Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

AI-BASED PHISHING SITE IDENTIFICATION USING SVM AND LIGHT GBM

Mr. K. JAYA KRISHNA, T. VENKATA PRASAD
#1 Associate Professor Department of Master of Computer Applications
#2 Pursuing M.C.A
QIS COLLEGE OF ENGINEERING & TECHNOLOGY
Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272

ABSTRACT

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as using light gbm and svm algorithm.

Keywords: URL, SVM, Light GBM, Cyber security, phishing website.

INTRODUCTION

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world. Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualizes and associations have lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection challenge while using artificial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and

their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition. Opposite to our recognition system sinceit needs only six features completely extracted from the URL as input. In this paper, after a summary of these Feld key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last but not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

LITERATURE SURVEY

Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning: Integrating Stacked Autoencoder and SVM

• Author: Naga Sushma Allur

• Merits: Proposes a hybrid model combining stacked autoencoder for dimensionality reduction with SVM for classification, achieving improved accuracy and reduced false positives in phishing detection.

• Demerits: The complexity of the model may lead to longer training times and higher computational costs. • Reference: Journal of Science & Technology, 2020jst.org.in

URL Based Detection of Phishing Using Random Forest Algorithm and SVM • Authors: Usha Kiruthika, Kowshik Varma, Sunvith Bangarraju

• Merits: Focuses on URL-based features for phishing detection, utilizing Random Forest and SVM algorithms to classify websites efficiently.

• Demerits: Limited to URL features; may not capture content-based phishing tactics.

• Reference: International Journal of Advanced Science and Technology, 2020SERSC

Implementation of Phishing Detection using SVM

• Authors: A. Christy Jeba Malar, R. Kanmani, Vijayavarman R, Praveen Kumar R, Poorna Bharathi G

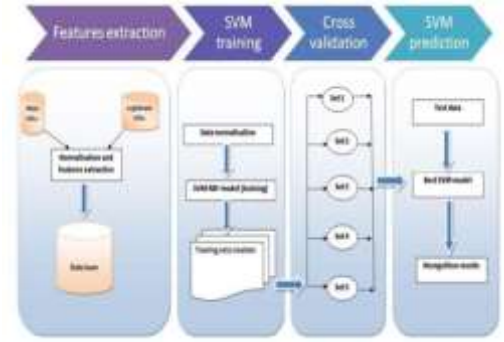
• Merits: Employs various SVM kernels to detect phishing websites based on lexical and host properties, achieving over 90% accuracy.

• Demerits: The model's performance heavily depends on the quality and diversity of the training dataset. • Reference: TEST Engineering & Management, 2020Test Magazine

4. Phishing URLs Detection Using Naives Bayes, Random Forest, and LightGBM Algorithms • Authors: Cik Feresa Mohd Foozy, Muhammad Amir Izaan Anuar, Andi Maslan, Husaini Aza Mohd Adam, Hairulnizam Mahdin • Merits: Compares multiple algorithms, including LightGBM, for phishing URL detection, providing insights into their performance across different datasets. • Demerits: The study's scope is limited to URL-based features, potentially overlooking other phishing indicators. • Reference: International Journal of Data Science, 2024ijods.org

5. Phishing Website Detection Using Advanced Machine Learning Techniques • Authors: Nitin N. Sakhare, Jyoti L. Bangare, Radhika G. Purandare, Disha S. Wankhede, Pooja Dehankar • Merits: Integrates multiple machine learning algorithms, including LightGBM, to analyze URL structures and content patterns for phishing detection. • Demerits: The complexity of the model may lead to challenges in real-time deployment. • Reference: International Journal of Intelligent Systems and Applications in Engineering, 2020IJISAE

SYSTEM ARCHITECTURE:



RELATED WORKS

Phishing is a prevalent cyber threat involving fraudulent websites that mimic legitimate ones to steal user credentials and sensitive data. AI-based methods have significantly enhanced phishing site detection by enabling real-time, automated, and intelligent identification of malicious web patterns. Researchers have developed a wide range of techniques using machine learning (ML), deep learning (DL), and hybrid approaches.

1. Traditional Blacklist and Heuristic Methods

Early phishing detection systems relied on blacklists (e.g., Google Safe Browsing, PhishTank) and rule-based heuristics.

Zhang et al. (2007) introduced CANTINA, a content-based heuristic method using TF-IDF to detect phishing websites.

Merits: Easy to deploy and interpret.

Demerits: Cannot detect zero-day phishing websites or cleverly disguised domains.

2. Machine Learning-Based Approaches

AI-powered systems have significantly improved the detection of unknown phishing websites using feature-based classification.

Abdelhamid et al. (2014) applied Decision Trees and Naïve Bayes classifiers to identify phishing using lexical and URL features.

Ma et al. (2009) explored URL-based detection using features like domain age, length, and use of special characters, showing high detection rates with Random Forest and SVM models.

Rathore et al. (2018) used feature reduction techniques with ensemble classifiers to increase speed and reduce false positives.

Merits: Can detect previously unseen phishing domains.

Demerits: Requires effective feature engineering and large labeled datasets.

3. Deep Learning Approaches

Deep learning techniques can automatically learn features from URLs or page content, improving accuracy without manual feature extraction.

Bahnsen et al. (2017) developed a character-level LSTM model for detecting phishing URLs, which effectively captured sequence patterns in domain names.

Mohammad et al. (2021) implemented a CNN-based model on screenshots of phishing sites, using image-based detection.

Li et al. (2020) combined BERT-based embeddings with website metadata for phishing classification.

Merits: High accuracy and adaptability to new attack patterns.

Demerits: Require more computational resources and data.

4. Hybrid and Ensemble Techniques

Hybrid systems combine multiple models or feature sets to improve robustness and reduce false positives.

Jain & Gupta (2018) proposed a hybrid approach combining URL, domain, and visual features using Random Forest and Gradient Boosting.

Sahoo et al. (2019) developed an ensemble of SVM, Naïve Bayes, and KNN classifiers using a meta-classifier for final prediction.
Merits: More accurate and resilient.
Demerits: Complex to train and tune.
5. Feature Engineering and Selection
Effective phishing detection often relies on carefully selected features:
Le et al. (2019) focused on lexical features (e.g., URL length, use of digits, special characters).
Marchal et al. (2016) introduced web traffic-based features like bounce rate and domain popularity to enhance detection.
Merits: Improves interpretability and model efficiency.
Demerits: Feature relevance may change over time.
6. Real-Time and Edge-Based Systems
Zhou et al. (2021) implemented lightweight models for real-time phishing detection on browsers and mobile apps.
Shirazi et al. (2020) proposed browser extensions using deep learning to flag phishing attempts in real time.
Merits: Practical for user protection on the fly.
Demerits: Trade-offs between speed and model depth.
7. Datasets and Evaluation
Commonly used datasets include:
PhishTank
UCI Phishing Website Dataset
OpenPhish
URLNet Dataset
SYSTEM ANALYSIS
Existing System
Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers.
PROPOSED SYSTEM:
Phishing attacks have evolved in terms of sophistication and have increased in sheer number in recent years. This has led to corresponding developments in the methods used to evade the detection of phishing attacks, which pose daunting challenges to the privacy and security of the users of smart systems. This study uses LightGBM and features of the domain name to propose a machine-learning-based method to identify phishing websites and maintain the security of smart systems. Domain name features, often known as symmetry, are the property wherein multiple domain-name-generation algorithms remain constant. The proposed model of detection is first used to extract features of the domain name of the given website, including character-level features and information on the domain name. The features are filtered to improve the model's accuracy and are subsequently used for classification. The results of experimental comparisons showed that the proposed model of detection, which integrates two types of features for training, significantly outperforms the model that uses a single type of feature. The proposed method also has a higher detection accuracy than other methods and is suitable for the

real-time detection of many phishing websites.

IMPLEMENTATION

Modules

1. URL Acquisition Module

Function: Collects and monitors URLs from various sources for analysis.

Sources:

Real-time user input (e.g., browser extension)

Email links

Public phishing databases (e.g., PhishTank, OpenPhish)

Crawled web data

Tools: Web scrapers, APIs

2. Feature Extraction Module

Function: Extracts useful features from each URL and webpage.

Types of Features:

Lexical Features: URL length, number of digits/special characters, presence of IP addresses.

Domain Features: Age, WHOIS data, registration country.

Content-Based Features: Presence of login forms, suspicious keywords, iframes, pop-ups.

Visual Features (optional): Screenshots of websites for image-based classification.

3. Preprocessing Module

Function: Cleans and formats raw data before feeding it into the model.

Tasks:

Encode categorical variables (e.g., domain registrar)

Normalize numeric features

Remove duplicates and irrelevant entries

Handle missing values

4. Feature Selection Module (Optional)

Function: Reduces dimensionality and selects the most predictive features.

Techniques:

Recursive Feature Elimination (RFE)

Information Gain

PCA or L1-Regularization

5. Classification Engine

Function: Predicts whether a URL is phishing or legitimate.

Approaches:

Machine Learning Models: Random Forest, Decision Trees, SVM, Naïve Bayes

Deep Learning Models:

CNNs for visual or textual patterns

RNNs/LSTM for sequential URL analysis

BERT or Transformers for contextual understanding of web content

6. Model Training and Evaluation Module

Function: Trains the AI models and evaluates performance.

Tasks:

Train/test/validation data split

Cross-validation

Hyperparameter tuning

Performance metrics (Accuracy, Precision, Recall, F1-Score, AUC)

7. Real-Time Detection Module

Function: Applies the trained model in real-time to detect phishing attempts.

Deployment Platforms:

Browser extension

Email gateway filter

Cloud-based threat detection API

8. Alert and Response Module

Function: Responds to detected phishing threats.

Actions:

Block or warn users

Redirect to a warning page

Notify administrators

Add URL to blacklist

9. Logging and Analytics Module

Function: Logs analysis results and generates performance reports.

Includes:

URL logs

Detection results

False positives/negatives

User interaction reports

10. Feedback and Retraining Module (Optional)

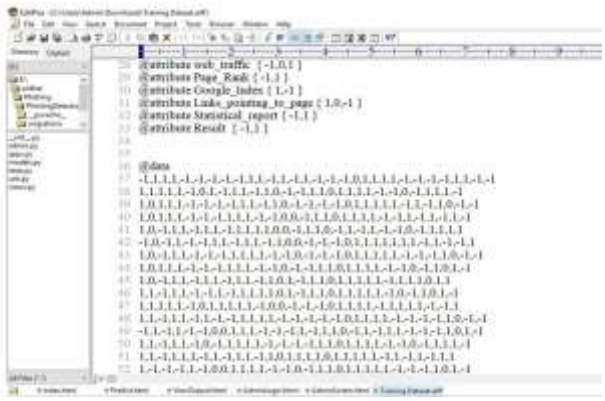
Function: Incorporates user or system feedback to improve the model.
Mechanism:
User reports false positives/negatives
Add new labeled samples to dataset
Retrain periodically to handle new phishing techniques
11. Security and Privacy Module
Function: Ensures user data and detection results are protected.
Features:
Data encryption
Secure logging
GDPR-compliant data handling
Methodology
The methodology consists of several stages including data acquisition, feature engineering, model development, and real-time classification.
1. Data Collection
Sources:
Public phishing datasets: PhishTank, OpenPhish, UCI Phishing Websites Dataset
Legitimate websites: Alexa Top Sites, Dmoz Directory
Data Format:
Raw URLs
Webpage content (HTML, scripts)
WHOIS and DNS information
Screenshots (for visual analysis)
2. Data Preprocessing
Cleaning:
Remove duplicates and broken links
Validate label consistency (phishing vs legitimate)
Text Normalization:
Lowercase URLs
Strip special characters or encode them for sequence models
Balancing Dataset:
Use oversampling (e.g., SMOTE) or undersampling to handle class imbalance between phishing and legitimate sites.
3. Feature Extraction
This stage transforms raw data into numerical representations suitable for AI models.
A. URL-based Features:
Length of URL, use of HTTPS
Number of subdomains, special characters, IP address usage
Presence of shortening services (e.g., bit.ly)
B. Domain-based Features:
Domain age
WHOIS registration details
Presence in blacklists
C. Content-based Features:
HTML form analysis (e.g., password fields)
JavaScript redirection or iframe usage
Keyword frequency (e.g., “login”, “verify”, “secure”)
D. Visual-based Features (Optional):
Page screenshots converted to image features using CNNs
4. Model Building
Choose and train appropriate models for phishing classification.
A. Machine Learning Models:
Algorithms: Decision Trees, Random Forest, Naïve Bayes, SVM, XGBoost
Feature Selection: Recursive Feature Elimination (RFE), Information Gain
B. Deep Learning Models:
Character-Level LSTM/CNN: Processes raw URL strings as sequences
Transformer/BERT Models: Analyze content and metadata contextually
Image-Based CNNs: Used if using webpage screenshots as input
5. Model Training & Evaluation
Dataset Split:
Training: 70%, Validation: 15%, Testing: 15%
Performance Metrics:
Accuracy
Precision, Recall, F1-score
ROC-AUC (to measure overall detection capability)
Validation Techniques:
k-Fold Cross-validation

Confusion matrix analysis
6. Phishing Detection System Design
Integrate the trained model into a real-time system.
Real-Time Classification:
Input: User-submitted or scraped URL
Output: Phishing or Legitimate
Response Actions:
Show warning message
Block access or redirect
Notify administrator/log event
7. Feedback Loop and Retraining
User Feedback:
Collect false positive/false negative reports from users
Model Update:
Retrain periodically with new labeled data to adapt to evolving phishing tactics
8. Security & Privacy Considerations
Anonymize user-submitted URLs if stored
Secure API endpoints and prediction services
Comply with data regulations (e.g., GDPR)
RESULTS AND DISCUSSION

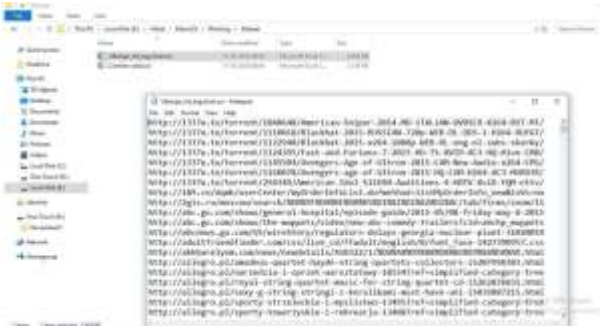
Fig 1

Fig 2

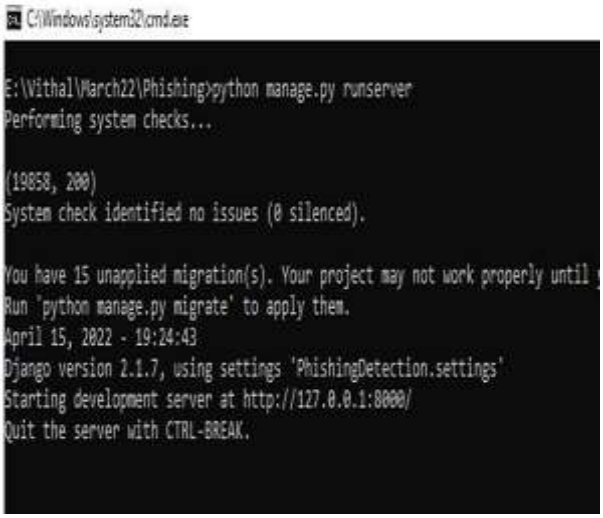
FUTURE SCOPE A
Implementation
In this project we are implementing SVM and Light GBM machine learning algorithms o detect phishing website URLS. We are training all these algorithms with normal and phishing URLS and build a trained model and this train model will be applied on new TEST URL to detect whether its normal or phishing URL. In this project you asked to use UCI machine learning phishing dataset but this dataset contains only 0's and 1's values like below screen



From above dataset ML algorithms can get trained but we can't understand anything so I am using REAL WORLD URL dataset which contains normal and phishing URLS like below screen



In above screen you can see our dataset contains 2 folders called benign (phishing URLS) and valid (normal URL) and this are real world URLS and we will train all algorithms with above dataset and then when we input any test URL then ML model will predict as normal or phishing To run this project double click on 'run.bat' file to start python DJANO server like below screen

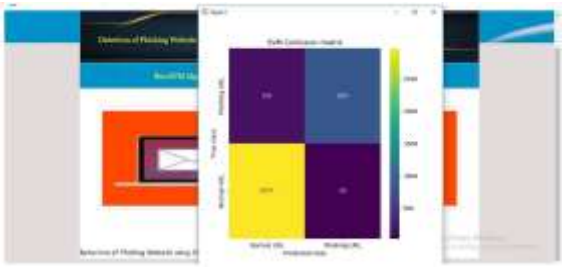


In above screen DJANGO webserver started and now open browser and enter URL http://127.0.0.1:8000/index.html and press enter key to get below output

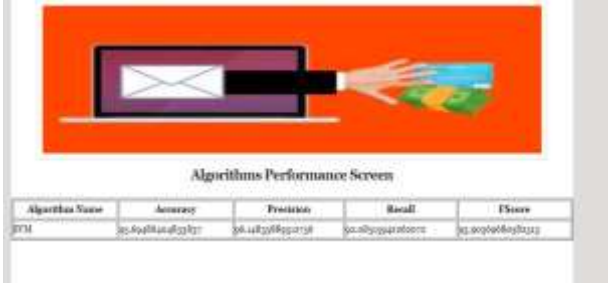
In above screen click on 'Admin Login Here' link to get below login screen

In above screen enter username and password as 'admin' and 'admin' and then press button to get below output

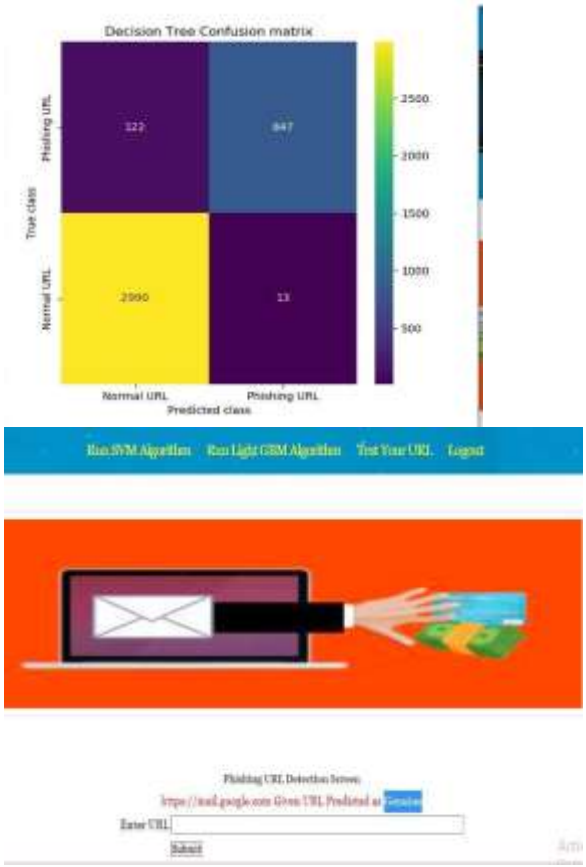
In above screen click on 'Run SVM Algorithm' link to train SVM algorithm and get below output



In above screen we can see SVM confusion matrix where x-axis represents predicted class and y- axis represents TRUE class and we can see SVM predict 2977 records correctly as NORMAL and only 145 are incorrect prediction and it predict 824 records as PHISHING URL and only 26 are incorrect prediction and now close above graph to get below output



In above screen with SVM we got 95% accuracy and now click on 'Run Light GBM Algorithm' link to get below output



In above screen in blue colour text we can see given URL predicted as GENUINE (normal) and now test other URL. Similarly now I will enter Google.com in below screen

CONCLUSION AND FUTURE ENHANCEMENTS

The features of the domain name used here can be obtained only by using known strings of domain names without obtaining information related to user privacy, such as traffic in the network. Features of the domain name can be divided into two categories according to the acquisition method: features of the characters used in the domain name and features of information on the domain name. The features of information on the domain name can be obtained through the corresponding website or other query websites to this end, whereas the features of the characters used in the domain name can be obtained through a local feature-extraction algorithm without visiting the website.

REFERENCES

[1] Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)

[2] Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.

[3] Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)

[4] R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

[5] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)

[6] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL- Based Phishing Detection.13 June 2021

[7] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021

[8] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.

[9] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.

[10] Valid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

[11] A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.
[12] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.
[13] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-14

Mr. K. Jaya Krishna is an Associate Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai, and his M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). With a strong research background, he has authored and co-authored over 90 research papers published in reputed peer-reviewed Scopus-indexed journals. He has also actively presented his work at various national and international conferences, with several of his publications appearing in IEEE-indexed proceedings. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. TALLAPALEM VENKATA PRASAD has received his MCA (Masters of Computer Applications) from QIS college of Engineering and Technology Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272 affiliated to JNTUK in 2023-2025