Machine Learning Based Approach For Phishing Detection

Chandana BH[1], Divya MH[2], Moulya C[3], Navya P[4], Mr. Hemanth Kumar K[5]
[1, 2, 3, 4] Dept of Information Science and Engineering
[5]Asst. Professor, Dept of Information Science and Engineering
[1, 2, 3, 4, 5] East West Institute of Technology, Bengaluru.

Abstract- *In modern days with regards to the significant growth of world wide web, Phishing is one of the most usual web threats which is one of the internet security problem that target only human vulnerabilities rather than software. The online customers are attacked with sophisticated techniques by phishing attackers. This is the process of obtaining sensitive information from online such as usernames and passwords. Hence it is necessary to develop a anti phishing system which is real time, fast and a intelligent phishing detection solution. In this paper we develop a reliable detection system that detects the online real time phishing websites that are developed by the attackers to deceive the customers. Our method uses machine learning algorithm to detect the phishing websites that automatically notifies the user when it detects a phishing website. This type of web attack starts with a fraudulent email or other communication as a weapon that is designed to lure a victim. The message is made to look as though it comes from a trusted sender, and also the appearances of their malicious websites are similar to the trusted website.*

Keywords- phishing, machine learning, vulnerabilities, sophisticated, fraudulent.

I. INTRODUCTION

In present days' web as become an integral part of 21st century which is growing rapidly and place an important role in every individual life. The web is a platform for supporting a wide range of criminal enterprises such as financial fraud and as a vector for propagating malware. Nowadays, many criminals focus on cyber space to find their victims with some specific ideas like phishing. The security thieves have started to target the personal information which have become a major security problem. This as lead the users to have no trust to provide their private information to the internet.

In this project, we offer an intelligent system for detecting phishing websites. The system is based on a machine learning method, particularly supervised learning. We have selected the Logistic Regression technique due to its good performance in classification. Our focus is to pursue a higher performance classifier by studying the features of phishing website and choose the better combination of them to train the classifier.

This type of web attack starts with a fraudulent email or other communication as a weapon that is designed to lure a victim. The message is made to look as though it comes from a trusted sender, and also the appearances of their malicious websites are similar to the trusted website.

All the researches work as a standalone solution where the patient or the end user needs to either physically deploy the wearable devices or access the solution in their personal laptop or mobile application. None of these solutions are been made available over the cloud using asa-service model thus extending the availability of the solution across the globe.

II. LITERATURE SURVEY

A. CRI APPROACH TO ADDRESS PHISHING

Every research related to phishing attacks has been tremendously focused on solution leaving the crime and problem uncharted. The literature also identified that providing the solution without exploring the problem is not the way to manage this threat. Therefore, a CRI approach is proposed to explore the crime factor, review on prevention techniques and investigate the gaps. It is essential to implement a CRI approach to help the future research by adding a new source of literature.

The aim is to support the new researcher and strengthen the phishing crime literature review, the CRI approach will eventually result in a holistic anti-phishing literature review framework Crime has increased within the explosion of Information technology, internet services and digital equipment as criminals have used those tools and environments on the cyber space as well as in the real world. Typical cyber-crime is internet crime, such as credit card fraud, spoof website, virus spreading and network intrusion

and hacking. This paper emphasize on one of the most concerned cyber-crime i.e. Phishing. Currently, millions of Internet users communicate either personal or business levels are providing an opportunity to phisher to deceive them easily.

B. PHISHING ATTACK FRAMEWORK

"Phishing attacks use a combination of social engineering and technical spoofing techniques to persuade users into giving away sensitive information (e.g., using a web form on a spoofed web page) that the attacker can then use to make a financial profit.". In addition, Banday and Qadri have discussed the different types of threats to online systems which includes, Vishing, Malware, Trojan Horse, etc. and how these threats affect an online user, professional companies like banks and eBay, and the infrastructure of other organizations. There are two major segments of phishing: First, a potential victim receiving a phishing email and the second, where the victim is driving to the spoofed website via a fake email.

• Phishing Email

Criminals generally use a social technique rather than technical tricks in their phishing emails to fool their end-users. For instance, the conveying urgency is a well-known method used by criminals to misdirect people's attention. In some cases, the criminals pretend to be a system administrator warning people about a new attack and urging them to install the attached path.

• Spoof website

Most phishing attacks try to convince people to go to a fake site where personal information can be collected. To host a fake site, Phisher's use free web space, use a compromised machine or register a new domain.

III. EXISITING SYSTEM

Several anti-phishing approaches have been developed, that they use blacklist and machine learning basic techniques. Unfortunately, these approaches cannot prevent all types of phishing attacks, especially zero day once. Blacklist is a well- known method to detect phishing websites where it contains a list of phishing url and domain names. If the user clicks on the website, immediately its url would be checked against the Blacklist database. A user that visits a new phishing websites is vulnerable until that website is added to Blacklist. By using detection algorithms for detecting new phishing websites as its users phishing properties to examine those websites that are not in the Blacklist.

IV. PROPOSED SYSTEM

In this project, we offer an intelligent system for detecting phishing websites. The system is based on a machine learning method, particularly supervised learning. We have selected the Logistic Regression technique due to its good performance in classification. Our focus is to pursue a higher performance classifier by studying the features of phishing website and choose the better combination of them to train the classifier.

The Web has become a platform for supporting a wide range of criminal enterprises such as spam advertised commerce, financial fraud and as a vector for propagating malware. The rapid growth of Word Wide Web and the Internet-based technologies change the users from traditional shopping to electronic commerce.

Nowadays, many criminals focus on cyberspace to find their victims with some specific tricks such as phishing. The phishing attack is a fraudulent attempt or identity theft in which the attacker deceives victims to use a malicious website, the look and feel of which is identical to the legitimate one. attempt to obtain their victims' crucial information such as passwords, account details, credit card numbers, usernames, passwords, etc. In other words, phishing is an example of social engineering techniques being used to deceive users. Users are often lured by communications purporting to be from trusted parties such as social websites, auction sites, banks, online payment processors or IT administrators.

This type of web attack starts with a fraudulent email or other communication as a weapon that is designed to lure a victim. The message is made to look as though it comes from a trusted sender, and also the appearance of their malicious websites is similar to the trusted website.

We propose to develop an application which can predict the vulnerability of a phishing website given basic data like url length, domain length etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

We propose to collect relevant data pertaining all elements related to our field of study, train the data as per the proposed algorithm of machine learning and predict how strong is there a possibility for a website to be a phishing website.

We propose to develop an application which can predict the vulnerability of a phishing websites given basic

data like url length, domain length etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

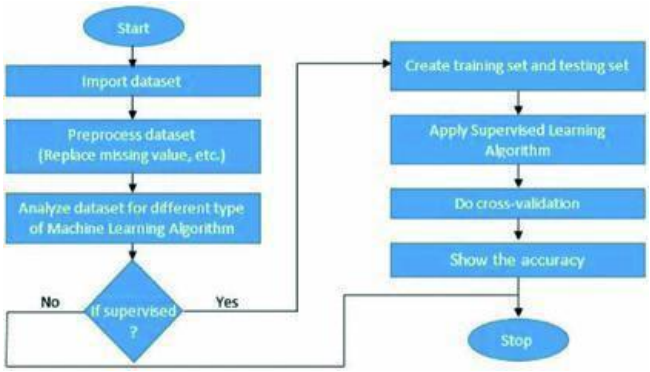

Fig 1. System Model

Various divisions in the project:

- **Users Profile Operations**

Here, the end users can perform various operations on their profiles. Firstly, the users can register a new account and thus getting an access to the portal. And then the users can login to their accounts using the registered email ID and password to access various other divisions in the portal. The users can then choose to update their profile by providing the new values to the fields they have provided during the registration phase, or the user can wish to change their password by providing their old password and new password. The user can also opt to delete their accounts in case they wish to no longer access our portal. The user can also logout from the portal to make sure the session created for them during login is terminated.

- **Implementation of Logistic Regression Algorithm**

This module implements the following Logistic Regression machine learning algorithm to detect if the inputted URL is a phishing site or not Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical. For example,

To predict whether an email is spam (1) or (0)

Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence inrealime.

Training and Testing the model for accuracy Here, the model will be trained using the datasets and tested for finding the accuracy of the model. Optimization will be done to improve the accuracy if needed. In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms work by making data driven predictions or decisions, through building a mathematical model from input data.

The data used to build the final model usually comes from multiple datasets. In particular, three data sets are commonly used in different stages of the creation of the model.

The model is initially fit on a training dataset, that is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model.

The model (e.g. a neural net or a naive Bayes classifier) is trained on the training dataset using a supervised learning method (e.g. gradient descent or stochastic gradient descent). In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label).

The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset.

Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation. Successively, the fitted model is used to predict the responses for the observations in a second dataset called the validation dataset.

The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyper parameters (e.g. the number of hidden units in a neural network). Validation datasets can be used for

regularization by early stopping: stop training when the error on the validation dataset increases, as this is a sign of overfitting to the training dataset. This simple procedure is complicated in practice by the fact that the validation dataset's error may fluctuate during training, producing multiple local minima. This complication has led to the creation of many ad-hoc rules for deciding when overfitting done.

- **Implementation of RESTful APIs for exposing the model to other apps/clients**

Here, the APIs will be developed so that the existing applications can re-use the model we developed in the second module. Representational state transfer (REST) is a software architectural style that defines a set of constraints to be used for creating Web services. Web services that conform to the REST architectural style, called RESTful Web services, provide interoperability between computer systems on the Internet. RESTful Web services allow the requesting systems to access and manipulate textual representations of Web resources by using a uniform and predefined set of stateless operations. Other kinds of Web services, such as SOAP Web services, expose their own arbitrary sets of operations. "Web resources" were first defined on the World Wide Web as documents or files identified by their URLs. However, today they have a much more generic and abstract definition that encompasses everything or entity that can be identified, named, addressed, or handled, in any way whatsoever, on the Web. In a RESTful Web service, requests made to a resource's URI will elicit a response with a payload formatted in HTML, XML, JSON, or some other format. The response can confirm that some alteration has been made to the stored resource, and the response can provide hypertext links to other related resources or collections of resources. When HTTP is used, as is most common, the operations (HTTP methods) available are GET,

HEAD, POST, PUT, PATCH, DELETE, CONNECT, OPTIONS and TRACE. By using a stateless protocol and standard operations, RESTful systems aim for fast performance, reliability, and the ability to grow by reusing components that can be managed and updated without affecting the system as a whole, even while it is running.

The term representational state transfer was introduced and defined in 2000 by Roy Fielding in his doctoral dissertation. Fielding's dissertation explained the REST principles that were known as the "HTTP object model" beginning in 1994, and were used in designing the HTTP 1.1 and Uniform Resource Identifiers (URI) standards. The term is intended to evoke an image of how a well-designed Web application behaves: it is a network of Web resources (a virtual state-machine)

- **User Interface design for the model**

Here, the front end interface will be designed so that the end users can interact with the model with ease. User interface design (UI) or user interface engineering is the design of user interfaces for machines and software, such as computers, home appliances, mobile devices, and other electronic devices, with the focus on maximizing usability and the user experience. The goal of user interface design is to make the user's interaction as simple and efficient as possible, in terms of accomplishing user goals (user-centered design). Good user interface design facilitates finishing the task at hand without drawing unnecessary attention to itself. Graphic design and typography are utilized to support its usability, influencing how the user performs certain interactions and improving the aesthetic appeal of the design; design aesthetics may enhance or detract from the ability of users to use the functions of the interface.

The design process must balance technical functionality and visual elements (e.g., mental model) to create a system that is not only operational but also usable and adaptable to changing user needs. Interface design is involved in a wide range of projects from computer systems, to cars, to commercial planes; all of these projects involve much of the same basic human interactions yet also require some unique skills and knowledge. As a result, designers tend to specialize in certain types of projects and have skills centered on their expertise, whether it is a software design, user research, web design, or industrial design.

- **Cloud based deployment process of the model**

Here, the model will be deployed on a cloud server to make the solution accessible across the geographical areas. For the cloud deployment process, we use either of Amazon web service or the Google Cloud.

V. SYSTEM ARCHITECTURE

The below figure shows a general block diagram describing the activities performed by this project.

The entire architecture has been implemented in nine modules which we will see in high level design and low level design in later chapters.

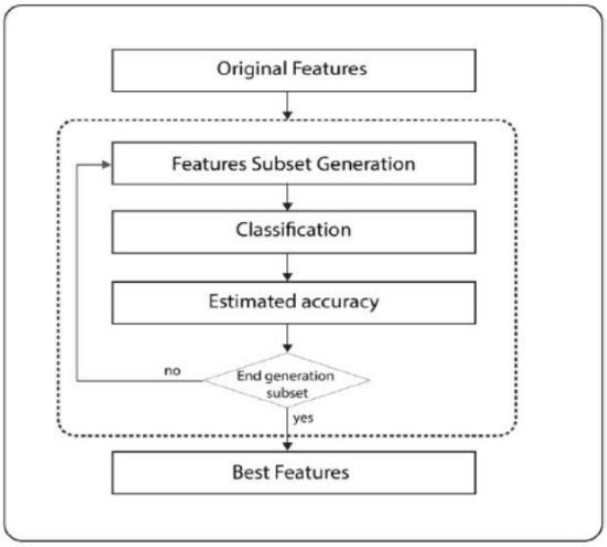

Fig 2. System Architecture

Major divisions in this project are:

A. DATA ACCESS LAYER:

Data access layer is the one which exposes all the possible operations on the data base to the outside world. It will contain the DAO classes, DAO interfaces, POJOs, and Utils as the internal components.

All the other modules of this project will be communicating with the DAO layer for their data access needs.

B. ACCOUNT OPERATIONS:

Account operations module provides the following functionalities to the end users of our project.

- Register a new seller/ buyer account
- Login to an existing account
- Logout from the session
- Edit the existing Profile
- Change Password for security issues
- Forgot Password and receive the current password over email
- Delete an existing Account

Account operations module will be re-using the DAO layer to provide the above functionalities.

C. IMPLEMENTATION OF LOGISTIC REGRESSION ALGORITHM:

This module implements the following Logistic Regression machine learning algorithm to detect if the inputted URL is a phishing site or not.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications Logistic Regression is used when the dependent variable(target) is categorical.

For example,

To predict whether an email is spam (1) or (0)

Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

D. TRAINING AND TESTING THE MODEL FOR ACCURACY:

Here, the model will be trained using the datasets and tested for finding the accuracy of the model. Optimization will be done to improve the accuracy if needed. In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data.

The data used to build the final model usually comes from multiple datasets. In particular, three data sets are commonly used in different stages of the creation of the model.

The model is initially fit on a training dataset, that is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model. The model (e.g. a neural net or a naive Bayes classifier) is trained on the training dataset using a supervised learning method (e.g. gradient descent or stochastic gradient descent). In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label).

The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

Successively, the fitted model is used to predict the responses for the observations in a second dataset called the validation dataset. The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyper parameters (e.g. the number of hidden units in a neural network).

Validation datasets can be used for regularization by early stopping: stop training when the error on the validation dataset increases, as this is a sign of overfitting to the training dataset. This simple procedure is complicated in practice by the fact that the validation dataset's error may fluctuate during training, producing multiple local minima. This complication has led to the creation of many ad-hoc rules for deciding when overfitting has truly begun.

Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. If the data in the test dataset has never been used in training (for example in cross-validation), the test dataset is also called a holdout dataset.

E. IMPLEMENTATION OF RESTFUL APIS FOR EXPOSING THE MODEL TO OTHER APPS/CLIENTS:

Here, the APIs will be developed so that the existing applications can re-use the model we developed in the second module. Representational state transfer (REST) is a software architectural style that defines a set of constraints to be used for creating Web services. Web services that conform to the REST architectural style, called RESTful Web services, provide interoperability between computer systems on the Internet. RESTful Web services allow the requesting systems to access and manipulate textual representations of Web resources by using a uniform and predefined set of stateless operations. Other kinds of Web services, such as SOAP Web services, expose their own arbitrary sets of operations.

In a RESTful Web service, requests made to a resource's URI will elicit a response with a payload formatted in HTML, XML, JSON, or some other format.

The response can confirm that some alteration has been made to the stored resource, and the response can provide hypertext links to other related resources or collections of resources. When HTTP is used, as is most common, the operations (HTTP methods) available are GET,

HEAD, POST, PUT, PATCH, DELETE, CONNECT, OPTIONS and TRACE.

By using a stateless protocol and standard operations, RESTful systems aim for fast performance, reliability, and the ability to grow by reusing components that can be managed and updated without affecting the system as a whole, even while it is running.

The term representational state transfer was introduced and defined in 2000 by Roy Fielding in his doctoral dissertation. Fielding's dissertation explained the REST principles that were known as the "HTTP object model" beginning in 1994, and were used in designing the HTTP 1.1 and Uniform Resource Identifiers (URI) standards.

The term is intended to evoke an image of how a well-designed Web application behaves: it is a network of Web resources (a virtual state-machine) where the user progresses through the application by selecting resource identifiers are http://www.example.com/articles/21 and resource operations such as GET or POST (application state transitions), resulting in the next resource's representation (the next application state) being transferred to the end user for their use.

F. USER INTERFACE DESIGN FOR THE MODEL

Here, the front end interface will be designed so that the end users can interact with the model with ease. User interface design (UI) or user interface engineering is the design of user interfaces for machines and software, such as computers, home appliances, mobile devices, and other electronic devices, with the focus on maximizing usability and the user experience. The goal of user interface design is to make the user's interaction as simple and efficient as possible, in terms of accomplishing user goals (user-centered design).

Good user interface design facilitates finishing the task at hand without drawing unnecessary attention to itself. Graphic design and typography are utilized to support its usability, influencing how the user performs certain interactions and improving the aesthetic appeal of the design; design aesthetics may enhance or detract from the ability of users to use the functions of the interface. The design process must balance technical functionality and visual elements (e.g., mental model) to create a system that is not only operational but also usable and adaptable to changing user needs.

Interface design is involved in a wide range of projects from computer systems, to cars, to commercial planes; all of these projects involve much of the same basic human interactions yet also require some unique skills and knowledge. As a result, designers tend to specialize in certain types of projects and have skills centered on their expertise, whether it is a software design, user research, web design, or industrial design.

G. CLOUD BASED DEPLOYMENT PROCESS OF THE MODEL

Here, the model will be deployed on a cloud server to make the solution accessible across the geographical areas. For the cloud deployment process, we use either of Amazon web service or the Google Cloud.

Here is the overall representation of the project:

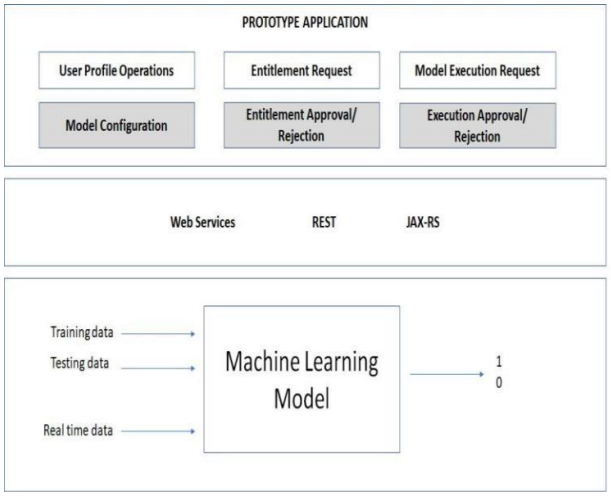

Fig 3. Cloud Based Deployment Process Model

VI. IMPLEMENTATION

MODEL IMPLEMENTATION

This module implements the following Logistic Regression machine learning algorithm to detect if the inputted URL is a phishing site or not

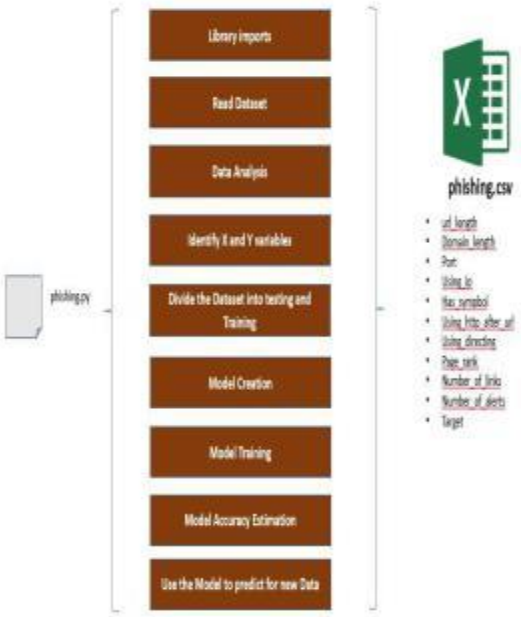

Fig 4. Model implementation

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,

To predict whether an email is spam (1) or (0)

Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

REST IMPLEMENTATION

This module provides an end point to the outside world so that the third party applications can invoke it by sending the new patients' data**.**

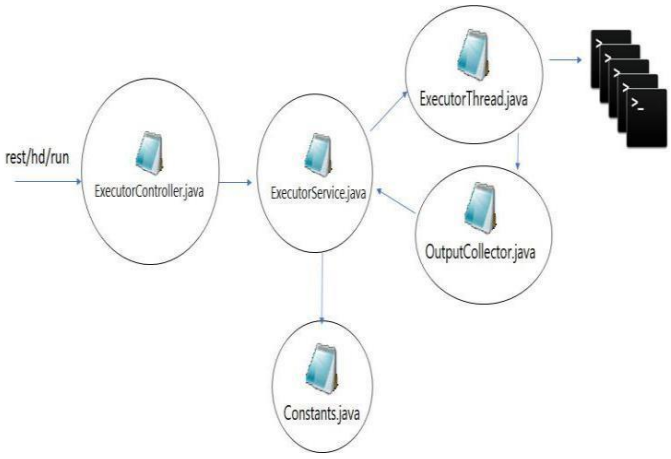

Fig 5. REST Implementation

The end point upon receiving the data, will then invoke the seven python programs by sending the data as a command line parameters. Each of these python program will either give output as 1 or 0. The output 1 indicates there's a disease and the output 0 indicates there's no disease. The end point will then analyse the output from each of the models. The end point will count the number of models providing the output 0 and then the number of models providing the output 1. Whichever is having a majority, the model will consider that result as the actual result and then responds back the same to the third party application who invoked this. This REST web service endpoint is exposed as a POST method.

PROTOTYPE – USER OPERATIONS

This module implements the basic user profile operations on the prototype application.

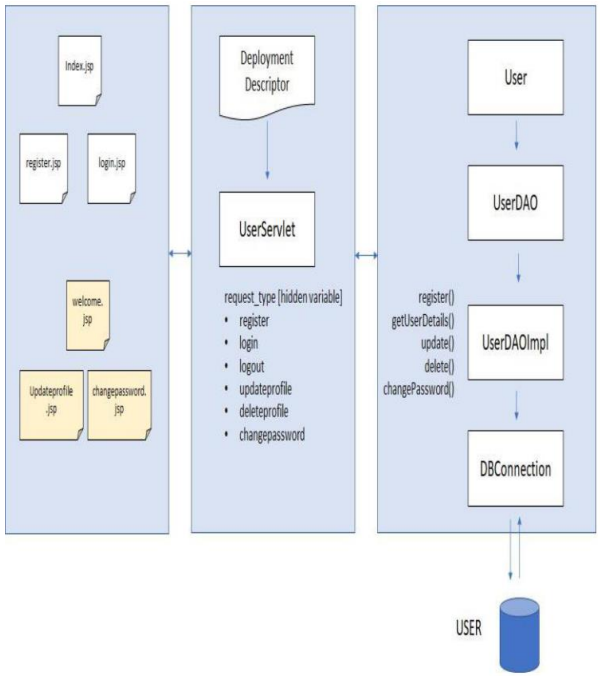

Fig 6. Prototype – User operation

The user profile operations include creating a new account, logging in to the existing account, logging out, editing the profile, changing the password, and deleting the profile if not needed anymore.

This application is also deployed on the cloud server so that this can be accessed by anyone across the globe using the IP address of this cloud server. The implementation is done using the J2EE architecture and for the database needs we have used SQLITE3 Apart from the user profile operations, this module also include other operations that can be executed by the users. This includes – Requesting for the entitlement, requesting for the prediction, and retrieving the output of the executed request. The users can't execute the prediction request by himself/herself. It has to go through the admin.

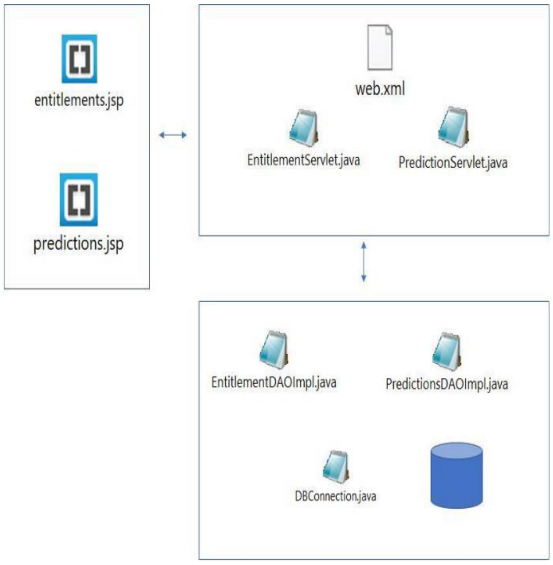

Fig 7. Prototype – User operation

PROTOTYPE – ADMIN OPERATIONS

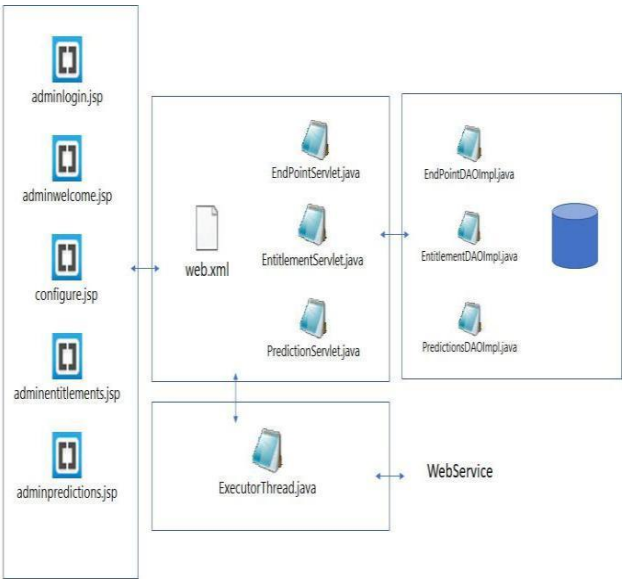

Fig 8. Prototype – Admin Operations

This module implements the Admin operations in our prototype application. The admin operations includes approving or rejecting the entitlement requests, approving or rejecting the prediction requests, and visualizing the output of the executed requests. The admin of the project will have the full control on who executes the prediction algorithm and against which data. Any user who wants to run the algorithm against their new data will first have to get it approved from the admin

IMPLEMENTATION OF PROTOTYPE APPLICATION - USER PORTAL

This module implements the basic user profile operations on the prototype application.

The user profile operations include creating a new account, logging in to the existing account, logging out, editing the profile, changing the password, and deleting the profile if not needed anymore.

This application is also deployed on the cloud server so that this can be accessed by anyone across the globe using the IP address of this cloud server. The implementation is done using the J2EE architecture and for the database needs we have used SQLITE3.

Apart from the user profile operations, this module also include other operations that can be executed by the users. This includes – Requesting for the entitlement, requesting for the prediction, and retrieving the output of the executed request. The users can't execute the prediction request by himself/herself. It has to go through the admin.

IMPLEMENTATION OF PROTOTYPE APPLICATION - ADMIN PORTAL

This module implements the Admin operations in our prototype application. The admin operations includes approving or rejecting the entitlement requests, approving or rejecting the prediction requests, and visualizing the output of the executed requests.

The admin of the project will have the full control on who executes the prediction algorithm and against which data. Any user who wants to run the algorithm against their new data will first have to get it approved from the admin.

VII. RESULT

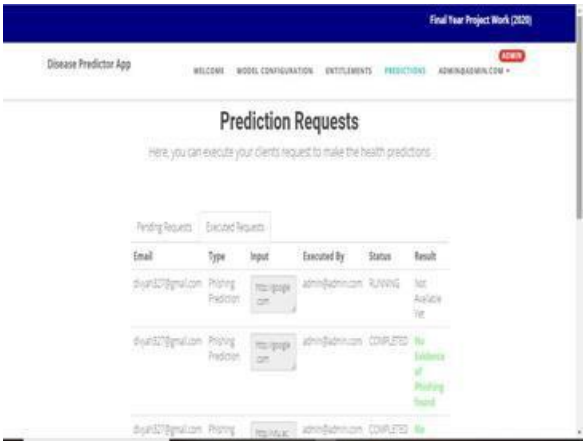

Fig 9. Screenshot

This is the final result user can get from the admin. Admin should approve the request and then admin has to give the result whether the website is phishing or not.

VIII. CONCLUSION

In this paper we offer a intelligent system for detecting phishing websites. The system is based on a machine learning method, particularly supervised learning. We have selected the logistic regression technique due to its good performance in classification our focus is to pursue a higher performance classifier by studying the features of phishing website and choose the better combination of them to train the classifier.as a result ,we conclude our paper with accuracy of 99% .

IX. ACKNOWLEDGEMENT

Any achievement, be it scholastic or otherwise does not depend solely on the individual efforts but on the guidance, encouragement and cooperation of intellectuals, elders and friends. A number of personalities, in their own capacities have helped us in carrying out this project.

We would like to take this opportunity to thank them all. First and fore most we would like to thank Mr. Hemanth Kumar K, Asst. Professor, EWIT, for his moral support and valuable suggestions and expert advice. towards completing our project work. and constantly guiding me to organize the project in a systematic manner.

We thank our Parents, and all the Faculty members of Department of Information science & Engineering for their constant support and encouragement. Last, but not the least, we would like to thank our peers and friends who provided us with valuable suggestions to improve our project.

REFERENCES

[1] AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-PhishingLanding-Page- Soc.html [Oct 30, 2017].

[2] ”Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money” Internet: https://www.kaspersky.com/about/pressreleases/2017 financial-threats-in2016. Feb 22, 2017 [Oct 30, 2017].

[3] Y. Zhang, J. I. Hong, and L. F. Cranor, ”Cantina: A Content-based Approach to Detecting Phishing Web Sites,” New York, NY, USA, 2007, pp. 639-648.

[4] M. Blasi, ”Techniques for detecting zero day phishing websites.” M.A. thesis, Iowa State University, USA, 2009.

[5] R. S. Rao and S. T. Ali, ”PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.

[6] E. Jakobsson, and E. Myers, Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft. Wiley, 2006, pp.2–3.

[7] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, ”Detecting phishing web sites: A heuristic URL-based approach,” in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.

[8] Z. Zhang, Q. He, and B. Wang, ”A Novel Multi-Layer Heuristic Model for AntiPhishing,” New York, NY, USA, 2017, p. 21:1-21:6.

[9] N. Sanglerdsinlapachai and A. Rungsawang, ”Web Phishing Detection Using Classifier Ensemble,” New York, NY, USA, 2010, pp. 210-215.

[10] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ”CANTINA+: A FeatureRich Machine Learning Framework for Detecting Phishing Web Sites,” ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 21:1-21:28, Sep. 2011.

[11] R. M. Mohammad, F. Thabtah, and L. McCluskey, ”Predicting phishing websites based on self-structuring neural network,” Neural Comput & Applic, vol. 25, no. 2, pp. 443-458, Aug. 2014.

[12] Pradeepthi K V and Kannan A, ”Performance study of classification techniques for phishing URL detection,” in 2014 Sixth International Conference on Advanced Computing (ICoAC), 2014, pp. 135-139.

[13] S. Marchal, J. Franois, R. State, and T. Engel, ”PhishStorm: Detecting Phishing With Streaming Analytics,” IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458-471, Dec. 2014.

[14] A. Sirageldin, B. B. Baharudin, and L. T. Jung, ”Malicious Web Page Detection: A Machine Learning Approach,” in Advances in Computer Science and its Applications, Springer, Berlin, Heidelberg, 2014, pp. 217-224.

[15] R. Verma and K. Dyer, ”On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers,” New York, NY, USA, 2015, pp. 111-122.

[16] H. H. Nguyen and D. T. Nguyen, ”Machine Learning Based Phishing Web Sites Detection,” in AETA 2015: Recent Advances in Electrical Engineering and Related Sciences, V. H. Duy, T. T. Dao, I. Zelinka, H.- S. Choi, and M. Chadli, Eds. Cham: Springer International Publishing, 2016, pp. 123-131

[17] M. A. U. H. Tahir, S. Asghar, A. Zafar, and S. Gillani, ”A Hybrid Model to Detect 76 Phishing-Sites Using Supervised Learning Algorithms,” in 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 2016, pp. 1126-1133