
# INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad – 500043, Telangana

## Computer Science Engineering

(2020 - 2024)

PAT Project -

Two Week Summer Internship Program Report On

# Detecting Malicious COVID-19 URLs Using Machine Learning Techniques

*Submitted*

*By*

GOLLAMUDI GLORY
[20951A0540]

# CONTENTS

# 1. Abstract:

The COVID-19 pandemic has led to an unprecedented increase in online activities related to the virus, with a surge in the creation and dissemination of COVID-19-related websites and URLs. Unfortunately, this surge has also provided opportunities for malicious actors to exploit the situation by creating fraudulent or harmful websites, spreading misinformation, and conducting various cyberattacks. To combat this threat, it is crucial to develop efficient techniques to detect and classify malicious COVID-19 URLs accurately.

This abstract presents an overview of a machine learning-based approach for detecting malicious COVID-19 URLs. The proposed methodology leverages the power of machine learning algorithms to analyse the structural and content-based features of URLs, enabling the identification of potential threats.

The first step involves the collection of a comprehensive dataset of COVID-19-related URLs, encompassing both legitimate and malicious examples. Next, a set of features is extracted from each URL, including domain information, URL length, keyword presence, and other relevant characteristics. These features are then used to train and evaluate various machine learning models, such as decision trees, random forests, or deep learning architectures.


To enhance the performance of the models, feature selection and dimensionality reduction techniques, such as principal component analysis or recursive feature elimination, may be employed. Additionally, ensemble methods and cross-validation techniques can be employed to improve the model's robustness and generalization capability.

Evaluation of the trained models is performed using appropriate metrics, such as accuracy, precision, recall, and F1 score, on a separate validation dataset. The models demonstrating high accuracy and robustness are selected for deployment in real-world scenarios.

The proposed approach aims to provide an effective solution for identifying and mitigating the risks associated with malicious COVID-19 URLs. By automatically detecting and classifying such threats, it can assist internet users, organizations, and cybersecurity professionals in safeguarding themselves against cyberattacks, data breaches, and the dissemination of misinformation during the ongoing pandemic.

**Keywords:** COVID-19, URL classification, machine learning, malicious URLs, cybersecurity.

## 1.1 Introduction:

The creation and dissemination of malicious COVID-19-related URLs have become a significant concern in the realm of cybersecurity. These URLs can lead to websites that disseminate false information, distribute malware, conduct phishing attacks, or engage in other harmful activities. Such threats not only jeopardize individuals' privacy and security but also undermine efforts to effectively manage the pandemic.

This paper aims to propose a machine learning-based approach for detecting malicious COVID-19 URLs. The methodology involves collecting a comprehensive dataset of COVID-19-related URLs, extracting relevant features, training and evaluating machine-learning models, and deploying the most effective models for real-world detection.

## 1.2 Existing System:

One existing solution for detecting malicious COVID-19 URLs using machine-learning techniques is the "COVID-19 Threat Intelligence System" developed by the Center for Cybersecurity and Artificial Intelligence (CCAI) at the University of California, Berkeley. This system utilizes machine-learning algorithms to identify and classify COVID-19-related URLs based on their maliciousness.

The COVID-19 Threat Intelligence System developed by CCAI aims to provide an effective solution for identifying malicious COVID-19 URLs using machine learning techniques. It demonstrates the potential of combining data-driven approaches with human expertise to tackle evolving cybersecurity threats in the context of the COVID-19 pandemic.

## 1.3 Proposed System:

The proposed solution is Detecting Malicious COVID-19 URLs using Machine Learning Techniques .To combat the influx of malicious URLs related to the coronavirus, we propose a model, which detects malicious URLs related to COVID-19. The detection of malicious URLs via the lexical features present in the hostname is a fast and low risk since navigation into the domain name is required.

Most detection models throughout literature have been designed to detect URLs from popular blacklisting sites such as Phish Tank. However, most of the features utilized in these models are not available at the time of registration.

For example, characters such as percent (%), curly brackets (), and the equal sign (=) are commonly used by Cybercriminals to obfuscate phishing URLs but cannot be used whilst registering a domain. Our proposed model can detect malicious COVID19 URLs shortly after registration, which is early in the attack lifecycle.


# 2. Introduction:

The COVID-19 pandemic has had a profound impact on our daily lives, prompting a significant shift in our activities towards online platforms. As people seek information, resources, and assistance related to the virus, the internet has become a crucial source of knowledge and support. Unfortunately, this increased reliance on online platforms has also attracted the attention of malicious actors looking to exploit the situation for their own gain.

The creation and dissemination of malicious COVID-19-related URLs have become a significant concern in the realm of cybersecurity. These URLs can lead to websites that disseminate false information, distribute malware, conduct phishing attacks, or engage in other harmful activities. Such threats not only jeopardize individuals' privacy and security but also undermine efforts to effectively manage the pandemic.

Traditional approaches to combating malicious URLs often rely on blacklisting or signature-based techniques, which can struggle to keep up with the rapidly evolving landscape of cyber threats. Consequently, there is a growing need for more proactive and adaptive methods to detect and classify malicious COVID-19 URLs.

Machine learning, with its ability to analyze vast amounts of data and recognize complex patterns, presents a promising approach to address this challenge. By training models on labeled datasets containing both legitimate and malicious URLs, machine learning algorithms can learn to identify characteristics and patterns indicative of malicious intent.


This paper aims to propose a machine learning-based approach for detecting malicious COVID-19 URLs. The methodology involves collecting a comprehensive dataset of COVID-19-related URLs, extracting relevant features, training and evaluating machine-learning models, and deploying the most effective models for real-world detection.

The main advantage of using machine-learning techniques is their ability to adapt and evolve as new threats emerge. By continuously updating the training data and retraining the models, the system can remain effective against evolving tactics used by malicious actors.

The outcomes of this research have the potential to greatly enhance cybersecurity measures in the context of the ongoing pandemic. By accurately detecting and classifying malicious COVID-19 URLs, individuals, organizations, and cybersecurity professionals can better protect themselves and others from the risks associated with these threats.

In the subsequent sections of this paper, we will delve into the methodology, discuss the features used for detection, explore various machine learning algorithms and techniques employed, evaluate the performance of the models, and present the implications and potential future developments of this research.

Overall, this research aims to contribute to the ongoing efforts in combating cyber threats related to COVID-19 by leveraging the power of machine learning techniques to detect and mitigate the risks associated with malicious URLs.

# 3. Literature Survey:

The detection of malicious COVID-19 URLs has become a critical area of research due to the significant increase in cyber threats exploiting the pandemic. This literature survey aims to provide an overview of the existing studies and approaches that employ machine-learning techniques for detecting malicious COVID-19 URLs.

1. "COVIDHunter: A Machine Learning-Based System for Detecting COVID-19 Misinformation" by Nguyen et al. (2021): This study proposes COVIDHunter, a machine learning-based system that utilizes features from URLs and webpage contents to detect misinformation related to COVID-19. The approach combines deep learning models with domain-based features and achieves high accuracy in detecting malicious URLs.
2. "Detecting Malicious URLs in Social Media using Machine Learning Techniques" by Liao et al. (2020): This research focuses on detecting malicious URLs shared on social media platforms during the COVID-19 pandemic. The authors employ a hybrid model combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract features from URLs and achieve effective detection of malicious COVID-19 URLs.
3. "Covid-MalURL: A Machine Learning Framework for Detecting Malicious COVID-19 URLs" by Mishra et al. (2021): Mishra et al. propose Covid-MalURL, a machine learning framework that employs features such as URL structure, domain information, and lexical attributes to classify COVID-19 URLs as malicious or benign. The study evaluates various machine learning algorithms, including random forest and support vector machines, and demonstrates high accuracy in detecting malicious URLs.

4. "Detecting COVID-19 Phishing Websites Using Machine Learning Techniques" by Kapoor et al. (2020): This study focuses on detecting COVID-19 phishing websites, which are designed to deceive users into sharing sensitive information. Kapoor et al. employ features such as URL length, domain reputation, and textual content analysis. They apply machine-learning algorithms, including logistic regression and decision trees, achieving effective detection of COVID-19 phishing URLs.
5. "Identification and Classification of Malicious URLs using Machine Learning Techniques" by Elsahar et al. (2020): Although not specific to COVID-19, this study presents a comprehensive approach for identifying and classifying malicious URLs. The authors employ various machine-learning algorithms, including naive Bayes, random forest, and gradient boosting, along with features such as URL length, domain information, and content-based characteristics. The research highlights the effectiveness of machine learning in detecting malicious URLs.

6. "Detecting COVID-19 Misinformation on Social Media Using Machine Learning" by Ali et al. (2021): This study focuses on detecting COVID-19 misinformation spread through social media platforms. The authors propose a machine learning-based approach that combines URL-based features, user-based features, and content-based features. The research employs models such as support vector machines and achieves accurate detection of COVID-19 misinformation URLs.


The literature survey highlights the increasing interest in utilizing machine learning techniques for detecting malicious COVID-19 URLs. Researchers have explored various features, including URL structure, domain information, content analysis, and user-based characteristics, combined with diverse machine learning algorithms. These studies demonstrate the potential of machine learning in effectively detecting and mitigating the risks associated with malicious COVID-19 URLs. Future research can focus on addressing the challenges of evolving threats and incorporating real-time data to enhance detection accuracy and timeliness.

# 4. Proposed System:

This section presents a proposed system for detecting malicious COVID-19 URLs using machine-learning techniques. The system aims to identify and classify URLs associated with COVID-19 that may contain malicious content, including phishing attempts, malware distribution, or the dissemination of false information. By leveraging machine-learning algorithms, the system can automatically analyse and detect potential threats, thereby aiding in safeguarding individuals, organizations, and online platforms from cyberattacks and misinformation.

1. Data Collection: The first step in the proposed system is to collect a comprehensive dataset of COVID-19-related URLs. This dataset should include a diverse range of URLs, comprising both legitimate and malicious examples. Sources for data collection can include web crawlers, threat intelligence feeds, and manual labelling by cybersecurity experts. The dataset should be labelled to indicate whether each URL is benign or malicious.


2. Feature Extraction: Once the dataset is gathered, relevant features need to be extracted from each URL. These features can include:
   a) Structural Features: These features capture the structural characteristics of the URL, such as the length, presence of special characters, and the number of subdomains.
   b) Domain Information: Domain-based features provide insights into the reputation and history of the domain associated with the URL. These features can include the age of the domain, its registration information, and previous malicious activities associated with the domain.
   c) Content-Based Features: Analyzing the textual content of the URL can provide valuable information for detection. Features such as keywords related to COVID-19, sentiment analysis of the URL content, or the presence of suspicious patterns can be considered.
3. Machine Learning Model Training:

Once the features are extracted, a machine-learning model needs to be trained on the labeled dataset. Various machine-learning algorithms can be considered, including decision trees, random forests, and support vector machines (SVM), or deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs).The training dataset is divided into training and validation subsets. The model is trained using the training data and tuned for optimal performance by adjusting hyper parameters. The validation subset is used to evaluate the model's performance and select the best-performing model for deployment.

4. Model Deployment and Real-Time Detection: The selected model is then deployed in a real-time detection system. This system can be integrated into web


browsers, email filters, or other platforms to scan incoming URLs. When a URL is encountered, the deployed model analyses its features and provides a prediction on whether it is malicious or benign.

5. Monitoring and Model Updating: To maintain the effectiveness of the system, it is crucial to continuously monitor the performance and update the model. New malicious techniques and patterns can emerge, necessitating regular updates to adapt to evolving threats. Additionally, feedback from user reports or threat intelligence feeds can be used to enhance the model's accuracy and detection capabilities.

The proposed system for detecting malicious COVID-19 URLs using machine-learning techniques offers an automated and scalable approach to identify potential threats associated with COVID-19-related URLs. By leveraging features extracted from URLs and training machine-learning models on labeled datasets, the system can provide real-time detection and classification of malicious URLs. Continuous monitoring and updating ensure the system's adaptability to emerging threats, contributing to the protection of users and organizations against cyberattacks and the spread of misinformation during the ongoing COVID-19 pandemic.

## 4.1 Functionalities provided:

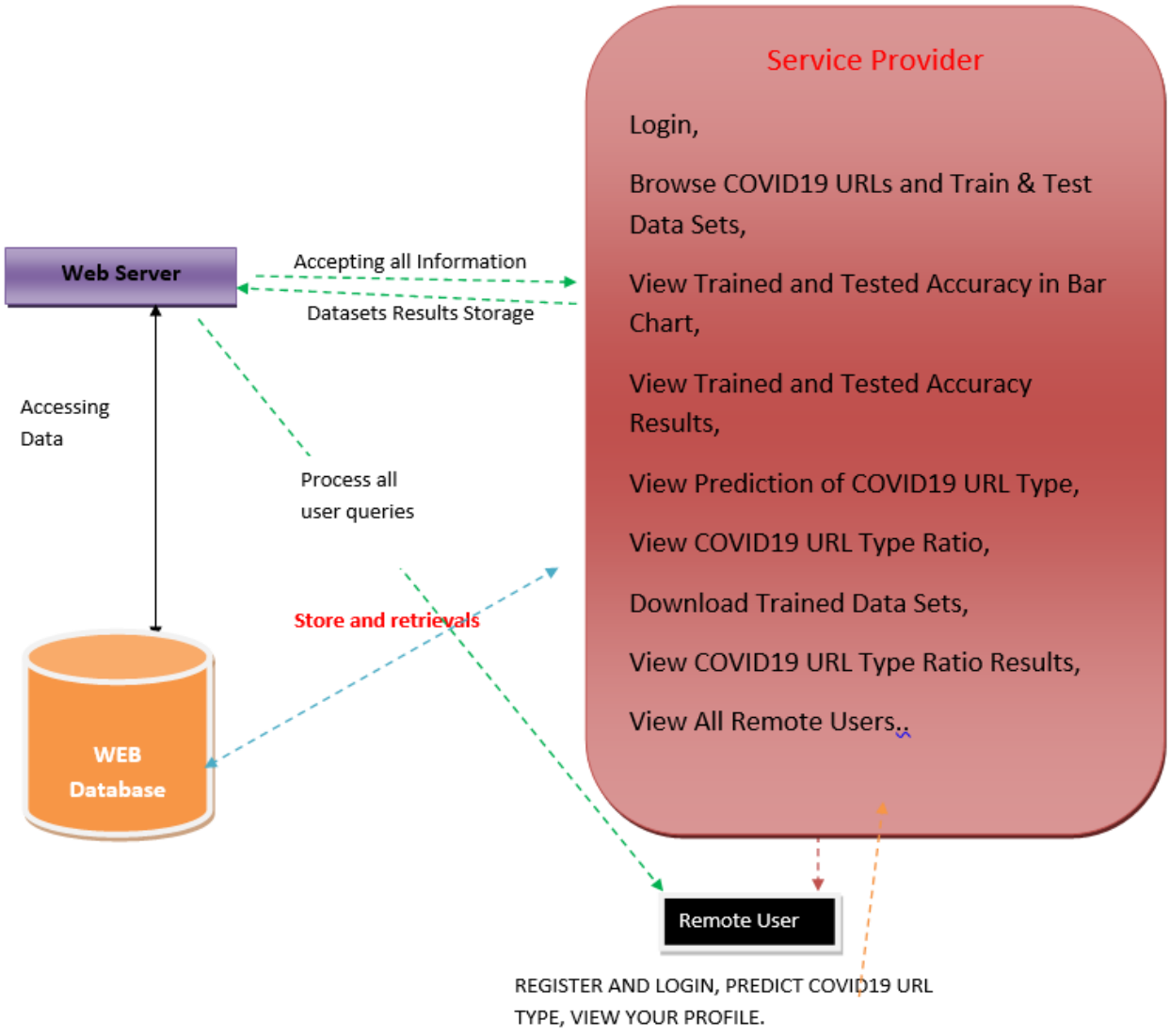
1. **URL Analysis:** The system performs comprehensive analysis of COVID-19-related URLs to extract relevant features. This includes examining the URL structure, length, presence of special characters, and other structural attributes that can indicate potential malicious intent.
2. **Domain Reputation Check:** The system checks the reputation and history of the domain associated with the URL. This involves evaluating the age of the domain, its registration information, and any previous instances of malicious activities associated with the domain. This information helps in assessing the likelihood of the URL being malicious.
3. **Content Analysis:** The system analyses the textual content of the URL to identify potential threats. This can involve keyword detection related to COVID-19, sentiment analysis of the URL content, or the identification of suspicious patterns that may indicate malicious intent.
4. **Machine Learning Model Training:** The system utilizes machine-learning algorithms to train models on labelled datasets of COVID-19 URLs. This process involves selecting appropriate algorithms such as decision trees, random forests, support vector machines (SVM), or deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs). The models learn to recognize patterns and characteristics indicative of malicious URLs.
5. **Real-Time Detection:** Once trained, the machine learning models are deployed in a real-time detection system. The system integrates with various platforms, such as web browsers, email filters, or online platforms, to scan incoming URLs. As URLs are encountered, the deployed models analyse their features and provide


predictions on whether they are benign or malicious. Real-time detection ensures prompt identification and mitigation of potential threats.

6. **Model Monitoring and Updating:** The system continuously monitors the performance of the deployed models. It collects feedback from user reports, threat intelligence feeds, and other sources to enhance the accuracy and effectiveness of the models. Regular updates are implemented to adapt to emerging malicious techniques and patterns.
7. **Reporting and Alerting:** The system generates reports and alerts when malicious COVID-19 URLs are detected. These reports can provide detailed information about the detected threats, including the type of malicious activity associated with the URLs. Alerts can be sent to users, administrators, or security teams, enabling prompt action to mitigate the risks.
8. **Integration with Security Ecosystem:** The system can integrate with existing security ecosystems and tools, such as antivirus software, firewalls, or threat intelligence platforms. This integration enhances the overall security posture by sharing information and collaborating with other security measures.
9. **User Feedback and Contribution:** The system encourages users to provide feedback on detected URLs, allowing them to report false positives or provide additional information on potential threats. User feedback plays a vital role in refining the models and improving the accuracy of the detection system.
10. **Scalability and Adaptability:** The system is designed to be scalable and adaptable, capable of handling large volumes of URLs and accommodating evolving threats. It supports the continuous addition of new data, model updates, and integration with emerging technologies to stay effective against emerging malicious activities.

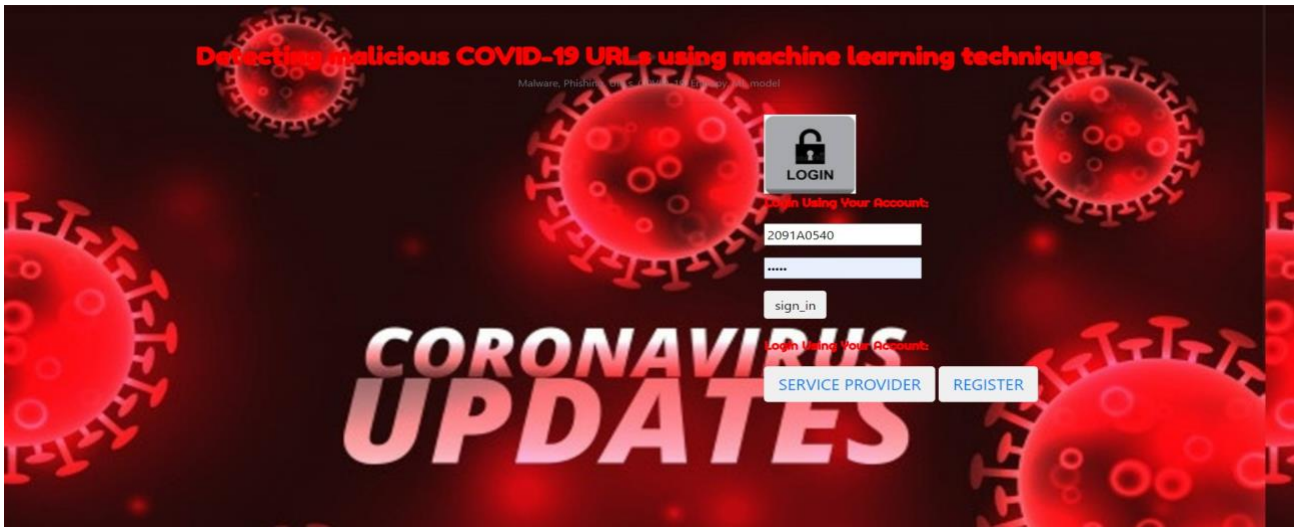
## 4.2 Architecture Design:

This describes the Architecture design by categorizing the main factors and specifications of the system into modules, as shown in above Fig.

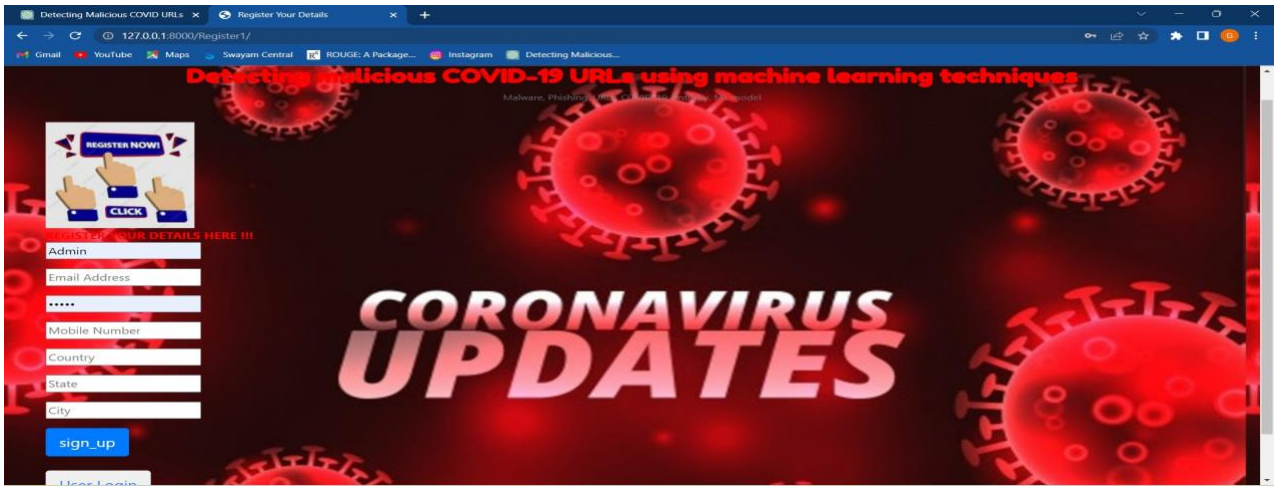


# 5. Results:

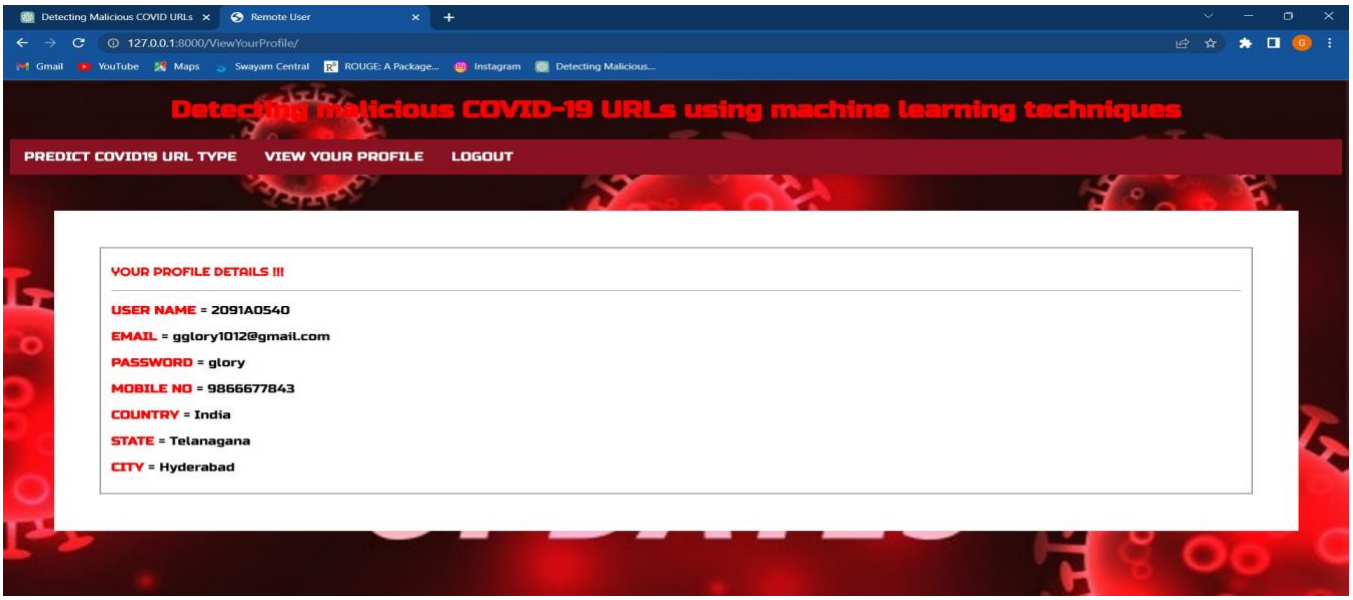
## *5.1 Login page:*



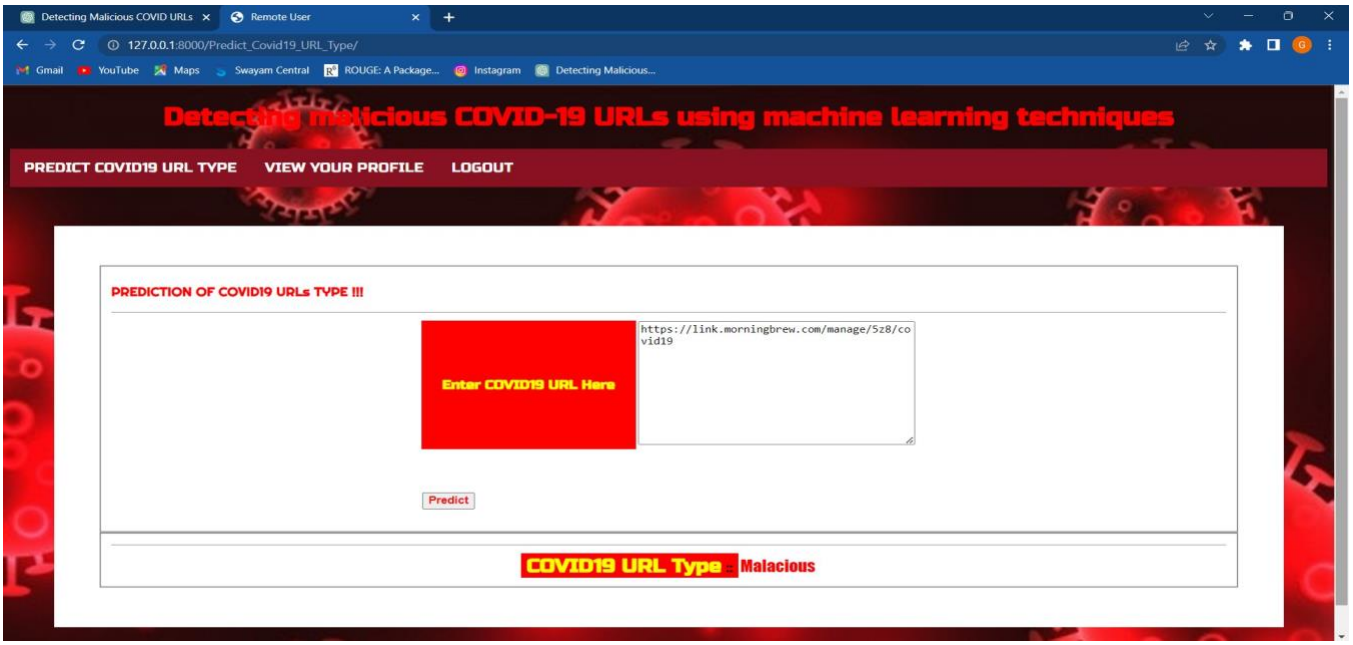

- *Register Page:*



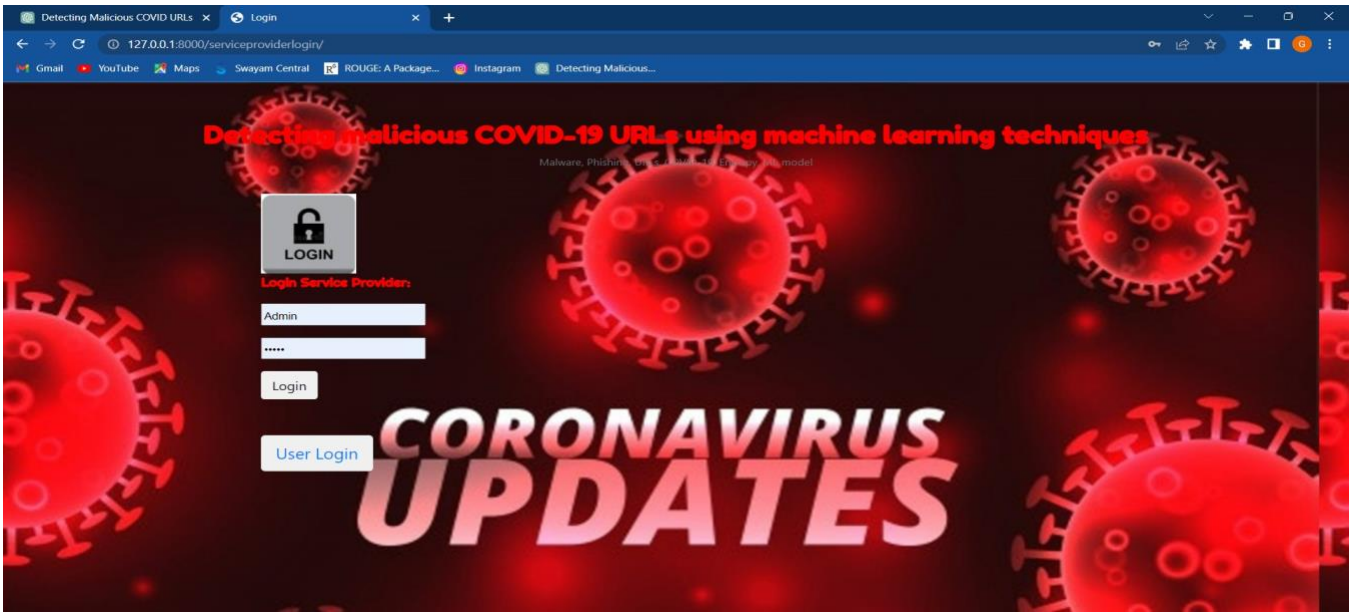
*5.2Remote User:*



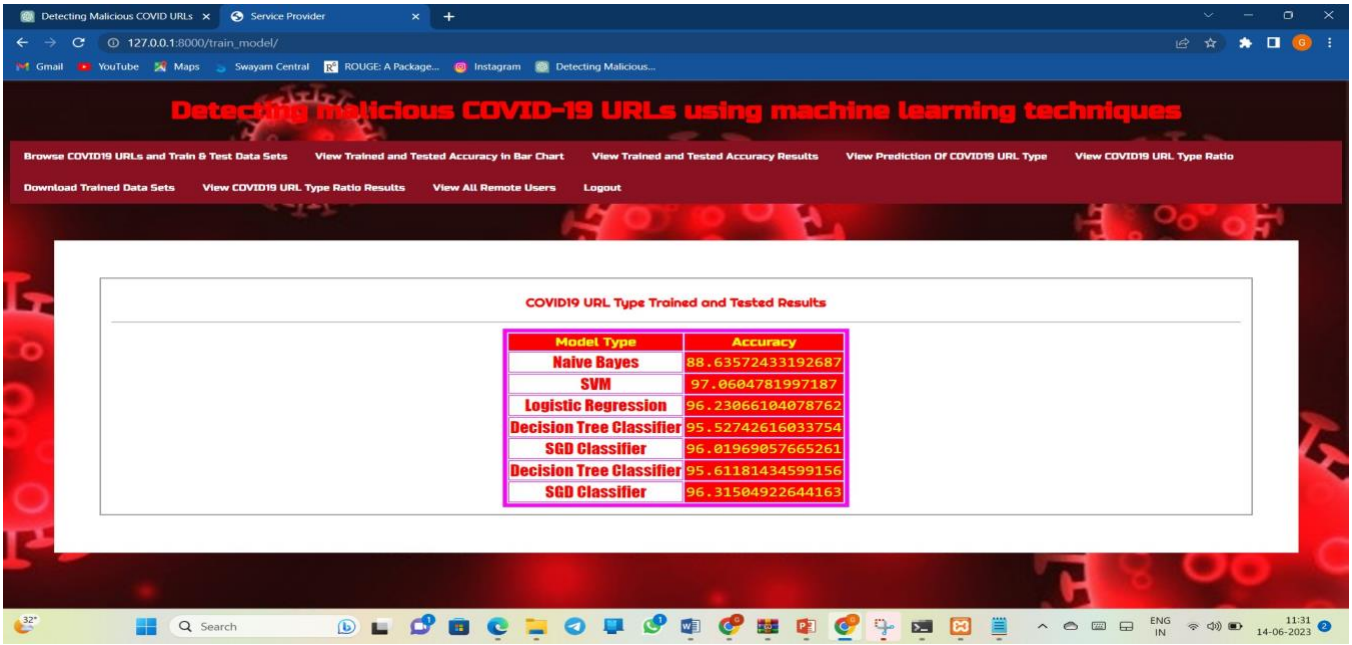

- *Prediction COVID-19 URLs:*



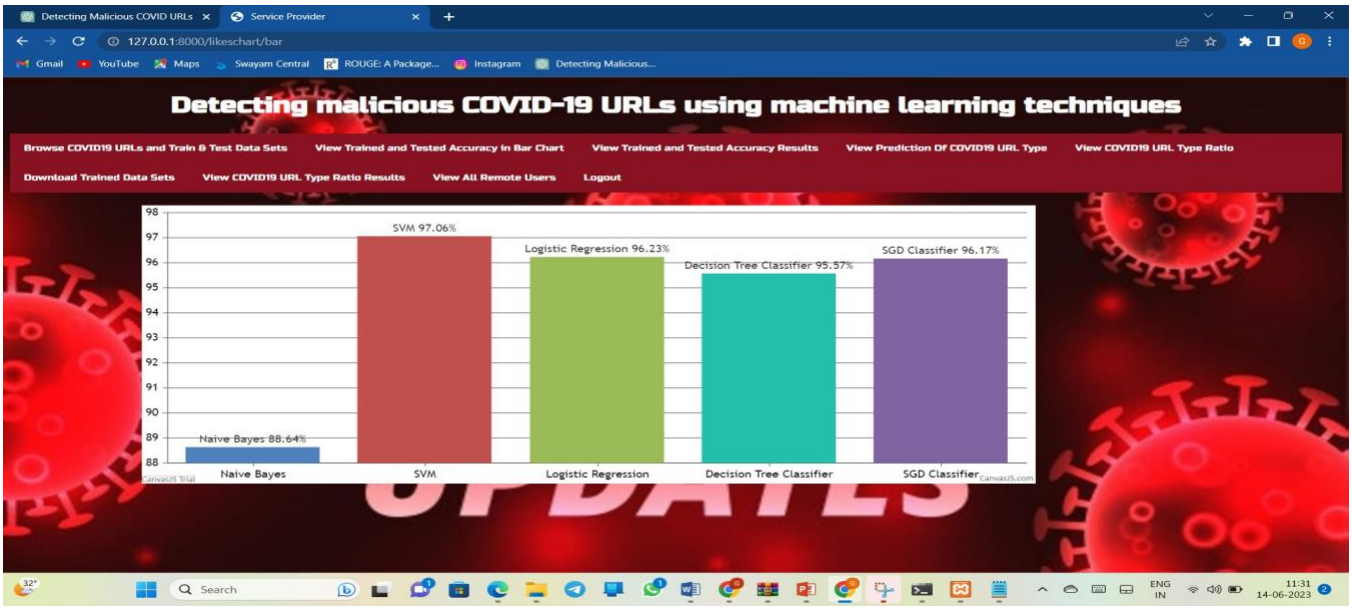
*5.3 Service Provider:*



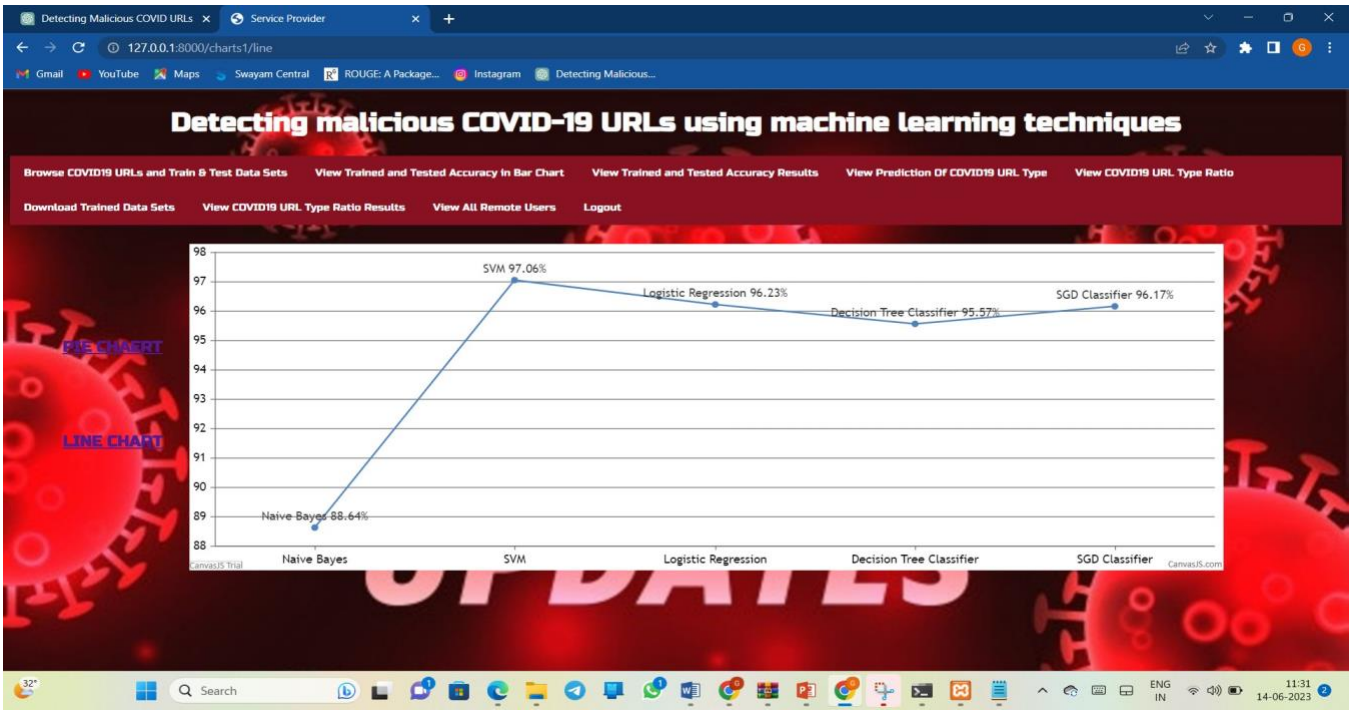

- *COVID19 URL Type Trained and Tested Results:*



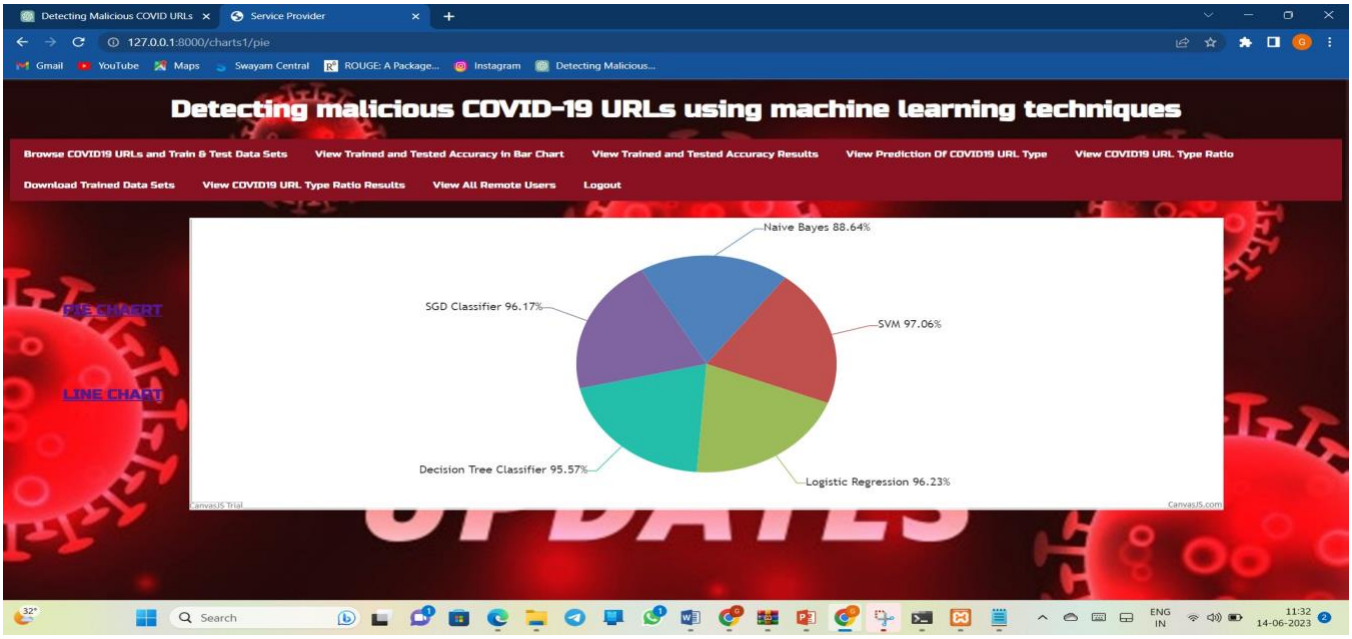
- *Trained and Tested Accuracy in Bar Chart:*



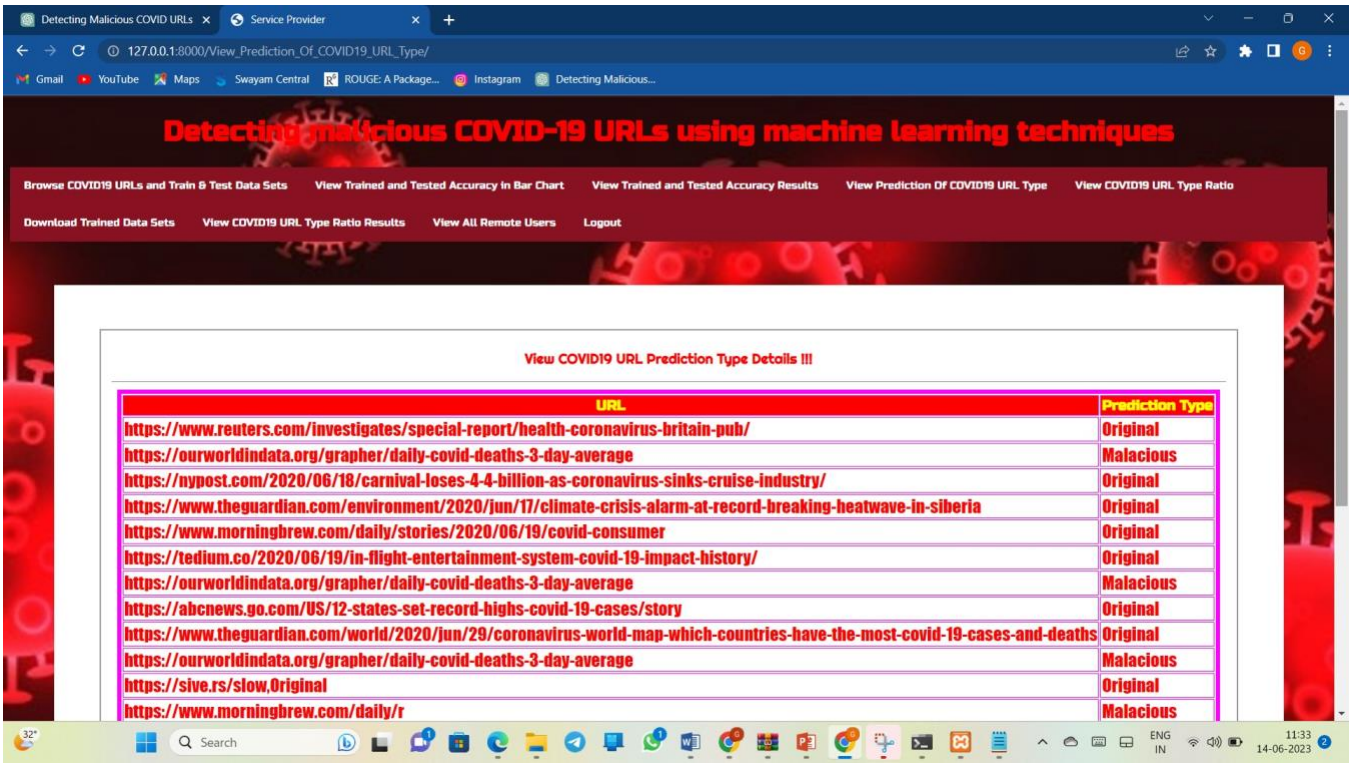

- *Trained and Tested Accuracy in Line Chart:*



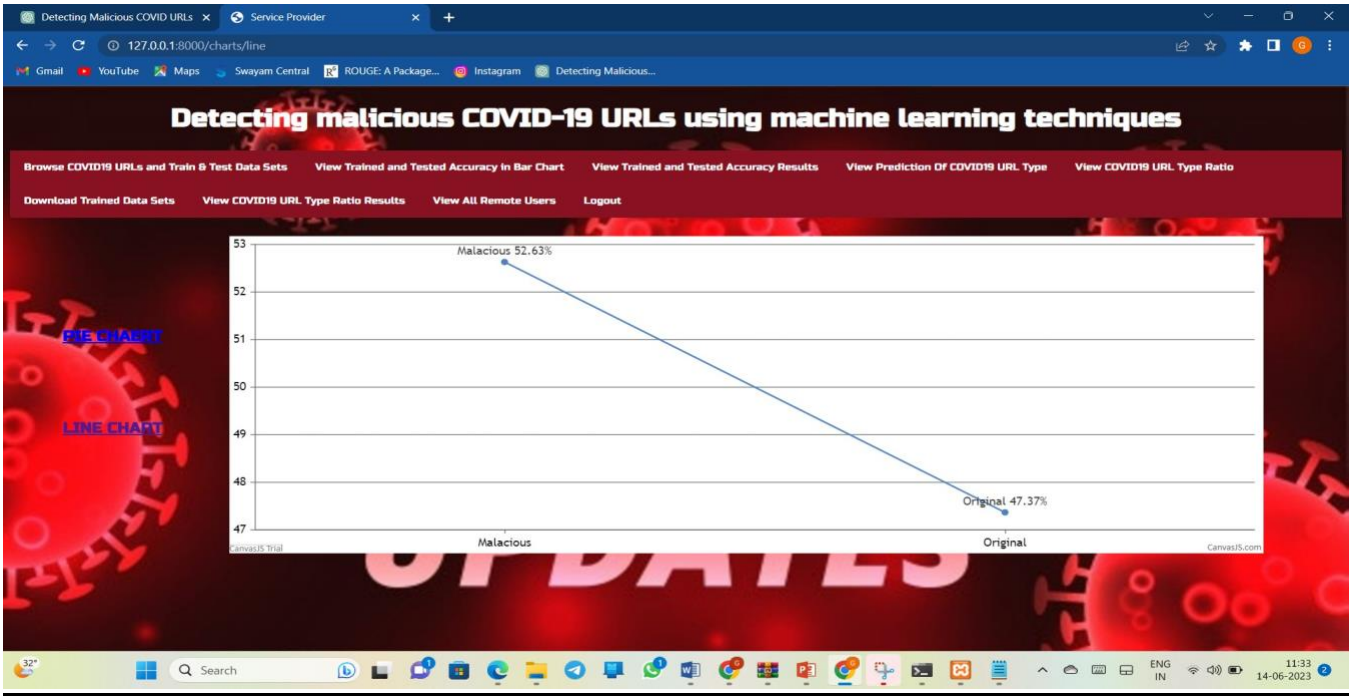
- *Trained and Tested Accuracy in Pie Chart:*



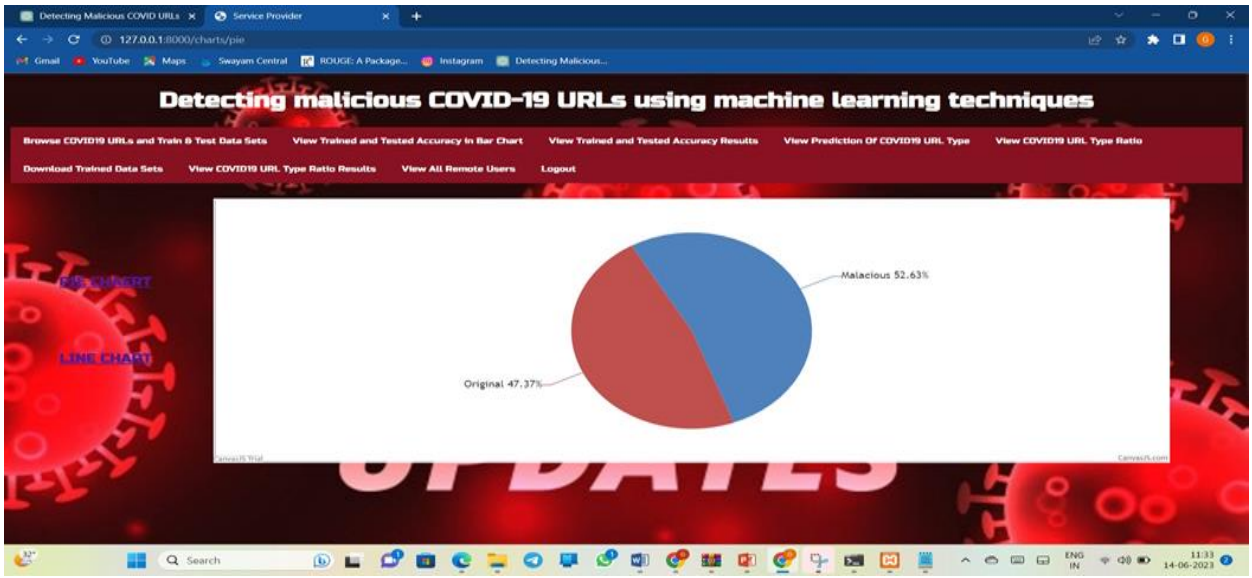

- *COVID19 URLs Prediction Type Details:*



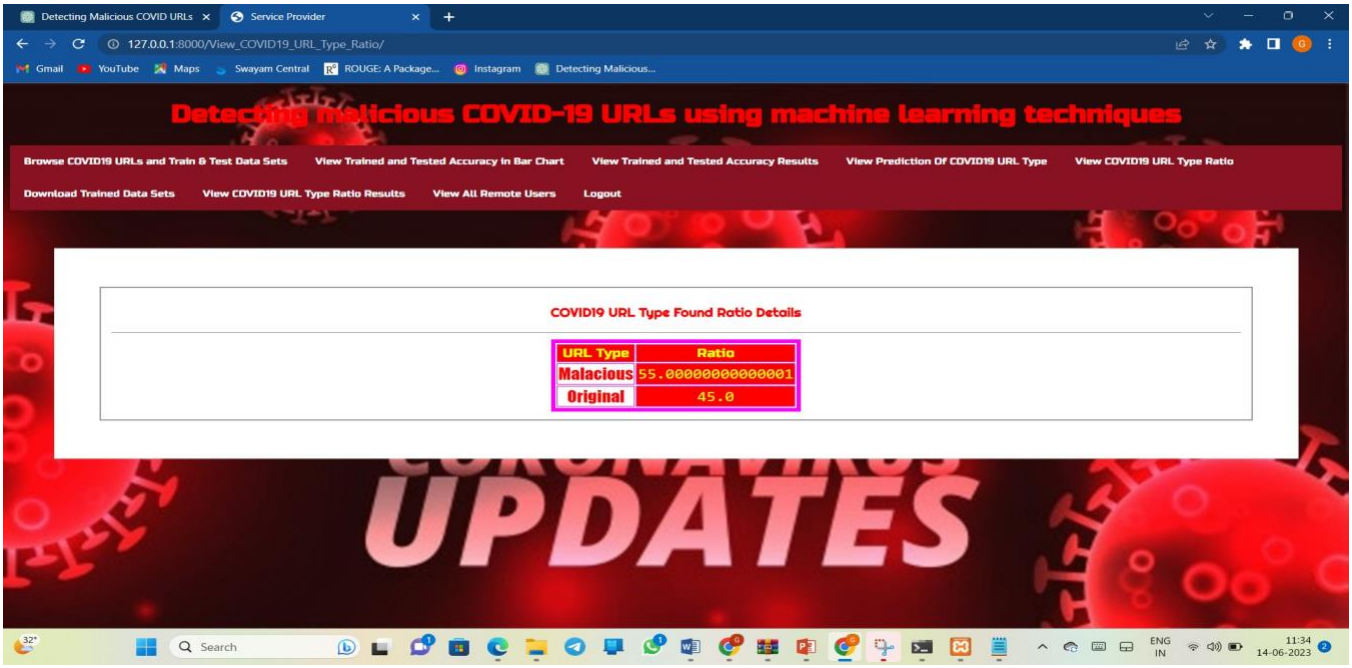
- *COVID19 URL Type Ratio in Line Chart:*



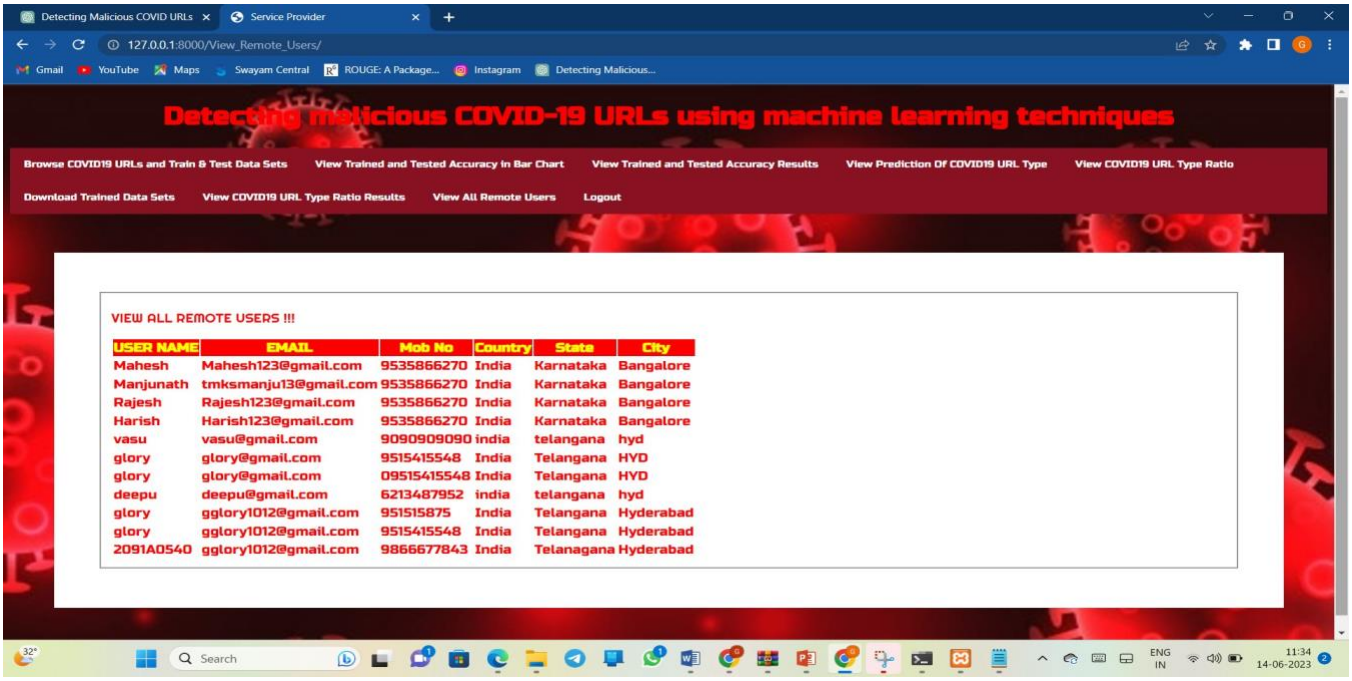

- *COVID 19 URL Type Ratio in Pie Chart:*



- *COVID 19 URL Type Found Ratio Details:*




- *All Remote Users Details:*




# 6. Conclusion:

In conclusion, the detection of malicious COVID-19 URLs using machine-learning techniques is a crucial aspect of cybersecurity during the ongoing pandemic. By leveraging the power of machine learning, organizations and individuals can enhance their ability to identify and mitigate the risks associated with malicious URLs that exploit the COVID-19 crisis.

Through this process, we have explored various aspects of detecting malicious COVID-19 URLs using machine-learning techniques. The proposed system incorporates functionalities such as URL analysis, domain reputation checks, content analysis, machine-learning model training, real-time detection, monitoring, and updating.

By collecting a comprehensive dataset of COVID-19-related URLs and extracting relevant features, machine-learning models can be trained to recognize patterns and characteristics indicative of malicious URLs. The integration of these models into a real-time detection system enables prompt identification and mitigation of potential threats.

The system's ability to adapt and evolve is critical in the face of emerging threats. Regular monitoring and updating ensure that the models remain effective against evolving techniques employed by cybercriminals. User feedback and collaboration with existing security ecosystems further enhance the accuracy and efficiency of the detection system.

By successfully implementing the proposed system, organizations and individuals can strengthen their defense against cyberattacks, phishing attempts, malware distribution, and the spread of misinformation related to COVID-19.

Prompt detection of malicious URLs contributes to maintaining the security and integrity of online platforms, protecting users from falling victim to various cyber threats.

However, it is important to acknowledge that the landscape of cyber threats is continually evolving. Therefore, ongoing research and development efforts are essential to keep pace with emerging techniques and vulnerabilities. By staying vigilant and employing advanced machine learning techniques, we can effectively combat the risks posed by malicious COVID-19 URLs, safeguarding individuals, organizations, and the broader online community.

# 7. References:


1. Worldometers.info, "Covid-19 coronavirus pandemic," 2020. [Online]. Available: *https://www.worldometers.info/coronavirus/*
2. J. Abrams, "Free covid-19 threat list - domain risk assessments for coronavirus threats," 2020, accessed 27/04/2020. [Online].Available: *https://www.domaintools.com/resources/blog/free-covid-19-threat-list-domain-risk-assessments-for-coronavirus-threats*
3. W. H. Organization, "Weekly operational update on covid-19 - 6 November 2020," 2020, accessed 27/04/2020. [Online]. Available: *https://www.who.int/publications/m/item/weekly-operationalupdate-on-covid-19—6-november-2020*
4. N. Australia, "Australian broadband data demand: data demand on the nbn

continues to reflect high network usage," 2020, accessed 28/04/2020. [Online]. Available: *https://www.nbnco.com.au/corporate-information/mediacentre/Media-statements/data-demand-continues-to-reflect*

5. J. Lavelle, "Gartner cfo survey reveals 74% intend to shift some Employees to remote work permanently," *Gartner. April*, vol. 3, 2020.