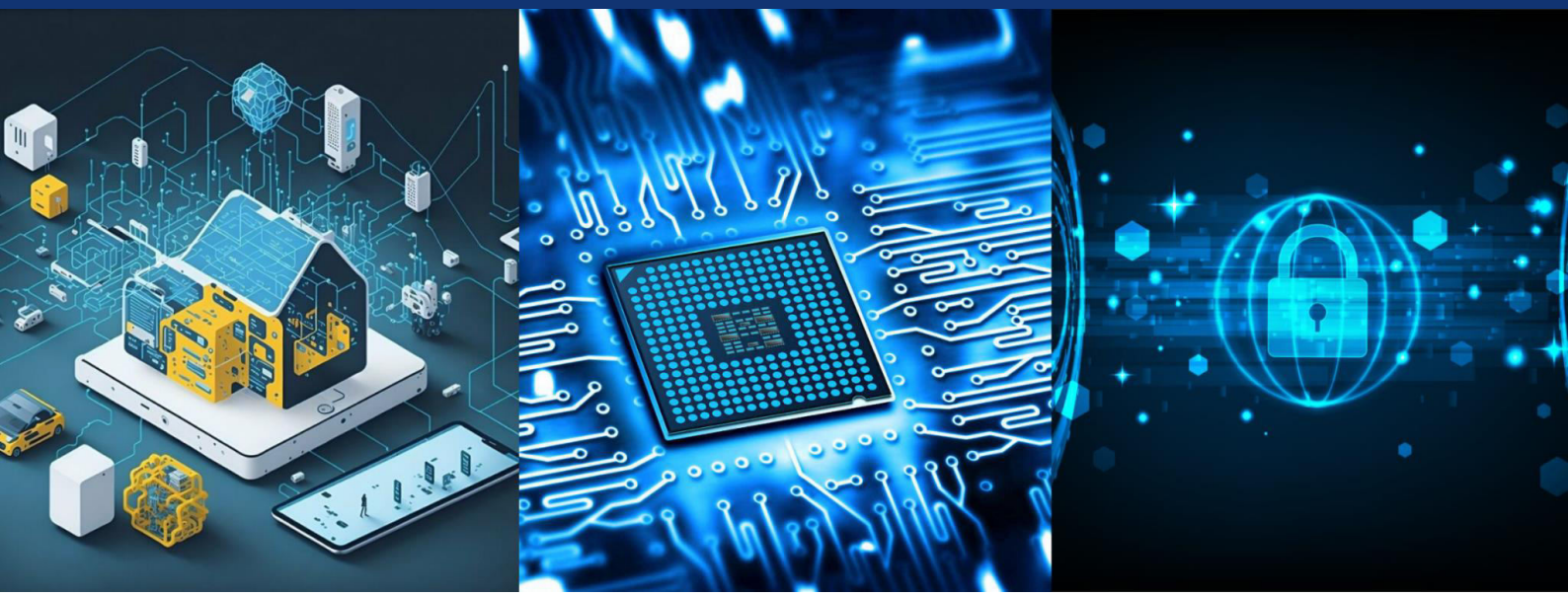
IJIRCCE

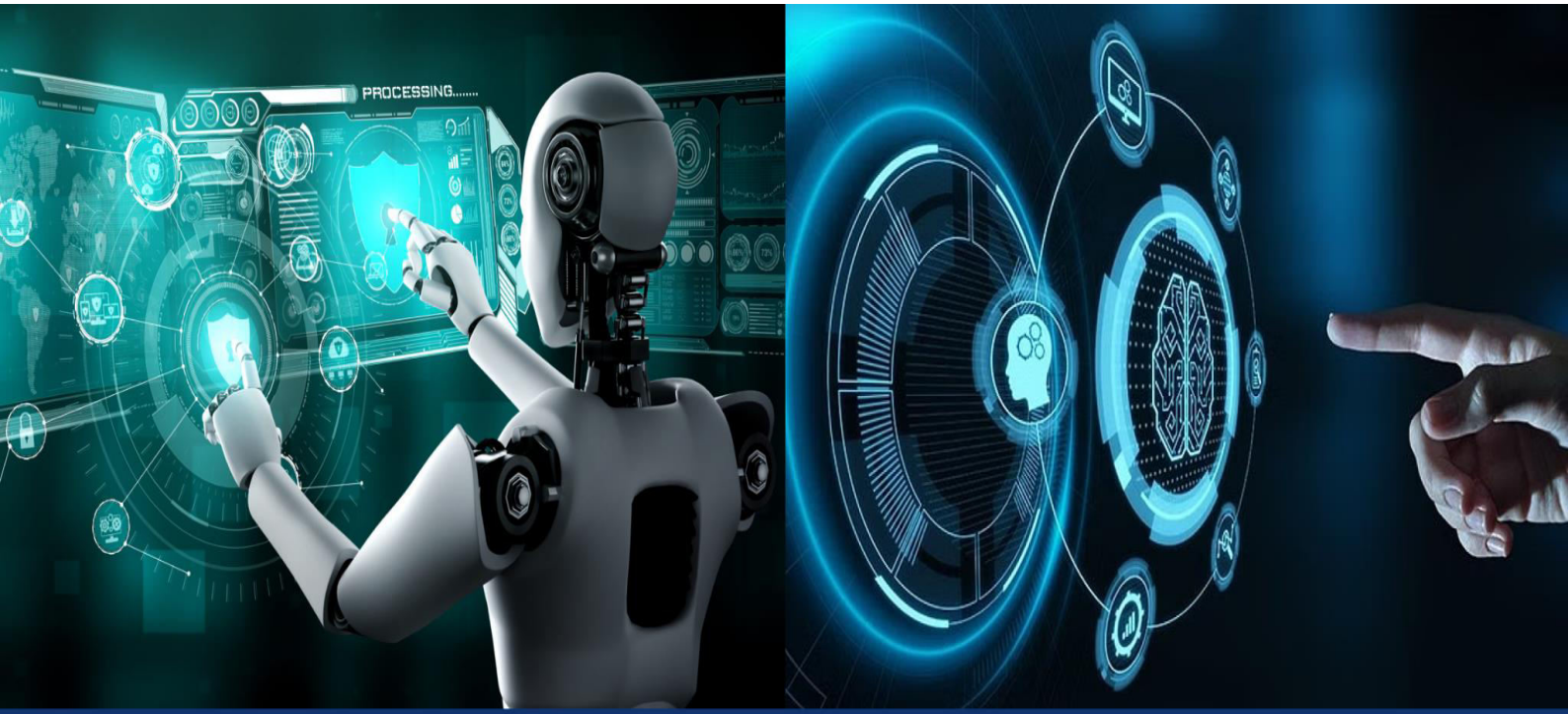
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Cyber Hacking Breaches Prediction and Detection using Machine Learning

M. Hariesh [1], Ms.C.Vanessa[2]

PG Scholar, Dept. of Master of Computer Applications, R.V.S. College of Engineering, Dindigul, Tamil Nadu, India[1]

Assistant Professor, Dept. of Master of Computer Applications, R.V.S. College of Engineering, Dindigul, Tamil Nadu, India[2]

ABSTRACT: In an era of increasingly sophisticated cyber-security threats, the development of robust prediction and detection systems to protect against cyber hacking breaches has become essential. This study presents an effective approach that employs Machine Learning techniques, namely Logistic Regression, Random Forest, and Support Vector Classifier, to predict and detect potential cyber threats. The proposed system is developed using a carefully curated dataset comprising 11,430 URLs with 87 extracted features. Notably, the dataset is balanced, containing an equal distribution of phishing and legitimate URLs. The primary objective of the system is to accurately identify cyber threats while minimizing false positives. Through comprehensive training and evaluation, the experimental results demonstrate the strong performance and reliability of the proposed approach. This work represents a significant contribution to cyber hacking breach prediction and detection by leveraging Machine Learning to enhance security measures, enabling organizations to strengthen their cyber-security defenses against evolving cyber threats and attacks.

I. INTRODUCTION

The rapid growth of internet-based services has significantly increased the exposure of users and organizations to cyber threats, among which phishing attacks remain one of the most prevalent and damaging forms of cybercrime. Phishing websites are designed to impersonate legitimate entities in order to deceive users into disclosing sensitive information such as login credentials, financial data, and personal details. Due to their evolving nature and increasing sophistication, traditional rule-based and blacklist-driven detection mechanisms have proven insufficient in providing effective protection against such attacks. To address these limitations, recent research has increasingly focused on the application of Machine Learning (ML) and Deep Learning (DL) techniques for phishing detection. These data-driven approaches leverage structural, lexical, and behavioural features of URLs to learn patterns that distinguish malicious websites from legitimate ones. Studies have demonstrated that ML-based classifiers such as Logistic Regression, Random Forest, Support Vector Machines, and ensemble models can significantly enhance detection accuracy compared to conventional methods. However, the effectiveness of these models largely depends on the quality of feature selection and the ability to generalize to newly emerging phishing strategies.

Feature selection and dimensionality reduction play a critical role in improving phishing detection performance by eliminating redundant and irrelevant attributes, reducing computational complexity, and minimizing false positives. Several researchers have shown that combining feature selection with ensemble learning techniques yields superior classification accuracy and robustness. Additionally, the use of balanced datasets has been identified as a crucial factor in preventing biased learning and ensuring reliable performance evaluation in phishing detection systems. Despite these advancements, achieving a scalable, adaptive, and highly accurate phishing detection framework remains a challenging task. Many existing solutions either rely on limited feature sets or fail to maintain high accuracy when exposed to unseen phishing patterns. Motivated by these challenges, this paper proposes a smart cyber breach monitoring and detection system using Machine Learning models, including Logistic Regression, Random Forest, and Support Vector Classifier. The proposed approach utilizes a balanced dataset of URLs and applies rigorous feature ranking and selection techniques to enhance classification accuracy and robustness.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. EXISTING SYSTEM

The existing Internet provides a great opportunity for attackers to engage in criminal activities such as online fraud, malicious software, computer viruses, ransom ware, worms, intellectual property rights, denial of service attacks, money laundering, vandalism, electronic terrorism, and extortion. Hacking is a major destroyer of the Internet through which any person can hack computer information and use it in different ways to harm others. Immorality, which harms moral values, is a major issue for the younger generations. Therefore, an awareness of these websites is necessary. Viruses can damage an entire computer network and confidential information by spreading to multiple computers. It is not suitable to use unauthorized websites on the internet.

Disadvantages of Existing Systems

1. Existing phishing detection methods provide low accuracy, resulting in incorrect classification of malicious and legitimate websites.
2. Traditional approaches are unable to effectively support accurate and reliable decision-making for cyber threat detection.
3. The system struggles to identify newly emerging phishing attacks and unknown threat patterns, reducing overall security performance.
4. It offers limited protection against cybercrimes such as phishing, malware, ransom ware, and other online attacks, increasing security risks.

III. PROPOSED SYSTEM

We have derived best possible features and reduced the

- dimensionality of feature subset that can be used for
- phishing detection through the feature ranking using
- the combination of Random Forest algorithm, XGBoost
- algorithm and correlation matrix with heatmap.

We have evaluated performance of the implemented clas-

- sifiers and among them we have proposed the best hybrid
- classifier consisting of SVM, Decision Tree, Random
- Forest and XGBoost to attain higher accuracy.
- Our proposed hybrid classifier will help the Internet users
- verify authentic websites, thereby mitigating the risk of
- phishing websites and ensuing secure online usage

We have derived best possible features and reduced the dimensionality of feature subset that can be used for phishing detection through the feature ranking using the machine learning are Random Forest algorithm, Support Vector Classifier, Decision Tree and Logistic Regression algorithm. We have evaluated performance of the implemented classifiers and among them we have proposed the best machine learning consisting of logistic regression to attain higher accuracy. Our proposed hybrid classifier will help the Internet users verify authentic websites, thereby mitigating the risk of phishing websites and ensuing secure online usage. This project enhancement is to predict whether the URL is phishing or legitimate. The model used the relation between the features and the label to predict the phishing.

Advantages of Proposed System

1. The proposed system provides higher accuracy in detecting phishing websites using advanced machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Classifier.
2. It utilizes optimized feature selection and dimensionality reduction techniques, which improve detection performance and reduce computational complexity.
3. The system can effectively distinguish between legitimate and phishing URLs, helping users identify authentic websites and avoid cyber threats.
4. It enhances online security by reducing false positives and enabling reliable, scalable, and efficient phishing detection.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. SYSTEM ARCHITECTURE

The system architecture illustrates the workflow of the phishing URL detection system using machine learning. Initially, the user provides a URL through the graphical interface. The input data is then processed through feature extraction and data preprocessing stages, where relevant URL features are collected, cleaned, and transformed. The processed data is used to train machine learning models such as Decision Tree, Random Forest, Support Vector Classifier (SVC), and Logistic Regression. The trained model analyzes the input URL and predicts whether it is a phishing or legitimate website. Finally, the prediction result, probability score, and performance metrics are displayed to the user, enabling accurate and efficient phishing detection.

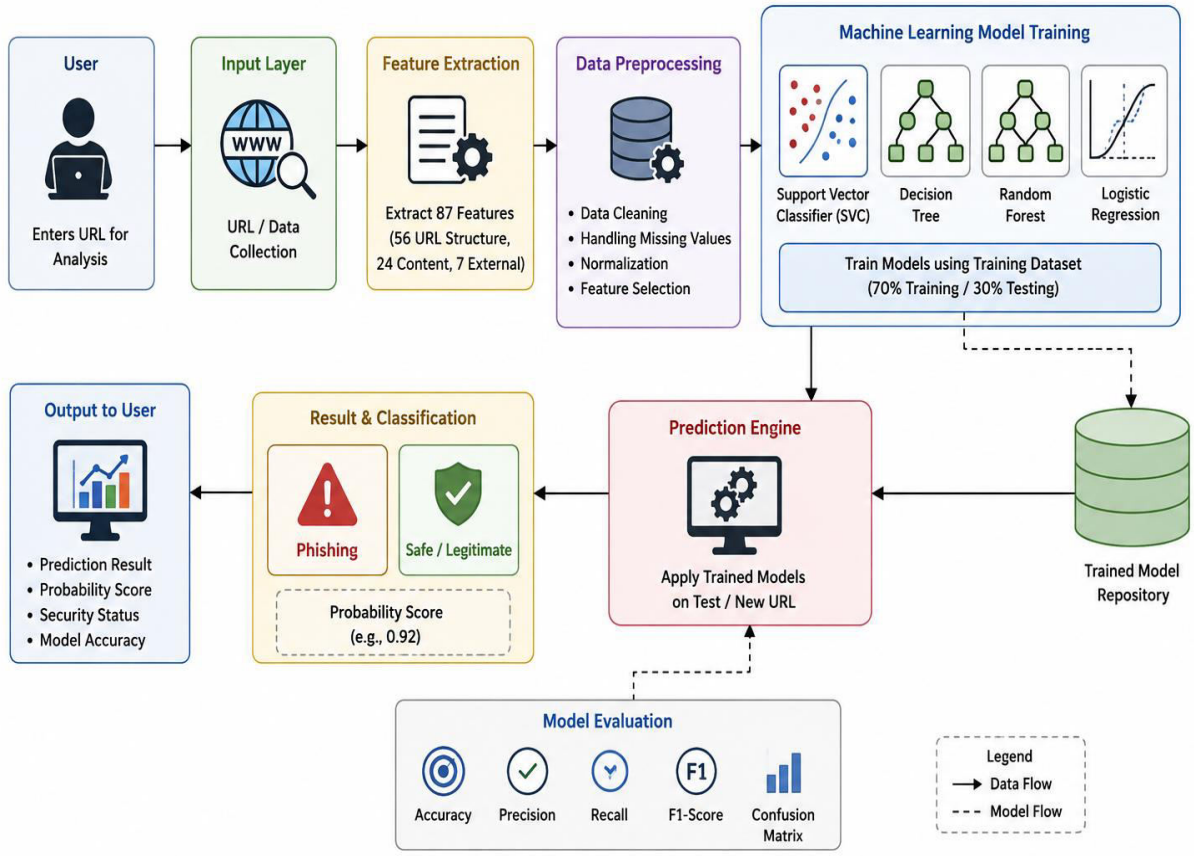


Fig No. 4.1 System Architecture of cyber hacking breaches prediction

V. MODULE DESCRIPTION

1. Data Exploration and Preprocessing
2. Exploratory Data Analysis (EDA)
3. Model Training and Evaluation
4. Model Comparison and Visualization
5. Cross-Validation and Test Size Sensitivity
6. Prediction

1. Data exploration and pre-processing

This module focuses on loading, examining, and preparing the dataset for machine learning analysis. The dataset is imported using Python libraries such as Pandas, and its structure is analyzed through summary statistics and

information functions. Missing or inconsistent values are identified and handled to improve data quality. Data transformation techniques are applied to convert categorical attributes into numerical values, making them suitable for machine learning algorithms. Proper pre-processing ensures that the dataset is clean, consistent, and ready for further analysis.

2. Exploratory data analysis (EDA)

This module is responsible for understanding the patterns, trends, and relationships present in the dataset. Various visualization techniques and statistical methods are used to analyze feature distributions and correlations. Correlation matrices and heat maps help identify strong relationships among variables. Feature selection methods are applied to remove irrelevant or redundant attributes, improving model performance. EDA provides valuable insights into the data and helps in making informed decisions during model development.

3. Model training and evaluation

This module focuses on building and evaluating machine learning models for phishing detection. The dataset is divided into training and testing subsets using train-test splitting techniques. Multiple algorithms such as Random Forest, Decision Tree, Logistic Regression, and Support Vector Classifier (SVC) are trained using the prepared dataset. The performance of each model is evaluated using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and classification reports. This process helps determine the effectiveness of each algorithm.

4. Model comparison and visualization

This module compares the performance of different machine learning algorithms and presents the results through visual representations. Training accuracy, testing accuracy, and validation scores are compared using bar charts and graphs. Feature importance analysis is performed to identify the most influential attributes affecting predictions. Visualization techniques help users understand model behaviour and performance differences more effectively. This module supports the selection of the most suitable model for phishing detection.

5. Cross-validation and test size sensitivity

This module evaluates the stability and reliability of machine learning models under different conditions. Cross-validation techniques are applied to assess model performance across multiple data splits and reduce the risk of over fitting. Different test sizes, such as 20%, 30%, 40%, and 50%, are examined to study their impact on prediction accuracy. The results are analyzed and compared to identify the optimal training and testing ratio. This module ensures that the selected model performs consistently on unseen data.

6. Prediction

This module is responsible for predicting whether a given URL is legitimate or phishing using the trained machine learning model. New input URLs are processed through the same pre-processing and feature extraction stages used during training. The trained model analyzes the extracted features and generates a prediction result along with a confidence score. The system displays the classification outcome and relevant performance information to the user. This module enables real-time phishing detection and enhances cyber-security protection.

VI. RESULTS AND DISCUSSION

The results of the phishing detection using machine learning (ML) show promising accuracy levels. The analysis reveals that the implemented ML algorithms, such as decision tree, random forest, SVC, and Logistic Regression, perform well in phishing detection. Comparing the performance of these algorithms, Decision tree, Random forest, SVC, and Logistic Regression stand out as efficient predictors of phishing. The model shows potential in accurately predicting. The web-based application developed for phishing prediction is user-friendly and can facilitate early intervention and improved outcomes for individuals with phishing detection.

ML Algorithm	Accuracy
Random Forest	80%
Support Vector Classifier	84%
Decision Tree	89%
Logistic Regression	91%

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

OUTPUT AND SCREENSHOTS

	url	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and
0	http://www.crestonwood.com/router.php	37	19	0	3	0	0	0	0
1	http://shadetreetechnology.com/V4/validation/a...	77	23	1	1	0	0	0	0
2	https://support-appleId.com.secureupdate.duila...	126	50	1	4	1	0	1	2
3	http://rgipt.ac.in	18	11	0	2	0	0	0	0
4	http://www.iracing.com/tracks/gateway-motorspo...	55	15	0	2	2	0	0	0

5 rows × 89 columns

Fig No.6.1: Dataset Preview and Feature Overview

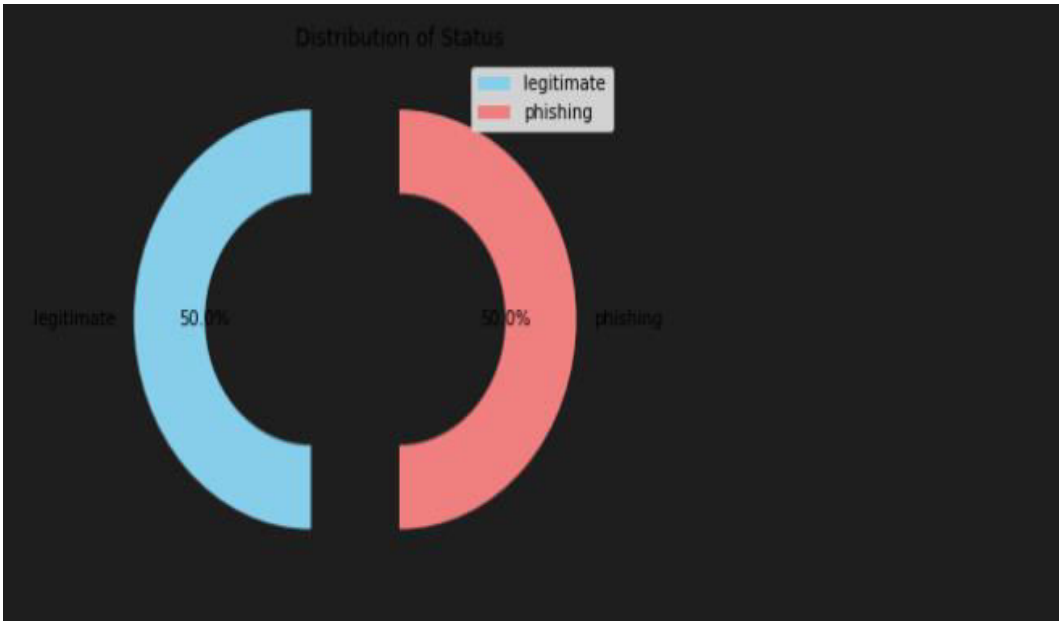


Fig No.6.2: Distribution of Legitimate and Phishing URLs

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

```
Training Accuracy : 0.9657022865142324
Logistic Regression Accuracy : 0.9132260321903429

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.92      0.91      0.91      1425
        Good       0.91      0.92      0.91      1433

    accuracy                           0.91      2858
   macro avg       0.91      0.91      0.91      2858
weighted avg       0.91      0.91      0.91      2858
```

Fig No.6.3: Logistic Regression Classification Report and Accuracy Results

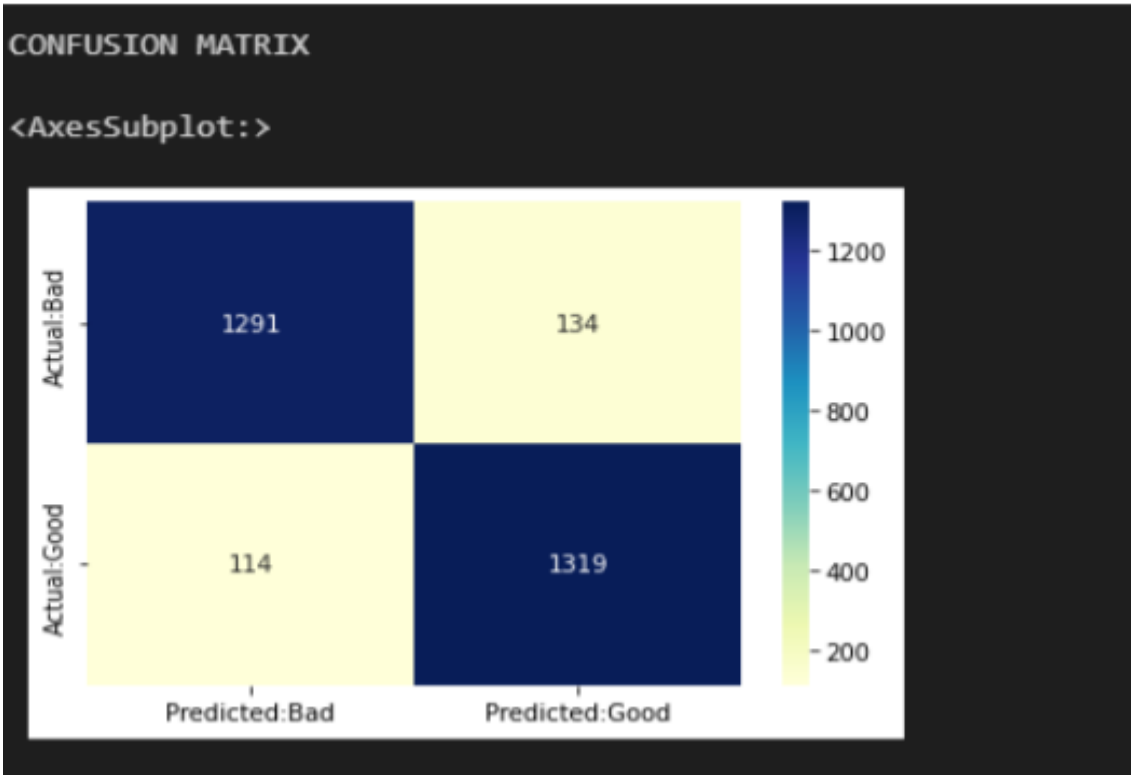


Fig No.6.4: Confusion Matrix of Logistic Regression Model

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

```
Training Accuracy : 0.8525431637890807
SVC Model Accuracy : 0.8414975507347796

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.81      0.86      0.83      1314
        Good       0.88      0.82      0.85      1544

    accuracy                           0.84      2858
   macro avg       0.84      0.84      0.84      2858
weighted avg       0.84      0.84      0.84      2858
```

Fig No.6.5: SVC Classification Report and Accuracy Results

```
CONFUSION MATRIX

<AxesSubplot:>

Training Accuracy : 0.8142790480634624
Training Accuracy : 0.8142790480634624
Random Forest Accuracy : 0.8023093072078377

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.98      0.72      0.83      1912
        Good       0.63      0.97      0.76       946

    accuracy                           0.80      2858
   macro avg       0.81      0.84      0.80      2858
weighted avg       0.86      0.80      0.81      2858
```

Fig No.6.6: Confusion Matrix of SVC Model

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

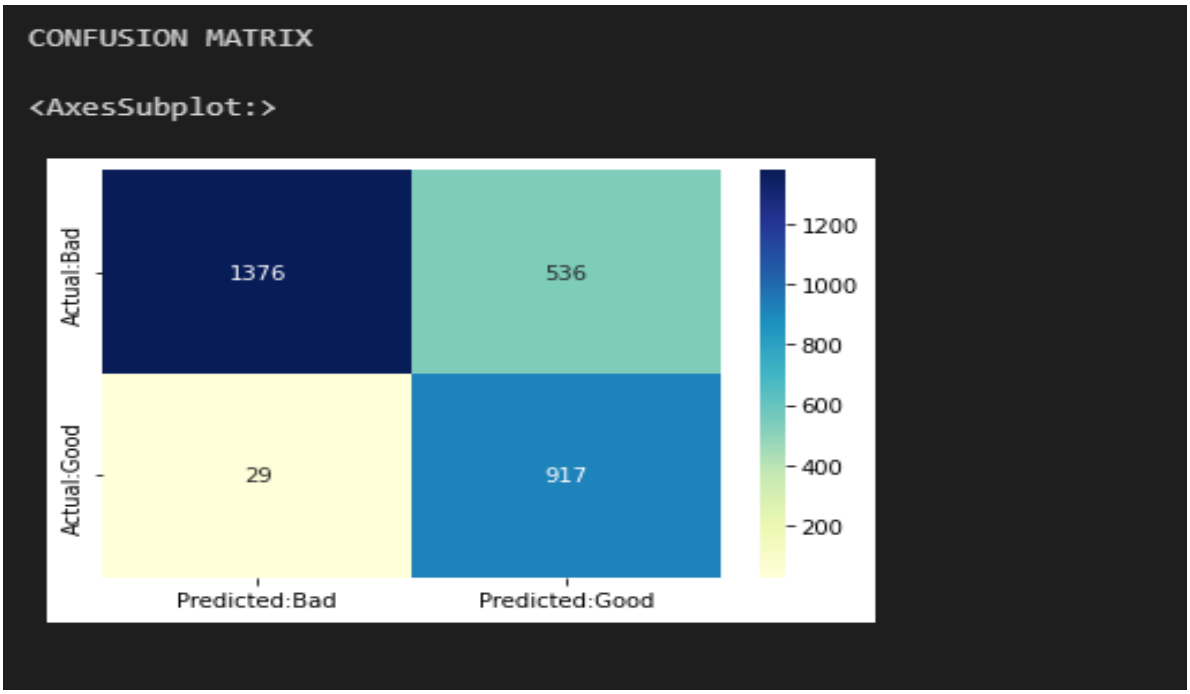


Fig No.6.8: Confusion Matrix of Random Forest Model

```
Training Accuracy : 1.0
Decision Tree Accuracy : 0.8922323303009098

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.89      0.89      0.89      1409
        Good       0.89      0.90      0.89      1449

    accuracy                           0.89      2858
   macro avg       0.89      0.89      0.89      2858
weighted avg       0.89      0.89      0.89      2858
```

Fig No.6.9: Decision Tree Classification Report and Accuracy Results

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406077

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

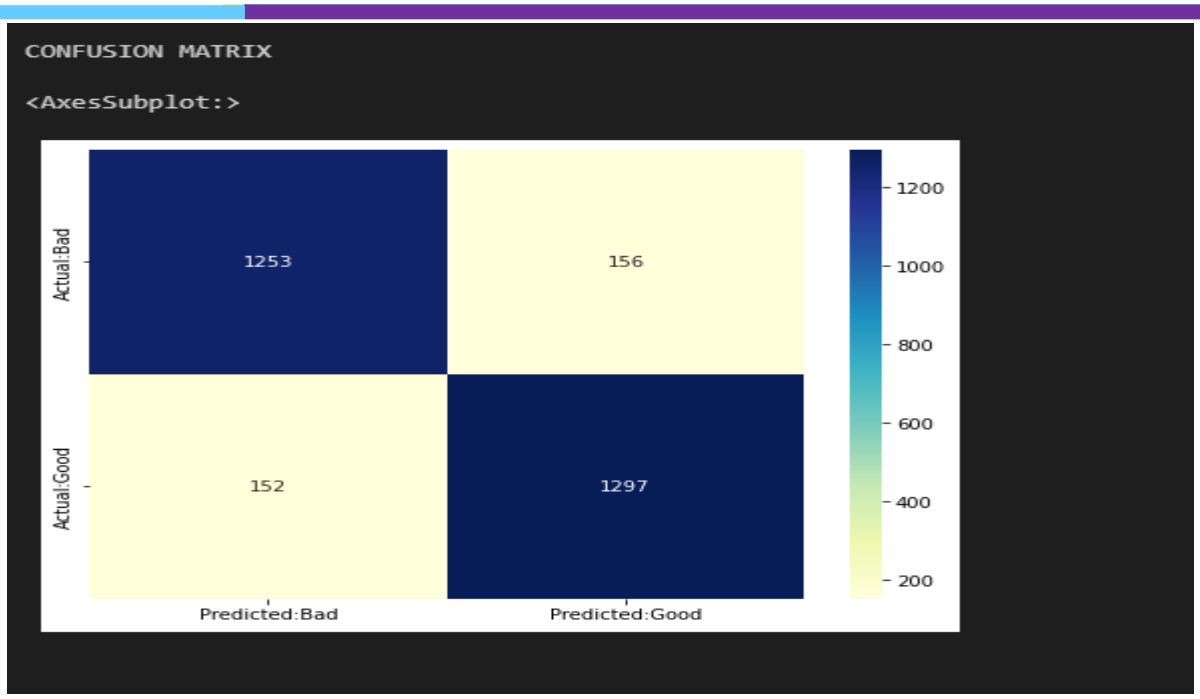


Fig No.6.10: Confusion Matrix of Decision Tree Model

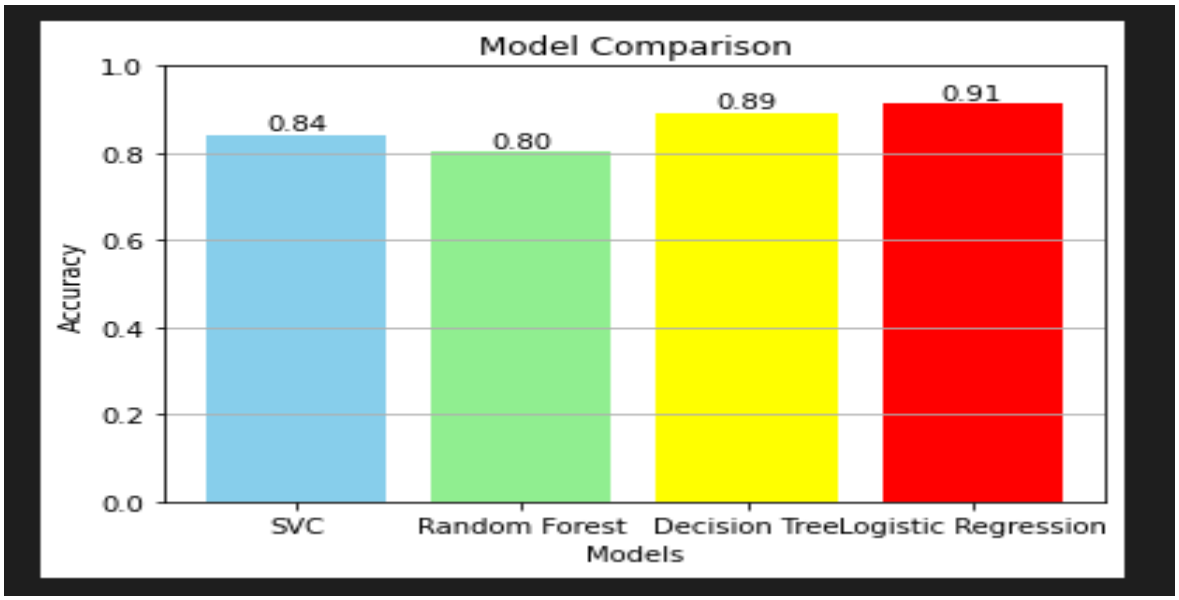


Fig No.6.11: Model Comparison

VII. CONCLUSION

The **Cyber Hacking Breaches Prediction and Detection Using Machine Learning** project successfully demonstrates the use of machine learning techniques for identifying phishing websites and potential cyber threats. By utilizing algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Classifier (SVC), the system effectively analyzes URL-based features and classifies websites as legitimate or phishing. Data pre-processing, feature extraction, and model evaluation techniques significantly improve the prediction performance. The

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

experimental results show that machine learning-based approaches can provide accurate, reliable, and scalable solutions for phishing detection. This system helps users avoid malicious websites and enhances cyber-security by providing timely and accurate threat identification.

VIII. FUTURE ENHANCEMENT

The proposed system can be further enhanced by integrating advanced deep learning techniques such as Convolution Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to improve the accuracy and efficiency of phishing website detection. Real-time threat intelligence feeds can be incorporated to identify newly emerging phishing attacks and cyber threats dynamically. The system can also be extended as a browser extension or mobile application, enabling users to receive instant security alerts while browsing the internet. Additionally, future versions may include the detection of other cyber threats such as malware, ransom ware, spam websites, and fraudulent online activities. The implementation of automated feature selection, continuous model updates, and cloud-based deployment can further improve scalability, performance, and real-time threat monitoring capabilities.

REFERENCES

[1] Sahingoz, O.K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345–357.

[2] Rao, R.S., & Pais, A.R. (2019). Detection of phishing websites using machine learning approaches. International Journal of Information Security, 18(1), 87–99.

[3] Mohammad, R.M., Thabtah, F., & McCluskey, L. (2015). Phishing websites features. International Journal of Cyber-Security and Digital Forensics, 4(1), 1–14.

[4] Aburrous, M., Hossain, M.A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing website detection system using fuzzy data mining. Expert Systems with Applications, 37(12), 7913–7921.

[5] Jain, A.K., & Gupta, B.B. (2018). Towards detection of phishing websites on client-side using machine learning based approach. Telematics and Informatics, 35(5), 1528–1538.

[6] Adebowale, M.A., Lwin, K.T., Sanchez, E., & Hossain, M.A. (2019). Intelligent web-phishing detection and protection scheme using integrated features of Images, Frames and Text. Expert Systems with Applications, 115, 300–313.

[7] Basnet, R., Mukkamala, S., & Sung, A.H. (2014). Detection of phishing attacks using machine learning. International Journal of Information Security and Privacy, 8(4), 45–62.

[8] Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S.C., & Tiong, W.K. (2018). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences, 484, 153–166.

[9] Verma, R., & Das, A. (2017). What's in a URL: Fast feature extraction and malicious URL detection. Proceedings of the ACM Conference on Data and Application Security and Privacy, 55–63.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details