

A Comprehensive Exploration and Evaluation of Supervised Machine Learning Algorithms for Detecting SQL Injection Attack on Web Applications

Prof. Nilima D.Bobade¹ and Dr.S.S.Sherekar²

¹Prof. Ram Meghe Institute of Technology & Research Badnera, Amravati, India

²S.G.B.A.U. Amravati, Amravati, India

Abstract—Structure Query Language Injection is one of threats to the web applications. SQL Injection attack compromise confidentiality, integrity and availability of data in the database. By implementing SQL Injection, attacker gain full control on the database of the application, ultimately on overall web application. Web applications that do not appropriately validate the user's input will result in vulnerable to SQL Injection attack. SQL Injection attack occurs when attacker is successful to insert 'malicious query' through the user's input and it executed on the backend database.

This paper describes typical SQL Injection attack, Experimental setup, Machine learning approach for detection of SQL Injection attack, Analysis of various classifiers.

Keywords-SQL Injection Attack, Detection, Analysis.

I. INTRODUCTION

Web applications uses three tier client server architecture consists of three important parts client, server and database. The clients request to the server and receive response from the server. Server is responsible for processing of the client request. Database stores crucial information. SQL Injection is a technique that exploits and compromises the security at third layer i.e at database layer of a web application.

Most usually, any web application is prone to SQL Injection if any of the following vulnerabilities are present in the web application .When user do not apply proper filtration, sanitization and validation techniques on input data from the user.

There are basically seven types of SQL Injection attack: Tautologies, Illegal/logically incorrect queries, Union Query, Stored Procedure, Piggy -Backed Queries, Inference and Alternate Encodings. The Inference attack further classified into Blind Injection and Timing attacks. [1]

II. TYPICAL SQL INJECTION

In SQL Injection attack, the attacker inserts malicious input through input and gain access to the database and take control on database and can damage database in various ways such as retrieval of sensitive data or unauthorized manipulation of database.[2]

In the following typical SQL Injection attack tautological attack at login phase is demonstrated. On the Login page the Username Nilima OR 1=1 is provided. These specially crafted username and password will become a part of dynamic SQL Query and will be fired on a database. A malicious user would like to gain control over database.

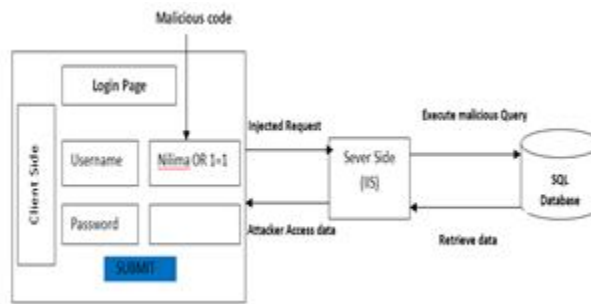


Figure 1: Typical SQL Injection Attack

The main contribution of this paper is as follow

- 1) Experimental setup for detection of SQL Injection attack using machine learning approach.
- 2) Proposed solution based on machine learning algorithm.
- 3) Comparative Analysis of machine learning algorithms to detect SQL Injection attack based on different parameters.

III. RELATED WORK

Morufu Olalere et al. proposed model that uses Naïve Bayes based pattern recognition for detection and categorization of SQL injection attack. Experiment and its result are demonstrated detection and categorization and performance of proposed model compared with performance of other existing techniques.[3]

Ines Jamel et al. presented general idea of SQL Injection attack and its classification. To detect and prevent SQL Injection attack based on ontology and machine learning is proposed.[4]

Saima Saleem et al. proposed detection and prevention of SQL Injection attack using machine learning based model and web server logs. The model discovered specific log is malicious log or normal log and also indicate the attack type. Apache WAMP server, private client server network architecture was deployed for gathering of web server logs. For vulnerable web application DVWA is used. [5]

Mehraj Begum presented extraction of SQL Injection Patterns with the help of existing parsing and tagging techniques. The model trained with Patter based tags using Multilayer Perception that perform classification of query with accuracy of 94.4%. [6]

Yazeed Abdul Malik et al. designed and investigate model to improve the efficiency for detection of SQL Injection attack with the help of machine learning techniques by extraction of semantic features also used tenfold approach for evaluation and validation of the proposed detection model. [7]

Chamundeswari et al. proposed system focus on predicting the SQL Injection attack from the given source at a particular time on a given server with applications deployed on it. JMeter tool is used for prediction. To simulate logs data Apache JMeter is used. Using Apache Jmeter author process, extract features and done classification which is given input to a Model to predict SQL Injection Attack. [8]

Akinsole Jide et al. focused on predictive analytics to detect and prevent web application from various SQL Injection attack types, classes. For improvement of rising security breaches the author addressed static and dynamic analysis approaches using supervised machine learning. WEKA tool is used in the study. [9]

IV. EXPERIMENTAL SETUP

To perform the experiment to detect SQL Injection using machine Learning. We need to install

- 1) Python
- 2) Anaconda
- 3) Jupyter Notebook

V. METHODOLOGY

The main aim of the proposed model is to detect SQL Injection attack. The complete process is performed in four stages.

- 1) The first stage we focus on collecting the dataset that contains proper SQL Injection attack queries.
- 2) The second stage deals with data cleaning.
- 3) Third stage deals with training the model.

4) The fourth stage is application of Machine Learning algorithm

A. Dataset

The most significant part in detecting a SQL Injection attack is collecting meaningful dataset that contains malicious i.e SQL Injection attack queries. The standard SQL injection dataset downloaded from Kaggle. Kaggle is one of the most popular platforms for data scientists and machine learning practitioners to improve their skills, collaborate with others and solve real world data challenges. The SQLiV dataset contains malicious and normal SQL query so that the propose model will properly comprehend and differentiate between normal and malicious query. [10]

Description of dataset

The description of dataset is as follows

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4200 entries, 0 to 4199
Data columns (total 2 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   Sentence    4187 non-null   object
1   Label       4200 non-null   int64
dtypes: int64(1), object(1)
memory usage: 65.8+ KB
```

Figure 2: Description of dataset

Feature Extaction

The dataset consist of two feature 1) Sentence 2) Label. Sentence column will contain malicious and normal SQL Query While Label column will contain 0,1.

B. Data cleaning

For data cleaning we removed duplicate rows and null values from the dataset. After cleaning data in the dataset contain total 4187 queries as follows.

TABLE I. DESCRIPTION OF LABELS

Sr.No	Label	Description	Count	Ratio
1	0	It represents the normal SQL Query	3060	73.08%
2	1	It represents the malicious query	1127	26.92%
Total queries			4187	100%

C. Training & Testing

Break the data into training set and testing set to prepare to train and test the model. For training and testing we used 80% data for training and 20% data for testing.

D. Machine Learning Algorithms

The proposed model used the following machine learning algorithms

- 1) Logistic Regression
- 2) AdaBoost Classifier
- 3) Random Forest Classifier
- 4) Naïve Baye's Classifier
- 5) SVM Classifier

To preserve the experimental aspect of our model, every parameter for the algorithm kept to default.

E. Evaluation Results

Figure 3 shows a sample confusion matrix and classification report of Logistic Regression algorithm used in this research effect.

```

Confusion Matrix:
[[609  4]
 [ 57 170]]
Classification Report:
              precision    recall  f1-score   support

     0       0.91      0.99      0.95      613
     1       0.98      0.75      0.85      227

 accuracy      0.93
 macro avg     0.95
 weighted avg  0.93

```

Figure. 3 confusion Matrix and classification report for Logistic Regression algorithm

Table II shows the confusion matrix that describes the essential metrics to evaluate the machine learning classifiers.

TABLE II: CONFUSION MATRIX

	Predicted as Normal Query	Predicted as Malicious Query
Normal Query	TP	FN
Malicious Query	FP	TN

TP is the True Positive rate represents the number of normal query correctly predicted as normal query. TN is the True Negative rate. It represents the number of malicious query correctly predicted as malicious query. FN is the False Negative rate. It represents the number of incorrectly predicted normal query as malicious query. FP is the False Positive rate. It represents the number of incorrectly predicted malicious query as normal query. Following Table II Summarizes the evaluation results on different parameter used by this machine Learning algorithms.

TABLE III. EVALUATION RESULTS

Algorithm	PARAMETERS			
	Accuracy	f1-Score	Precesion	Recall
Logistic regression	0.92	0.84	0.97	0.74
AdaBoost Classifier	0.9	0.78	0.96	0.66
Random Forest	0.92	0.88	0.78	1
naive Bayes	0.98	0.97	0.94	1
SVM	0.93	0.76	0.97	0.62

F. Parameters for Performance Evaluation

The model was evaluated using Accuracy, F1-score, Precision and Recall parameters and calculated using following formulas.

- 1) Accuracy= $(TP+TN)/(TP+TN+FP+FN)$
- 2) F1-Score= $2*(Precision*Recall)/(Precision + Recall)$
- 3) Precision= $TP/(TP+FP)$
- 4) Recall= $TP/(TP+FN)$

VI. RESULTS AND DISCUSSION

According to the experiments that we conducted, we come to conclusion that the proposed system is efficiently detected SQL Injection attack queries from normal and malicious SQL Injection attack Queries. As the dataset contains all types of SQL Injection attack queries, model is used to detect all types of SQL Injection attack queries efficiently.

To evaluate the performance of our proposed model we applied machine learning algorithms and parameters and results are depicted in the above table II. We separated the results in different figures where in figure every graph

represents different classification metrics such as Accuracy, Precision, Recall ,F1-Score.The results depicted in below figures 4-7.

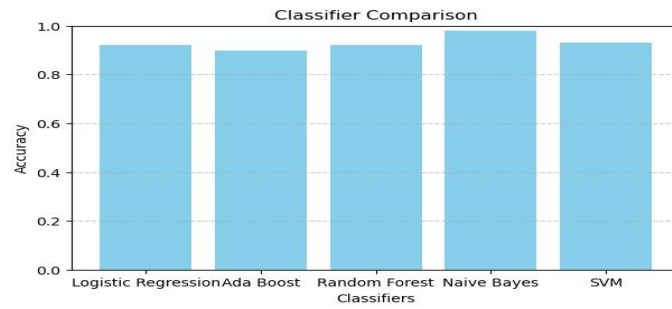


Figure.4 Comparison based on Accuracy parameter

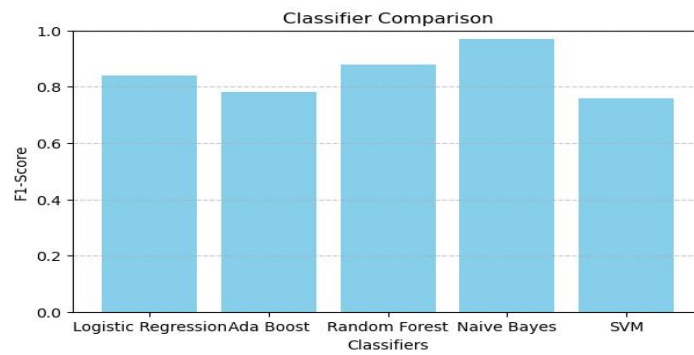


Figure. 5 Comparison based on F1-Score parameter

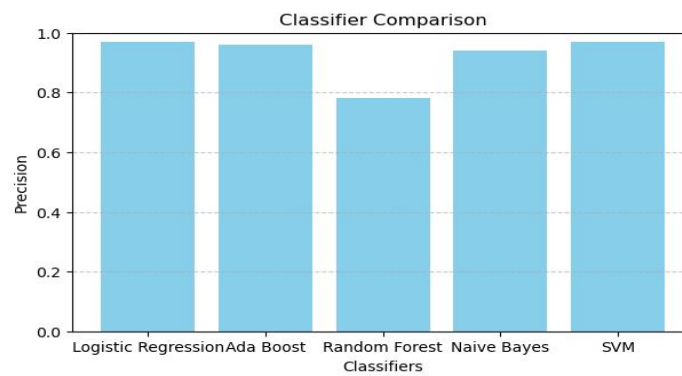


Figure. 6 Comparison based on Precision parameter

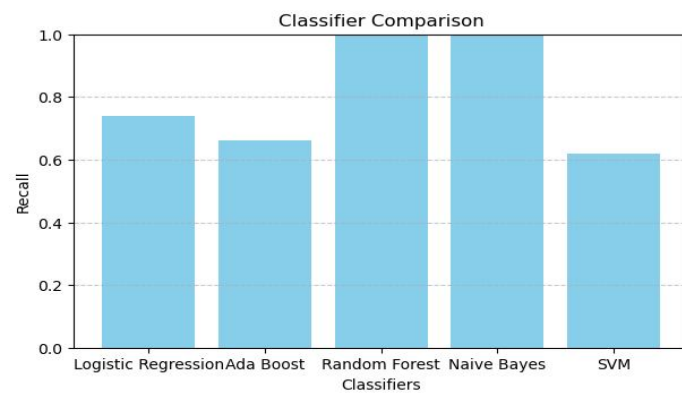


Figure.7 Comparison based on Recall parameter

VII. COMPARATIVE ANALYSIS

The comparative analysis for the research that has been made on SQL Injection attack is depicted in the graph below (Fig.8). The analysis shows that Naïve Bayes is efficient than Logistic Regression, AdaBoost, Random Forest, SVM as its shows highest accuracy (98%), highest F1-Score (97%), Recall is 1 and precision is 94%.

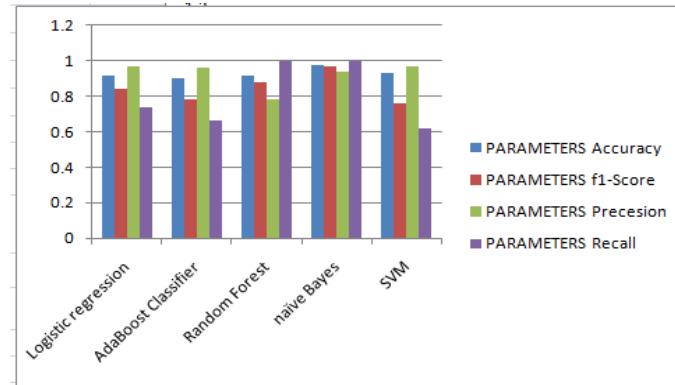


Fig.8 Comparative analysis of LR, AB, RF, NB, SVM classifier on four different parameters

IX. CONCLUSION AND FUTURE WORK

In this research, complete experimental setup to run different machine learning algorithms to detect SQL Injection attack was developed. Research results can be used by research community for working on detection of SQL Injection attack. Accuracy, F1-Score, Precision and Recall parameters of various machine learning algorithm used in research were determined.

In Future, we can also apply techniques to larger dataset and find the effects on performance parameters of various machine learning algorithms. Research has the potential to be extended in future by adding more machine learning algorithms with Bagging, Boosting and Stacking.

REFERENCES

- [1] Zainab S.Alwan,Manal F.Younis, "Detection and Prevention of SQL Injection Attack: A Survey",2017,International Journal of Computer Science and Mobile Computing ,,vol.6,Issue 8,pg.5-17
- [2] Nischay T., Hareesha H, Rajesh V, "Prevention and Detection of SQL Injection Attacks,s",International Journal of Innovative Research in Science,Engineering and Technology ,2018,ISSN(Online):2319-8753 ISSN(Print):2347-6710
- [3] Morufu Olalere,Raji Abdullahi Egidigbo,Ojieniyi Joseph, Ismalia Idris ,Rasheed Gbenga Jimoh , " A Naïve Naves Based Pattern Recognition Model for Detection and Categorization of Structured Query Language Injection Attack",2018,International Journal of Cyber-Security and Digital Forensics(IJCSDf)7(20:189-199,The society of Digital Information and Wireless Communication(SDIWC),ISSN:2305-001
- [4] Ines Jemal,Omar Cheikhrouhou,Habib Hamam and Adel Mahfoudhi, "SQL Injection Attack Detection and Prevention Techniques Using Machine Learning",2020,International Journal of Applied Engineering Research ISSN 0973 4562,volume 15,Number 6,pp 569-580. <http://www.ripublication.com>
- [5] Saima Saleem ,Muhammad Sheeraz,Dr.Muhammad Hanif ,Dr.Umar Farooq, " Web Server Attack Detection using Machine Learning ",2020,IEEE,978-1-7281-6840-1/20/\$31
- [6] Meharaj Begum A,Michael Arock , "",2021 "International conference on computing ,communication, and Intelligent Systems(ICCIS) ",ISBN:978-1-7281-8529-3/21/\$31.00©2021 IEEE
- [7] Yazeed Abdulmalik , "An Improved SQL Injection Attack Detection Model Using Machine Learning Techniques",2021,International Journal of Innovative Computing 11(1) 53-57,ISSN 2180-4370,<https://doi.org/10.11113/ijic.v11n1.300>
- [8] Chamundeswari Arumugam1,Varsha Bhargavi et al., " Prediction of SQL Injection Attacks in Web Applications",2019,©Springer NatureSwitzerland AG2019:ICCSA2019,LNCS11622,pp 496-505,2019,https://doi.org/10.1007/979-3-030-24305-0_37
- [9] Akinsola,Jide ,Awodele et al., "SQL Injection Attacks Predictive Analytics Using Supervised Machine Learning Techniques",International Journal of Computer Applications Technology and Research,Volume 9-Issue 04,139-149,2020,ISSN:2319-8656,www.ijcat.com
- [10] Kaggle,<https://www.kaggle.com/code/princeroy15/sql-injection-detection-by-machine-learning/input?select=sqli.csv>