iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** IV **Month of publication:** April 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.60932

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue IV Apr 2024- Available at www.ijraset.com*

# Applied Machine Learning Predictive Analytics to SQL Injection Attack Detection and Prevention

Bhaskar P[1], Shashikala K[2], Susma Swaraj V[3], Gayathri A[4], Madhavi G[5], Venkata Juhitha C[6]
[1]*Assistant professor,* [2, 3, 4, 5, 6]*Student, Department of Computer Science Engineering, Santhiram Engineering College, Andhra Pradesh, India*

***Abstract:*** ***The back-end database is pivotal to the storage of the massive size of big data Internet exchanges stemming from cloud-hosted web applications to Internet of Things (IoT) smart devices. Structured Query Language (SQL) Injection Attack (SQLIA) remains an intruder's exploit of choice on vulnerable web applications to pilfer confidential data from the database with potentially damaging consequences. The existing solutions of mostly signature approaches were all before the recent challenges of big data mining and at such lacks the functionality and ability to cope with new signatures concealed in web requests. An alternative Machine Learning (ML) predictive analytics provides a functional and scalable mining to big data in detection and prevention of SQLIA. Unfortunately, lack of availability of readymade robust corpus or data set with patterns and historical data items to train a classifier are issues well known in SQLIA research. In this project, we explore the generation of data set containing extraction from known attack patterns including SQL tokens and symbols present at injection points. Also, as a test case, we build a web application that expects dictionary word list as vector variables to demonstrate massive quantities of learning data. The data set is pre-processed, labelled and feature hashing for supervised learning. The trained classifier to be deployed as a web service that is consumed in a custom dot NET application implementing a web proxy Application Programming Interface (API) to intercept and accurately predict SQLIA in web requests thereby preventing malicious web requests from reaching the protected back-end database. This project demonstrates a full proof of concept implementation of an ML predictive analytics and deployment of resultant web service that accurately predicts and prevents SQLIA with empirical evaluations presented in Confusion Matrix (CM) and Receiver Operating Curve (ROC).***
***Keywords: SQLIA, SQLIA analytics, SQL Injection, SQLIA big data, SQLIA hashing.***

## I. INTRODUCTION

SQL injection is an attack technique that exploits a security vulnerability occurring in the database layer of an application. Hackers use injections to obtain unauthorized access to the underlying data, structure, and DBMS. By an SQL injection attacker can embed a malicious code in a poorly-designed application and then passed to the backend database. The malicious data then produces database query results or actions that should never have been executed. By using an SQL Injection vulnerability, given the right circumstances, an attacker can use it to bypass a web application's authentication and authorization mechanisms and retrieve the contents of an entire database. SQL Injection can also be used to add, modify and delete records in a database, affecting data integrity. To such an extent, SQL Injection can provide an attacker with unauthorized access to sensitive data. SQL injection is a code injection technique, used to attack data-driven applications, in which malicious SQL statements are inserted into an entry field for execution (e.g. to dump the database contents to the attacker). SQL injection must exploit a security vulnerability in an application's software.

## II. LITERATURE SURVEY

### *A. Enhancing SVM-Based SQLIA Prediction with Comprehensive Data Sets and Pre-Processing Techniques*

A critical aspect of improving SVM-based SQLIA prediction lies in the quality of data sets and the effectiveness of text pre-processing techniques. This section delves into strategies for generating robust training data, incorporating diverse patterns, and optimizing text pre-processing to enhance the accuracy of SVM classifiers in predicting SQLIA.Data augmentation techniques can further enrich the training data for SVM-based SQLIA prediction. Methods such as oversampling minority classes, synthetic data generation through techniques like SMOTE (Synthetic Minority Over-sampling Technique).

### *B. Challenges in Data Engineering for SVM-Based SQLIA Mitigation*

One of the primary hurdles in utilizing Support Vector Machine (SVM) machine learning for SQL Injection Attack (SQLIA) mitigation lies in the realm of data engineering.

©IJRASET: All Rights are Reserved 

While SVM holds promise for bolstering security measures, its effectiveness is heavily contingent upon the quality of data preprocessing and feature extraction. Existing SVM-based approaches frequently encounter shortcomings in adequately processing textual data, a critical aspect in detecting SQLIA. The lack of comprehensive text preprocessing techniques often leads to the inability to accurately capture the nuanced patterns indicative of SQL injection attempts. Consequently, these inadequacies undermine the efficacy of SVM classifiers in combatting SQLIA, perpetuating the vulnerability of web applications to such attacks. Addressing these challenges necessitates a concerted effort to refine data engineering practices, with a focus on robust text preprocessing methodologies tailored specifically to the intricacies of SQL injection detection. By enhancing the quality of input data and optimizing text preprocessing techniques, SVM-based approaches can be better equipped to identify and mitigate SQLIA, thereby fortifying the security posture of web applications against malicious exploitation.

### *C. Gap in ML Application for Predicting SQLIA in Big Data Contexts*

To date, there has been a significant oversight in discussing the application of machine learning (ML) for predicting SQLIA within the realm of big data. The focus on patterns and text pre-processing within the Multi-Aspect Multi-Layer (MAML) architecture remains unexplored territory, despite the potential for significant improvements in prediction accuracy.

## III. EXISTING SYSTEM

Existing systems for the SQL language syntax closely resembles plain English and the SQLIA keywords are also in plain text. Therefore, the SQLIA problem in a big data context is a plausible candidate for predictive analytics of a supervised learning model trained via both known historical attack signatures and safe web requests patterns. The attack signatures at injection points will contain patterns of SQL tokens and symbols as SQLIA positive while valid web requests would take the form of expected data from the application. In this project, we build a predictive analytics web application with quantities of learning data to train a classifier. The learning data are labelled vector matrix, or features of both patterns of dictionary word list (SQLIA negative) and SQL tokens (SQLIA positive).

## IV. PROPOSED SYSTEM

In proposed system, we make a prophetic analytics web operation with amounts of learning data to train a classifier. The literacy data are labelled vector matrix, or features of both patterns of dictionary word list (SQLIA negative) and SQL commemoratives (SQLIA positive). The beneficials this project makes give a representative data set that suffer point mincing to train a supervised literacy model enforcing Support Vector Machine (SVM) algorithm that directly predicts SQLIA thereby precluding vicious web requests from reaching the target back- end database. It also offers a environment of SQLIA discovery the big data internet. Also, this project presents a evidence of conception of a working prototype using ML algorithms of Two- Class Support Vector Machine (TCSVM) enforced on Microsoft Azure Machine Learning (MAML) to prognosticate SQLIA. This methodology also forms the subject of the empirical evaluation in Receiver Operating Curve (ROC).

## V. SYSTEM ARCHITECTURE

The trained model exposed as a web service. The web service is called in a custom built dot NET application for this research named NETSQLIA for an ongoing SQLIA detection and prevention. Critical to the deployment in every new domain, the administrator or system expert need to feed the data engineering or text pre-processing module with a new rule that matches the patterns present in the new data set which triggers the retraining of the classifier to adapt to a new environment

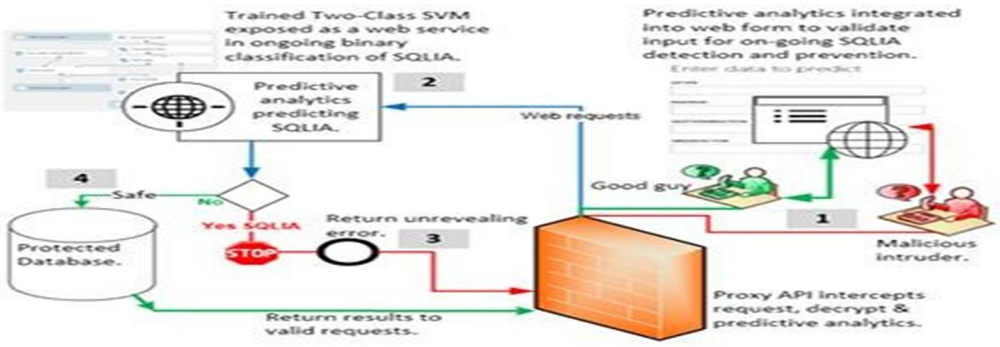



Figure 1:A custom application is consuming the trained SVM web service for ongoing SQLIA detection and prevention

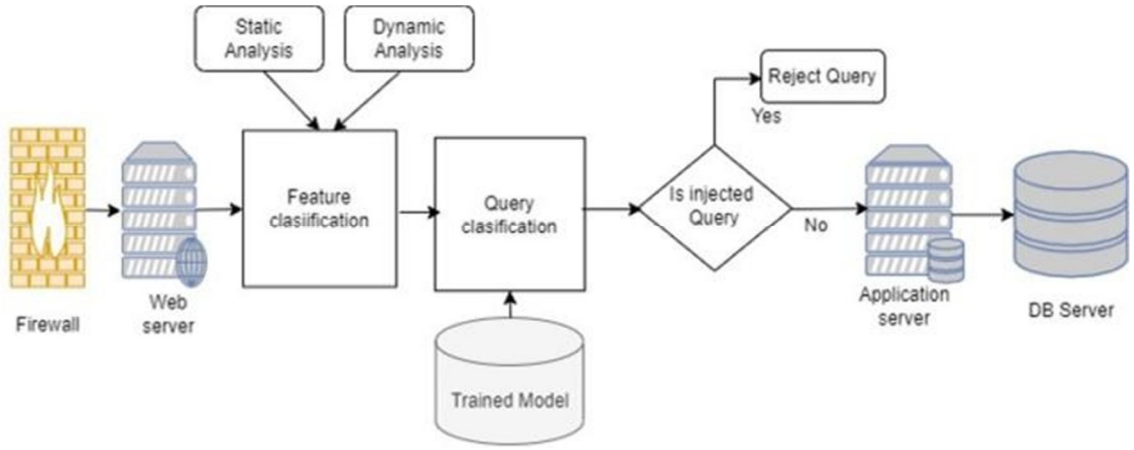
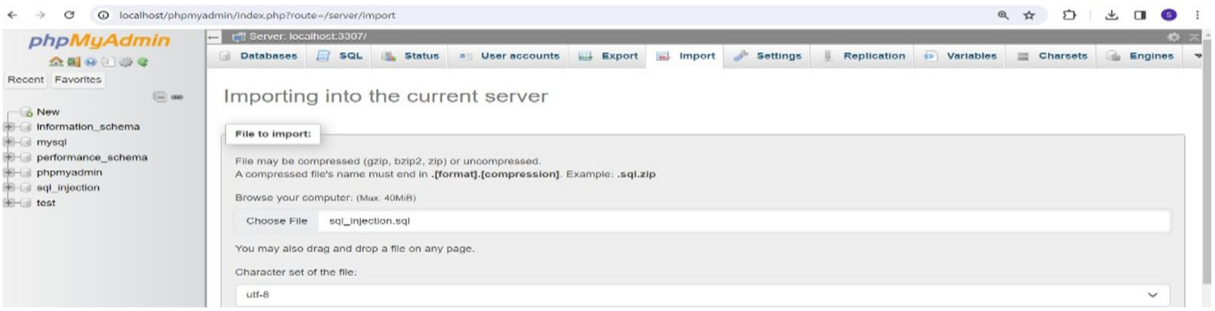



Figure 2: System Architecture

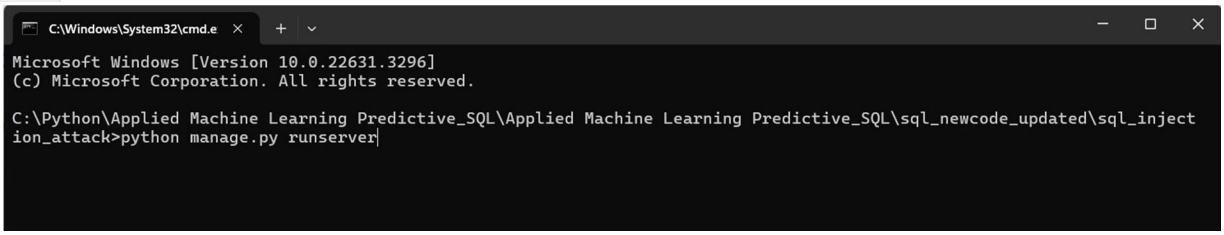
## VI. STEPS TO IMPLEMENT PROPOSED MODEL

1) *Test Preprocessing:* This stage involves R Scripting and regular expression pattern matching applying the procedure integrated to train a model for ongoing detection and prevention. In a real domain application, the data set items are expanded with patterns of both valid requests and bad requests. There were 362,603-row items after text pre-processing of parsing data set for: patterns, duplicates, normalized to lower cases and the removal of the missing words. The data set is sampled as to provide an even distribution of row items (records). The imbalanced data set (majority negatives over positives) were corrected with Synthetic Minority Over-Sampling Technique (SMOTE) to have a dataset items of 725206 split equally, 362,603-row items of attack/respondent (positives) and 362,603-row items of non-attack/non-respondent (negatives). These actions improve both the trained model recall and precision.
2) *SQL injection Point:* A web proxy API is the most suitable to intercept requests originating from any injection mechanisms. Injection mechanisms can originate from any Web page forms e.g. login screen; second-order injection by concealing a Trojan horse for the attack at a later date; exploiting web- enabled server variables to gain access to the back-end database; and, through cookies that have stored state information used to obtain unauthorized access to the back-end database. SQLIA types are techniques an intruder would employ at injection points in any combination to carry out an attack that includes: Tautology; Invalid/Logical Incorrect; Union; Piggybacked; Store procedure; Time-based; and, Alternate encoding obfuscation. SQLIA types provide an extract for the SQLIA positive in data set items during labelling.
3) *Proxy Filters:* This method intercepts web requests at a proxy for SQLIA detection and prevention having the advantage of being able to decrypt obfuscated internet traffic for thorough analysis. We propose a SQL parsing tree which uses a combination of proxy and SQL parser tree for SQL syntax sequence alignment. The model proposed in this paper uses proxy API to backhaul web requests for predictive analytics of incoming web requests for SQLIA negatives and positives.
4) *Classifying Attacks:* Here, we compare the classification performance of SVM with other popular machine learning algorithms. We have selected several popular classification algorithms. For all algorithms, we attempt to use multiple sets of parameters to maximize the performance of each algorithm. Using SVM algorithms classification for malware bag-of-words weightage.

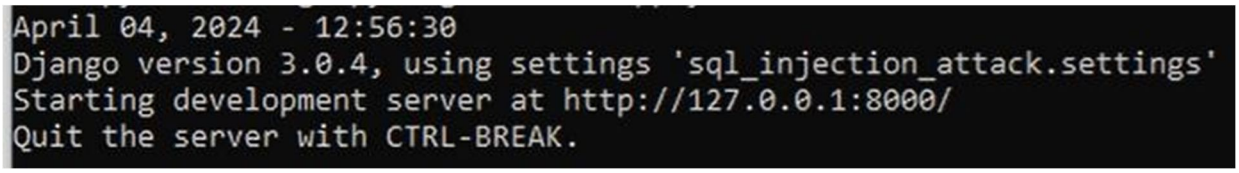
## VII. EXPERIMENTAL RESULT



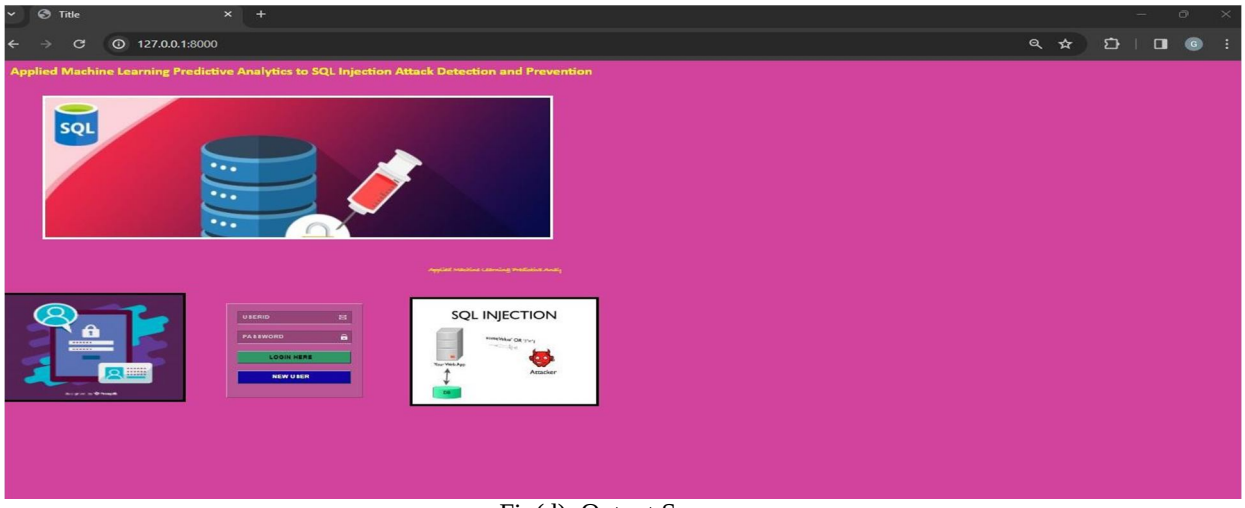

Fig(a): Import the SQL Source File



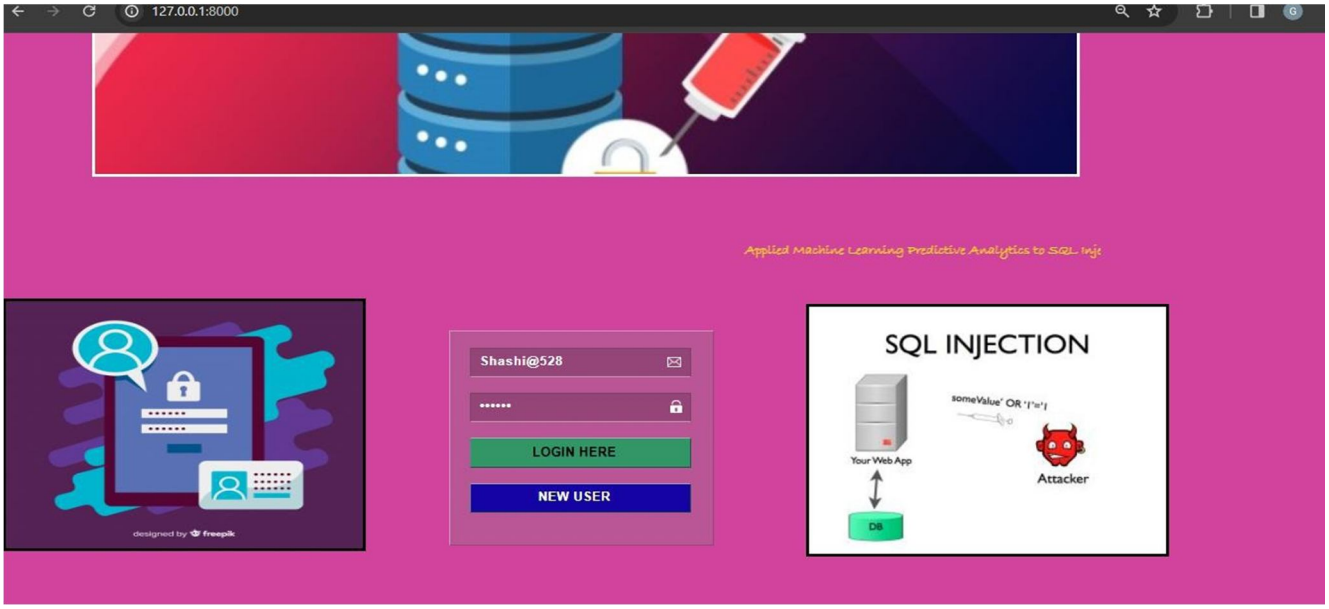

Fig(b): Run Server



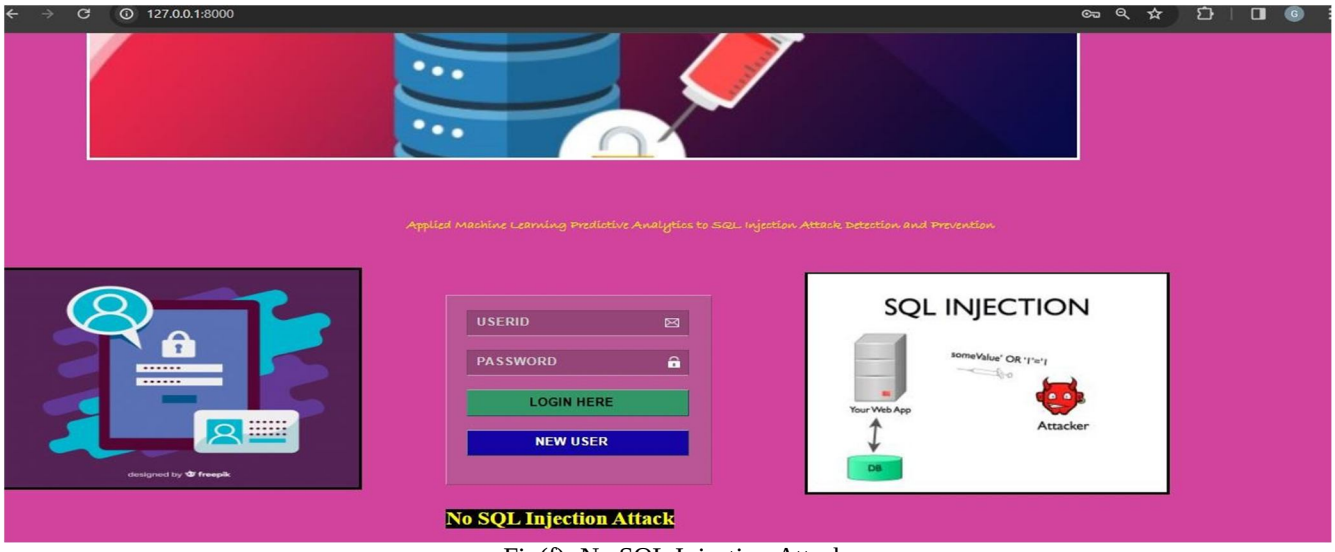

Fig(c): Server URL Generation



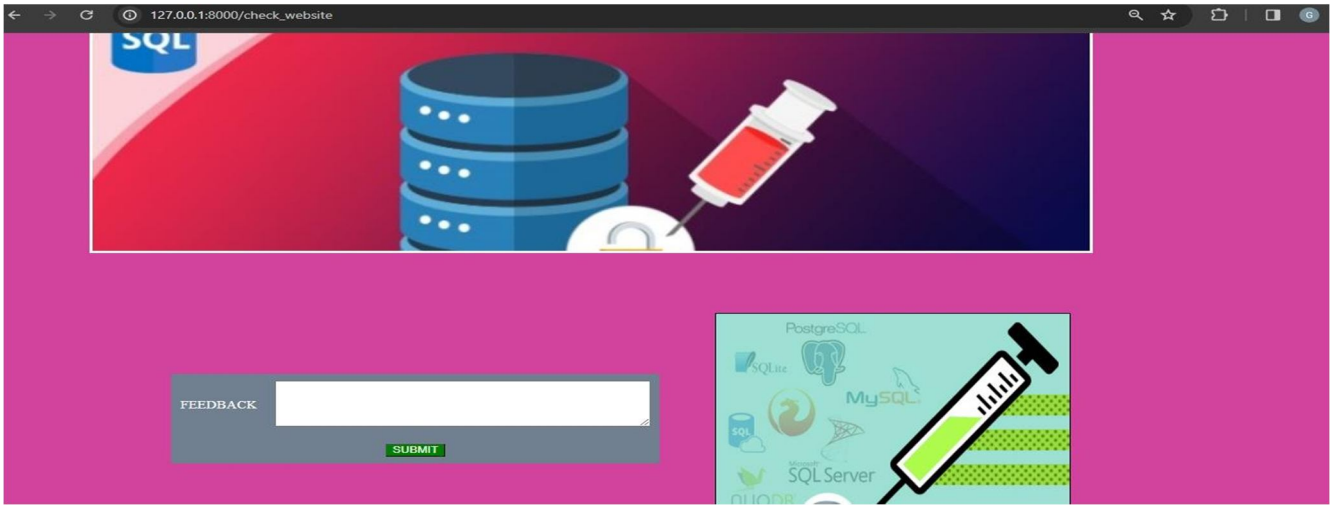

Fig(d): Output Screen



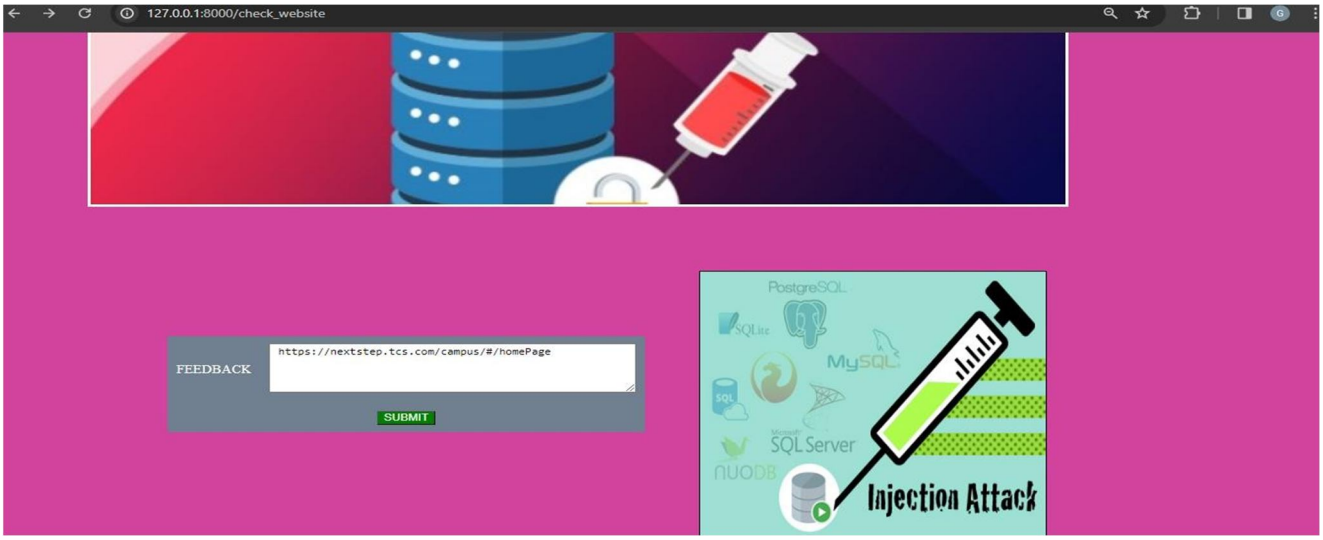

Fig(e): User Login Page




Fig(f): No SQL Injection Attack




Fig(g): Check Website




Fig(h): Choose Website




Fig(i): Feedback about Website

## VIII. CONCLUSION

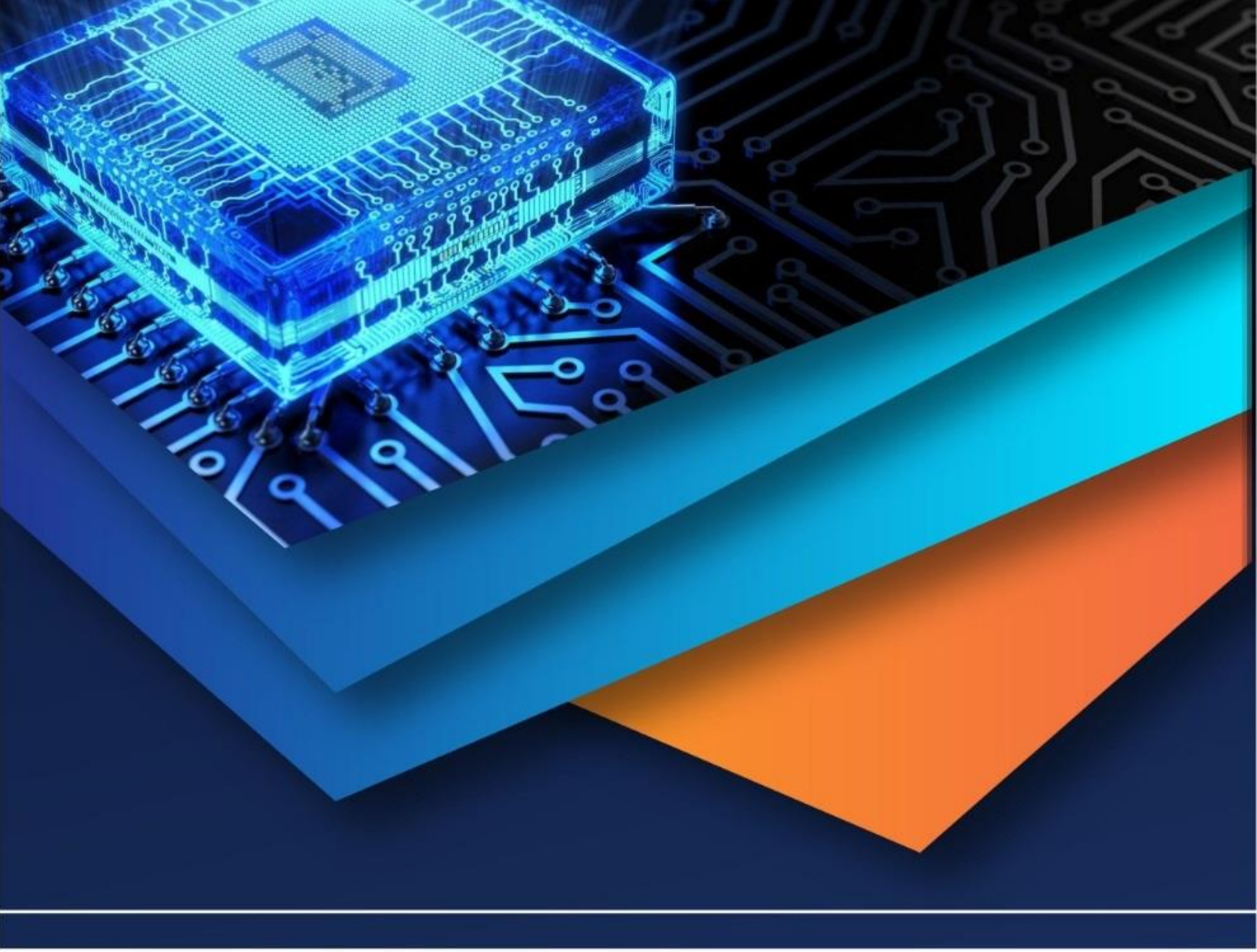
In this Project we demonstrated that applied predictive analytics to SQLIA detection and prevention in big data context with an excellent result that is empirically evaluated in the confusion matrix and the ROC graph. In benchmarking this project against existing works, the methodology proposed here is functional in a big data context which is lacking in existing works before now on SQLIA to our knowledge. Future work involves employing multi-class classifier to identify and group the different SQLIA types as they are predicted.

## IX. ACKNOWLEDGEMENT

We are deeply thankful to Mr..P.Bhaskar for his support and encouragement. He provided us with the necessary resources and environment to pursue academic excellence and we are able to complete this research under his mentorship successfully.

## REFERENCES

[1] Prediction Of Covid-19 Infection Based on Lifestyle Habits Employing Random Forest Algorithm FS Mahammad, P Bhaskar, A Prudvi, NY Reddy, PJ Reddy journal of algebraic statistics 13 (3), 40-45.
[2] Machine Learning Based Predictive Model for Closed Loop Air Filtering SystemP Bhaskar, FS Mahammad, AH Kumar, DR Kumar, SMA Khadar, ... Journal of Algebraic Statistics 13 (3), 609-616
[3] Devi, M. S., Mahammad, F. S., Bhavana, D., Sukanya, D., Thanusha, T. S., Chandrakala, M., & Swathi, P. V. (2022).” Machine Learning Based Classification and Clustering Analysis of Efficiency of Exercise Against Covid-19 Infection.” Journal of Algebraic Statistics, 13(3), 112-117.
[4] Devi, M. M. S., & Gangadhar, M. Y. (2012).” A comparative Study of Classification Algorithm for Printed Telugu Character Recognition.” International Journal of Electronics Communication and Computer Engineering, 3(3), 633-641.
[5] Devi, M. S., Meghana, A. I., Susmitha, M., Mounika, G., Vineela, G., & Padmavathi, M. MISSING CHILD IDENTIFICATION SYSTEM USING DEEP LEARNING.
[6] Kumar, M. S., Harika, A., Sushama, C., & Neelima, P. (2022). Automated Extraction of Non-Functional Requirements From Text Files: A Supervised Learning Approach. Handbook of Intelligent Computing and Optimization for Sustainable Development, 149-170.
[7] Devi, M. S., Poojitha, M., Sucharitha, R., Keerthi, K., Manideepika, P., & Vasudha, C. Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language.
[8] B.Krishna Naga Deepthi, Dr.M.V.Subramanyam,” Analysis And Optimization Of Power And Area Of Domino Full Adder And Its Applications”, Iosr Journal Of Electronics And Communication Engineering, Vol.10,No.3,Pp.55-63,2015.
[9] Y.Murali Mohan Babu, Dr.M.V.Subramanyam,M.N. Giri Prasad,” A New Approach For Sar Image Denoising”, International Journal Of Electrical And Computer Engineering, Vol.5,No.5,Pp.984-991,2015. (Scopus Indexed)
[10] Ch.Nagaraju, Dr.Anil Kumar Sharma, Dr.M.V.Subramanyam,” A Review On Ber Performance Analysis And Papr Mitigation In Mimo Ofdm Systems”, International Journal Of Engineering Technology And Computer Research, Vol.3,No.3,Pp.237-238, June, 2015.
[11] D.Lakshmaiah, Dr.M.Subramanyam, Dr.K.Satya Prasad,” Design Of Low Power 4- Bit Cmos Braun Multiplier Based On Threshold Voltage Techniques”, Global JOURNAL OF RESEARCH IN ENGINEERING, VOL.14(9),PP.1125-1131,2014.
[12] R Sumalatha, Dr.M.Subramanyam, “Image Denoising Using Spatial Adaptive Mask Filter”, Ieee International Conference On Electrical, Electronics, Signals, Communication &Amp; Optimization (Eesco-2015), Organized Byvignans Institute Of Information Technology, Vishakapatnam, 24 Th To 26th January 2015. (Scopus Indexed)
[13] P.Balamurali Krishna, Dr.M.V.Subramanyam, Dr.K.Satya Prasad, “Hybrid Genetic Optimization To Mitigate Starvation In Wireless Mesh Networks”, Indian Journal Of Science And Technology,Vol.8,No.23,2015. (Scopus Indexed)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)