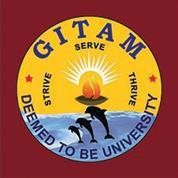
**GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

**(Deemed to be University)**

**Bengaluru-561203**



**“DETECTING PHISHING WEBSITES USING AI-ML*”***

**Department of Electronics and Communication Engineering**

Submitted by:

**NAMES REGISTER NUMBERS**

**Kodidela Dinesh Naidu 322010403006**

**Boksam Nithin Varma 322010403031**

**Mange Amith Das 322010402033**

**R Nikhil Kumar 322010402004**

**Under the Guidance of**

**Dr.Ramesha M**

Assistant professor, Department of Electronics and Communication Engineering,

GITAM School of Technology

GITAM

(DEEMED TO BE UNIVERSITY)

(Estd. u/s 3 of the UGC act 1956)

NH 207, Nagadenehalli, Doddaballapur taluk,

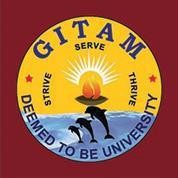
Bengaluru-561203

Karnataka, INDIA.

**GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

**(Deemed to be University)**

**Bengaluru - 561203**



Department of Electronics and Communication Engineering

Certificate

*This is to certify that the project titled “Detecting phishing websites using AI-ML” is the bona fide work carried out by Kodidela Dinesh Naidu (322010403006), Boksam Nithin Varma (322010403031),Mange Amith Das(322010402033),R Nikhil Kumar(322010402004),the students*

*of B Tech (ECE) of GITAM Deemed to be University, Bengaluru campus during the academic year 2023-2024, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology (Electronics and communication Engineering ) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.*

*Signature of the Guide Signature of HOD*

|  |  |
| --- | --- |
| **Dr.RAMESHA M**  Assistant Professor  Department of ECE, GST |  |
|  |  |
|  |  |

**Dr.PRITHVI SEKHAR**

Head of the Department

GITAM School of Technology

## DECLARATION

We **Kodidela Dinesh Naidu, Boksam Nithin Varma, Mange Amith Das and R Nikhil Kumar** students of 8th semester B.Tech in Electronics and communication Engineeringfrom GITAM (Deemed to be University),Bangalore, hereby declare that the dissertation work entitled **“Detecting phishing website using AI-ML ”** has been carried out under the guidance of  **DR RAMESHA M,** Assistant Professor Department of, GITAM (Deemed to be University), Bangalore, in the partial fulfilment of the requirement of the degree of the **BACHELOR OF TECHNOLOGY IN ELECTRONICS AND COMMUNICATION ENGINEERING****OF GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM).**We declare

that we have not submitted this dissertation either inpass or in full to any other University for the award of any degree.

#### Signature of the Students

**Place:** Bengaluru

**Date:**

# ABSTRACT

Phishing is the most generally perceived and most unsafe assaults among cyber-crimes. The purpose of these assaults is to take the records utilized by people and relationship to lead trades. Phishing destinations involve stand-out notice signs among their substance and web program fundamentally based realities. This major project proposes a novel approach to tackle this challenge by leveraging Artificial Intelligence (AI) and Machine Learning (ML) techniques. The primary objective is to develop an intelligent system capable of accurately identifying phishing websites, thereby enhancing cybersecurity.

Random Forest is a versatile machine learning algorithm that combines the strength of multiple decision trees to make accurate predictions. By building an ensemble of diverse trees and leveraging random feature selection and bagging techniques, Random Forest effectively handles complex datasets and mitigates the risk of overfitting. Its ability to provide reliable predictions and feature importance analysis makes it a popular choice for various classification and regression tasks in real-world applications.

## II

## TABLE OF CONTENTS

|  |  |
| --- | --- |
| **Title** | **Page No.** |
| Declaration | I |
| Abstract | II |
| Table of Contents | III |
| List of Figures | IV |
|  |  |
|  |  |
| 1. INTRODUCTION 2. Literarture Survey | 6  9 |
| 1. Problem Identification and Objectives   3.1 Solution for the problem statement   1. Methodology    1. Data Pre-processing 2. Overview of Technologies   5.1.Hardware and Software Requirements  5.2.Hardware Requirements  5.3.Software Requirements  6 Implementation   * 1. Coding      1. Data Pre-processing      2. Splitting      3. Features Extraction      4. Classification of URLs using Random forest  1. Results    1. Iteration 1    2. Iteration 2 2. Conclusion 3. References     III | 13  13  15  16  19  19  22  22  23  23  23  23  24  26  28  28  29  30  31 |
|  |  |
|  |  |
|  |  |

**LIST OF FIGURES**

**Figure No. Page No.**

|  |  |
| --- | --- |
| 1. Flow chart of the proposed system | 18 |
| 2. Block diagram of Random Forest Prediction | 27 |
| 3. Input screen of Iteration1 | 29 |
| 4. Output Screen of Iteration 1 | 29 |
| 5. Input screen of Iteration 2 | 30 |
| 6. Output Screen of Iteration 2 | 30 |

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER I**

**INTRODUCTION**

Phishing is a very in famous method of operation in cybercrime activities. Phishers began to make money and started doing this in a very organized fashion. These phishers have used a range of techniques to conduct threats on unsuspecting users such as VOIP, online messages, authentic false websites, spoofed connections and legitimate looking URLs.

Moreover, attackers raise security inquiries to answer. Once users answer those queries, they are just forced into phishing attacks without their knowledge. Machine learning algorithms are also used to detect the phishing websites as they are one of the most powerful techniques for this application. Attacks on social engineering can be a large-scale security hazard accustomed to revealing personal and sensitive information by simply tricking users into giving them out without being detected. In keeping with, net spoofing or phishing method is one amongst social engineering attacks. The spoofed links are placed on widespread websites or sent through email to the victim. URLs of phishing websites usually contain some distinctive characteristics which build completely unique from URLs of the legitimate website.

Phishing attacks may be launched via causing Associate in Nursing mail that seems to be sent from a public or personal organization to users by attackers. Various ways like blogs, forums and file sharing may be used by attackers for phishing. The systems of intelligence offer clean edges in deciding to inform the security professionals. A phishing attack is once a criminal tries to send Associate in Nursing electronic mail or the Uniform Resource Locator dissimulation to be something or somebody he’s not, to urge sensual information from the victim. The type of thievery and fraud that would come about by simply exploiting the small print of some organizations or someone’s account which can’t extremely be imaginary.

DETECTING PHISHING WESITES USING AI-ML

A URL may be an international reference through the address of a document within the WWW, and its servers because the basic means to find a doc on the web. One of the most downsides with machine learning-based approaches to this downside is that only few other training information sets usually contains phishing URLs which can be obtained within the right property.

The Internet plays a vital role in everyone’s lives. It has become a convenient and valuable mechanism to support public transactions such as e-commerce and e-banking. Fraudsters attempt to deviate user by giving them their info supported by exploiting human vulnerabilities instead of package vulnerabilities. Several researchers used the URL and linked it to current blacklists containing lists of fake websites they have created.

The US-CRET commonly known as United States Computer Emergency Readiness Team formulated phishing as a kind of social engineering that takes electronic mails and sensible information from an individual or any kind of organization by being as the most trustworthy entity. This is how most of the entities steal the most sensible information from an individual or any organization. The solution to overcome from phishing attacks within the organizations is by educating the employees on how to identify phishing emails or links so that it will be helpful for the organization to detect the phishing attacks.

The major problem with the usage of machine learning-based approaches are only less amount of data sets are made available with finite URLs in the public domain. The usual machine learning algorithm that are used for classifying the Uniform Resource Locators are Naive Bayes classifier, SVM, neural network and decision tree. One of the best approaches is by using a blacklist of fake URLs predicated by the groups of anti-viruses. The drawback of this blacklist strategy being that the blacklist cannot be finished due to new fake URLs appearing instantly.

Phishing attacks can not only be done by sending an electronic mail that seems to be coming from a trusted device or any private organizations. The other ways of carrying out phishing attacks include methods such as file conversion, in the context of blogs and forums, which are often used by attackers to capture personal details from organizations and other individuals. Moreover, there are various ways to stop the phishing attacks and that includes

DETECTING PHISHING WESITES USING AI-ML

providing adequate education, legal solution and also technical solutions. In the present days, majorly information and the communication devices are used in a way which is highly dense with the details.

Many people such as researchers have been using machine learning approaches as it is a subject of computer science and even a sub dividend of the artificial intelligence. Artificial intelligence executes various tasks and is also able to learn or move in an intelligent way. Machine learning has two types of learning and there are supervised learning and unsupervised learning. Supervised learning is nothing but providing a training model with a set of measuring tasks which are associated with certain targets and after when the current model is instructed to produce a new target with unknown details. Now talking about unsupervised learning in this type of learning there will no data given in prior and hence it generates new data without giving certain instructions or training.

Few authors found a method to determine the types of attacks occurred through phising by using machine learning and natural language processing. This technique helps to find a syntactic kind of analysis of the text to determine malicious goals. This NLP technique Helps to decipher every sentence and senses verbal word employment. There is another tool called anti-phishing toolbar and it is mainly used for upholding the authenticity of the Uniform Resource Locator. This toolbar is performed with scripts of python and also a dataset Nazario phishing electronic mail is utilized. After the performance, the results are obtained with precision 98% and 95% respectively.

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER II**

**LITERATURE SURVEY**

R Khiruthiga, et al(sep 2019) has expressed their significant approach for the detection of websites that phish, using a strong and reliable machine learning algorithm that will be increasingly efficient in detecting such phishing websites. From a few of the algorithms used namely Random forest, Decision tree and Gradient boosting the effective algorithm was the random forest which has given 98.4% accuracy, 98.59% recall and 97.70% accuracy. There is also a toolbar that exists for anti-phishing which is used to verify any URL and also to verify the authenticity of the corresponding URL. Few authors also created the extension that could be added to Google Chrome to detect such phishing websites which used a variety of machine learning algorithms to perform the function.

Amani Alswailem, et al(jul 2019)has expressed that phishing has become a horrendous problem for many organizations. However, this can occur in two ways such as one is by receiving suspicious websites which leads to fraudulent sites and second by clicking on the irrelevant links that redirect the user to a website that phishers use to make fraudulent advances. Back in the early ’90s, the phishers used to create a fraudulent website with a false identity for many online services providers.

Mamun, who researched this horizon found that the random forest algorithm has the highest accuracy when compared among the results of the KNN algorithm and decision tree classifier. The focus was on the four ways of the malevolent use of these URLs namely the categories defacement, phished, spam and malware.

Vaibhav Patil, et al (aug 2018)has expressed their view by bringing up the new approach that is by using the blacklist and whitelist approach to detect phishing websites. Minaxi Gupta implemented the blacklist approach to identify phishing websites. However, by utilising a well-suited matching algorithm and in addition to using heuristics it is very easy to identify phishing websites.

DETECTING PHISHING WESITES USING AI-ML

Heuristics help in by creating new URLs with the combination of parts from known phishing websites in the blacklist that is available to us after the matching algorithm calculates the score of URL and if the score results of the calculation are more than the corresponding acceptable value then it can safely label the particular URL of the website as a phishing website.

Arun Kulkarni, et al(2019) has expressed Machine Learning techniques that detect phishing URLs which are based on some URL features and they are host-based and lexical features. The role played by the host-based is to identify the characteristics of the website such as the location, who regulates it and when was the site installed. On the other side, lexical features help in the description of the textual properties of a URL, of which URLs are simply text strings that majorly contain a protocol, hostname, and a path of the web. To categorize the phishing URLs, it is being classified into three types of websites namely as phishing websites, suspicious websites and legitimate websites; and this is named as a multi-based classifier by Abdelhamid which is precisely used for detecting phishing URLs.

As indicated by Institute of Research Engineers and Doctors, USA, phishing location procedures are fundamentally separated into two methodologies and they are boycott based and heuristic-based methodologies. The boycott based methodology is utilized to keep up a database rundown of addresses (URLs) of those locales that are named as vindictive. In the event that a client demands a site and if that is remembered for this rundown, the association consequently gets blocked. The boycott based methodology incorporates points of interest of simple usage and a low bogus positive rate [5]; nonetheless, it has an imperfection that it can't distinguish phishing destinations which are not referenced in the database, and that incorporates brief locales.

As per the International Journal of Advanced research and imaginative thoughts in instruction (IJARIIE) diary paper, the Multi-Label Classifier based Associative Classification (MCAC) information mining approach is additionally one of the procedures that is explicitly utilized for recognizing phishing sites. The affiliated arrangement calculation recognizes phishing sites with average precision.

DETECTING PHISHING WESITES USING AI-ML

MCAC involves three principle steps which are Rule disclosure, classifier building and class task. Be that as it may, going to the initial step of this calculation, rules are found and extricated by emphasizing over the preparation informational index (verifiable sites traits or information gathered from numerous sources). This progression is tied in with, converging of any of the subsequent guidelines that have a similar forerunner (left hand side) that happens and furthermore connected with various classes to deliver the multi- mark rules. Alongside this, excess principles are likewise wiped out. The resultant of the subsequent advance is the classifier which suits both single and multi-mark rules. The last advance comprises of testing the classifier on a testing informational collection to process its exhibition. In the forecast procedure, the standard in the classifier bunch which coordinates the test information properties is regularly terminated to figure its sort (class). The MCAC calculation produces decides further that rules are ordered by utilizing arranging calculations. The significant imperfection that MCAC confronted was discovering trouble in deciding least certainty and least help when there is a lot of information and later on there came progressively advanced calculations to supplant this which were increasingly precise and had lesser time multifaceted nature. Utilize different lexical and host-based highlights of the URL for grouping, however bar website page content. In the event that appropriately computerized, this procedure can manage the cost of the low order overhead of boycotting while at the same time offering far more prominent exactness. In this proposed framework manufactured a URL characterization framework that utilizes a live feed of named URLs from an enormous web mail supplier, and that gathers highlights for the URLs continuously

As a result of the colossal proportion of posts (more than billions) by means of online systems administration media, genuinely checking each post to get the spams is unbelievable. flexible powerful learning approach proposed to truly affirm anyway numerous spams as could be permitted This structure has a couple of focal points thus gathering spam practices in casual association, Introducing both picture and substance features and relational association features to exhibit spam practices Integrating with our GAD batching estimation to manage gigantic extension data and Introducing a versatile unique adapting approach to manage recognize existing spams with limited human

undertakings, and perform online unique making sense of how to perceive spams consistently.

DETECTING PHISHING WESITES USING AI-ML

Weibo Chu et al study the evaluation the adequacy of AI based phishing affirmation utilizing essentially lexical and space highlights, which are open in any case, when the phishing Webpages are blocked off. Different phishing Webpages are momentary, routinely under 20 hours, and URLs may change now and again (energetic flux).in this appraisal work essentially lexical highlights and locale highlights are utilized in our phishing affirmation. These highlights are speedily accessible without getting to the Webpage, and as such can be utilized whether the phishing URLs are not, presently open. Getting to a sketchy Web Page may bring extra dangers since the present phishing Webpages may contain vindictive code, for example, Secure bank Phishing Trojan. In his examination work displays that phishing identifier is altogether productive even with the reduced sorts of discriminative highlights, with affirmation rates superior to 98% with fake positive rates at 0.64% or less.

Other author proposed a sketchy URL obvious proof structure for use in easygoing affiliation conditions subject to Bayesian depiction aggressors may utilize SNSs as vehicles, utilizing traded off client records to post messages that contain noxious URLs. Social affiliation clients generally trust the data that their accomplices submit in posts and feeds; thusly, they become the difficulties of social structure assaults. Noxious URLs utilized in easygoing affiliations may utilize the trust and social affiliations which take in the wake of phishing goals in spam; along these lines.

Wei Xu et al. made procedures to see spammers in easygoing affiliations, and amassed their messages in gigantic spam battles. To amass the information about spamming improvement, make a giant and unmistakable approach of "nectar profiles" on three massive individual to singular correspondence districts, and note the sort of contacts and messages that they got then dissected the gathered information and perceived curious lead of clients who showed up at that profile.

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER III**

**PROBLEM IDENTIFICATION AND OBJECTIVES**

Phishing is one of the most common and dangerous attacks among cybercrimes. The motive of this attack is to hack the user data by accessing the credentials that's employed by people and any of the organizations. This project aim is to sight the accuracy of a phishing web site supported the options that we've got taken. It tells however correct the phishing web site is, and vulnerable the web site is for users in order that it refer to exploiting weakness of users. we tend to talk about 3 methodologies for criminologist work phishing sites. Beginning stage is by breaking down changed alternatives of PC address, second stage is to check the authenticity of site by knowing any place the site is being facilitated and WHO unit is overseeing it, the third stage is that approach utilizes visual look based generally examination for checking validity of site. we tend to assemble utilization of Machine Learning strategies and calculations for investigation of those totally various alternatives of PC address and site.

### 3.1 Solution for the problem statement

This URL approach is intended to see phishing locales is by induction totally various parts from the URL and processing a measurement for each component. At that point, the page positioning will be joined with the accomplished measurements to make your psyche up whether the sites are phishing sites. The outcomes indicated that the procedure will see over hour precision of phishing sites. This framework is for foreseeing phishing URLs by creating rules of affiliation rule mining. We tend to utilized the irregular woodland recipe to pick better-known information from visit thing set properties that were separated from the dataset. Some others furthermore utilized another equation that performs on concealed data to get the precision of affiliation administers, that could be a prognosticative apriori that draws in the presumption and furthermore the help strategies that are estimated in its exactness, as opposed to from the earlier, that exclusively mark decides that have the self- importance strategy.

DETECTING PHISHING WESITES USING AI-ML

Subsequently, they given significant alternatives of the URL that recognize if it's phishing or authentic. The arranged model spotlights on unmistakable the phishing assault upheld checking phishing sites choices. With regards to scarcely any hand-picked alternatives might be acclimated separate among authentic and satirize sites. These hand-picked choices ar a few like URLs, space personality, security and coding, ASCII content document, page vogue and substance, net location bar and social human issue.

This investigation centers exclusively around URLs and name choices. Choices of URLs and area names are checked exploitation numerous rules like logical control Address, long URL address, including a prefix or addition, diverting exploitation the picture "//", and URLs having the picture "@".These choices ar investigated utilizing a lot of rules in order to distinguish URLs of phishing site pages from the URLs of genuine sites.

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER IV**

**METHODOLOGY**

ML methods that set up phishing URLs for the most part valuate a URL upheld some element or set of alternatives removed from it. There are 2 general assortments of alternatives which will be extricated from URLs, explicitly have based choices and lexical choices. Host-based alternatives portray qualities of the site, as any place it's put, World Health Organization oversees it, and used to be the situating placed in. as another option, lexical choices depict the issue properties of the URL. Since URLs are only content strings which will be separated into subparts along with the convention, hostname, and way, a framework will evaluate a site's authenticity bolstered any blend of these components.

In this, we will in general legitimize the best approach to locate the phishing site exactness by abuse Random Forest algorithmic standard. By considering the different parts and alternatives of PC address we can see the precision of the site that is phished.

In our framework we will in general square measure making utilization of Random backwoods strategy procedure to beat the issues that square measure been for quite some time confronted once misuse existing system.URL approach is wanted to locate phishing locales is by etymologizing very surprising parts from the PC address and processing a measurement for each component.

At that point the page positioning will be joined with the accomplished measurements to go to a choice whether the sites square measure phishing sites.

The phishing sites have sure attributes and examples that might be considered as their alternatives.

All out assortment of phishing sites thirty-six any place three square measure new choices.

1. Highlights might be separated from PC address.
2. Highlights might be separated from page content.

DETECTING PHISHING WESITES USING AI-ML

1. Highlights might be separated from page rank.

We utilize the amount of information email and passwords for phishing sites. This data is touchy, phishing sites to take.

We Notices that a larger than average assortment of phishing site doesn't utilize the submit button. In this way, we will in general mull over it as highlight as phishing site.

## 4.1 DATA PREPROCESSING:

1. Processing information is tough in Machine Learning algorithms. Such knowledge can’t incline within the cubic centimetre rule. we tend to should pre-process the info and use it. The dataset that the collected for our project consists of few web site links. Before we tend to extract the options from the info, we've got pre-processed the info.
2. At first import the Pandas and NumPy library packages. NumPy means that ‘Numerical Python’ or ‘Numeric Python’. This give quick calculations on arrays and matrices. Since we want confusion matrix within the project NumPy is needed.
3. Pandas is associate degree open offer library built on prime of NumPy providing superior, easy-to-use data structures and knowledge analysis tools for the Python linguistic communication. to create a tabular data of our dataset. With rows and columns in step with the choices that we've got an inclination to extract.
4. we tend to had splitted the info in step with components of uniform resource locator that are protocol (http), a hostname(www.xyz.com), and a file name(index.html). once

cacophonous the uniform resource locator into components we tend to mix into one frame exploitation concatenation that Pandas library is useful.

1. By currently we've got thirty options that ar helpful to find the phishing web site. Among them we tend to took 5 options to examine whether or not the web site is phished or not.
2. URL length.
3. Using “@” symbol.
4. Using “//” symbol.
5. Using “- “symbol.

DETECTING PHISHING WESITES USING AI-ML

1. Sub and multi sub domain

Currently allow us to see however every feature is employed to search out the phished web site.

* **URL length:** the standard universal resource locater length has been calculated. If the number of universal resource locater characters is capable fifty four or bigger than fifty four then universal resource locater has been classified as phishing.

**Rule:** If the universal resource locater length < cardinal cardinal legitimate, universal resource locater length ≥ fifty four and ≤ seventy 5 cardinal suspicious, otherwise cardinal phishing.

* **Using “@” image:** It’s been same that succeeding a locality of"@" image in uniform resource locater is unnoticed by the browser. it has been same that resultant a district of "@" image in uniform resource locater is typically the vital address.

**Rule:** If uniform resource locater is containing @ image cardinal phishing, otherwise cardinal legitimate.

* **Using “//” symbol:** The user is additionally directed to a distinct scientific discipline system mistreatment “//”in uniform resource locater. If uniform resource locater starts with “HTTP” then “//” image ought to be among the sixth position. If uniform resource locater starts with “HTTPS” then “//” image ought to be among the seventh position. **Rule:** The position of last prevalence of "//" in uniform resource locater > seven cardinal phishing, otherwise→ legitimate.
* **Using “-” image:** The dash image is not utilized within the legitimate uniform resource locater. throughout this suggests users suppose that they are using a legitimate website.

**Rule:** If (underscore)"-" image exists in name only cardinal phishing, otherwise cardinal legitimate

**Sub and multi sub domain**: "[www."](http://www/) and country code at intervals the uniform resource surveyor area unit neglected. The remaining points area unit counted at intervals the uniform resource surveyor. If Dots in Domain 0.5 is equal to1 then it's Legitimate and if the Dots in Domain 0.5 area unit a try of the it's Suspicious otherwise phishing data processor.

DETECTING PHISHING WESITES USING AI-ML

Rule: sort of dots in domain = one 54 legitimate, sort of dots in domain = a try of 54 suspicious, otherwise 54 phishing.

We tend to build the classification of universal resource locator, exploitation Random Forest technique as within the following steps.

1. Split information columns that contains the options, that we tend to desire verify the phishing web site.
2. take a look at the splitted information to predict the universal resource locator whether or not it's phished or not, if yes=0 and if no=1.
3. once the second step, produce a random forest classifier ‘clf’.
4. Train the classifier to require the coaching options.
5. Execute the ultimate classifier.

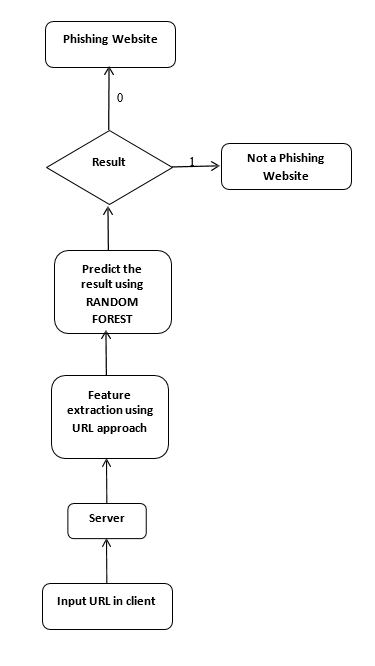


Fig. 1 Flow chart of the proposed system.

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER V**

**OVERVIEW OF TECHNOLOGIES**

### Hardware and Software Requirements

A hardware necessities list is commonly amid a hardware compatibility list (HCL), particularly just in case of in operation systems. code necessities affect process code resource necessities Associate in Nursing conditions that require to be put in on a pc to produce best functioning of an application. These necessities or conditions area unit usually not enclosed within the code installation package and wish to be put in singly before the code is put in

1. NumPy
2. Pandas
3. Random Forest Method
4. Operating System
5. Python
6. Scikit-learn

### 1 NumPy:

NumPy might be a universally handy exhibit handling bundle. It gives a prevalent level cluster item, and instruments for working with these exhibits. it's the basic bundle for logical registering with Python. It contains various choices along with these indispensable ones:

* A ground-breaking N-dimensional exhibit object
* Sophisticated (broadcasting) capacities
* Tools for integration C/C++ and FORTRAN code
* Useful polynomial math, Fourier rebuild, and arbitrary assortment capacities

DETECTING PHISHING WESITES USING AI-ML

Other than its undeniable logical uses, NumPy likewise can be utilized as Associate in Nursing prudent multi-dimensional instrumentality of nonexclusive information. eccentric datatypes are frequently sketched out misuse NumPy that grants NumPy to flawlessly and rapidly coordinate with a huge sort of databases.

### 2 Pandas

Pandas is partner ASCII content document Python Library giving unrivalled data control and examination apparatus exploitation its ground-breaking data structures. Python was significantly utilized for data munging and planning. It had little commitment towards data examination. Pandas illuminated this drawback. exploitation Pandas, we will achieve 5 regular strides inside the procedure and investigation of information, regardless of the starting point of information load, get ready, control, model, and break down. Python with Pandas is utilized in an exceedingly enormous choice of fields just as instructive and business spaces just as fund, financial matters, Statistics, investigation, and so on.

### 3 Random forest method

Random forest technique or random call forests square measure associate degree ensemble learning technique for classification, regression and alternative tasks that operates by constructing a large number of call trees at coaching time and outputting the category that's

the mode of the categories (classification) or mean prediction (regression) of the individual trees.

The same random forest algorithmic rule or the random forest classifier will use for each classification and also the regression task. Random forest classifier can handle the missing values. once we have additional trees within the forest, random forest classifier won’t over match the model. will model the random forest classifier for categorical values conjointly

Random forest may be a supervised learning algorithmic rule that is employed for each classification additionally as regression. However, but it's chiefly used for classification issues. As we all know that a forest is formed of trees and additional trees suggests that additional strong forest. Similarly, random forest algorithmic rule creates call trees on information samples and so gets the prediction from every of them and at last selects the

DETECTING PHISHING WESITES USING AI-ML

most effective answer by suggests that of vote. it's associate degree ensemble technique that is healthier than one call tree as a result of it reduces the over-fitting by averaging the result.

#### Calculation:

1. To start with, start with the choice of arbitrary examples from a taken dataset.
2. Next, this calculation will develop a choice tree for each example. At that point it will get the expectation result from each choice tree.
3. Casting a ballot will be performed for each anticipated outcome.
4. At last, select the most casted a ballot expectation result as the last forecast outcome.

### 4 Operating System

Windows ten may be a series of private pc in operation systems made by Microsoft as a part of its Windows NGO family of in operation systems. it's the successor to Windows eight.1, and was discharged to producing on Gregorian calendar month fifteen, 2015, and generally discharged for retail sale on Gregorian calendar month twenty-nine, 2015.

Windows ten receives new builds on AN current basis, that are obtainable at no further price to users, additionally to further check builds of Windows ten that are obtainable to Windows Insiders.

### Python

Python is A taken, elevated level, general counterfeit language made by Guido van Rossum and beginning released in 1991, Python's style reasoning underscores code lucidness with its remarkable utilization of genuine whitespace. Its language develops and object- arranged methodology mean to help software engineers compose clear, legitimate code for little and enormous scope comes. Python is powerfully composed and trash gathered.

DETECTING PHISHING WESITES USING AI-ML

**6 Scikit – learn**

Scikit-learn gives an assortment of directed and unaided learning calculations by means of an even interface in Python. it's authorized underneath a lenient rearranged BSD permit and is appropriated underneath a few UNIX disseminations, empowering instructional exercise and modern use. The library is made upon the SciPy (Scientific Python) that must be placed in before you'll have the option to utilize scikit-learn.

This stack incorporates:

* NumPy: a library for controlling multi-dimensional clusters and frameworks. It also has a top to bottom aggregation of numerical capacities for action shifted estimations.
* SciPy: partner degree framework comprising of arranged libraries for completing specialized figuring assignments.
* Pandas: data structures and examination
* Extensions or modules for SciPy care traditionally named scikits. All things considered, the module gives learning calculations and is known as scikit-learn.

### Hardware Requirements

* + - RAM: 4GB and Higher
    - Processor: Intel i3 and above
    - Hard Disk: 500GB: Minimum

### Software Requirements

* + - OS: Windows or Linux
    - Python IDE: python 3.5
    - Setup tools and pip to be installed for 3.6 and above
    - Language: Python

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER VI**

**IMPLEMENTATION**

### 6.1 Coding

In detecting the accuracy of phished websites, we undergo few steps. Data pre-processing, splitting the data, features extraction and classification.

If we go in detail about these steps:

### Step 1:

#### Data pre-processing:

* + - * This dataset contains few website links (Some of them are legitimate websites and a few are fake websites)
      * Pre-Processing the data before building a model and also Extracting the features from the data based on certain conditions.
      * Import NumPy and pandas which are required for data pre-processing.
      * Load the data.

### Step 2:

#### Splitting:

* We need to part the information as indicated by parts of the URL.
* A run of the mill URL could resemble: [http://www.example.com/index.html,](http://www.example.com/index.html) which demonstrates a convention (http), a hostname (www.example.com), and a record name (index.html).
* Here we partitioned the convention from the whole URL, yet need it to be isolated it in independent sections.
* Domain name section can be further sub separated into space names just as sub-area names

DETECTING PHISHING WESITES USING AI-ML

* Similarly, address section can likewise be further sub partitioned into way, question string,record.

### Step 3:

#### Features Extraction: Feature-1

**1. Long URL to Hide the Suspicious Part**

If the length of the URLs exceeds or equals to 54 characters, then the URL will be classified as a probable phishing website

0 --- indicates legitimate

1 --- indicates Phishing

2 --- indicates Suspicious

#### Feature-2

**2. URL’s that have the AT “@” Symbol**

Using the AT “@” symbol in the URL fools the browser into ignoring everything after the AT “@” symbol and the real address is hidden behind the AT “@” symbol.

IF {Url Having @ Symbol→ Phishing Otherwise→ Legitimate}

1. --- indicates legitimate
2. --- indicates Phishing

#### Feature-3

1. **Redirection with double-slash “//”**

The occurrence of “//” in the URL path itself indicates the user will be taken to a completely different website. An example of this kind of URL is:

DETECTING PHISHING WESITES USING AI-ML

“http://www.legitimate.com/[/http:](http://www.phishing.com/)/[/www.phishing.com”](http://www.phishing.com/). We examine the location where the double- slash “//” occurs in the URL. We find that if the URL starts with “HTTP”, that indicates that the double-slash “//” should occur in the URL’s sixth position. However, if the URL employs “HTTPS” then the double-slash “//” should occur in the URL’s seventh position.

IF {The Position of the Last Occurrence of "//" in the URL > 7→ Phishing Otherwise→ Legitimate

1. --- indicates legitimate
2. --- indicates Phishing

#### Feature-4

**4. Addition of a Prefix or Suffix, set-apart by a dash (-) to the Domain**

The dash symbol is very rarely seen in URLs that are real. Phishers usually add prefixes or suffixes disjoint by the dash symbol (-) to domain names that are real, so that users feel that they are on a legitimate website.

For example, [http://www.Confirme-paypal.com/](http://www.confirme-paypal.com/).

IF {Domain Name Part Includes (−) Symbol → Phishing Otherwise → Legitimate

1 --> indicates phishing 0 --> indicates legitimate

#### Feature – 5

**5.Sub-Domain and Multi Sub-Domains**

Legitimate URL links/paths must have two dots in the URL so that we do not need to type “www” every time. If the number of dots is three then the URL will be classified as “Suspicious” because it will have one sub-domain. In any case that the dots exceed three, it can be classified as a potential phishing website because it may have multiple sub- domains.

DETECTING PHISHING WESITES USING AI-ML

1. --- indicates legitimate
2. --- indicates Phishing
3. --- indicates Suspicious

### Step 4:

#### Classification of URLs using Random forest:

**Random Forest Classifier**

Random forest is a supervised learning algorithm that we have used for both purposes of classifications and regression. Similarly, the random forest algorithm makes decision trees with the available data samples and then comes up with predictions from each one of them and in the end selects the best possible answer by taking a vote. It is an ensemble method which is more useful just one decision tree as it removes the disadvantage of over-fitting by getting results by means of averaging.

Advantages of the Random Forest algorithm:

Compared to a variety of classification methods, three major advantages are pointed out by the author.

1. Random forest algorithm removes the problem of over-fitting, for application of classification.
2. Random forest can be used for the likes of both regression as well as classification
3. Feature engineering, the process of identifying most sought-after features from a dataset used for training; can use the random forest algorithm.

DETECTING PHISHING WESITES USING AI-ML

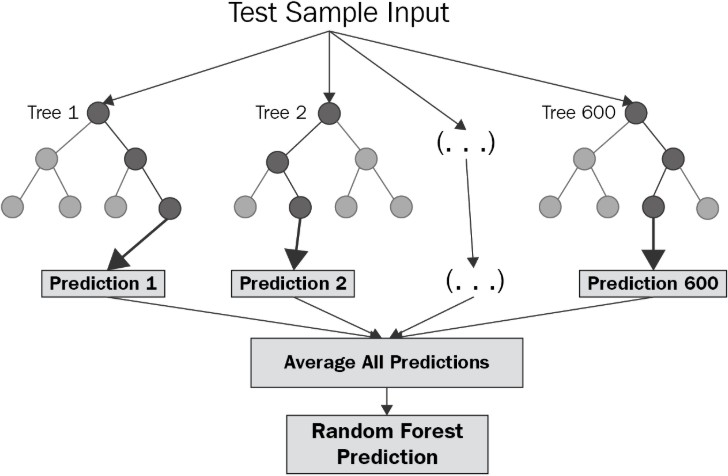


Fig 2: Block diagram of Random forest prediction

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER VII**

**RESULTS**

### ITERATION 1:

### Input: <https://newgstudent.gitam.edu/Home>

### To process the input link [https://newgstudent.gitam.edu/Home](https://newgstudent.gitam.edu/Home%20) using the Random Forest method, you would typically follow these steps:

### Data Collection:

### Use web scraping techniques to gather data from the provided link. This could involve extracting HTML content, parsing the relevant information, and converting it into a structured format suitable for analysis.

### Feature Extraction:

### Extract meaningful features from the collected data. For a web page, features could include text content, HTML tags, URL structure, metadata, etc.

### Preprocess the extracted features by cleaning, tokenizing, and transforming them into a suitable format for model input.

### Model Loading:

### Load the pre-trained Random Forest model that has been trained on relevant data, possibly containing features extracted from similar web pages.

### Feature Engineering:

### Ensure that the features extracted from the input link match the format expected by the Random Forest model. This may involve encoding categorical variables, scaling numerical features, etc.

### Prediction:

### Use the loaded Random Forest model to predict the target variable of interest. In the context of web page analysis, this could be binary classification (e.g., phishing vs. legitimate) or multi-class classification (e.g., different types of web pages).

### Output Interpretation:

### Interpret the model predictions to determine the outcome for the input link. This could involve analyzing the predicted class probabilities, feature importance scores, or any other relevant metrics provided by the Random Forest model.

### Display Result:

### Finally, present the model prediction or classification outcome for the input link to the user. This could be in the form of a simple binary outcome (e.g., phishing or not phishing) or more detailed analysis based on the model's predictions.

### 

### Fig 3: input screen of Iteration 1

### Output:

### 

### Fig 4: Output screen of Iteration 2

DETECTING PHISHING WESITES USING AI-ML

### ITERATION 2 :

### Input[: http://ix-event.com.tr/log/SN?id=1281279520](:%20http:/ix-event.com.tr/log/SN?id=1281279520)

### 

### Fig 5: Input screen of Iteration 2

### Output:

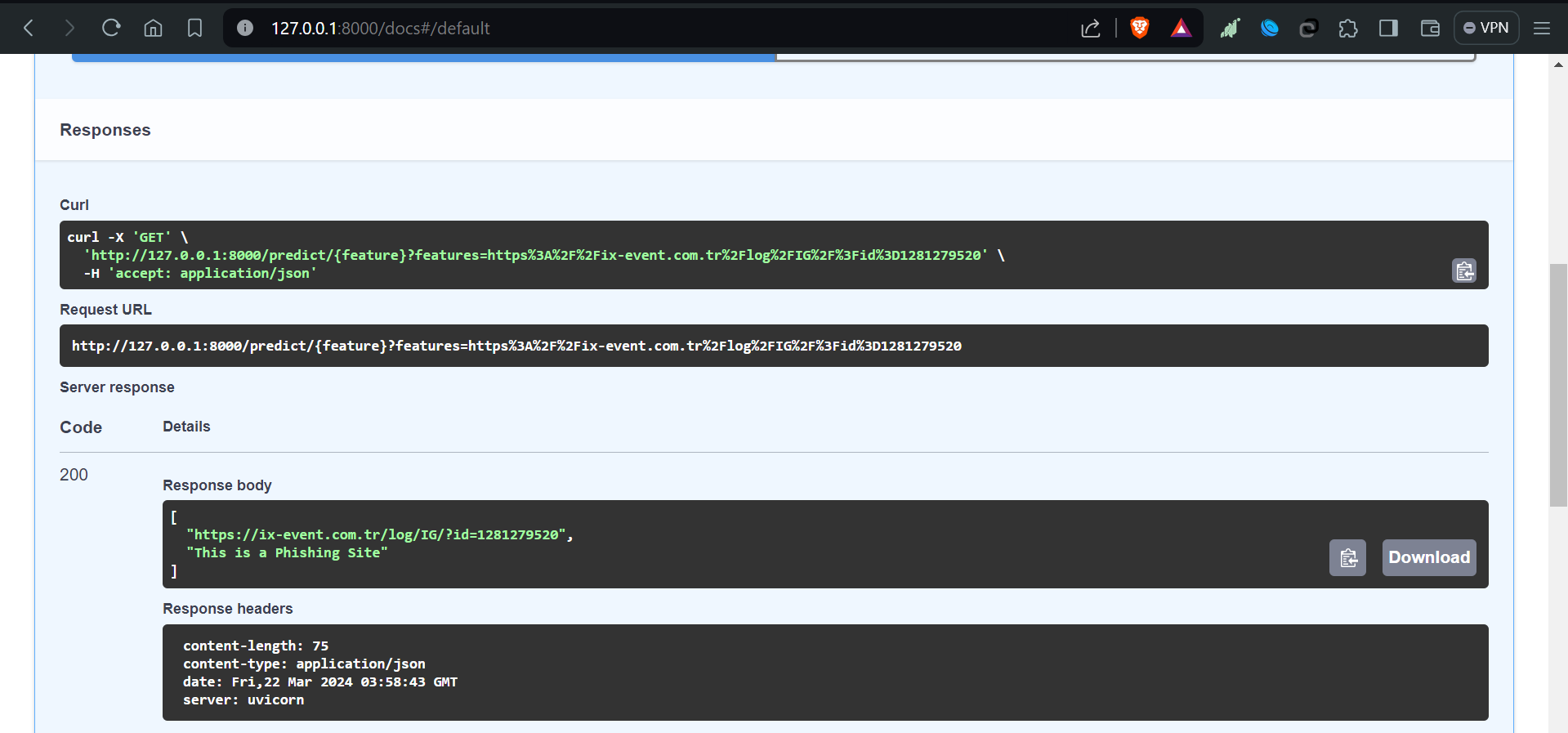


Fig 6: Output screen of Iteration 2

DETECTING PHISHING WESITES USING AI-ML

**CHAPTER VIII**

**CONCLUSION**

The system that is proposed allows people to use the internet while having a safe experience and exercising secure communication. It allows the users to refrain from revealing private details that must not be made public. Providing internet users with this system as an extension allows increases the ease of use. In this concept, we adapt the Random forest algorithm to avail maximum accuracy. The only demerit of this system’s approach is detecting the phishing websites using only five features to verify. This drawback can be eliminated by introducing a feature that is of higher quality and held high in regards of priority to be fed to the machine learning algorithm which would give us results with greater accuracy.

The scope to work on this proposed system of detecting phishing websites is almost infinite. As long as phishers come up with newer ways to fool the common internet browsing user, the patterns have to be detected and noted for further classification. Just like the five features that we have used, if more accurate and complex patterns can be detected beforehand, the phishing websites can be detected not just earlier but before they have a full-fledged attack on the users that do not know what to predict. The users can have a safe experience without compromising their details and being victims of large-scale online frauds and phishing.

REFERENCES

[1]. R. Kiruthiga, D. Akila, “Phishing Websites Detection Using Machine Learning”, 2019 International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.

[2]. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, and Dr.Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning”, IEEE 2019.

[3]. Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, and Prof. S. P. Godse, “Detection and Prevention of Phishing Websites using Machine Learning Approach”, IEEE 2018.

[4]. Hemali Sampat, Manisha Saharkar, Ajay Pandey , and Hezal Lopes, “Detection of Phishing Website Using Machine Learning”, International Research Journal of Engineering and Technology (IRJET) eISSN: 2395-0056 Volume: 05 Issue: 03 | Mar-2018 [www.irjet.net](http://www.irjet.net/) p- ISSN: 2395-0072.

[5]. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, and Siri Chandana Guttha, “Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques”, International Journal of Innovative Technology and Exploring Engineering(IJITEE) ISSN: 2278-3075, Volume-8 Issue-8S2, June 2019.

[6]. Arun Kulkarni , Leonard L. Brown, “Phishing Websites Detection using Machine Learning”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.

[7]. G. Jaspher Willsie Kathrine, Paradise Mercy Praise, A. Amrutha Rose, and Eligious Kalaivani. C, “Variants of phishing attacks and their detection techniques”, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN:978-1-5386-9439-8.

[8]. Ebubekir Buber, Önder Demir, Ozgur Koray Sahingoz, “Feature Selections for the Machine Learning based Detection of Phishing Websites” IEEE 2017.

[9]. ZUHAL TASDEMIR1 (Member, IEEE), OLIVER PERIC2 , DAVIDE SACCHETTO3 (Member, IEEE),GEORG ERNEST FANTNER2 , YUSUF LEBLEBICI3 (Fellow, IEEE),

AND B. ERDEM ALACA1,4 (Member, IEEE), “Monolithic Fabrication of Silicon Nanowires Bridging Thick Silicon Structures”, IEEE 2015.

[10]. Maher Aburrous, M. A. Hossain, Keshav Dahal, and Fadi Thabtah, “Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies”, IEEE 2010.

[11]. 1Shraddha Parekh, 2Dhwanil Parikh, 3Srushti Kotak, and 4Prof. Smita Sankhe, “A new method for Detection of Phishing Websites: URL Detection”, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant – Part Number: CFP18BAC-ART; ISBN:978-1-5386- 1974-2.

[12]. ABDULGHANI ALI AHMED, NURUL AMIRAH ABDULLAH, “Real Time Detection of

Phishing Websites”, IEEE 2016.