections describe the dataset employed, as well as the preprocessing done on it, followed by the design for each of the models and the evaluation strategy considered for training and testing.

**3.2. Dataset**

**3.2.2. Dataset Description**

This study uses two separate datasets for training and testing purposes:

* **Training Dataset (SQLIV3\_cleaned2.csv)**:
  + Contains **30,614 SQL queries**.
  + Each entry is composed of a Sentence (the SQL query) and a Label (0 for benign, 1 for SQL Injection).
  + Label distribution:
    - **Normal (Label = 0)**: 19,268 queries
    - **Malicious (Label = 1)**: 11,346 queries



**(SQLIV3\_cleaned2.csv)**

* **Testing Dataset (sqliv2\_utf8.csv)**:
  + Contains **33,760 SQL queries**.
  + Structured in the same format with Sentence and Label columns.
  + Label distribution:
    - **Normal (Label = 0)**: 22,305 queries
    - **Malicious (Label = 1)**: 11,455 queries

**3.2.3. Data Preprocessing**

Before training, both datasets were cleaned by removing duplicate queries. These queries were then vectorized using appropriate text processing techniques such as **TF-IDF** or **token embeddings**, depending on the model.

**3.3 Models Implemented**

This section on models presents the various models that have been developed for detecting SQL injections. A number of machine learning models and deep learning models were trained and evaluated with varying architecture and hyperparameters. The aim is to see how traditional techniques and modern, state-of-the-art techniques differ in input to achieve results on the same dataset.

For each model, we describe the structure, the main hyperparameters used during training, and the results obtained. These models are as follows:

Support Vector Machine (SVM)

Logistic Regression (LR)

Multilayer Perceptron (MLP)

Simple Neural Network (SN)

Recurrent Neural Network (RNN)

Long Short-Term Memory (LSTM)

BERT Transformer Model (dedicated in Chapter 4)

**3.3.1 Support Vector Machine (SVM)**

We used a Support Vector Machine with a linear kernel (`kernel='linear'`) and regularization parameter C = 1. The input queries were transformed into vectors by applying TF-IDF with 3000 features as the maximum. We dropped duplicates and split the dataset into 80% training and 20% testing sets prior to model training.

Evaluation:

- Accuracy: 98.27%

- Precision: 100%

- Recall: 99.13%

- F1-score: 99.13%

The model performed impressive detection of SQL injection attacks with minimal false positives. The model was accurately able to classify **98.27%** of the malware queries and 99.80% of the clean queries with great generalization and strength.

**3.3.2 Logistic Regression (LR)**

We trained a Logistic Regression classifier over TF-IDF features to detect SQL injection payloads. The classifier was initialized with solver='liblinear' and penalty='l1'. It gave 98.12% accuracy and 99.73% precision on the test set.

|  |
| --- |
| # Train a logistic regression model with L1 regularization  lrc = LogisticRegression(solver='liblinear', penalty='l1')  lrc.fit(X\_train, y\_train)  # Predict labels on the test set  y\_pred = lrc.predict(X\_test) |

On 11,424 malicious payloads (label=1), it detected 95.36% of them with 100% precision and F1-score of 97.63%. On 22,303 benign inputs (label=0), it achieved 99.52% accuracy, 100% precision, and F1-score of 99.76%.

These results suggest that the model is very effective and precise at detecting SQL injection attacks with very minimal false positive.

**3.3.3 Multilayer Perceptron (MLP)**

We trained a Multi-Layer Perceptron (MLP) with three hidden layers (512, 256, 128 units, ReLU activation) and a sigmoid output layer. TF-IDF with 3000 features was used for vectorization. Training was done with SGD (learning rate = 0.01) for 27 epochs using binary cross-entropy. The training achieved **99.44% accuracy** and a final loss of **0.0306**.

Before evaluation, duplicates were removed from the test set to ensure clean input.

|  |
| --- |
| model = Sequential([  Dense(512, input\_dim=3000, activation='relu'),  Dense(256, activation='relu'),  Dense(128, activation='relu'),  Dense(1, activation='sigmoid')  ])  model.compile(loss='binary\_crossentropy', optimizer=SGD(learning\_rate=0.01), metrics=['accuracy'])  history = model.fit(vectorizer.fit\_transform(train['Sentence']).toarray(), train['Label'].values, epochs=27, batch\_size=32) |

|  |
| --- |
| test\_x = vectorizer.transform(test.dropna(subset=['Sentence']).drop\_duplicates(subset='Sentence')['Sentence']).toarray()  predicted\_classes = (model.predict(test\_x) >= 0.5).astype(int) |

Evaluation:

- Accuracy 98.80%

- Precision: 100%

- Recall: 98.80%

- F1-score: 99.37%

**3.3.4 Simple Neural Network (SN)**

we considered a simple neural network model with one output layer and sigmoid activation. Vectorized the input sentences using TF-IDF (3000 features). Trained the model using SGD (learning rate = 0.01) for 27 epochs with binary cross-entropy loss.

|  |
| --- |
| model = Sequential()  model.add(Dense(1, input\_dim=3000, activation='sigmoid'))  model.compile(loss='binary\_crossentropy', optimizer=SGD(learning\_rate=0.01), metrics=['accuracy']) |

During evaluation, duplicates were removed from the test set to ensure clean inputs. The model was first tested on malicious inputs (Label = 1), detecting **9539** SQL injection queries with:

* Accuracy: 83.50%
* Precision: 100.00%
* Recall: 83.50%
* F1-score: 91.01%

It was then tested on benign inputs (Label = 0) with 99.93% accuracy and a recall of 99.93%, showing excellent performance at rejecting false positives.

**3.3.5 Recurrent Neural Network (RNN)**

To discover sequential patterns in SQL injection attempts, we built a Recurrent Neural Network (RNN) upon tokenized and padded input sentences, retaining SQL syntax and casing by turning off character filters and preserving case sensitivity.

The architecture consists of an embedding layer (256 dimensions), followed by two stacked SimpleRNN layers (with 256 and 128 units respectively), a Dense layer (64 units with ReLU activation), and a final sigmoid output layer for binary classification. The model was trained for 20 epochs using the Adam optimizer (learning rate = 0.0001) and binary cross-entropy as the loss function.

|  |
| --- |
| model = Sequential([  Embedding(input\_dim=15000, output\_dim=256, input\_length=max\_len),  SimpleRNN(256, return\_sequences=True),  Dropout(0.3),  SimpleRNN(128),  Dense(64, activation='relu'),  Dense(1, activation='sigmoid')  ]) |

During training, the model achieved a final loss of **0.1444** and an accuracy of **95.83%**.

On the test set, the final accuracy was **95.83%**, based on **11,424 unique SQL injection attempts** after removing duplicates.

**Evaluation on SQL injection queries (Label = 1):**  
 Accuracy: **99.43%**  
Precision: **100.00%**  
 Recall: **99.43%**  
 F1 Score: **99.71%**

**Evaluation on benign queries (Label = 0):**  
 Accuracy: **93.98%**  
 Precision: **100.00%**  
 Recall: **93.98%**  
 F1 Score: **96.90%**

**3.3.6 Long Short-Term Memory (LSTM)**

To further enhance the detection of sequential patterns in SQL injection scenarios, we implemented a Long Short-Term Memory (LSTM) network on tokenized and padded input sentences, allowing SQL syntax to remain intact while preserving case sensitivity by having character filters turned off and case sensitivity on.

The architecture is built up of an embedding layer (256 dimensions), two stacked LSTM layers (the first has 256 units, the second has 128 units), Dense layer (64 units ReLU activation), and a final sigmoid output layer for binary classification. The model was trained for 30 epochs using the Adam optimizer (learning rate = 0.0001) and used binary cross-entropy as our loss function.

|  |
| --- |
| model = Sequential([  Embedding(input\_dim=15000, output\_dim=256, input\_length=max\_len),  LSTM(256, return\_sequences=True),  Dropout(0.3),  LSTM(128),  Dense(64, activation='relu'),  Dense(1, activation='sigmoid')  ]) |

The model achieved a final loss of 0.0481, and accuracy of 99.45% at the end of training.   
  
On the test set, the final accuracy 99.45% based on 11,424 unique SQL injection attempts with duplicates removed.   
  
5Evaluation on SQL Injection Queries (Label = 1)  
  
  
Accuracy: 99.44%   
  
Precision: 100.00%  
  
Recall: 99.44%  
  
F1 Score: 99.72%  
  
  
Evaluation on Normal Queries (Label = 0)  
  
  
Accuracy: 99.46%  
  
Precision: 100.00%  
  
Recall: 99.46%  
  
F1 Score: 99.73%a  
  
This shows that the model is effective in classifying normal queries correctly but with high precision and recall. In a real-world situation, it is important for the system to be able to distinguish between legitimate queries and malicious queries.