Survey on Phishing Websites Detection using

Machine Learning

Abstract:

Phishing is a widespread method of tricking unsuspecting people into disclosing personal information by using fake websites. Phishing website URLs are designed to steal personal information such as user names, passwords, and online banking activities. Phishers employ webpages that are visually and semantically identical to legitimate websites. As technology advances, phishing strategies have become more sophisticated, necessitating the use of anti-phishing measures to identify phishing. Machine learning is an effective method for combating phishing assaults. This study examines the features utilised in detection as well as machine learning-based detection approaches.

Phishing is popular among attackers because it is easier to persuade someone to click on a malicious link that appears to be legitimate than it is to break through a computer's protection measures. The malicious links in the message body are made to look like they go to the faked organisation by utilising the spoofed organization's logos and other valid material. We'll go through the characteristics of phishing domains (also known as fraudulent domains), the qualities that distinguish them from real domains, why it's crucial to detect them, and how they can be discovered using machine learning and natural
language processing techniques.

Keywords: Phishing, personal information, machine learning, malicious links, and phishing domain characteristics are all terms that come up when people think of phishing.

I INTRODUCTION

Phishing has become a major source of concern for security professionals in recent years since it is relatively easy to develop a phoney website that appears to be identical to a legitimate website.

Although experts can detect bogus websites, not all users can, and as a result, they become victims of phishing attacks. The attacker's main goal is to steal bank account credentials. Because of a lack of user awareness, phishing assaults are becoming more successful. Because phishing attacks take advantage of user flaws, it is difficult to mitigate them, but it is critical to improve phishing detection tools. Phishing is a type of wide extortion in which a malicious website imitates a genuine one-time memory with the sole purpose of obtaining sensitive data, such as passwords, account details, or MasterCard numbers. Despite the fact that there are still some anti-phishing programming and strategies for detecting possible phishing attempts in messages and typical phishing content on websites, phishes devise fresh and crossbred procedures to get around public programming and frameworks. Phishing is a type of fraud that combines social engineering with access to sensitive and personal data, such as passwords and open-end credit unpretentious components by assuming the characteristics of a trustworthy person or business via electronic correspondence. Hacking uses spoof messages that appear legitimate and are instructed to originate from legitimate sources such as financial institutions, online business goals, and so on, to entice users to visit phoney destinations via links provided on phishing websites.

II LITERATURE SURVEY

Huang et al., (2009) proposed frameworks to distinguish phishing using page section similitude, which breaks down universal resource locator tokens to create forecast precision phishing pages typically keep their CSS look similar to their target pages. This strategy was suggested by Marchal et al., (2017) to differentiate The analysis of authentic site server log knowledge is required for phishing websites. An off-the-shelf programme or the detection of a phishing website. Free has a number of distinguishing characteristics, including high precision, complete autonomy, and beautiful language-freedom, speed of choosing, flexibility to dynamic phishing, and flexibility to advance in phishing methods.

By extracting website URL features and analysing subset based feature selection methods, Mustafa Aydin et al. suggested a classification algorithm for phishing website identification. For the detection of phishing websites, it uses feature extraction and selection approaches.
Alphanumeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis, and Rank Based Analysis are five separate analyses of the retrieved features about the URLs of the sites and the built feature matrix. The majority of these elements are textual properties of the URL, with others relying on third-party services. PhishStorm is an automated phishing detection system developed by Samuel Marchal et al. that can analyse any URL in real time to identify probable phishing sites. To protect consumers from phishing content, Phish Storm is presented as an automatic real-time URL phishingness evaluation system. PhishStorm can be used as a Website reputation evaluation system that delivers a phishingness score for URLs.
Fadi Thabtah et al. compared a huge variety of machine learning approaches on real phishing datasets using various metrics. The goal of this comparison is to highlight the benefits and drawbacks of machine learning predictive models, as well as their real effectiveness when it comes to phishing assaults. Covering approach models are more appropriate as antiphishing solutions, according to the experimental data.
Muhemmet Baykara et al. proposed the Anti Phishing Simulator, which provides information about the phishing detection challenge and how to detect phishing emails. The Bayesian algorithm adds spam emails to the database. Phishing attackers utilise JavaScript to insert a valid URL into the address bar of the browser. The study recommends using the e-mail text as a keyword only for advanced word processing.

A. Using the IP Address

Users can be sure that someone is attempting to steal their personal sensitive information if an IP address is used instead of the domain name in the URL, such as "http://125.98.3.123/fake.html." As demonstrated in the following link, the IP address is sometimes converted into hexadecimal code.
“http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html”.
Rule: Phishing if the domain part has an IP address; else, legitimate.

B. Long URL to Hide the Suspicious Part

Phishers can disguise the suspicious component of the URL in the address bar by using lengthy URLs.
For example, http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f 416dbe46b773a5e/?cmd= home&dispatch=11004d58f
5b74f8dc1e7c2e8dd4105e8@phishing.website. html
We calculated the length of URLs in the dataset and produced an average URL length to assure the accuracy of our analysis.
The results revealed that if the URL is more than or equivalent to 54 characters, it is considered as phishing. We discovered 1220 URLs with lengths of 54 or more when evaluating our dataset, accounting for 48.8% of the entire dataset size.
If the URL length is 75, it is legal; otherwise, it is phishing.
We were able to improve the accuracy of this feature rule by updating it with an approach based on frequency.

C. Adding a Domain Prefix or Suffix Separated by (-)

In genuine URLs, the dash symbol is frequently used. Phishers frequently append prefixes or suffixes to domain names, separated by (-), to give the impression that they are dealing with a legitimate website. Consider the website http://www.Confirmepaypal.com/.
If the domain name part has the (-) symbol, it is phishing; otherwise, it is legitimate.

D. Submitting Information to Email

A web form allows a user to submit sensitive personal information that is sent to a server for processing. A phisher may send the user's data to his personal email address. A server-side scripting language, such as PHP's "mail()" function, might be used to accomplish this. The "mailto:" function is another client-side function that might be used for this purpose. If you use the "mail()" or "mailto:" functions to submit user information, you're phishing. Legitimate.

E. Use of a Pop-up Window

It's unusual to come across a legitimate website that requests personal information via a pop-up window. This function, on the other hand, has been employed on certain genuine websites, and its main objective is to alert consumers about fraudulent activity or broadcast a welcome statement, but no personal information was requested through these pop-up windows. If the popup window contains text fields, the rule is: Otherwise, legitimate phishing.

III. MOTIVATION

Detecting and stopping phishing websites is always an important area of research. Different sorts of phishing strategies provide torrential and vital means for effective police work and, more importantly, for the protection of persons and organisations. In phishing, the uniform resource location is crucial. The Uniform Resource Locator (URL) is a key location via which websites are launched and pages are redirected to the next page via links. The vulnerable architecture in phishing (i.e., through the hyperlink) is redirecting the pages; the pages are redirected to the legitimate web site or the phishing site.

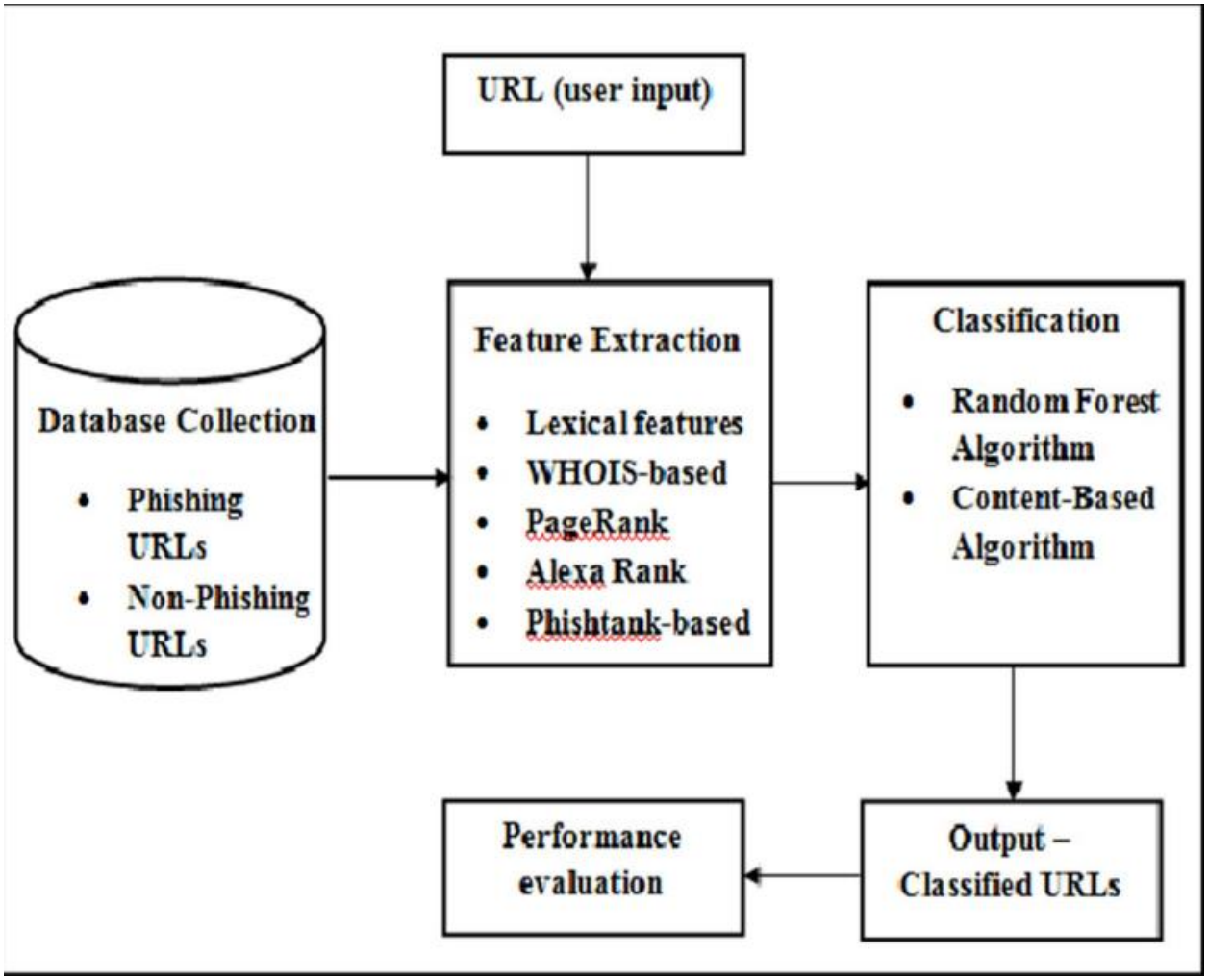
IV. EXISTING SYSTEM

To detect phishing sites, the existing system employs Classifiers, Fusion Algorithms, and Bayesian Models. Text and visual material can be classified by the classifiers. Text classifiers are used to categorise text material, whereas Image classifiers are used to categorise image content. The threshold value is calculated using a Bayesian model. The Fusion Algorithm uses the results of both classifiers to determine whether or not the site is phishing. Correct classification ratio, F-score, Matthews' correlation coefficient, False negative ratio, and False alarm ratio are used to evaluate the performance of different classifiers.

V. PROPOSED SYSTEM

This section explains the suggested phishing attack detection model. The suggested methodology focuses on detecting phishing attacks using the properties of phishing websites, the Blacklist, and the WHOIS database. Few criteria can be utilised to distinguish between real and faked web pages, according to experts. URLs, domain identification, security & encryption, source code, page style and contents, web address bar, and social human component are only a few of the features that have been chosen. This research is limited to URLs and domain name characteristics. . These characteristics are examined using a set of rules in order to distinguish phishing webpage URLs from authentic website URL

VII .SYSTEM MODEL



VII. IMPLEMENTATION

A. Algorithms Used

Three machine learning classification model Decision Tree, Random forest and Support vector machine has been selected to detect phishing websites.

1) Decision Tree Algorithm

One of the most extensively used algorithms in the field of machine learning. The decision tree algorithm is simple to comprehend and apply. The decision tree starts by selecting the best splitter from the available qualities for categorization, which is referred to as the tree's root.

The algorithm keeps building the tree until it reaches the leaf node.

Each internal node of the tree represents an attribute, and each leaf node represents a class label. The gini index and information gain approaches are employed in the decision tree algorithm to determine these nodes.

2) Random Forest Algorithm

Random forest algorithm is based on the notion of decision tree algorithm and is one of the most powerful algorithms in machine learning technology. The random forest algorithm generates a forest of decision trees. A large number of trees means a high level of detection accuracy.

The bootstrap method is used to create trees. To generate a single tree, the bootstrap approach selects characteristics and samples from the dataset at random with replacement. Random forest algorithm, like decision tree algorithm, chooses the best splitter for classification from randomly picked features. Random forest algorithm also uses gini index and information gain methods to determine the best splitter. This technique will be repeated until the random forest has produced n trees.

Each tree in the forest forecasts the goal value, after which the algorithm calculates the votes for each forecasted target. Finally, the random forest method uses the anticipated target with the most votes as the final prediction.

3) Support Vector Machine Algorithm

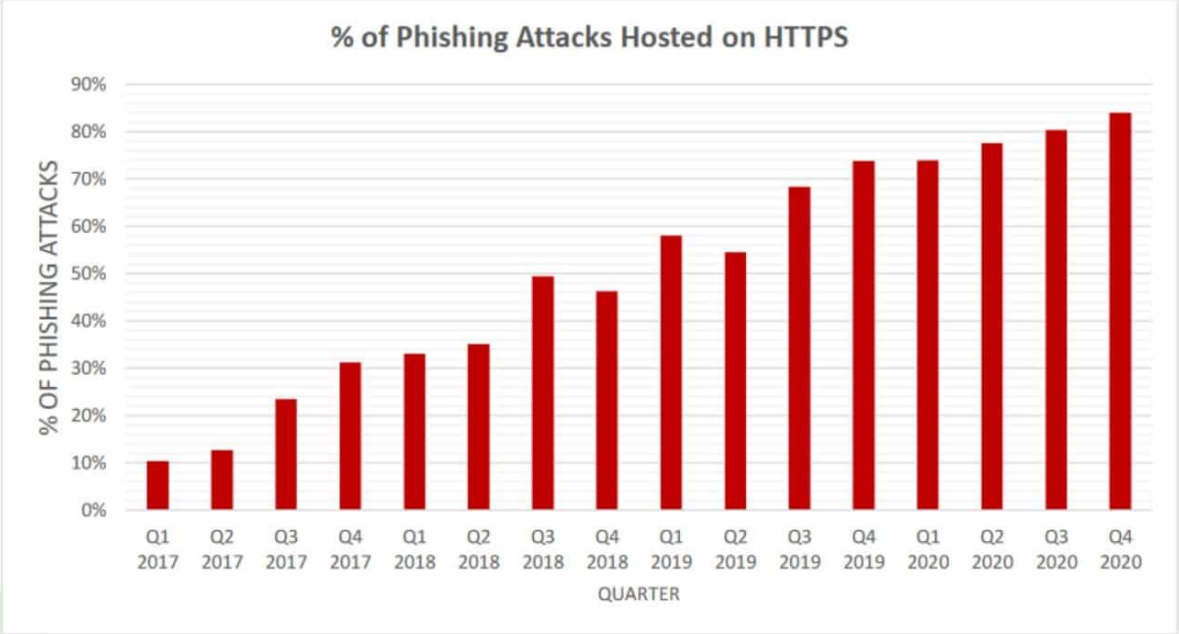
Another useful approach in machine learning is the support vector machine. Each input item is displayed as a point in n-dimensional pace in the support vector machine algorithm, and the algorithm creates a separating line for classification of two classes, which is known as a hyperplane.

The support vector machine looks for the closest points, which are called support vectors, and then constructs a line linking them. The support vector machine then creates a separation line that is perpendicular to and bisects the connecting line. The margin should be as large as possible in order to accurately classify data. The margin is the distance between the hyperplane and support vectors in this case. Because it is impossible to separate complicated and nonlinear data in real life, the support vector machine employs a kernel approach that converts lower dimensions space to higher dimensional space to tackle this difficulty.

VIII. SSL is no longer an indicator of a safe site.

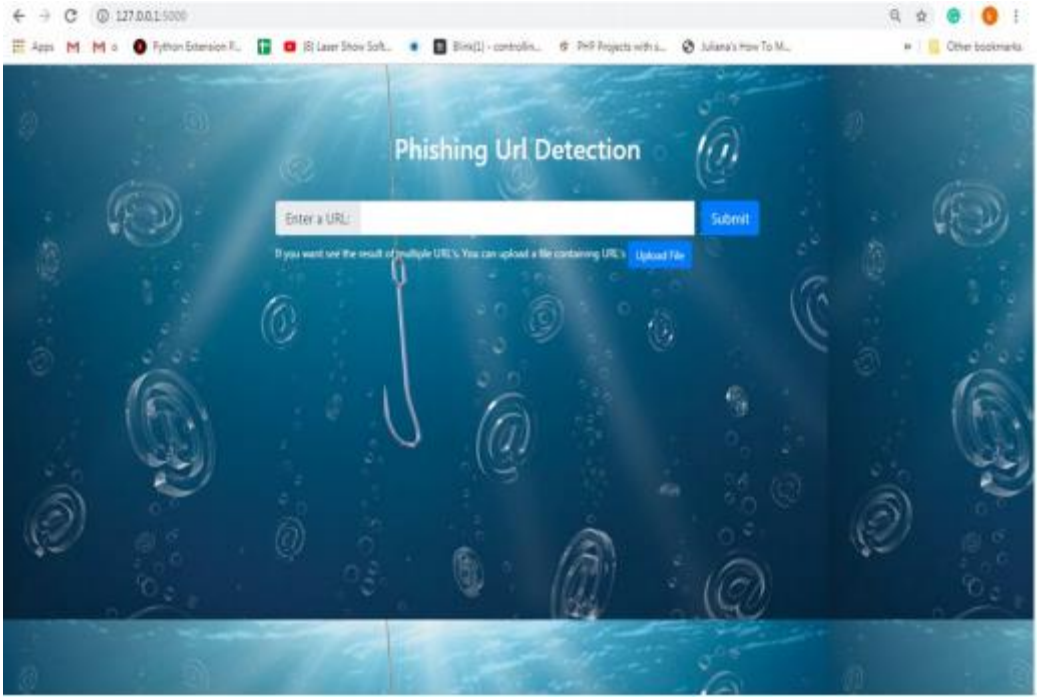
For many years, one of the primary tips for avoiding phishing sites has been to examine URLs carefully and avoid sites that don't have an SSL certificate. "HTTPS" in the URL (versus "HTTP") signifies that a site has an SSL certificate and is protected by the HTTPS encryption protocol.

However, this is **no longer a good tactic for recognizing dubious sites**. As reported by APWG, a whopping 84 percent of phishing sites examined in Q4 of 2020 used SSL. This continues the long-running trend of increasing around 3% every quarter.

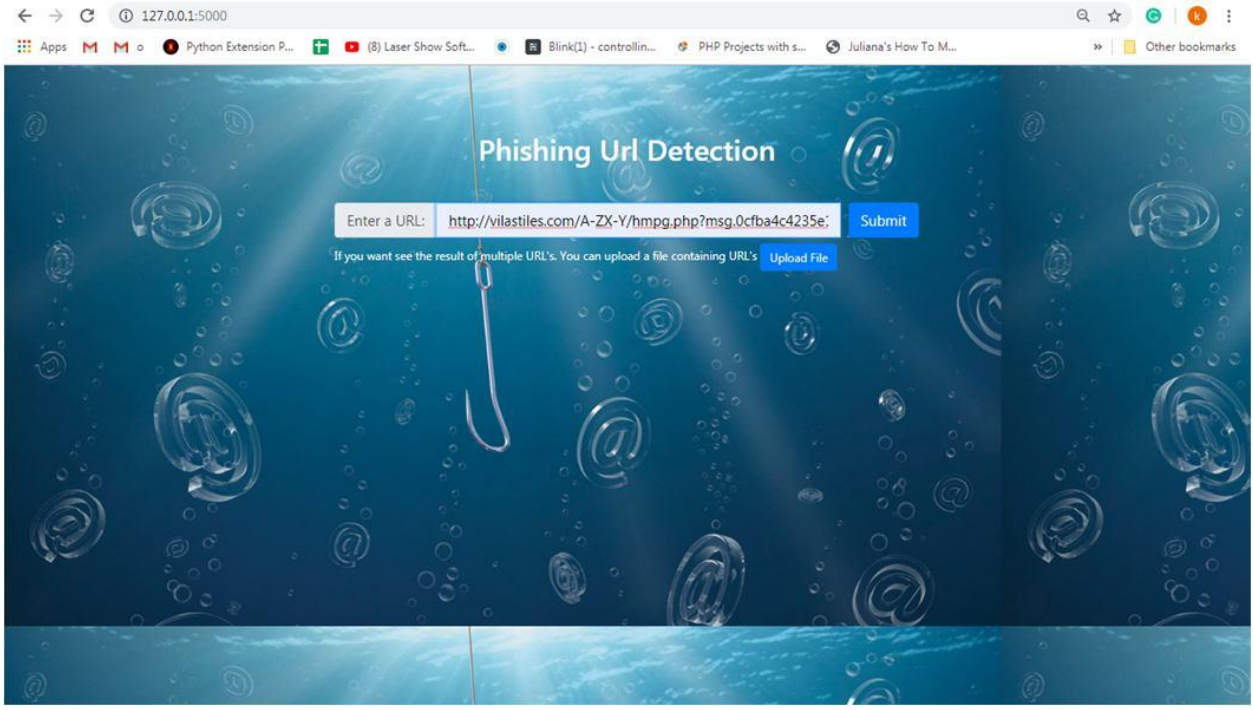


IX. RESULTS

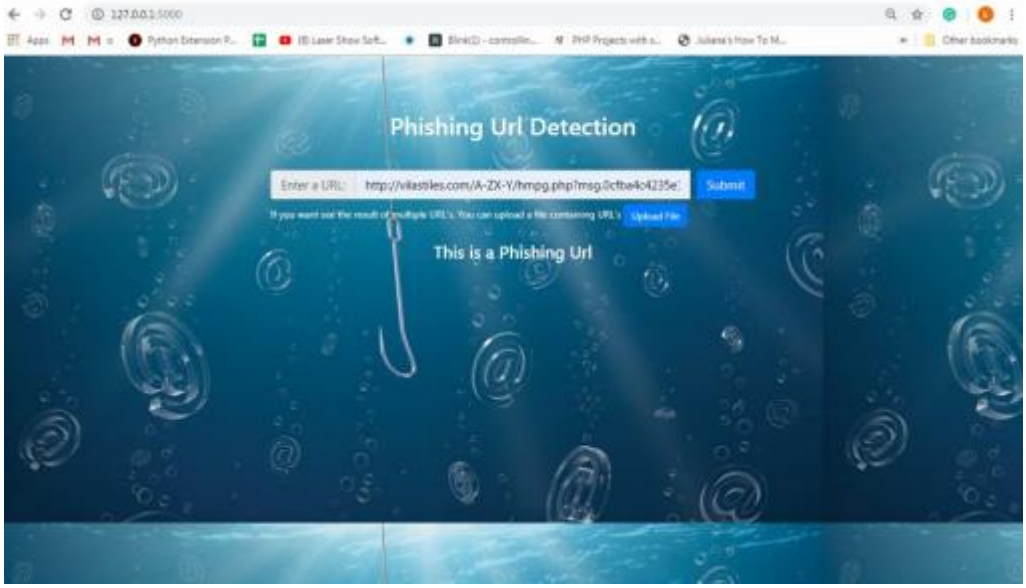
A. Home Page



B. Input URL



C. Final Result Page



X. CONCLUSION

Education awareness is the most significant strategy to protect users from phishing attacks. Internet users should be aware of all security recommendations made by professionals. Every user should also be taught not to mindlessly

follow links to websites where sensitive information must be entered. Before visiting a website, make sure to check the URL. In the future, the system could be upgraded to automatically detect the web page and the application's compatibility with the web browser. Additional work can be done to distinguish fraudulent web pages from authentic web pages by adding certain additional characteristics. In order to detect phishing on the mobile platform, the PhishChecker programme can be upgraded into a web phone application.

XI. FUTURE SCOPE

We'll look into the links between phishing sites and hosting and DNS registration providers in more detail. We'll also look at other features like Content Security Policies, certificate authorities, and TLS fingerprinting that can be used. In addition, we will compare SVMs and neural networks to other machine learning techniques such as random forest classifiers for speed and accuracy.