Combined Strategy For Accurate Prediction of Phishing Emails

Bharath Kumar B R[1]*, Dr. Mohan Aradhya[2]

[1]*PG student, Master Of Computer Applications, RV College Of Engineering, Bengalore

[2]Associate Professor, Master Of Computer Applications, RV College Of Engineering, Bengalore

***Abstract*—** Phishing is a common and serious threat in the digital world, where attackers deceive people into sharing sensitive information, like passwords or credit card numbers, by pretending to be legitimate entities. This research focuses on developing a practical method for accurately detecting phishing emails using two machine learning algorithms: the Stochastic Gradient Descent (SGD) classifier and the Gradient Boosting classifier. The process starts by receiving an email and then separating the main content from any URLs within the email. Then the email content undergoes preprocessing, this cleaned text is then broken down into tokens, or individual words, and transformed into numerical values using a technique called Term Frequency-Inverse Document Frequency (TF-IDF). This transformation allows the content to be analyzed by the SGD classifier, which has been trained to identify phishing content. The model is highly effective, achieving an accuracy of 98% in distinguishing between phishing and safe emails. At the same time, the URLs extracted from the email are analyzed separately. The Gradient Boosting classifier is used for this purpose, focusing on features commonly associated with phishing, such as the length of the domain, the presence of numbers or special characters, and other suspicious patterns within the URL. This classifier also demonstrates strong performance, with a 97% accuracy in identifying phishing URLs. After both the content and URLs are analyzed, the results are combined to make a final decision about the email. If there is Phishing signal in either the content or any of the URLs then the whole email will be labelled as phishing. This two-step approach gives a far better and more efficient detection mechanism, guaranteeing that even if a small segment of the email looks safe, the other is scanned hence reducing the chances of missing a specific phishing attempt. This approach presents more of a thoroughly account for all these techniques are incorporated here which makes it extremely difficult for the phishing attacks on email systems to succeed. The paper also explains the techniques applied in the process of distinguishing the content of the email and the URLs, how the models were trained and validated, and, in general, reveals the efficiency of this approach in terms of protecting against phishing attacks.

Keywords*—Phishing detection, email security, stochastic gradient descent, gradient boosting, URL analysis, email content classification, cybersecurity.*

I. INTRODUCTION

Phishing in the modern world can be established as one of the most spread and constantly evolving threats with scammers trying to lure people into providing their passwords, credit card numbers, or any other sensitive information. These attacks are usually conducted via emails that appear to be originating from such organizations as banks, social networks, relatives or office colleagues. These emails may contain quite convincing words and phrases which make the receiver feel like they have to click on a particular link, or download a certain attachment, or input his or her details in a certain web link which is in fact a fake link. As phishing techniques become sophisticated, traditional signature-based approach of checking for phishing by checking for certain keywords are unable to work. Such techniques are inadequate today because the tactics of phishers have evolved to use such elements as URL disguise or

imitation of popular brands. There is need to come up with more advanced and reliable methods of detecting phish since the consequences of being phished are dire; it costs identities, money as well as grants unauthorized access to important information. This paper focuses on an integrated approach using machine learning techniques in enhancing the accuracy of the phishing detection, as proposing a panacea to the aforementioned conventional methods.

The strategy outlined in this paper incorporates two primary machine learning algorithms: are the Stochastic Gradient Descent (SGD) classifier, and Gradient Boosting classifier. Such algorithms have been chosen because they give good performance with large datasets and they can improve their performance step by step. The methodology begins with the screening of the content of the email and the URLs contained therein or any other content present in the email in which case the content is first separated. This segregation of the email content and URLs enable the method to focus on the shape of the text and the dynamic URLs that used by the phishers more often. Preprocessing is carried out on the content of the received e-mail, and here the content is purged of any material that is unnecessary to include in the analysis such as stop words and special characters. This cleaning process makes the text standardized and also makes the job of the machine learning models easier in that they are able to look for patterns which are characteristic of phishing. The cleaning process of the text is subsequently followed by preprocessing where it is converted to numerical data using the method called Term Frequency-Inverse Document Frequency (TF-IDF) technique, which highlights the importance of some words in the context of the email. This transformed data is then pass through the SGD classifier which has the ability of differentiating between the phishing emails and genuine emails. It proves that SGD classifier is very efficient and produces an accuracy of 98%, therefore, makes it as reliable tool to detect the phishing content with the help of textual features only.

At the same time, the URLs extracted from the email are subject to another analysis with the Gradient Boosting classifier. This classifier aims at the structural aspects of the URLs, such as length of the domain name, the usage of numbers and/or special characters in the URL and the general construction of the URL. These elements are essential in detecting phishing because attackers, as a rule, work on the look-alike URLs of genuine websites with minor differences. The Gradient Boosting classifier has been trained with the set of recognized phishing and safe URLs; in such a way the classifier is capable to identify suspicious patterns and could potentially alert on possible threats. The Gradient Boosting classifier has a 97% accuracy level to carry burdens of another line of protection that checks on the content of the email and then ensures that the URL is scanned for phishing symptoms. This two step analysis greatly reduces the possibility of a phishing mail being flagged as not suspicious, and hence improving the overall security of an email system.

The final step of this strategy entails combining the results from the content of the email and the analysis of the URL to arrive at a comprehensive conclusion of the safety of the email. If either the content or any of them URLs is detected to be phishing by any of its specific classifiers then it is grouped as phishing mail. This approach parousing the both textual and structural counterparts can provide more accurate identification of a given email as a phishing one, or that could contain mnemonic hints of a phishing email. This paper also discusses the practical implementation of these machine learning algorithms; the importance of pre-processing steps; and methods of categorizing the content of emails from URLs as well as the training and testing methodology of the models. Through these points, the research of this work offers a workable and economic detection of phishing that can be implemented in real life email security systems. Furthermore, the paper discusses the flexibility of the proposed method, underlining its potential for updates and enhancements as phishing strategies continue to evolve. This flexibility is essential in sustaining a high level of defense against an ever-changing array of cyber threats, ensuring the safety of users in their online communications.

II. LITERATURE REVIEW

It has been found that the domain of phishing email detection has received a fair amount of attention in various studies which may cover different aspects of the problem. Below is a review of selected literature relevant to the development of automated phishing detection systems:

A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework

This paper[1] presents a comparison of various approaches to the detection of phishing with emphasis on the optimum URL features. It emphasizes on the selection of appropriate features, its source as well as the usage of updated detection models in order to maintain their efficiency. The authors put forward the strategy for creating antiphishing systems that correspond to the current state of phishing threats. The paper is therefore useful for the creation of more effective phishing detection measures; the author reiterates that updating is of paramount importance.

Anti-phishing Prevention Measure for Email Systems

This paper[2] presents the proposed MLAPT machine learning framework for improving the security of email by detecting certain followers of phishing such as interface mimicry and domain confirmation problems. It outlines a clear format of avoiding the attacks because it gives out signs that might mean that one is likely to be a victim of phishing attack. The unified solution is a derivative of the integration of technical and behavioral analysis in the determination of the best method of identifying the adversaries' phishing initiatives. This one is intended for coping with the dynamics in phishing styles, to make sure that the specific countermeasures continue to work.

An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage

Based on the[3] analysis of the previous research results, the training system in this study aims to enhance the security perception and prevent loss of information to phishing attacks. The system is very economical and most data are saved in local computers in contrast to public hosts this would minimize the cases of hacking. The training conducts the user through basic phishing types and provides him with practical tasks on the conducted materials. As can be noted from the paper, the need for constant updating and training is crucial in ensuring nothing adverse and particularly, phishing attacks, happens.

Phishing Attack Detection Using Convolutional Neural Networks

This research work[4] proposes a method of analyzing emails employing Random Forest for conducting feature selection and employing Convolutional Neural Networks (CNN) for identifying phishing emails. This approach yield a high accuracy rate of 98%. As low as 68% in the ability to identify phishing emails from other normal ones. CNN is employed to extract intricate patterns in the given email data and Random Forest yields to an enhanced efficient model by selecting only those vital features. The paper only focuses on selecting perfect features in improving the accuracy of detection and that why this method is useful for even large and diverse datasets.

Email Anti-Phishing Detection Application

In this study, an[5] anti-phishing tool is designed to detect the URLs of phishing websites included in emails with the aid of a decision tree. It has the primary function to safeguard user information by dissecting the URLs and compiling phishing reports. This paper describes the structure of the system and explains in particular how the decision tree is employed for the classification task. The application is compatible with the current systems for managing emails and underlines the necessity of the prompt identification of phishing threats. Some of the ways solving with this tool are shown through different test scenarios in order to showcase the usefulness of the tool.

Convolutional Neural Network Optimization for Phishing Email Classification

The current paper focuses[6] on enhancing Convolutional Neural Networks (CNNs) in the classification of phishing emails. The optimized model yielded good levels of performance, and these are accuracy of 98%, precision of 98,wramid recall, etc. 125% recall, and 98. 269% precision. One of the key contributions of the study is to understand how different CNN parameters have to be tuned for improving detection. Results are obtained using real-life email datasets for evaluation of the proposed model for the detection of phishing emails. However, the approach is scalable to allow extension to accommodate more numbers of comparable words in large-scale implementations. The results indicate the improvement over the traditional approaches that makes it suitable for use in phishing detection.

Detecting Phishing Attacks Using Convolutional Neural Network and LSTM

This work presents a model[7] that considers CNN for spatial feature extraction, and LSTM for the sequential patterns of emails' contents. This makes the model have high accuracy rate of 98 percent. Percentage of increase in the scores when using the software to detect the phishing attacks was 89%. We employ CNNs to extract features of text email data and LSTMs for the sequence of the word. It has the ability to detect elaborate phishing patterns that are not easily detected by other models Through the use of CNN and LSTM the model is highly efficient for this approach and is highly robust in several phishing incidences.

Study of Machine Learning and Deep Learning Algorithms for the Detection of Email Spam Based on Python Implementation

The present paper investigates[8] various ML and DL methods for email spam identification the best of which is the Deep CNN with an accuracy of 99%. The study also analyses the performance of the algorithms such as k-Nearest Neighbors (kNN), Na Ìµve Bayes, Bidirectional Long Short-Term Memory (BiLSTM), and Deep Convolutional Neural Network (Deep CNN). Therefore, it identifies Deep CNN as more appropriate to address the intricate spammers' behavior. The paper describes how different types of spam should be approached with a view to choosing the right model and gives a python implementation. Here the results also show how Deep CNN can be effectively applied in the identification of spam.

The accuracy of several methods used to evaluate email content is compared in the chart. The SGD Classifier outperforms the rest with an astounding accuracy rating of 98.57%. This high accuracy suggests that phishing emails can be identified by the SGD Classifier with great effectiveness. In accurately identifying these kinds of risks, it performs better than other models. The graph illustrates how well the SGD Classifier performs email analysis in comparison to its rivals. Because of its exceptional accuracy, it is a dependable option for phishing detection.

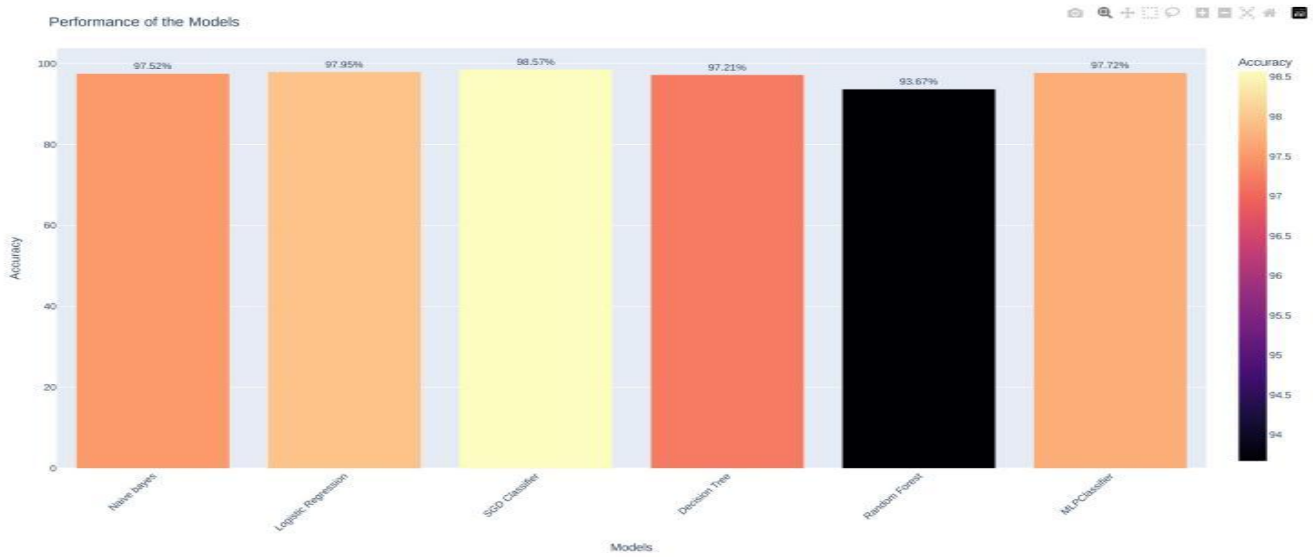

Figure.1 Various Models Performance for email content analysis

III. PROPOSED METHODOLOGY

The suggested method for spotting phishing emails aims to create a strong and precise system by combining content analysis with URL classification. The system's structure, as illustrated in the diagram, has several essential parts that work together to spot phishing emails with great accuracy. The process starts with Gmail Integration using IMAP, which retrieves emails, including both metadata and content. However, this part will focus on the main components involved in the analysis rather than the specifics of the IMAP integration. The core of the phishing detection system relies on Email Content Analysis and URL Classification. The system puts the retrieved email content through two main steps:

Email Content Analysis with SGD Classifier: The data captured from the Email content is cleaned and preprocessed in order to eliminate unnecessary artifacts like stop words and special characters. After that, the cleaned text is transformed into numerical data with the help of a method named Term Frequency-Inverse Document Frequency (TF-IDF). The text passed through this process is taken to the Stochastic Gradient Descent (SGD) classifier which is used to identify patterns of phishing. The classifier takes under consideration the content of a message and assigns a phishing or a safe label to the given e-mail depending on the probability of it being phishing.

URL Extraction and Classification using Gradient Boosting: At the same time, URLs contained in the body of the email are identified and processed independently. URL extraction is concerned with the identification of all the link in an email, which is followed by the evaluation of the characteristics like the length of domain, presence of special characters and other linked structures. These features are important because the phishing URLs often look like the actual URL but with slight difference for the purpose of deceiving the recipient. The Gradient Boosting classifier which was trained with an experiment dataset of phishing as well as legitimate URLs, assesses these features and assigns each URL to be either phishing or safe.

The Inbox Email Classification module merges the findings from content analysis and URL classification. An email gets labeled as phishing if its content or any of its URLs raise red flags. This two-pronged strategy boosts precision by looking at both text and structure that might point to a phishing try.

Users see the end result as a simple "phishing" or "safe" tag for the email. This method not makes phishing detection more accurate but also makes the system more trustworthy by blending two different analysis techniques. The system's building-block design allows it to grow and adapt, so it can be updated and fine-tuned as phishers come up with new tricks.

This suggested approach gives a full answer to the problem of spotting phishing. It brings together the best parts of machine learning-based content study and URL sorting to offer strong protection against phishing attacks.

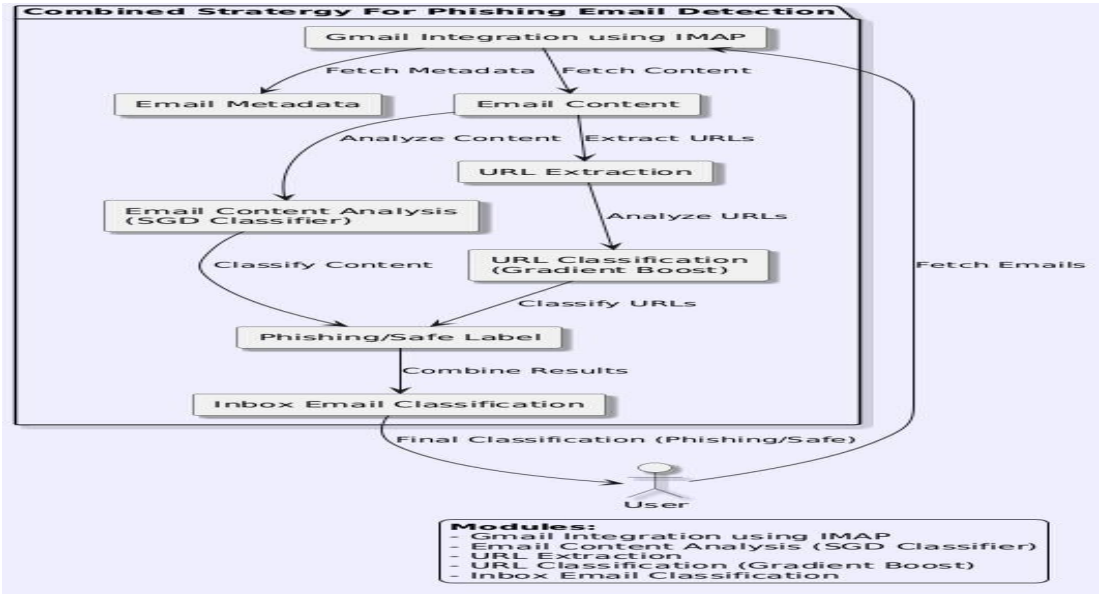

Figure.2 Architecture Diagram of Proposed work

IV. ALGORITHM

Input: Email Data
Output: Phishing or Non-Phishing Classification

- Data Collection
- Data Preprocessing
- Feature Extraction
- Model Training and Evaluation
- Final Classification

A. Data Collection

Gathering data forms the base for putting together datasets to train and test models. The SGD Classifier, which looks at email content, uses a dataset of 16,000 emails in total. This set has a balance of 10,000 safe emails and 6,000 phishing emails. Having a varied dataset is key, as it helps the model see many different types of real and fake email patterns. This allows it to learn the small differences between the two kinds.

The Gradient Boosting Classifier, which helps analyze URLs, uses a dataset of 11,000 web addresses. This collection has 6,000 safe URLs and 5,000 phishing URLs. The web addresses in this set show different types of phishing tricks. Some copy real websites with small changes, while others use short or hidden links. Getting and organizing a top-notch dataset plays a key role. The models' success depends on how varied and true-to-life the training data is.

B. Data Preprocessing

After acquiring the data, the process that comes next is Data Preprocessing where the actual dataset is prepared for analysis. It is critical in the process to reduce the risk of feeding the models with dirty data, data that has not been preprocessed, and normalized. In the preprocessing of email content features, the first step is text cleaning. Before that, it refers to the process of stripping extra white spaces, eliminating all non-text elements and HTML tags. The objective is to preserve all the textual content of the e-mail possible, but only that which has semantic significance to it. After this, the process of stop word removal is done; that involves the elimination of general words such as ‘and’, ‘the’, ‘is’, among others. These words are usually not helpful for classification and may in fact contaminate a model.

Then the lemmatization process is carried out to transform words to the base form, so that all the inflections of the word (for example, ‘run’, ‘ran’, ‘running’) are processed under one feature. This is succeeded by tokenizing where the text is divided into individual words or tokens which can then be classified independently by the machine learning algorithm.

On URLs, the preprocessing usually entails converting all the URLs expressed in the dataset to standardized format. This includes removing protocols (e.g., "http://”, “https://”), converting all symbols to the lowercase allowing person to avoid capital sensitivity problems and extracting the domains and paths for the further analysis. Encoding problems are also checked at URLs and URLs are decoded if needed. The above steps proposing uniformly pre-processed data reduces variations in feature set of the model and in turn helps the model to easily get patterns of phishing.

C. Feature Extraction

Feature Extraction is the phase where the preprocessed data is transformed into a format that the machine learning models can understand. For email content, the TF-IDF (Term Frequency-Inverse Document Frequency) method is utilized. TF-IDF transforms the text into numerical features that represent weight or relevance of each word of the email or what we can term as email word importance. It gives higher weights to the words that are often used in a particular e-mail, but rarely used in all the e-

mails; thus, it enables the model to identify the phishing indicators using the frequency of word usage. When dealing with URLs, the process of extracting features is centered around several key elements:

Domain characteristics: This aspect looks at the domain's size, the use of numbers or hyphens, and if it appears to be a recognized legitimate domain.

Path characteristics: These elements scrutinize the URL path for any suspicious patterns, like lengthy or convoluted paths, or paths that contain deceptive terms (for example, "login", "secure").

Special character analysis: The frequency and presence of special characters such as "?", "=", "&", and "%", which are often employed in phishing URLs to disguise the true purpose of the link.

IP address utilization: Identifying URLs that rely on IP addresses rather than domain names, a tactic frequently used in phishing attempts.

These elements are meticulously chosen to equip the Gradient Boosting Classifier with the essential data to accurately differentiate between phishing and safe URLs.

D. Model Training and Evaluation

Following feature extraction, the subsequent step is Model Training and Evaluation. This phase involves training the machine learning models on the datasets that have been prepared and assessing their effectiveness through various performance indicators.

SGD Classifier for Email Content Analysis:

The training process of the SGD Classifier when it comes to the analysis of the content of emails starts with the conversion of the raw email text into a numerical form using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method. This step is important sine it transforms textual data into feature space where every word gets a numerical value indicative of its importance in the particular email against the background of the entire database. These features are then passed through SGD Classifier which is a linear model that uses stochastic gradient descent to optimize its weight vectors and bias to minimize the loss function; most especially the logistic loss function that is implemented for binary classification of tasks such as detecting phishingurls. The refinement of the model during training by batch processing of data also have the advantage of handling a big data set while avoiding over-fitting by escaping the local minima.

Validation study of the SGD Classifier is done under a set that was not incorporated in the training process. Methods of model evaluation including accuracy, precision, recall and F1-Scores measurements are used to determine the performance of the proposed model. Accuracy gives an overall picture of the performance of the model but in cases of phishing detection, precision and the recall gives better understanding. Precision determines the likelihood of expert identified emails to be actually phishing reducing on false positives while recall rates capture the capability of correctly identifying actual phishing threats thus no potential threats are left out. In situations were there is a skewed number, the F1-Score which is the average of precision and recall scores gives a balanced measure of the models performance.

There are three significant parameters that are adjusted during the evaluation of the SGD Classifier, where they are learning rate, the regularization parameter, and the number of iterations. It also uses cross-validational techniques in which the sample is divided into subsamples in an effort to evaluate its proficiency regardless the particular samples on which it was trained. This all-encompassing method to train and assess the SGD Classifier makes it strong and dependable. As a result, it works well to tell apart phishing emails from real ones in everyday use.

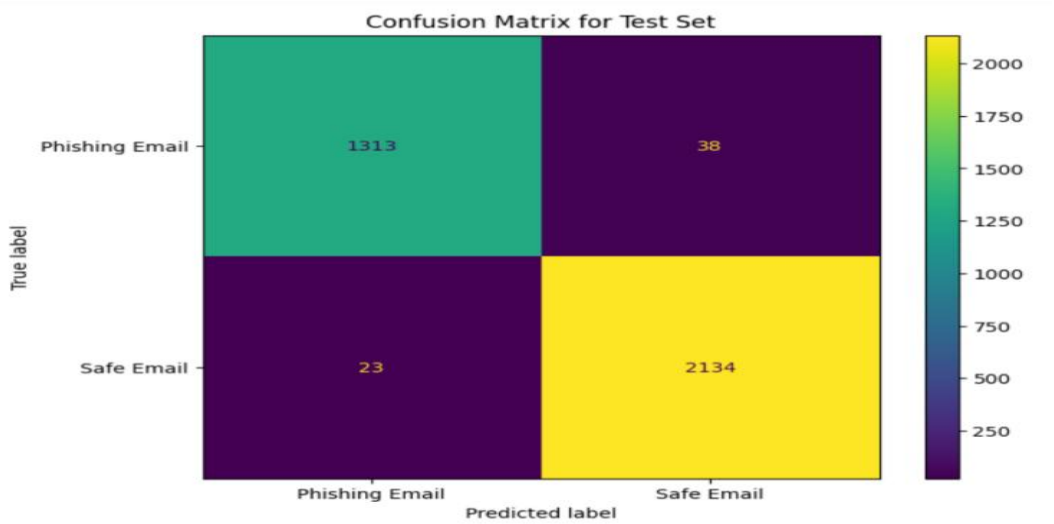

Figure.3 Confusion matrix Diagram of SGD Classifier

	precision	recall	f1-score	support
Phishing Email	0.98	0.98	0.98	1351.0
Safe Email	0.98	0.99	0.99	2157.0
accuracy	0.98	0.98	0.98	0.98
macro avg	0.98	0.98	0.98	3508.0
weighted avg	0.98	0.98	0.98	3508.0

Figure.4 Classification report for SGD Classifier

Gradient Boosting for URL Analysis:

When it is implemented in URL analysis for phishing detection, Gradient Boosting Classifier's training process is arranged systematically to enhance the classifier's capability in the differentiation of the safe URLs from the phishing ones gradually. The training starts with input of a set of URLs with each URL described by a set of features related to its structure and lexical composition. Some of these features may include the number of characters in the URL, or the use of such characters as '@,' '#,' '$,' '&', the number of appearances of specific words most of the time used in phishing, among others. The basic concept of the Gradient Boosting algorithm is to first build a weak learner – often an un-pruned decision tree of only few levels – for making preliminary prediction with these features. The residuals, or more simply errors, from this first model, namely the difference between the predicted labels and the actual labels are then used to train the second tree in the sequence. successively each new tree is designed to learn from the mistakes of the previous tree, which makes it possible to invest all efforts in classifying the most complex URLs. This is repeated again and again with each new tree, and the trees added into the ensemble, and the process gradually improves the model's prediction of the loss function and optimization is through gradient descent.

The process of evaluating the Gradient Boosting model is, in fact, one of the most important phases since the model trained has to be tested for its performance. It is in this phase that the model undergoes the validation stage using new set of data that has not been used in the training phase in order to test its forecast capability. Precision, recall, F1-Score and, Accuracy are generally used to evaluate the performance of the execution. Accuracy gives the overall performance of the model that how much of it is correct, where as precision and recall gives deep look into the model as how well it is able to identify phishing urls without going wrong too often. Accuracy is very critical in this regard because linking a

genuine URL as the phishing one may interfere with normal activities. If there's any doubt that the model is not capturing all the possible phishing threats, then Recall confirms it. Precisely, Since both precision and recall are rather sensitive to skewed distribution or the distribution of classes in general, the F1-Score that is the harmonic mean of precision and recall presents a balanced picture of the report-card like performance.

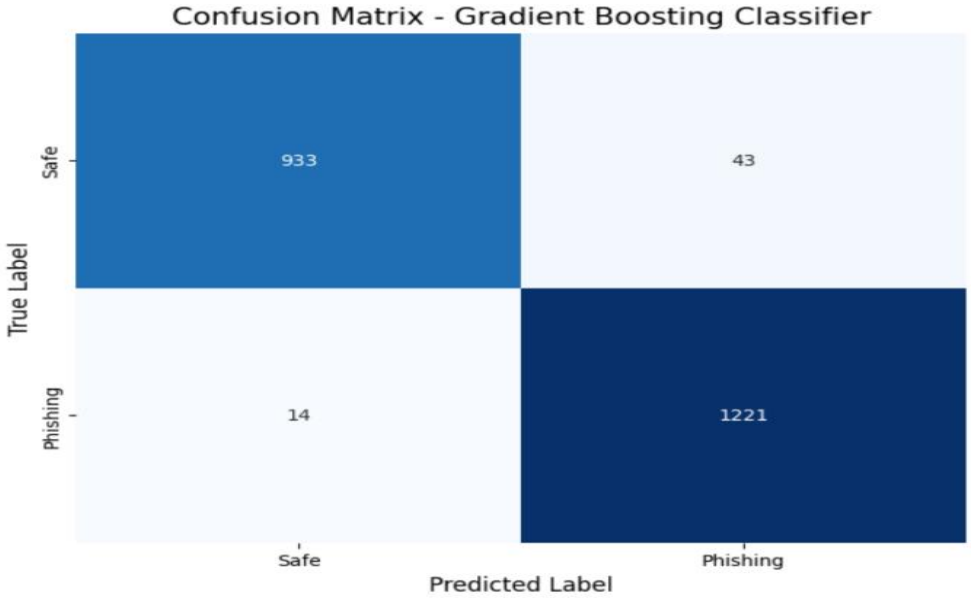

Figure.5 Confusion matrix Diagram of Gradient Boosting Classifier

There exists possibility of optimizing the Gradient Boosting Classifier even more through the act of hyperparameter tuning in relation to URL analysis. Splitting criteria for deciding the structure of individual decision trees, the number of decision trees, and the size of the trees are managed to avoid both high bias and high variance. Also during the assessment phase methods such as cross validation are used to check that the model is well generalized to other data that you have not used in training. In cross-validation, the data sample is divided into subsets and the model is trained on different combinations of these subsets so that an assurance of the stability of the technique used in arriving at the result can be ensured. Again, going through such intense training and evaluation minimizes the possibility of the Gradient Boosting Classifier performing inaccurately when used in the actual predictive mission of finding out phishing URLs from real-word URLs.

	precision	recall	f1-score	support
Phishing URL	0.9852164730728616	0.9559426229508197	0.970358814352574	976.0
Safe URL	0.9659810126582279	0.9886396761133	0.977190876350540	1235.0
accuracy	0.9742198100407056	0.9742198100407056	0.9742198100407056	0.9742198100407056
macro avg	0.9755987428655448	0.9723032952810778	0.973774845351557	2211.0
weighted avg	0.9744721068982471	0.9742198100407056	0.974175004568534	2211.0

Figure.6 Classification report for Gradient Boosting Classifier

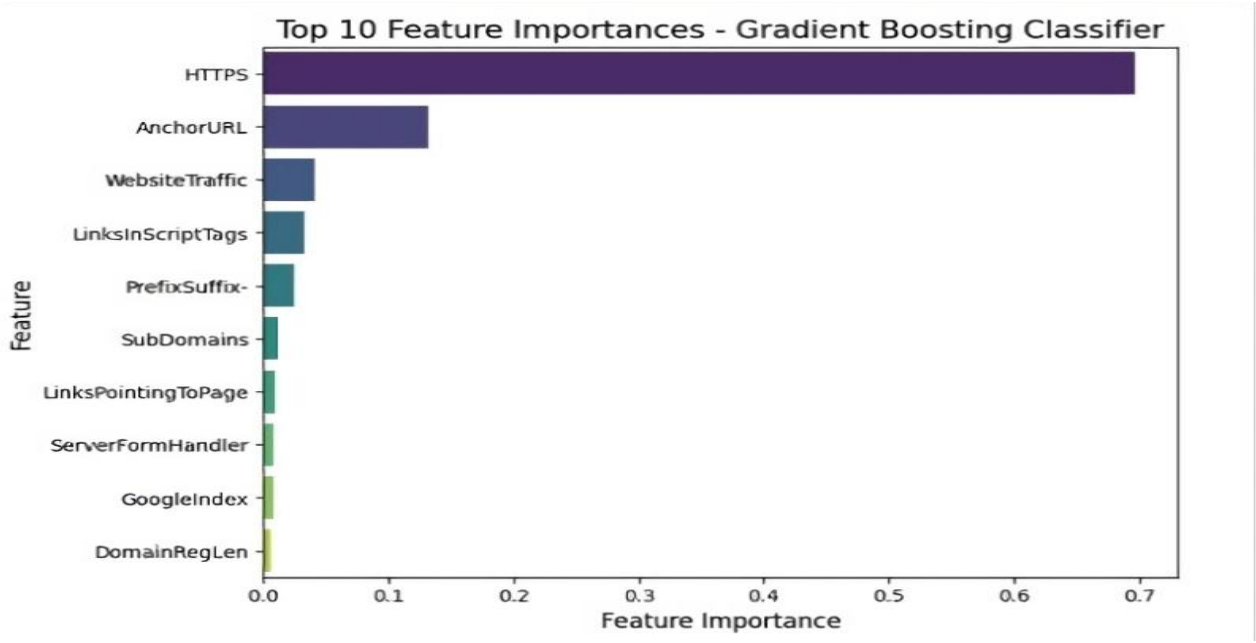

Figure.7 Feature Importance Diagram

E. Final Classification

The last step in the classification process merges the outcomes from both the analysis of content and the categorization of URLs. Following the processing of the email's content and URLs by each model, their results are merged to determine the email's status. Should either the content or any of the URLs be identified as phishing, the email is categorized as phishing. This unified method guarantees that even if one part of the email is deemed safe, the other can still spot possible dangers, thus minimizing the chance of missing any threats. This ultimate verdict is then shown to the user, with phishing emails being highlighted for additional review. This two-tiered strategy not only enhances the precision of detecting phishing emails but also creates a strong defense system that can evolve to new phishing strategies as they arise. By integrating content analysis with URL classification, this system provides an all-encompassing approach to tackling the issue of phishing emails, ensuring that users are shielded from various phishing attacks.

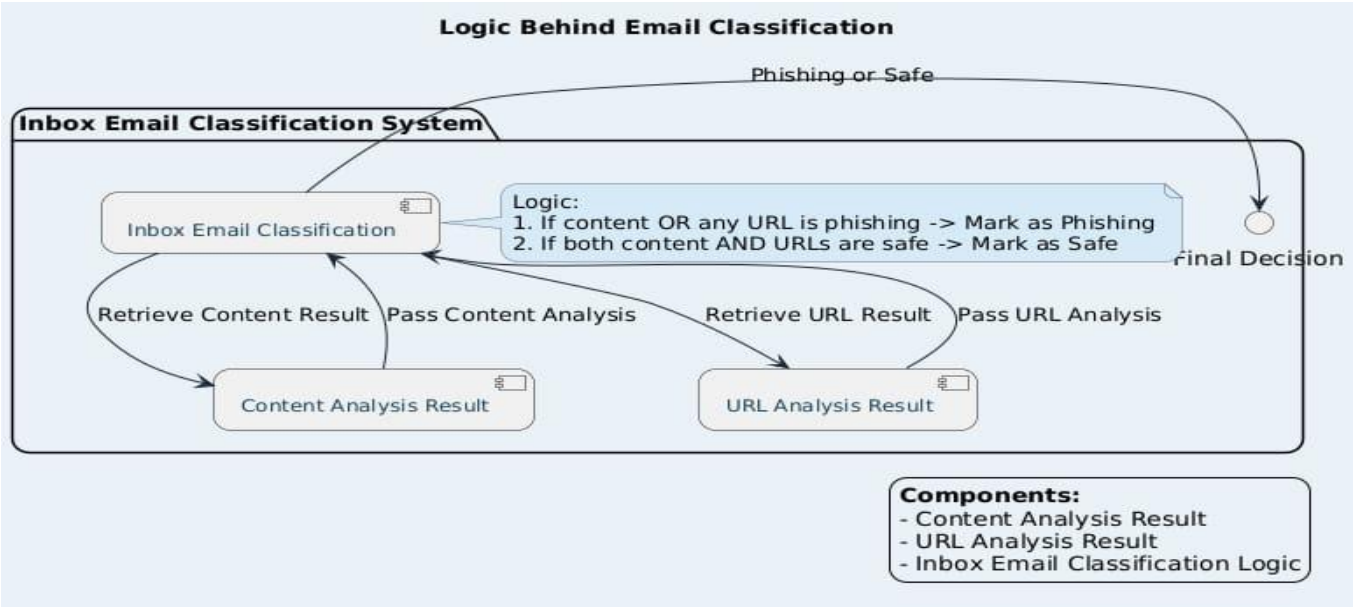

Figure.8 Logic Behind Email Classification

V. The Overall Outcomes

The work as a whole was able to design a unified framework for the detection of fraudulent emails and it was able to use both the text content and the URL of an email message as feature vectors for classification. The Stochastic Gradient Descent (SGD) classifier that has been trained from a data-set that consisted of 16000 emails was equally successful and accurate up to 98% in distinguishing between safe and phishing emails which proves the hypothesis that is laid down in content-based detection. At the same time, the Gradient Boosting classifier, trained on 11,000 URLs received 97% of accuracy and, thus, it is important to focus on the URLs analysis when the decision about the phishing attempts will be made. By combining these two classifiers, The BLP1 classifier along with the BLP2 classifier form a one system that offers deep protection against phishing attacks. The complete content analysis of the emails under consideration in combination with the URLs decreases the rate of false negative to minimum allowing for detecting the most intricate kinds of phishing. This double-layered approach increases the general dependability and flexibility of the detection system in order to address new phishing approaches. Finally, the results of this study offer a strong and indeed easily implementable approach to prevent phishing attacks, or in any case of detecting them, providing users with a wider layer of protection against various forms of phishing attacks, thereby contributing to the improvement of cybersecurity in communication.

VI. Future Work

However, there are many ideas for the future improvement and further exploration of the given current approach that shows the accuracy of detection of phishing emails. Another possibility of development is the connection with the continuous data feed, so that the system will notice the phishing attempts in the process and protect the users on the fly. Also, enlarging the data set with more sets of updated and newer techniques of phishing can improve the performance of the model. A further interesting path for future research is thus the scenario of training the system with more sophisticated NLP techniques like transformers or deep learning models, which might provide higher capacity of the system to comprehend the more intricate content of the emails. The authors also state that experimenting with the different configuration of several classifiers to form an ensemble of classifiers could also be effective in having better detection of malware and a minimized number of false positives. Furthermore, improving the capability of the system on the identification of such scams through various modes of communication including social media and instant messaging may not be limited to emails. Finally, designing the system in a way that will minimize user-interaction with the detected phishing material and provide basic instructions and/or precautions to be taken by the user, in simple language, will make the system more effective and actually implementable in everyday real-life scenarios. These future enhancements would serve to make a more sophisticated and effective way of detecting phishing attacks in order to have improved and flexible defense from the constantly changing threats.

VII. Conclusion

This paper proposed a novel and efficient method to detect phishing emails from the other. normal emails through content analysis and URL classification by utilizing machine learning approach. The SGD classifier resulted in a predictor value of 98% when the data used was the content of the emails and the GB classifier which had the URL data scored 97% predictor value. The inclusion of these two models in the system made it have a strong ability to combat against phishing sites as opposed to achieving high false negatives basing only on the textual content of the sites and URLs without the other's support. Apart from raising accuracy of detection, the two-tier system was also a great success as it adjusted to new strategies used by phishers. With regard to content and URL-based threats the described methodology provides for more accurate and efficient approach to identification of phishing, which in turn protect users from as many kinds of phishing as possible. The findings of this Research are beneficial for the progress of cybersecurity, especially in the field of email security because this method allows identifying, prioritizing, and eliminating risks associated with phishing, and enhances the

effectiveness of current solutions greatly. For future improvements, it would be possible to integrate a real-time identification of these problems, the inclusion of more comprehensive sets of data and employment of more complex natural language processing methods, for instance. These future directions could enhance the system to become more rigid and powerful and guarantee it will function effectively with advanced incidents of phishing attacks. Finally, this research builds the prerequisites for safer and more secure digital communication which will guard the users from one of the most significant threats of the world wide web.

VIII. REFERENCE

[1] S. Patil and S. Dhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 588-593, doi: 10.1109/ICACCS.2019.8728356.

[2] T. Ayodele, C. A. Shoniregun and G. Akmayeva, "Anti-phishing prevention measure for email systems," World Congress on Internet Security (WorldCIS-2012), Guelph, ON, Canada, 2012, pp. 208-211.

[3] M. Higashino, T. Kawato, M. Ohmori and T. Kawamura, "An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage," 2019 5th International Conference on Information Management (ICIM), Cambridge, UK, 2019, pp. 82-86, doi: 10.1109/INFOMAN.2019.8714691.

[4] S. S. S. P, S. Velpula, R. Parise, N. K. Vamsi and S. K. Chaitanya, "Phishing Attack Detection Using Convolutional Neural Networks," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 1381-1385, doi: 10.1109/ICACCS57279.2023.10113077.

[5] R. A. A. Helmi, C. S. Ren, A. Jamal and M. I. Abdullah, "Email Anti-Phishing Detection Application," 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), Shah Alam, Malaysia, 2019, pp. 264-267, doi: 10.1109/ICSEngT.2019.8906316.

[6] C. McGinley and S. A. S. Monroy, "Convolutional Neural Network Optimization for Phishing Email Classification," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 5609-5613, doi: 10.1109/BigData52589.2021.9671531.

[7] J. S and S. Eliyas, "Detecting phishing attacks using Convolutional Neural Network and LSTM," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 2789-2792, doi: 10.1109/ICACITE57410.2023.10183234.

[8] S. T. Singh Surinder Pal Singh, M. D. Gabhane and C. Mahamuni, "Study of Machine Learning and Deep Learning Algorithms for the Detection of Email Spam based on Python Implementation," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 637-642, doi: 10.1109/ICDT57929.2023.10150836.

[9] J. Ramprasath, S. Priyanka, R. Manudev and M. Gokul, "Identification and Mitigation of Phishing Email Attacks using Deep Learning," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 466-470, doi: 10.1109/ICACITE57410.2023.10182911.

[10] A. K. Sharma, R. K. Galav and B. Sharma, "A Comprehensive Survey of various Cyber Attacks," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-4, doi: 10.1109/ISCON57294.2023.10111998.