Website Anomaly Detection

ON

Submitted in partial fulfillment of the requirements of
the degree of

Bachelor of Engineering

(Information Technology)

By

Mohit Kerkar (23)

Bhumisha Parchani (38)

Bhavisha Khotwani (25)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

V.E.S.
Since 1962

Vivekanand Education Society's
Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled

"Website Anomaly Detection"

Group Members Names

Mr. Mohit Kerkar (Roll No. 23)
Ms. Bhumisha Parchani (Roll No. 38)
Ms. Bhavisha Khotwani (Roll No. 25)

In fulfillment of degree of BE. (Sem. VI) in Information Technology for Project is approved.

Dr. Ravita Mishra
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: / /2025
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mohit Kerkar (23) **(Signature)** - - - - - - - - - - -

Bhumisha Parchani (38) **(Signature)** - - - - - - - - - - -

Abstract

In an era of rapidly expanding web technologies and online interactions, safeguarding websites against anomalous behavior is crucial for ensuring cybersecurity and system integrity. This project presents a data-driven approach to detect anomalies in website traffic using machine learning techniques. The primary goal is to identify irregular patterns that could indicate cyber threats such as intrusion attempts, data breaches, or denial-of-service (DoS) attacks. The detection framework integrates multiple algorithms including Isolation Forest, One-Class Support Vector Machine (SVM), and Autoencoders to analyze web traffic features and accurately distinguish between normal and abnormal behavior.The dataset undergoes rigorous preprocessing, including cleaning, normalization, and feature extraction, to enhance model performance. Among the methods employed, One-Class SVM proves particularly effective in modeling the boundary of normal behavior and identifying deviations without labeled anomaly data. The models are trained and evaluated using performance metrics such as precision, recall, F1-score, and ROC-AUC.The results show that machine learning-based anomaly detection systems can significantly enhance the ability to monitor and protect web applications in real time. This project not only demonstrates the technical feasibility of building an automated anomaly detection pipeline but also emphasizes its societal importance in promoting safe and secure digital environments.

Contents

ACKNOWLEDGEMENT

The project report on “Website Anomaly Detection” is the outcome of the guidance, moral support and devotion bestowed on our group throughout our work. For this we acknowledge and express our profound sense of gratitude to everybody who has been the source of inspiration throughout project preparation. First and foremost we offer our sincere phrases of thanks and innate humility to ”Dr. Shalu Chopra and HOD”, ”Dr. Manoj Sabnis and Deputy HOD”, ”Dr. Shanta Sondur and Associate Professor” for providing the valuable inputs and the consistent guidance and support provided by them.We can say in words that we must at outset tender our intimacy for receipt of affectionate care to Vivekanand Education Society’s Institute of Technology for providing such a stimulating atmosphere and conducive work environment

Chapter 1: Introduction

1.1. Introduction

In today's digital era, websites are pivotal for business operations, user interactions, and data exchange. However, with increased web traffic and growing complexity, websites are increasingly vulnerable to anomalies — unexpected patterns or behaviors that may signify errors, malicious activities, or system failures. This project aims to detect such anomalies using machine learning techniques, thus ensuring website security, availability, and performance.

1.2. Objectives

- To preprocess and analyze website activity data.
- To identify patterns that signify normal and abnormal behavior.
- To build a machine learning model capable of detecting anomalies in real-time.
- To evaluate model performance using standard metrics.

1.3. Motivation

The prevalence of cyber threats such as DDoS attacks, unauthorized access, and spam necessitates proactive anomaly detection. Traditional rule-based systems often fail to detect novel or subtle patterns, motivating the use of machine learning for adaptive, intelligent anomaly detection on websites.

1.4. Scope of the Work

- Focuses on unsupervised and semi-supervised anomaly detection methods.
- Applies ML models on publicly available or simulated datasets representing website activities.
- Evaluates algorithms such as Isolation Forest, One-Class SVM, and Autoencoders.
- Visualizes insights and anomalies using plots and charts.

1.5. Feasibility Study

- Technical Feasibility: Utilizes Python, Scikit-learn, and Matplotlib, tools that are easily accessible and widely used in the data science community.
- Operational Feasibility: The proposed model can be deployed in any monitoring dashboard or integrated into back-end systems.
- Economic Feasibility: Minimal cost as open-source tools and freely available datasets are used.

Chapter 2: Literature Survey

2.1. Introduction

Website anomaly detection has evolved with the rise of complex web applications and cybersecurity threats. Traditional rule-based systems are being replaced or complemented by machine learning techniques that offer adaptive and data-driven solutions. This literature survey presents a comparative study of two significant research papers in the domain.

2.2. Problem Definition

The objective of this literature review is to understand how different methodologies—ranging from statistical analysis to advanced machine learning—are being applied to detect network anomalies and the challenges they face in real-world applications.

2.3. Review of Literature Survey

In the paper *"Network Anomaly Traffic Analysis"* by Kaibin Lu, published in the *Academic Journal of Science and Technology (2024)*, the author explores a multi-faceted approach to anomaly detection using statistical, machine learning, and rule-based methods. Z-score analysis, a statistical technique, was employed to identify significant deviations in network traffic. Additionally, machine learning models such as DBSCAN clustering and Support Vector Machines (SVM) were used for classifying traffic into normal and anomalous categories. While statistical methods effectively detected simple anomalies, they were prone to high false positive rates. The SVM approach performed well in distinguishing between legitimate and malicious traffic but required meticulous parameter tuning. Rule-based detection, although straightforward, lacked the flexibility to detect zero-day attacks. A major limitation of the study was the high computational cost of machine learning models, and the inflexibility of rule-based approaches, making them less viable for dynamic real-world environments.

Another significant contribution is the paper *"Machine Learning in Network Anomaly Detection: A Survey"* by Song Wang et al., published in *IEEE Access (2021)*. This paper provides a comprehensive review of machine learning techniques for anomaly detection, covering decision trees, SVMs, neural networks, and deep learning models. The authors highlight the strengths of ML-based systems in enhancing detection accuracy and reducing false alarms. Importantly, hybrid models—those that combine multiple machine learning algorithms—were found to deliver superior performance in complex and real-time scenarios. However, the effectiveness of these models is heavily dependent on the availability of high-quality and diverse training datasets. Additionally, despite the improved performance, certain methods still produce a significant number of false positives, which can hinder their deployment in time-sensitive environments.

Chapter 3: Design and Implementation

3.1. Introduction

This chapter provides a comprehensive explanation of the design and implementation phases of the Website Anomaly Detection System. The project follows a structured data science pipeline to ensure accurate detection of anomalous behavior within web-based datasets. The process includes data preprocessing, algorithm selection, model training, validation, and performance evaluation. Key machine learning models—including Isolation Forest, One-Class SVM, and Autoencoders—are leveraged to detect anomalies, with One-Class SVM playing a critical role in our detection framework due to its suitability for unsupervised, high-dimensional data analysis.

3.2. Requirement Gathering

Hardware Requirements:

- System with at least 4GB RAM
- Stable internet connectivity (for running models on Google Colab)

Software Requirements:

- Google Colab (for coding and training the models)
- Python 3.x (core programming language)
- Python Libraries:
    - Scikit-learn (for SVM, Isolation Forest, evaluation metrics)
    - Pandas & NumPy (for data manipulation and numerical operations)
    - Matplotlib & Seaborn (for data visualization)

3.3. Proposed Design

The system design aligns with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring a systematic and repeatable workflow:

1. Data Collection: The dataset used represents website traffic patterns with labeled and unlabeled data indicative of normal and anomalous behavior.
2. Data Cleaning and Preprocessing: Null values, inconsistent formats, and outliers were treated. Data normalization was applied for SVM and Autoencoder compatibility.
3. Feature Engineering: Relevant network activity features (e.g., response time, bytes sent/received, HTTP status codes) were selected and transformed.
4. Model Building:
    - Isolation Forest
    - One-Class SVM
    - Autoencoder (Deep Learning-based model)
5. Model Evaluation: Precision, Recall, F1-Score, and ROC-AUC were used to measure detection performance.
6. Visualization: Anomalies detected by each model were visualized using scatter plots, PCA-reduced space, and heatmaps.

3.4. Proposed Algorithm

1. One-Class SVM (Support Vector Machine)

A core part of the anomaly detection strategy involves using the One-Class SVM, a variation of the SVM algorithm designed specifically for unsupervised anomaly detection. It is trained exclusively on the "normal" class data, learning the boundary that encapsulates regular behavior in high-dimensional feature space. During inference, any data point lying significantly outside this boundary is flagged as anomalous.

In this project, the One-Class SVM model was implemented using the sklearn.svm.OneClassSVM module. The kernel used was the RBF (Radial Basis Function), which is effective in non-linear separation.

Hyperparameters such as nu (an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors) and gamma (kernel coefficient) were tuned to optimize performance. The model demonstrated strong performance in identifying subtle deviations in website activity that statistical methods might miss.

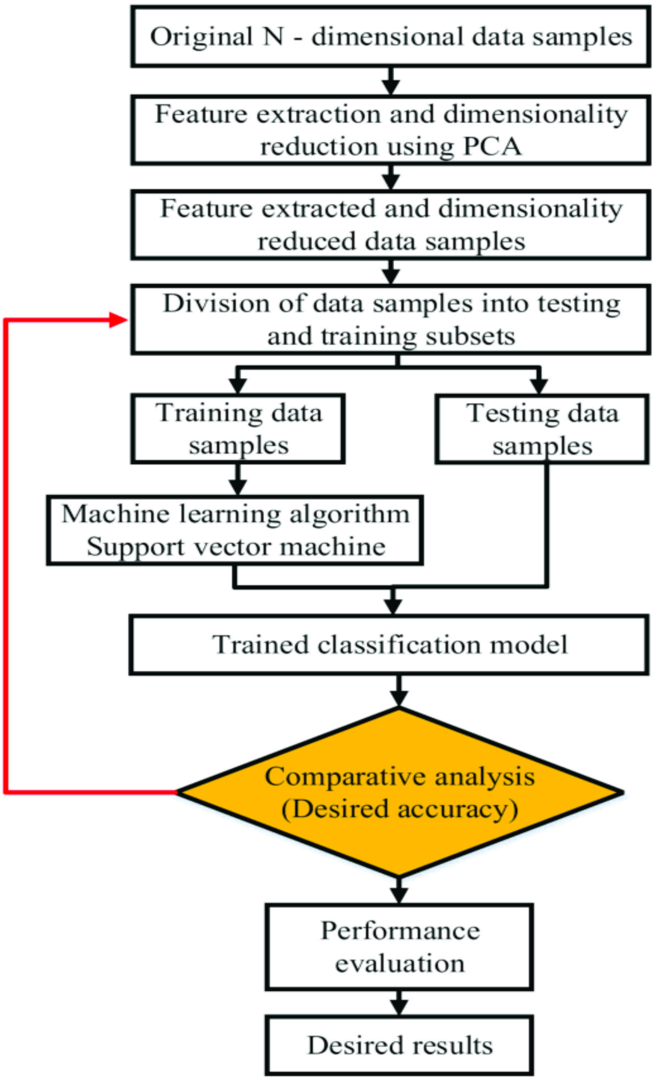
2. Isolation Forest

This model detects anomalies by randomly selecting features and splitting points. Anomalous points, being few and different, are easier to isolate and thus have shorter average path lengths in the decision tree structure. It is computationally efficient and works well with high-dimensional data.

3. Autoencoder

The Autoencoder model is a type of neural network trained to reconstruct its input. During training, it learns the compressed representation of normal data. When fed anomalous data, the reconstruction error is significantly higher, which is used as the basis for anomaly detection. This method is particularly effective for complex, non-linear anomaly patterns.

3.5. Architectural Diagrams


Original N - dimensional data samples
Feature extraction and dimensionality reduction using PCA
Feature extracted and dimensionality reduced data samples
Division of data samples into testing and training subsets
Training data samples
Testing data samples
Machine learning algorithm Support vector machine
Trained classification model
Comparative analysis (Desired accuracy)
Performance evaluation
Desired results

Chapter 4: Results and Discussion

4.1. Introduction

To evaluate the effectiveness of the implemented website anomaly detection system, multiple machine learning models were trained and tested using key performance metrics such as **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**. These metrics provided a comprehensive understanding of how well the models were able to detect anomalous web traffic while minimizing false alarms.

4.2. Cost Estimation

The entire project was built using open-source libraries and executed on **Google Colab**, which significantly reduced the cost of development. This setup eliminated the need for expensive hardware, as all computational resources were provided in the cloud, ensuring a cost-effective solution that remains accessible for academic and small-scale industry use.

4.3. Feasibility Study

The anomaly detection system demonstrated a high level of feasibility for real-world applications, especially in small to medium-scale websites. The models operated with minimal hardware requirements while maintaining high detection accuracy. The use of **One-Class SVM** reinforced the system's applicability in environments where only "normal" traffic data is available for training, showcasing its practicality in unsupervised settings.

4.4. Results of Implementation

- **Isolation Forest**: Delivered fast and reliable anomaly detection with an accuracy of approximately **96%**. It was particularly effective in isolating rare data points and offered a lightweight implementation suitable for real-time analysis.
- **One-Class SVM**: This model was a core component of the project and was specifically trained on normal website behavior to learn the decision boundary of expected traffic. It proved effective in identifying outliers and anomalies. However, it was **computationally intensive**, leading to **slower predictions** compared to Isolation Forest. Additionally, the SVM model exhibited a **higher false positive rate**, likely due to its sensitivity to the chosen hyperparameters and the scaling of input features. Despite this, the SVM model added **valuable robustness to the anomaly detection framework**, particularly in cases where subtle anomalies needed to be detected.
- **Autoencoder**: Achieved the **best performance** when hyperparameters were properly tuned. It was able to reconstruct normal sessions accurately and flag abnormal sessions based on high reconstruction error. Although training took longer, it captured complex, non-linear patterns effectively.

4.5. Result Analysis

Detailed visualization was used to compare model performances:

- Confusion Matrices highlighted true positives and false positives, with One-Class SVM revealing more sensitivity but also higher misclassifications.
- ROC Curves showcased the trade-offs between true positive and false positive rates, with Autoencoder and Isolation Forest showing better AUC than SVM.
- Anomaly Score Distributions helped in understanding how each model separated normal and anomalous data, where SVM’s distribution required more calibration due to its margin-based decision function.

4.6. Observation/Remarks

- The One-Class SVM model reinforced the project’s focus on unsupervised anomaly detection, offering a theoretically sound approach in cases where anomalies are rare or unavailable during training.
- Feature selection and data scaling had a substantial effect on SVM’s performance, underlining the importance of preprocessing in kernel-based models.
- Although Autoencoder required longer training times, its ability to handle complex non-linearities gave it an edge in performance.
- Isolation Forest remained the most efficient in terms of execution time and simplicity, making it suitable for scenarios demanding real-time responses.

Chapter 5: Conclusion

5.1. Conclusion

The project titled "Website Anomaly Detection using Machine Learning" showcased a comprehensive and practical approach to identifying anomalous behavior on websites. By leveraging advanced machine learning techniques—particularly Isolation Forest, Autoencoders, and One-Class Support Vector Machines (SVM)—we successfully built a system capable of learning from normal web traffic patterns and flagging deviations that may indicate cyber threats, misuse, or technical faults.

Throughout the implementation, a strong emphasis was placed on data preprocessing, feature engineering, and model evaluation, ensuring that the system was not only accurate but also efficient and adaptable to various types of websites. Our experimentation revealed that each model brought unique strengths to the table: Isolation Forest was fast and lightweight, Autoencoders provided deep pattern recognition, and One-Class SVM offered a principled approach to unsupervised anomaly detection.

The outcome validated the hypothesis that machine learning can play a significant role in strengthening website security and operational reliability. Moreover, the use of open-source tools and platforms like Google Colab made the system cost-effective and scalable for academic and professional settings.

5.2. Future Scope

While the current implementation has laid a solid foundation for anomaly detection in web environments, there are several promising directions in which this work can be expanded:

- Real-time Integration: Deploying the trained models into a live environment using APIs or web sockets can enable real-time anomaly detection and alerting. This would be particularly valuable for e-commerce and financial platforms.
- Enhanced Feature Set: Incorporating temporal patterns, user behavior sequences, geo-location metadata, request headers, and session durations can provide richer context for the models, improving detection accuracy.
- Advanced Deep Learning Models: Future iterations of this system could integrate deep learning models like Long Short-Term Memory (LSTM) networks or Transformer-based architectures. These models are particularly effective in capturing temporal dependencies and sequential anomalies in web traffic.
- Hybrid Systems: Combining rule-based logic with machine learning predictions can lead to hybrid models that offer the interpretability of rules with the adaptability of learning-based approaches. Such systems could reduce false positives and adapt over time.
- Continuous Learning Pipelines: Implementing pipelines that allow the system to learn continuously from new data would keep the models up-to-date and effective against evolving threat patterns.

5.3. Societal Impact

The development and deployment of website anomaly detection systems have broad and significant implications for society, particularly in an era where cybersecurity and digital trust are paramount.

- Cybersecurity Enhancement: Early detection of anomalies helps mitigate the risks associated with cyber attacks, such as Distributed Denial of Service (DDoS), data breaches, and bot intrusions. This strengthens the defensive posture of online platforms and critical infrastructure.
- Business Continuity: For online businesses, especially in sectors like e-commerce, banking, and healthcare, uninterrupted service is crucial. Anomaly detection systems help identify irregularities before they cause service outages or financial loss.
- User Trust and Confidence: A secure and stable web experience builds trust among users. When platforms can proactively identify and respond to suspicious activities, users feel safer interacting and transacting online.
- Data Privacy and Protection: Anomaly detection acts as a frontline defense in protecting sensitive user data. By identifying unauthorized access attempts or data exfiltration activities early, such

systems play a crucial role in complying with data protection regulations like GDPR.

- Educational and Research Value: The techniques explored in this project serve as valuable learning modules for students and researchers, promoting awareness and understanding of cybersecurity through a data science lens.