

# Assignment Report

# Information Security

---

Abdullah Shahid (i21-0326)

Muhammad Izen Ali (i21-0320)

27th March, 2025



# Malicious URL Classification

## 1. Introduction

The increasing use of the internet has led to a rise in malicious activities, often propagated through URLs. Traditional blacklist-based approaches struggle to keep up with newly generated malicious links. This project aims to classify URLs into five categories (benign, defacement, phishing, malware, spam) using various machine learning (ML) and large language model (LLM)-based approaches. The objective is to merge datasets, preprocess data, perform exploratory data analysis (EDA), extract meaningful features, and apply ML models for accurate classification.

## 2. Data Merging & Preprocessing

### 2.1 Dataset Integration

Two datasets were merged to create a comprehensive labeled dataset. The primary dataset contained four malicious URL categories, while a helper dataset was incorporated to introduce the fifth category.

### 2.2 Data Cleaning

- Handled missing values using appropriate imputation techniques.
- Removed duplicate and inconsistent records.
- Standardized the URL format and extracted relevant components.

### 2.3 Data Balancing

To mitigate class imbalance, oversampling techniques such as SMOTE and undersampling were applied, ensuring equal representation of all categories.

## 3. Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the dataset:

- Descriptive statistics were generated for URL structures.
- Data distributions were visualized using histograms and box plots.
- Correlations between URL attributes and malicious categories were analyzed.

### 3.1 Graphical Insights

- URL length distributions across categories.
- Frequency of special characters in malicious vs. benign URLs.
- Heatmaps displaying correlations between extracted features.
- Class distribution before and after balancing.
- Feature importance rankings from ML models.

## 4. Feature Extraction

### 4.1 Structural Features

- URL length, number of subdomains, and special character count were extracted.

### 4.2 NLP-Based Features

- TF-IDF, Word2Vec, and transformer-based embeddings were applied to capture textual context.



### 4.3 Sequence-Based Features

- Character-level sequence analysis was performed to detect patterns in malicious URLs.

## 5. Models

Three types of classification models were implemented:

### 5.1 Traditional ML Models

- Random Forest, Support Vector Machine (SVM), and XGBoost.

### 5.2 Deep Learning Models

- LSTMs and CNNs were applied for sequential URL analysis.

### 5.3 LLM-Based Models

- Fine-tuned transformer models such as BERT and GPT were utilized for URL classification.


## 6. Results & Analysis

### 6.1 Model Performance

- Accuracy, precision, recall, and F1-score were computed.
- Confusion matrices and ROC curves were generated for comparison.

### 6.2 Model Comparison

- Traditional ML models achieved high accuracy but struggled with complex patterns.

- 
- Deep learning models improved performance on textual features.
  - LLM-based models outperformed others, exceeding 90% accuracy due to contextual understanding.

## 7. Challenges & Future Work

### 7.1 Challenges

- Handling imbalanced datasets despite resampling techniques.
- Computational complexity of deep learning and LLM-based models.

### 7.2 Future Improvements

- Further fine-tuning of LLMs with domain-specific datasets.
- Exploring hybrid models combining traditional ML and deep learning techniques.
- Incorporating real-time URL detection for enhanced security.

## 8. Conclusion

This project successfully implemented various ML and LLM-based models to classify malicious URLs. The findings suggest that transformer-based models provide superior accuracy and contextual understanding, making them highly effective in detecting emerging threats.

