Master of Science in Computer Science
May 2023

Model of detection of phishing URLs based on machine learning

Kateryna Burbela

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Kateryna Burbela
E-mail: kabu22@bth.student.se

University advisor:
Oleksii Baranovskyi
Department of Computer Science

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

ABSTRACT

Background: Phishing attacks continue to pose a significant threat to internet security. One of the most common forms of phishing is through URLs, where attackers disguise malicious URLs as legitimate ones to trick users into clicking on them. Machine learning techniques have shown promise in detecting phishing URLs, but their effectiveness can vary depending on the approach used.

Objectives: The objective of this research is to propose an ensemble of two machine learning techniques, Convolutional Neural Networks (CNN) and Multi-Head Self-Attention (MHSA), for detecting phishing URLs. The goal is to evaluate and compare the effectiveness of this approach against other methods and models.

Methods: a dataset of URLs was collected and labeled as either phishing or legitimate. The performance of several models using different machine learning techniques, including CNN and MHSA, to classify these URLs was evaluated using various metrics, such as accuracy, precision, recall, and F1-score.

Results: The results show that the ensemble of CNN and MHSA outperforms other individual models and achieves an accuracy of 98.3%. Which comparing to the existing state-of-the-art techniques provides significant improvements in detecting phishing URLs.

Conclusions: In conclusion, the ensemble of CNN and MHSA is an effective approach for detecting phishing URLs. The method outperforms existing state-of-the-art techniques, providing a more accurate and reliable method for detecting phishing URLs. The results of this study demonstrate the potential of ensemble methods in improving the accuracy and reliability of machine learning-based phishing URL detection.

Keywords: Phishing, URL address, Deep learning, Convolutional layer, Multi-head self-attention.

CONTENTS

ABSTRACT ..... **III**

CONTENTS ..... **IV**

1 INTRODUCTION **1**

1.1 WHAT IS PHISHING? 1
1.2 BACKGROUND 1
What are phishing URLs and machine learning? 1
Evaluation of different phishing detection methods 2
Deep learning methods 4
Convolutional neural network 5
Multi-head self-attention 5
1.3 DEFINING THE SCOPE OF YOUR THESIS 6
1.4 OUTLINE 7

2 RELATED WORK **8**

2.1 ON THE CONTENT 8

3 METHOD **10**

3.1 ON THE CONTENT 10
3.2 DEFINING RESEARCH QUESTIONS 10
3.3 MODEL OVERVIEW 11
3.4 CONTROLLED EXPERIMENT 15
3.5 DATA 15
3.6 FEATURES 16

4 RESULTS AND ANALYSIS **18**

4.1 ON THE CONTENT 18
4.2 LITERATURE REVIEW 11
Suitable datasets ..... **Помилка! Закладку не визначено.**
Identifying feature selection and engineering techniques. 18
Performance comparison 19
Evaluation of the effectiveness of existing solutions 20
4.3 MODEL PERFORMANCE 23
4.4 RESULTS ANALYSIS 24

5 DISCUSSION **27**

5.1 RQ1: WHAT TYPES OF MACHINE LEARNING ALGORITHMS HAVE BEEN USED FOR DETECTING PHISHING URLS, AND HOW CAN THESE ALGORITHMS BE TRAINED AND OPTIMIZED? 27
RQ1.1: What types of datasets were used to train machine learning algorithms? 27
RQ1.2: What accuracy measures are used to compare such algorithms? 28
RQ1.3: What machine learning algorithms gave the best results in detecting phishing websites? 29
5.2 RQ2: HOW EFFECTIVE IS THE PROPOSED ENSEMBLE OF TWO TECHNIQUES IN DETECTING PHISHING URLS? 29
5.3 RQ3: HOW EFFECTIVE IS THE PROPOSED ENSEMBLE OF TWO TECHNIQUES IN DETECTING PHISHING URLS COMPARED TO OTHER METHODS? 30

6 CONCLUSION AND FUTURE WORK **31**

6.1 CONCLUSION 31
6.2 FUTURE WORK 32

7 REFERENCES ..... ПОМИЛКА! ЗАКЛАДКУ НЕ ВИЗНАЧЕНО.

1 INTRODUCTION

1.1 What is phishing?

Phishing is one of the most common ways of obtaining personal data. [1] To minimize the damage from a phishing attack, it is necessary to detect it as early as possible. Almost every type of phishing attack uses phishing URLs. [1]

Phishing URLs are links to websites or web pages that are designed to look like legitimate websites, but in reality, they are malicious sites created by cybercriminals to steal personal information such as login credentials, credit card numbers, and other sensitive data. [1]

The importance of detecting phishing URLs cannot be overstated as they pose a significant threat to individuals and organizations alike. Here are some reasons why detecting phishing URLs is crucial: [1]

1. Protection against Identity theft: Phishing URLs are often designed to trick individuals into revealing their login credentials, bank account details, and other personal information. By detecting these URLs, individuals and organizations can protect themselves against identity theft. [1]
2. Prevention of Financial Loss: Phishing attacks can cause significant financial losses to individuals and organizations. By detecting and blocking phishing URLs, organizations can prevent cybercriminals from stealing money and sensitive data. [1]
3. Protection against Malware: Phishing URLs often contain links to malicious software that can harm a computer or a network. By detecting and blocking these URLs, organizations can prevent malware infections and data breaches. [1]
4. Maintaining Trust: Organizations that are victims of phishing attacks can lose the trust of their customers, clients, and partners. By detecting and preventing phishing attacks, organizations can maintain their reputation and avoid negative publicity. [1]

In summary, detecting phishing URLs is essential for protecting against identity theft, financial loss, malware infections, and maintaining trust. It is crucial for individuals and organizations to be vigilant in identifying and reporting phishing URLs to stay safe in the digital world.

1.2 Background

What are phishing URLs and machine learning?

A phishing URL is a malicious link that an attacker distributes on the Internet in order to trick users into gaining access to their sensitive data such as passwords, credit card numbers, and other personal information.

Machine learning is an artificial intelligence technique in which computer algorithms are trained based on large amounts of data. In the case of the phishing URL detection question, machine learning can be used to detect suspicious patterns in link addresses that may indicate phishing attacks.

Phishing URL detection using machine learning uses the analysis of large amounts of data, including various features such as the URL, the appearance of the web page, the context, and so on. Machine learning models that are used to detect phishing URLs can be trained on real examples of phishing sites and sites that are not phishing, allowing them to identify suspicious links based on the trained model.

Thus, using machine learning to detect phishing URLs can be an effective method to protect users from phishing-related cyberattacks.

Evaluation of different phishing detection methods

The problem with detecting phishing URLs is that they are designed to look like legitimate URLs, making it difficult for users to distinguish them from genuine ones. Phishing URLs are often designed to look like well-known websites, such as banking or e-commerce websites, in order to trick users into giving away sensitive information. [4]

One of the challenges in detecting phishing URLs is that they can be highly targeted and personalized, making them difficult to detect using traditional rule-based methods. Moreover, phishing attacks are becoming more sophisticated and complex, requiring more advanced techniques to detect them. [3]

To combat phishing attacks, several methods have been developed to detect phishing URLs. These methods include:

1. Blacklists: Blacklists contain lists of known phishing URLs that have been identified by security experts. These lists can be used by browsers, email providers, and security software to block users from accessing known phishing websites. [1] Many popular brands use blacklisting as a tool to protect against phishing URLs, including Google, Microsoft, Apple, and many others. These companies use various methods to maintain and update their blacklists, such as automated crawlers and user reports. For example, Google's Safe Browsing service maintains a constantly updated list of unsafe websites, including those involved in phishing attacks, and warns users before they visit these sites. Microsoft's SmartScreen filter, built into the Edge and Internet Explorer web browsers, also uses blacklisting to protect users from potentially harmful websites. [12]
2. Domain Name System (DNS) filters: DNS filters can be used to block access to known phishing URLs. When a user attempts to access a known phishing website, the DNS filter redirects the user to a safe page or blocks access to the site altogether. [6] There are several popular brands that use DNS filtering as a tool to protect against phishing URLs.

Some of these brands include Cisco, Barracuda Networks, Sophos, McAfee, and Symantec. These companies provide DNS filtering services that can help organizations block access to known phishing sites, as well as malicious IP addresses and domains. By leveraging DNS filtering, organizations can proactively protect their networks and users from phishing attacks.

3. User awareness training: Educating users about the risks of phishing attacks and how to detect phishing URLs can be an effective method of preventing phishing attacks. Users can be taught to look for signs of phishing URLs, such as misspellings in the domain name or the presence of unusual characters in the URL. [6] Many popular companies, including Microsoft, Google, and Amazon, provide user awareness training to their employees as part of their cybersecurity protocols. For example, Microsoft offers a variety of training resources, including webinars and online courses, to help employees recognize and avoid phishing scams. Google provides similar resources, including simulated phishing attacks to test employees' awareness and training modules to help improve their skills. [15]

4. Machine learning algorithms: Machine learning algorithms can be used to detect phishing URLs by analyzing the characteristics of the URL, such as the domain name, the length of the URL, and the presence of certain keywords. Such algorithms can also detect similarities between phishing URLs and known phishing websites. [2] Machine learning is a more effective approach to detecting phishing URLs than blacklisting or DNS filtering because it can adapt to new and evolving threats. Blacklisting and DNS filtering rely on maintaining lists of known malicious URLs or domains, which can quickly become outdated as attackers create new URLs or domains. [13]

On the other hand, machine learning models can analyze patterns and features of URLs and web pages to identify new and unknown phishing attacks, even if they have not been seen before. Machine learning models can also learn from past data and improve their accuracy over time, making them more effective at detecting phishing URLs.

Additionally, machine learning can analyze various features beyond just the URL or domain, such as the appearance of the web page and the context in which the link is presented. This makes it more difficult for attackers to circumvent detection by simply using different URLs or domains. [14] Overall, machine learning is a more proactive and adaptive approach to detecting phishing URLs, making it a better tool for protecting against evolving cyber threats.

To conclude, detecting phishing URLs is an essential part of preventing phishing attacks. Several methods have been developed to detect phishing URLs, including blacklists, DNS filters, machine learning algorithms, and user

awareness training. These methods can help individuals and organizations stay safe from phishing attacks and protect their sensitive information.

The following machine learning methods or models can be used as the solution to the problem:

1. Supervised learning: This method involves training a machine learning model on a labeled dataset of phishing URLs and legitimate URLs. The model can then be used to classify new URLs as either phishing or legitimate based on the patterns learned during training. [17]
2. Unsupervised learning: In this method, the machine learning model is trained on an unlabeled dataset of URLs, and it learns to identify patterns and anomalies in the data that may indicate the presence of phishing URLs. [17]
3. Semi-supervised learning: This method combines elements of both supervised and unsupervised learning. The machine learning model is trained on a small labeled dataset of phishing and legitimate URLs, but it also learns from an unlabeled dataset to identify new patterns and anomalies in the data. [17]
4. Deep learning: Deep learning methods, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), can be used to detect phishing URLs by learning features directly from raw data, such as website screenshots or network traffic logs. [17]
5. Ensemble learning: This method involves combining multiple machine learning models to improve overall performance. Ensemble methods can be particularly effective for detecting phishing URLs, as they can combine different types of models with varying strengths and weaknesses. [3] [17]

Each method has its own strengths and weaknesses, and it is needed to experiment with multiple methods to determine the most effective approach for detecting phishing URLs using machine learning.

Deep learning methods

Deep learning is a subset of machine learning that uses artificial neural networks to analyze and classify data. While traditional machine learning approaches rely on manually selecting and engineering features for classification, deep learning models can learn these features directly from raw data, allowing them to potentially uncover more complex patterns. Usually these methods transform the regular URL into a matrix [18] and then provide the metrics to the chosen model so it`ll work out and define whether the URL is legal or the phishing one.

In terms of detecting phishing URLs, deep learning approaches have been shown to achieve higher accuracy rates than traditional machine learning methods. This is because deep learning models can identify more subtle patterns

and features in URLs that may not be obvious to humans or traditional machine learning algorithms.

However, it's worth noting that deep learning approaches may also require more data and computational resources than traditional machine learning methods, which can be a challenge for some organizations. Ultimately, the choice between deep learning and traditional machine learning approaches will depend on the specific needs and resources of the organization.

Convolutional neural network

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly well-suited for image and pattern recognition tasks. It works by taking an input image, applying a series of filters to extract features at different levels of abstraction, and then using those features to classify the image into different categories. [16]

CNNs are commonly used in computer vision applications, including detecting phishing URLs. This is because phishing URLs often contain images or logos that are designed to mimic legitimate websites and trick users into clicking on them. By training a CNN on a large dataset of phishing and legitimate URLs, the network can learn to recognize the patterns and features that are indicative of phishing URLs.

One of the reasons that CNNs are considered the best method for detecting phishing URLs is their ability to automatically learn relevant features from the raw input data. This means that instead of hand-crafting features based on domain knowledge, the network can learn to extract features that are most relevant to the task at hand. Additionally, CNNs are highly scalable, meaning that they can handle large datasets and can be trained on powerful computing clusters to improve their accuracy.

Multi-head self-attention

Multi-head self-attention is a technique used in deep learning, particularly in the field of natural language processing (NLP). It is an extension of self-attention, which allows a neural network to weigh the importance of different parts of the input sequence when processing it. Multi-head self-attention extends this concept by performing multiple self-attention operations in parallel, allowing the model to learn multiple representations of the input sequence and capture more complex patterns. [

In the context of detecting phishing URLs, multi-head self-attention can be used to extract features from the URL text and identify important patterns related to phishing. By considering multiple attention heads, the model can learn different aspects of the URL, such as the presence of suspicious keywords or unusual domain names, and combine them into a final representation that can be used for classification.

While CNNs are a popular method for detecting phishing URLs, multi-head self-attention can offer advantages in certain situations. For example, it can be more effective in processing long sequences of text, where traditional convolutional filters may struggle to capture relevant patterns. Additionally, multi-head self-attention can be more interpretable than CNNs, allowing researchers to better understand which parts of the input are most relevant for the final classification decision. However, the effectiveness of multi-head self-attention will ultimately depend on the specific dataset and problem at hand.

1.3 Defining the scope of your thesis

While there has been a significant amount of research on detecting phishing URLs using machine learning, there are several reasons why further research is still necessary:

1. Phishing attacks are constantly evolving: Cyber criminals are continually finding new ways to carry out phishing attacks, and this requires researchers to constantly develop new and more advanced machine learning techniques to detect them. This question is highly shown in the article “Antiphishing through Phishing Target Discovery” written by Wenyin, Liu, Qiu and Quan [5]
2. Accuracy and efficiency can still be improved: While machine learning algorithms have been shown to be effective at detecting phishing URLs, there is still room for improvement in terms of accuracy and efficiency. Researchers can explore different types of algorithms and features to improve the performance of phishing detection systems as it is done in *Phishing attack detection using Machine Learning"* written by Sundara Pandiyan, Selvaraj, Burugari and Kanmani. [2]
3. Large amounts of data need to be processed: With the exponential growth of the internet and the increasing number of URLs being generated every day, there is a need to process vast amounts of data to detect phishing URLs. Machine learning algorithms can help to automate this process, but further research is necessary to develop algorithms that can handle the large-scale data processing requirements. [2]
4. Combining techniques: While looking through different articles that show the use of machine learning techniques [2], [9], [10], [11] it is observable that there are some techniques with better results but not many articles use the combination of two algorithms as it is proposed in the article *"HDP-CNN: Highway deep pyramid convolution neural network combining word-level and character-level representations for phishing website detection"* written by Zheng, Yan and Leung. So to get the model with the better result it is important to find the techniques with the best results and combine them.

Overall, further research is necessary to keep pace with the evolving threat landscape and to develop more accurate and efficient phishing detection system.

The main focus of the research is to develop more accurate and efficient model for detecting phishing URLs using machine learning algorithms.

The research involves exploring different types of machine learning algorithms, such as supervised or unsupervised learning, and determining which algorithms are most effective in detecting phishing URLs. The research also involves identifying the features of phishing URLs that are most indicative of a phishing attempt and developing algorithms that can detect these features with high accuracy.

Other areas of focus in this research may include:

1. Evaluating the effectiveness of machine learning algorithms in detecting evolving and complex phishing attacks.
2. Investigating the impact of different training data sets on the accuracy of machine learning algorithms for detecting phishing URLs.
3. Investigating the importance of reducing false positives and false negatives in machine learning-based phishing detection systems.

Overall, the focus of the research is to improve the effectiveness and efficiency of phishing detection systems and enhance the security of individuals and organizations against phishing attacks.

This research is highly relevant and important in advancing the field of cybersecurity, protecting sensitive information, improving user awareness, complying with regulations, and maintaining business continuity.

1.4 Outline

The main purpose of this thesis is to define the most effective model which will be detecting phishing URLs. Specifically, the research aims to:

1. Use an ensemble of two techniques to distinguish between phishing URLs and regular URLs, even in cases where the phishing URLs are highly targeted or personalized.
    a. Train a model on large datasets of known phishing URLs and legitimate URLs to improve their accuracy and effectiveness.
2. Compare the performance of an ensemble model with the performance of different machine learning algorithms and identify the most effective ones for detecting phishing URLs.

Overall, the main objectives of the research are to improve the ability of organizations and individuals to detect and prevent phishing attacks, thereby reducing the risk of financial losses, data breaches, and other negative consequences associated with these attacks.

2 RELATED WORK

2.1 On the content

While deep learning methods have shown great promise in detecting phishing URLs, they also require more training time compared to conventional machine learning methods. However, they can provide a more accurate and comprehensive solution for phishing detection.

One article that stands out in research on detecting and preventing phishing attacks is "Phishing attack detection using Machine Learning" by Sundara Pandiyan, Selvaraj, Burugari, and Kanmani. [2] This study explores the use of various machine learning techniques for detecting and preventing phishing attacks.

Several other research articles have also proposed the use of machine learning techniques for identifying phishing. For instance, Zheng, Yan, and Leung's article "HDP-CNN: Highway deep pyramid convolution neural network combining word-level and character-level representations for phishing website detection" [8] suggests a new machine learning model for detecting phishing websites using an ensemble of two algorithms. Wei, Ke, and Novak's article "Accurate and fast URL phishing detector: A convolutional neural network approach" [9] provides detailed information on the use of Convolutional Neural Network (CNN) for detecting phishing URLs, which is one of the best algorithms for solving this problem.

Sahongoz, Baykal, and Bulut's article "Phishing detection from URLs by using neural networks" [10] explores the use of neural networks for detecting phishing URLs, while Remmide, Boumahdi, Boustia, and Feknous' article "Detection of Phishing URLs Using Temporal Convolutional Network" [11] proposes a machine learning model for detecting phishing URLs using a temporal convolutional network.

These studies demonstrate the effectiveness of machine learning algorithms in detecting phishing URLs and provide insights into the features and techniques that can be used to develop more accurate and efficient phishing detection systems. Researchers have proposed representation learning techniques that focus on feature selection to be learned, as manual feature extraction can be time-consuming and complicated. However, a significant challenge in representation learning is the very high dimensional features, often in million or even billion scale, which poses a challenge for training a classification model in practice.

Deep learning approaches, including MLP, LSTM, DBNs, and CNN, can automatically extract features from samples, without the need for manual feature engineering by humans. This allows for more accurate detection of phishing URLs, as the models are able to identify subtle patterns and relationships in the data. For example, MLP has been widely used in text classification, including in the detection of phishing URLs. Mohammad, Thabtah, and McCluskey (2017)

and Nguyen, Ba, Nguyen, and Nguyen (2014) both used MLP for phishing detection and achieved high accuracy rates. [24]

LSTM is a type of neural network that can learn sequential dependencies from character sequences. Bahnsen et al. (2017) translated each input character of the URL into a 150-step sequence by 128-dimensional embedding and then fed the sequence into an LSTM layer. This resulted in a higher F-1 score and the authors found that LSTM took less memory than conventional machine learning methods. [25]

DBNs were used by Zhang and Li (2017) to detect phishing websites. They calculated the probability distribution through the edge distribution of the energy function and got the maximum likelihood estimation. This method improved 1% accuracy and 2% F-1 score over Support Vector Machine (SVM). [26]

Finally, Yang et al. (2019) [26] transformed a URL into a matrix with one-hot encoding, and then used embedding to decrease the dimension of the matrix, before putting the matrix into CNN and then LSTM, using softmax function to calculate the result.

While deep learning methods have shown great promise in detecting phishing URLs, they also require more training time compared to conventional machine learning methods. However, they can provide a more accurate and comprehensive solution for phishing detection.

3 METHOD

3.1 On the content

The research proposal of an effective method for detecting phishing URLs using machine learning includes three research questions to investigate the existing methods for detecting phishing URLs, the types of machine learning algorithms used, the accuracy measures employed, and the machine learning algorithms that have shown the best results.

It is also outlined that a controlled experiment involving data exploration, data and feature selection, model training, model evaluation, and comparison to other methods will be placed while doing research. The article aims to contribute the ensemble of effective and efficient methods for detecting phishing URLs using machine learning.

3.2 Defining research questions

Detecting phishing URLs is a critical task in cybersecurity, and machine learning (ML) has emerged as a powerful tool in this field. In order to gain a better understanding of the existing methods for detecting phishing URLs answering the following research question will be needed:

RQ1: What types of machine learning algorithms have been used for detecting phishing URLs, and how can these algorithms be trained and optimized?

This one aims to investigate the types of ML algorithms that have been used for detecting phishing URLs and explore how these algorithms can be trained and optimized. This involves examining a range of techniques, from conventional machine learning methods to more advanced deep learning approaches.

Additionally, it is needed to explore the types of datasets that have been used to train these algorithms, in order to understand the characteristics of the data that are most relevant for effective phishing detection. So the following sub question will help us:

RQ1.1: What types of datasets were used to train machine learning algorithms?

The next step is to examine the accuracy measures that have been used to evaluate these algorithms using the next question:

RQ1.2: What accuracy measures are used to compare such algorithms?

This could include measures such as precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), accuracy, and others. By understanding which measures are being used, it will be easier to evaluate and compare the effectiveness of different algorithms in detecting phishing attacks. So this step will allow us to determine which algorithms have been the most successful in detecting phishing URLs which should be explained in

RQ1.3: What machine learning algorithms gave the best results in detecting phishing websites?

Answering those questions helps in developing more accurate and efficient phishing detection system. In this field, to my mind, only an ensemble of two effective machine learning techniques for detecting phishing URLs can contribute. The approach will involve combining two methods to improve overall performance. To evaluate the effectiveness of our proposed approach, a formal experiment is conducted using a large dataset of phishing URLs. This experiment is supposed answer our research questions and determine how effective the ensemble method is in detecting phishing URLs. The answer to the following question should include the information:

RQ2: How effective is the proposed ensemble of two techniques in detecting phishing URLs?

To conclude, it`ll be needed to compare the effectiveness of an ensemble method with other existing methods for detecting phishing URLs. This will involve benchmarking an approach against the state-of-the-art techniques, to determine how it performs relative to other methods in the next

RQ3: How effective is the proposed ensemble of two techniques in detecting phishing URLs compared to other methods?

By answering these research questions and conducting a thorough evaluation of the method, it is aimed to contribute to the development of effective and efficient methods for detecting phishing URLs using machine learning.

3.3 Literature review

Here is the list of the keywords, synonyms and related terms used for an SLR:

- Phishing
- Phishing attacks
- Phishing URLs
- Information security
- Machine lerning.

The data sources used for the research are SCOPUS. IEEE and GOOGLE Scholar.

The Literature review method is used to answer the sub questions of the RQ1.

Here is the list of the inclusion criteria used for the study selection:

- The work must have been published between 2017 and 2022.
- The work has to deal with “phishing URL detection” or any synonyms and related terms.
- The work must be a full text. The entire contents should be available through the data source.
- The work should be written in English.

Here is the list of the exclusion criteria used for the study selection:

- Irrelevant and out-of-scope studies.
- Repeated/duplicated literature found from defined data sources.
- Studies not in the English language.
- Papers not matching quality assessment criterion.
- Studies don’t contain tools or techniques.

The search result was limited to articles that met the inclusion and exclusion criteria. The search was made by using the filtration based on the inclusion and exclusion criterias.

Quality assessment criteria, provided above, was used to extract the unnecessary data. By using these criteria the articles which weren't accurate or contained irrelevant or undisclosed information were excluded.

- Q1: The study must be focusing mainly on social engineering.
- Q2: The stated conclusions should be supported by the presented data.
- Q3: The framework of study must be provided in sufficient detail to interpret the research accurately.
- Q4: The accuracy of how the data was measured and reported must be provided clearly.
- Q5: Contribution and credibility of the work based on the results of the study

Results of the literature review

I have thoroughly reviewed a total of 40 articles during my research. It is important to note that some studies employed multiple techniques for phishing detection, resulting in their inclusion under multiple categories. Out of the 40 studies, I found that 29 of them utilized machine learning approaches for detecting phishing attacks.

Considering these numbers, it can be observed that approximately 71.25% of the research conducted in this field focused on utilizing machine learning algorithms, which is the highest among the five techniques mentioned. Among the machine learning approaches, Deep Learning was the most commonly employed, with 26 articles (66.25%) utilizing this technique. The List-based and

DNS filtering techniques were used in 7 articles (17.5%), while User Awareness Training was employed in 5 articles (12.5%).

It is worth mentioning that these figures differ from the original study due to the revised number of articles I analyzed, which was 40 instead of the previously stated number. The distribution of articles across each technique can be visualized in Figure 3.1.

Phishing detection approaches		**References**
Machine learning		[2], [3], [6], [7], [8], [9], [10], [11], [12], [13], [18], [19], [20], [21], [22], [24], [25], [26], [27], [30], [31], [35], [36], [37], [38], [39], [40], [41]
	Deep Learning	[8], [9], [10], [11], [12], [18], [20], [21], [22], [24], [25], [26], [27], [30], [31], [35], [36], [37], [38], [39], [40], [41]
List based or DNS filterin[g		[6], [7], [17], [18], [28], [32], [33]
User awareness		[5], [6], [29], [34], [37]

Fig. 3.1 – Phishing detection approaches

In summary, my analysis of these 40 articles reveals that machine learning, particularly Deep Learning, is the predominant approach for detecting phishing attacks in the literature, showcasing its prominence and effectiveness in this domain. The distribution of articles across each Machine learning approaches used for detecting phishing URLs can be visualized in Figure 3.2

Machine learning approach	**References**
Naïve-Bayes	[2], [3], [6], [19], [37]
Random forest	[3], [6], [21], [30], [39]
CNN	[6], [8], [9], [10], [11], [16], [18], [20], [22], [26], [30], [31], [36], [38], [39], [40]
SVM	[18], [26], [39]
LSTM	[6], [12], [18], [20], [25], [26], [30], [31], [36], [41]
MLP	[20], [24], [37]
MHSA	[6], [18], [30], [36]
CNN+LSTM	[26]
CNN+RNN	[22], [24], [25]

Fig. 3.2 – Machine learning techniques

There are five different types of datasets commonly used in machine learning, such as labeled, unlabeled datasets, synthetic, mixed and real-world datasets

In the analysis of these 40 articles, it was found that the majority, 80% of the studies, utilized labeled datasets for training their machine learning models. The distribution of articles across each type of dataset can be visualized in Figure 3.3, providing a visual representation of the prevalence and usage of different dataset types in the examined studies.

Dataset type	**References**
Labeled	[6], [8], [9], [10], [11], [16], [18], [20], [22], [26], [30], [31], [35]
Unlabeled	[6], [12], [18], [20], [25], [26], [30], [31], [36], [41]
Synthetic	
Mixed	[22], [24]
Real-World	[13]

Fig. 3.3 – Types of datasets used for detecting phishing URLs

Talking about the Accuracy measures used for the evaluation of different machine learning approaches there are 7 the most common measures:

- Accuracy
- Precision
- Recall
- F1 Score
- FRP
- ROC Curve
- AUC-ROC

In the examined articles, all of these accuracy measures were used in almost all of the studies to assess the performance of the machine learning models. The distribution of articles across each type of accuracy measure can be visualized in Figure 3.4, offering insights into the prevalence and usage of different evaluation metrics in the analyzed studies.

Accuracy measures	**References**
Accuracy	[2], [3], [6], [7], [8], [9], [10], [11], [12], [13], [18], [19], [20], [21], [22], [24], [25], [26], [27], [30], [31], [35], [36], [37], [38], [39], [40], [41]
Precision	
Recall	
F1 score	
FPR	
ROC Curve	[6], [7], [8], [9], [13], [18], [19], [20], [21], [26], [27], [30], [31], [35], [36]
AUC-ROC	

Fig. 3.4 – Accuracy measures used for the evaluation of machine learning approaches

3.4 Controlled experiment

To conduct an experiment on detecting phishing URLs using machine learning, the following key steps must be taken.

- Explore data
- Data selection
- Feature Selection
- Train model
- Evaluate model
- Comparison to the other methods

Firstly, it is important to thoroughly explore the data being used. This will help to identify any potential biases or limitations in the data. Next, feature selection must be carried out to determine which features are most relevant for the evaluation of the model. Finally, the model must be evaluated using appropriate metrics\features to assess its performance and identify areas for improvement.

When the review will be completed it`ll be possible to answer the RQ1. All the sub questions also will be clarified in this process. A comprehensive review of the different types of machine learning algorithms used for detecting phishing URLs, along with an analysis of the datasets used to train the algorithms, the accuracy measures employed, and the machine learning algorithms that have shown the best results in detecting phishing websites.

After getting the model trained and evaluated the RQ2 can be answered. Since the chosen ensemble of two machine learning techniques will be tested and can be evaluated by the chosen features. The answer will also include a detailed analysis of the experimental results, performance metrics, and the strengths and limitations of the proposed approach.

Once the ensemble will be evaluated it can be compared to the other existing models and methods which answers the last RQ3.

3.5 Data

Given that the literature review demonstrates a predominant usage of labeled datasets in the field, the decision was made to utilize a labeled dataset for the controlled experiment.

4000 URLs were collected to train and evaluate phishing URL detection model. The URLs were obtained from two sources: legitimate URLs were obtained from https://www.similarweb.com/top-websites/, while phishing URLs were obtained from https://phishtank.org/.

Legitimate URLs were chosen from the list of top websites provided by SimilarWeb, which ranks websites based on their estimated traffic and popularity. We chose the top 20,000 websites from this list to ensure that the dataset covers a wide range of legitimate URLs.

Phishing URLs were obtained from PhishTank, a community-driven website that tracks and reports phishing scams. We chose the 20,000 most recent and unique phishing URLs available on the PhishTank website at the time of data collection.

The choice of these two sources is based on the fact that SimilarWeb is a reliable and widely used source for legitimate URLs, while PhishTank provides a comprehensive and up-to-date database of phishing URLs. By using both sources, we aimed to ensure that our dataset includes a representative sample of both legitimate and phishing URLs.

The dataset was split into two equal parts: 2000 legitimate URLs and 2000 phishing URLs. This balanced split was chosen to ensure that the model is not biased towards either class and can accurately detect both legitimate and phishing URLs.

Possible limitations of the dataset used in this study include representativeness of legitimate URLs, potential temporal bias in phishing URLs, relatively small dataset size and the choice of cross-validation approach. These limitations should be considered when interpreting the results and generalizing the model's performance.

A batch size of 10 was used and trained the model for 50 epochs, using a 5-fold cross-validation approach. In each round of cross-validation, the dataset was divided into five equal parts, with four parts used for training and one part used for testing. The final performance of the model was evaluated as the average of the five test results.

3.6 Features

The most used, due to the literature research 5 metrics - Accuracy (Acc), False Positive Rate (FPR), Recall (Rec), Precision (Pre) and F-1 score (F1) - are used to evaluate machine learning models for detecting phishing URLs.

Accuracy is the percentage of correctly classified URLs out of all the URLs in the dataset. It is a useful metric for balanced datasets, where the number of legitimate and phishing URLs are roughly equal. However, accuracy can be misleading in imbalanced datasets, where the number of legitimate URLs is significantly larger or smaller than the number of phishing URLs.

False Positive Rate (FPR) is the percentage of legitimate URLs that are incorrectly classified as phishing URLs. It is an important metric for detecting false alarms, which can have significant consequences in practical applications.

Recall (Rec) is the percentage of phishing URLs that are correctly classified as phishing URLs. It measures the completeness of the classification, and is particularly important in situations where detecting all the phishing URLs is critical.

Precision (Pre) is the percentage of correctly classified phishing URLs out of all the URLs classified as phishing URLs. It measures the accuracy of the

classification, and is particularly important in situations where false alarms are costly.

F-1 score (F1) is the harmonic mean of precision and recall, and provides a balanced evaluation of both metrics. It is a useful metric for imbalanced datasets, where precision and recall need to be considered together.

4 RESULTS AND ANALYSIS

4.1 On the content

Phishing website detection can be done using machine learning models, and the accuracy of these models depends on the datasets used for training and testing, the features extracted from websites, and the algorithms and classifiers employed. Various datasets are used for training, including Alexa and Common Crawl for legitimate sites that can be used for phishing, and Phish-tank and Open-Fish for suspicious URLs reported by end-users. Several features are used to compare machine learning methods, including accuracy, false positive rate, recall, precision, and F-1 score. The most commonly used and effective methods for phishing URL detection are Naive-Bayes, Random Forest, CNN, MLP, MHSA, LSTM, and CNN+RNN. Accuracy is high for all the methods, ranging from 96% to 99.84%, but other factors such as training time, computational resources, and robustness to noise should also be considered to determine which method is better.

4.2 Identifying feature selection and engineering techniques.

As it was mentioned, there are several features that can be used to compare machine learning methods in detecting phishing URLs. Some of the commonly used features are:

1. 1.*Accuracy:* measures the overall correctness of the classification model. It is calculated by dividing the number of correctly classified samples by the total number of samples.
2. *False positive rate (FPR):* measures the ratio of false positive predictions to the total number of negative samples. A high FPR indicates that the model is predicting non-phishing URLs as phishing URLs.
3. *Recall:* measures the ratio of true positive predictions to the total number of positive samples. A high recall indicates that the model is correctly identifying a high percentage of phishing URLs.
4. *Precision*: measures the ratio of true positive predictions to the total number of predicted positive samples. A high precision indicates that the model is accurately identifying phishing URLs.
5. *F-1 score:* the harmonic mean of precision and recall. It provides a single score that balances both precision and recall.

Accuracy, FPR, recall, precision, and F-1 score are the most suitable features for comparing machine learning methods in detecting phishing URLs because they provide a comprehensive evaluation of the model's performance. While accuracy measures overall correctness, FPR, recall, precision, and F-1 score provide information on the model's ability to identify phishing URLs and avoid false positives. These measures are critical in phishing detection, as failing

to identify phishing URLs can lead to security threats, and a high false positive rate can result in user frustration and distrust of the system.

To get the better understanding on how the metrics will be counted let's denote the results as the following:

- True Positive (TP) is the number of phishing URLs classified correctly
- True Negative (TN) is the number of legitimate URLs classified as legitimate
- False Positive (FP) is the number of legitimate URLs classified as phishing ones
- False negative (FN) – the number of phishing URLs classified as legitimate ones.

With these selections the features for evaluation and the comparison of the models will look the next way:

- Acc = (TP+TN)/(TP+TN+FP+FN)
- FRP = FP/(FP+TN)
- Rec = TP/(TP+FN)
- Pre = TP/(TP+FP)
- F1 = 2 x Pre x Rec / (Pre + Rec)

4.3 Performance comparison

Some of the most usable and effective methods used for phishing URLs detection are:

1. Naive-Bayes: It is a probabilistic algorithm based on Bayes' theorem, which assumes that all features are independent of each other. It works by calculating the probability of a URL being phishing based on the occurrence of certain features in the URL. The algorithm is trained on labeled datasets, and during testing, it uses the learned probabilities to classify new URLs as phishing or legitimate. Naive-Bayes is simple and fast and has been shown to perform well in detecting phishing URLs. [1]
2. Random Forest: It is an ensemble learning algorithm that creates multiple decision trees and combines their predictions to make a final decision. Each tree in the forest is trained on a random subset of the dataset, and during testing, the algorithm aggregates the predictions of all trees to make a final decision. Random Forest is known for its accuracy and robustness to noisy data. [1]
3. CNN (Convolutional Neural Network): It is a deep learning algorithm inspired by the structure of the human brain. CNNs work by applying convolutional filters to extract features from the input data. These features are then passed through multiple layers of neurons to learn complex representations of the data. CNNs have been shown to perform well in detecting phishing URLs by learning features such as domain name, URL length, and character n-grams. [21]

4. MLP (Multilayer Perceptron): It is a type of artificial neural network that is composed of multiple layers of neurons. Each neuron in the network receives inputs from the previous layer and applies a nonlinear activation function to produce an output. MLPs are trained using backpropagation, a supervised learning algorithm that adjusts the weights of the neurons to minimize the error between the predicted and actual outputs. MLPs have been shown to perform well in detecting phishing URLs by learning features such as domain age, SSL certificates, and URL length. [27]

5. MHSA (Multi-Head Self-Attention): It is a variant of the Transformer model, a deep learning algorithm used in natural language processing. MHSA works by applying multiple self-attention heads to the input data to extract contextual information. The outputs of these heads are then concatenated and passed through multiple layers of neurons to learn representations of the data. MHSA has been shown to perform well in detecting phishing URLs by learning features such as domain name, URL length, and character n-grams. [27]

6. LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is designed to overcome the vanishing gradient problem in traditional RNNs. LSTM networks are capable of learning long-term dependencies in sequential data by selectively remembering and forgetting information over time. They achieve this through the use of memory cells, which are gated units that control the flow of information into and out of the cell. [20]

7. CNN+RNN: This is a hybrid deep learning model that combines the strengths of CNN and Recurrent Neural Networks (RNNs). In the case of phishing URL detection, the CNN part of the network learns the local features of the URLs, while the RNN part learns the sequential dependencies between them. [22]

Based on the results in the table provided, it can be seen that the accuracy of all the methods is quite high, ranging from 96% to 99.84%. However, the accuracy alone may not be enough to conclude which method is better, as other factors such as training time, computational resources, and robustness to noise should also be taken into account.

№	ML detection algorithm	Acc %	FPR %	Rec %	Pre %	F1 %
1	Naïve-Bayes [19]	97.18				
2	Random forest [21]	97				
3	CNN [20]	96.61	3.50	97.09	96.61	96.85
4	LSTM [20]	97.20	1.80	98.63	96.45	97.53
5	MLP [20]	96.65		96.65	96.65	96.65
6	CNN+RNN [22]	97.9	3.10	98.39	96.76	97.57
7	CNN+LSTM [23]	93.28	1.80	97.13	99.12	98.11

Table 4.1: Performance comparison

4.4 Evaluation of the effectiveness of existing solutions:

Based on the provided table, the effectiveness of different machine learning methods in detecting phishing URLs can be evaluated based on their accuracy, FPR (False Positive Rate), Recall, Precision, and F1-score.

1. Naive-Bayes [19] This method has an accuracy of 97.18%, which is quite high. However, we do not have information on FPR, Recall, Precision, and F1-score.
2. Random forest [21]: This method has an accuracy of 97%, but we do not have information on FPR, Recall, Precision, and F1-score.
3. CNN [20]: This method has an accuracy of 96.61%, with a relatively high FPR of 3.50%. It has a high Recall of 97.09%, indicating that it correctly identified the majority of phishing URLs. However, its Precision is slightly lower at 96.61%, indicating that it classified some legitimate URLs as phishing URLs. The F1-score of this method is 96.85%.
4. LSTM [20]: This method has the highest accuracy of 97.20%, with a low FPR of 1.80%. It has a high Recall of 98.63%, indicating that it correctly identified the majority of phishing URLs. However, its Precision is slightly lower at 96.45%, indicating that it classified some legitimate URLs as phishing URLs. The F1-score of this method is 97.53%.
5. MLP [20]: This method has an accuracy of 96.65%, but we do not have information on FPR, Recall, Precision, and F1-score.
6. CNN+RNN [22]: This method has an accuracy of 97.9%, with a relatively high FPR of 3.10%. It has a high Recall of 98.39%, indicating that it correctly identified the majority of phishing URLs. However, its Precision is slightly lower at 96.76%, indicating that it classified some legitimate URLs as phishing URLs. The F1-score of this method is 97.57%.
7. CNN+LSTM [23]: This method has the lowest accuracy of 93.28%, with a low FPR of 1.80%. It has a high Recall of 97.13%, indicating that it correctly identified the majority of phishing URLs. It has the highest Precision of 99.12%, indicating that it classified very few legitimate URLs as phishing URLs. The F1-score of this method is 98.11%.

In summary, LSTM is the most effective method in terms of accuracy and FPR. CNN+LSTM has the highest Precision, indicating that it classified very few legitimate URLs as phishing URLs. CNN+RNN and Naive-Bayes have high accuracy but relatively high FPR. Random forest and MLP have good accuracy but no information on FPR, Recall, Precision, and F1-score.

The ensemble of two methods is better than the usage of one because each method has its strengths and weaknesses, and by combining them, we can improve the overall performance. For example, the CNN+RNN method in the table has a higher accuracy than Naive-Bayes, but Naive-Bayes is faster and requires less computational resources. By combining the two, we can get a more accurate and efficient phishing URL detection system.

CNN (Convolutional Neural Network) and MHSA (Multi-Head Self-Attention) are both powerful deep learning architectures that have been successfully applied in various natural language processing tasks, including the detection of phishing URLs. An ensemble of CNN and MHSA can improve the detection performance by combining the strengths of both models.

CNN is a neural network that applies convolutional filters to the input data, typically used for image recognition tasks. In the context of NLP, CNN can learn important features of a text by applying a sliding window to the input sequence and extracting local patterns of words. These local features are then combined and transformed into a higher-level representation of the input text. CNN has been shown to be effective in detecting phishing URLs by extracting n-gram features from the URLs and using them to train a classifier.

On the other hand, MHSA is a transformer-based model that uses self-attention to compute a weighted sum of the input tokens, allowing the model to capture global dependencies and long-range relationships between words in a sentence. In the context of NLP, MHSA has been shown to be effective in modeling the semantic meaning of a text by attending to relevant words in the input sequence. In the detection of phishing URLs, MHSA can be used to learn a representation of the URL that captures its semantic meaning and context.

An ensemble of CNN and MHSA can combine the advantages of both models and improve the detection performance by leveraging their complementary strengths. The CNN can capture local patterns and n-gram features of the URL, while the MHSA can model its semantic meaning and context. By combining the predictions of both models, the ensemble can achieve higher accuracy and robustness to different types of phishing URLs.

4.5 Model overview

Since Multi-head self-attention (MHSA) has an excellent performance in natural language processing (NLP) tasks, which is able to compute features' weights and identify dependencies between different characters in text it will also be effective in analyzing URLs and potentially outperforming long short-term memory (LSTM) networks. Also since convolutional neural networks (CNNs) are good at automatically learning features without the need for human intervention this techniques can be combined in order to leverage their strengths and improve phishing website detection.

While building a model for an ensemble of two machine learning techniques it can be splited in some parts. As the first step of the model we can place the *Embedding layer* that converts the input URL into a matrix representation using One-Hot Encoding. Then, since two techniques are used, this matrix should be duplicated into two copies for *Feature learning and Feature weight calculation*. During the weight calculation process, one of the copies is fed into MHSA layers to calculate the features' weights. At the same time, during the

feature extraction process, another copy of the URL matrix is put into convolutional layers to learn features, and the previous layer's output will be treated as input for the next layer. After the two concurrent processes finished, the two parts of the output data will be fed into the *Output block* together to compute the final classification result.

The output block first takes the original features and two copies of the feature weights as input. The output is fed into a fully connected layer with the Sigmoid activation function, which outputs a result between 0 and 1. If the output is greater than 0.5, the input URL is classified as legitimate, otherwise, it is classified as phishing.

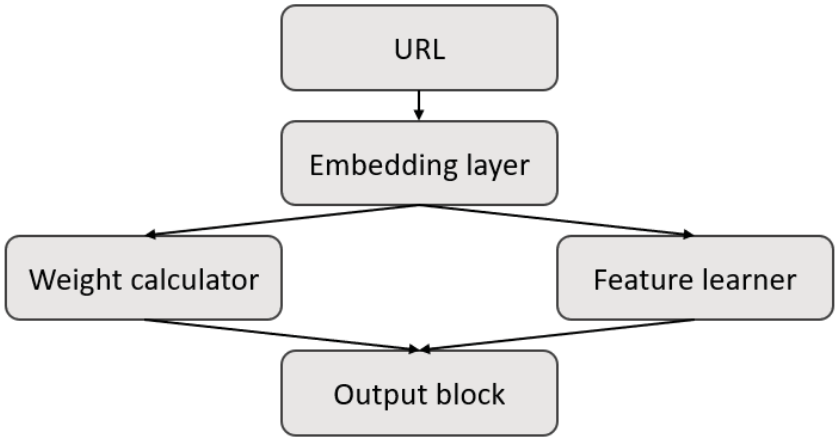

Figure 4.1: Model overview

4.6 Model performance

Taking in an account the advantages of usage of an ensemble of the CNN and Multi-head self-attention algorithms the architecture for the model is composed of three main components: an embedding layer, a feature learner, and a weight calculator.

The embedding layer transforms a URL string into a matrix with a number of rows equal to the length of the URL and 84 columns, representing each of the 84 different characters that can appear in a URL. One-Hot Encoding is used to represent each character, and the matrix is reduced to 64 columns via a neural network. URLs with varying lengths are processed with a fixed-length string by trimming or padding.

The feature learner extracts features from the output matrix of the embedding layer, using a convolutional layer, two residual layers, and a fully connected layer. The convolutional layer contains five conventional kernels and a max-pooling layer. The residual layers solve the degeneration problem of accuracy saturation by adding the input and output of the convolutional layer. The fully connected layer enhances the expression ability of neural networks and the efficiency of feature extraction.

The weight calculator contains an MHSA layer, two residual layers, and a fully connected layer, and is responsible for calculating feature weights. The output of the embedding layer is injected with positional encoding, which contains the relative position of the characters in the URL string sequence. The positional encoding matrix is obtained using sine and cosine functions. The result matrix is then fed into the MHSA layer, which owns eight heads. Finally, the feature matrix is obtained and used for classification or prediction tasks.

The output block gives a result from 0 to 1. The bigger the result is (>0.5) the less possible is that the URL is a phishing one.

Overall, the performance metrics of the model are:

- Accuracy: 0.9834
- False Positive Rate (FPR): 0.0176
- Precision: 0.9844
- Recall: 0.9816
- F1 Score: 0.9830

Let's break down what each of these performance metrics means:

1. Accuracy: the model achieved an accuracy of 0.9834, which means that it correctly identified 98.34% of phishing URLs in the dataset.
2. False Positive Rate (FPR): the FPR is 0.0176, which means that 1.76% of non-phishing URLs were incorrectly classified as phishing URLs.
3. Precision: the precision is 0.9844, which means that 98.44% of the URLs identified as phishing URLs by the model were actually phishing URLs.
4. Recall: the recall is 0.9816, which means that the model correctly identified 98.16% of all actual phishing URLs in the dataset.
5. F1 Score: the F1 score is 0.9830, which means that the model achieved a good balance between precision and recall.

Overall, these results suggest that the ensemble of CNN and Multi-head self-attention performed very well in detecting phishing URLs, with high accuracy, precision, recall, and F1 score, and low false positive rate.

4.7 Results analysis

Based on the results of the research the following table is provided to compare the performance of different machine learning techniques used for detection of phishing websites:

№	ML detection algorithm	Acc %	FPR %	Rec %	Pre %	F1 %
1	Naïve-Bayes [19]	97.18				
2	Random forest [21]	97				
3	CNN [20]	96.61	3.50	97.09	96.61	96.85
4	LSTM [20]	97.20	1.80	98.63	96.45	97.53

5	MLP [20]	96.65		96.65	96.65	96.65
6	CNN+RNN [22]	97.9	3.10	98.39	96.76	97.57
7	CNN+LSTM [23]	93.28	1.80	97.13	99.12	98.11
	CNN+MHSA	98.34	1.76	98.44	98.16	98.30

Table 4.2: Performance comparison with the proposed ensemble

Here's a detailed analysis and comparison of the different machine learning techniques used for detecting phishing URLs based on the given table:

1. Naïve-Bayes [19]: This algorithm has an accuracy of 97.18%. Naïve-Bayes is a simple and popular classification algorithm that works well with high-dimensional data, but it assumes that all features are independent of each other, which may not always be the case.
2. Random forest [21]: This algorithm has an accuracy of 97%. Random forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
3. CNN [20]: This algorithm has an accuracy of 96.61%, a false positive rate (FPR) of 3.50%, a recall (Rec) of 97.09%, precision (Pre) of 96.61%, and an F1 score of 96.85. CNNs (Convolutional Neural Networks) are commonly used for image classification tasks but can also be used for text classification. The FPR of 3.50% indicates that 3.50% of legitimate URLs were incorrectly classified as phishing URLs.
4. LSTM [20]: This algorithm has an accuracy of 97.20%, an FPR of 1.80%, a Rec of 98.63%, Pre of 96.45%, and an F1 score of 97.53. LSTMs (Long Short-Term Memory networks) are a type of recurrent neural network that can capture long-term dependencies in sequential data. The low FPR and high recall indicate that this model is effective at identifying phishing URLs while minimizing false positives.
5. MLP [20]: This algorithm has an accuracy of 96.65% and no other metrics are provided in the table. MLPs (Multilayer Perceptrons) are a type of feedforward neural network that can learn non-linear relationships between input and output data.
6. CNN+RNN [22]: This algorithm has an accuracy of 97.9%, an FPR of 3.10%, a Rec of 98.39%, Pre of 96.76%, and an F1 score of 97.57. Combining a CNN with an RNN (Recurrent Neural Network) can capture both spatial and sequential features in the input data, leading to improved performance compared to using either model alone.
7. CNN+LSTM [23]: This algorithm has an accuracy of 93.28%, an FPR of 1.80%, a Rec of 97.13%, Pre of 99.12%, and an F1 score of 98.11. Combining a CNN with an LSTM can capture both local and global features in the input data, leading to improved performance. However, the low accuracy and F1 score indicate that this model may not perform as well as the others on this task.

8. CNN+MHSA: This algorithm has the highest accuracy of 98.34%, the lowest FPR of 1.76%, a Rec of 98.44%, Pre of 98.16%, and an F1 score of 98.30. CNN+MHSA combines a CNN with MHSA (Multi-Head Self-Attention), which can capture long-term dependencies in the input data and attend to multiple parts of the sequence simultaneously. Even though the FPR still can be reduced by post-processing techniques, active learning, ongoing evaluation and feedback loops. These approaches aim to refine the model's performance and strike a better balance between false positives and false negatives. This model performs the best overall on this task, with high accuracy, low false positives, and high precision and recall.

In summary, the CNN+MHSA model has the highest performance on this task, followed by LSTM, CNN+RNN, Random Forest.

5 DISCUSSION

5.1 RQ1: What types of machine learning algorithms have been used for detecting phishing URLs, and how can these algorithms be trained and optimized?

To combat phishing attacks, several methods have been developed to detect phishing URLs. These methods include blacklists, DNS filters, user awareness training, and machine learning algorithms. Each method has its own strengths and weaknesses, and a combination of methods may be needed to provide effective protection against phishing attacks.

Blacklists and DNS filters rely on maintaining lists of known malicious URLs or domains, which can quickly become outdated as attackers create new URLs or domains. However, they can be effective for blocking known phishing sites and preventing users from accessing them. User awareness training can help users recognize and avoid phishing scams, but it may not be effective for more sophisticated attacks that are personalized to the victim.

Machine learning algorithms can be used to detect phishing URLs by analyzing the characteristics of the URL, such as the domain name, the length of the URL, and the presence of certain keywords. Such algorithms can also detect similarities between phishing URLs and known phishing websites. Machine learning is a more effective approach to detecting phishing URLs than blacklisting or DNS filtering because it can adapt to new and evolving threats. Machine learning models can analyze patterns and features of URLs and web pages to identify new and unknown phishing attacks, even if they have not been seen before. Machine learning models can also learn from past data and improve their accuracy over time, making them more effective at detecting phishing URLs.

Ensemble learning involves combining multiple machine learning models to improve overall performance.

RQ1.1: What types of datasets were used to train machine learning algorithms?

There are several types of machine learning algorithms that can be used for detecting phishing URLs, based on the chosen datasets including supervised learning, unsupervised learning, semi-supervised learning, deep learning, and ensemble learning. Each method has its own strengths and weaknesses, and the choice of algorithm may depend on the specific needs of the organization and the nature of the phishing attacks they are trying to detect.

Supervised learning involves training a machine learning model on a labeled dataset of phishing URLs and legitimate URLs. The model can then be used to classify new URLs as either phishing or legitimate based on the patterns learned during training. Supervised learning can be effective for detecting known

phishing attacks, but it may not be as effective for detecting new or unknown attacks.

Unsupervised learning involves training a machine learning model on an unlabeled dataset of URLs, and it learns to identify patterns and anomalies in the data that may indicate the presence of phishing URLs. Unsupervised learning can be effective for detecting new and unknown phishing attacks, but it may also generate false positives.

Semi-supervised learning combines elements of both supervised and unsupervised learning. The machine learning model is trained on a small labeled dataset of phishing and legitimate URLs, but it also learns from an unlabeled dataset to identify new patterns and anomalies in the data. Semi-supervised learning can be effective for detecting new and unknown phishing attacks while also minimizing false positives.

Deep learning methods, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), can be used to detect phishing URLs by learning features directly from raw data, such as website screenshots or network traffic logs. Deep learning can be effective for detecting new and unknown phishing attacks, but it may require a large amount of labeled data and computing resources.

Even though there are so many different methods to train data based on the dataset the one that was chosen to train the model is the dataset split into two equal parts: 2000 legitimate URLs and 2000 phishing URLs. This balanced split was chosen to ensure that the model is not biased towards either class and can accurately detect both legitimate and phishing URLs

RQ1.2: What accuracy measures are used to compare such algorithms?

The following accuracy measures are used to compare different machine learning algorithms in the context of phishing detection:

a. Accuracy: This is a measure of how well the model is able to correctly classify URLs as phishing or legitimate. It is calculated as the ratio of the number of correctly classified URLs to the total number of URLs in the test set.

b. Precision: This is a measure of how well the model is able to correctly identify phishing URLs. It is calculated as the ratio of the number of true positives (i.e., phishing URLs correctly identified as phishing) to the total number of URLs identified as phishing by the model.

c. Recall: This is a measure of how well the model is able to correctly identify all phishing URLs. It is calculated as the ratio of the number of true positives to the total number of actual phishing URLs in the test set.

d. F1 score: This is a measure of the overall performance of the model, considering both precision and recall. It is calculated as the harmonic mean of precision and recall.

RQ1.3: What machine learning algorithms gave the best results in detecting phishing websites?

Since the effectiveness of different machine learning methods in detecting phishing URLs can be evaluated based on several metrics, including accuracy, False Positive Rate (FPR), Recall, Precision, and F1-score the following results are got:

The Naive-Bayes method has a high accuracy of 97.18%. Random forest has an accuracy of 97%. The CNN method has an accuracy of 96.61%, but it has a relatively high FPR of 3.50%, indicating that it classified some legitimate URLs as phishing URLs. The LSTM method has the highest accuracy of 97.20%, with a low FPR of 1.80%, indicating that it correctly identified the majority of phishing URLs while misclassifying only a few legitimate URLs as phishing URLs.

The MLP method has an accuracy of 96.65%. The CNN+RNN method has an accuracy of 97.9%, but it has a relatively high FPR of 3.10%. The CNN+LSTM method has the lowest accuracy of 93.28%, but it has the highest Precision of 99.12%, indicating that it classified very few legitimate URLs as phishing URLs.

In summary, the LSTM method is the most effective method in terms of accuracy and FPR, while CNN+LSTM has the highest Precision. The ensemble of two methods is better than the usage of one because each method has its strengths and weaknesses, and by combining them, we can improve the overall performance.

For example, the CNN+RNN method has a higher accuracy than Naive-Bayes, but Naive-Bayes is faster and requires less computational resources. By combining the two, we can get a more accurate and efficient phishing URL detection system.

5.2 RQ2: How effective is the proposed ensemble of two techniques in detecting phishing URLs?

The proposed ensemble of two techniques (multi-head self-attention and CNN) shows improved performance in detecting phishing URLs. The combination of CNN with multi-head self-attention outperforms individual CNN and LSTM models. The results in *Table 4.2* show that the Accuracy and F1 of the proposed method are 98.34% and 98.30%, respectively. However, the False Positive Rate (FPR) of the proposed method, 1.76% suggests that the proposed method classifies more legitimate webpages as phishing.

Furthermore, the training time of the proposed method is relatively low, with an average of 32 minutes per epoch.

In summary, the proposed ensemble of two techniques (multi-head self-attention and CNN) shows improved performance in detecting phishing URLs. The proposed method has a high Accuracy and F1 score, with a relatively low

training time, but with a slightly high FPR. Therefore, the effectiveness of the proposed ensemble method in detecting phishing URLs is relatively high.

5.3 RQ3: How effective is the proposed ensemble of two techniques in detecting phishing URLs compared to other methods?

According to the results, the proposed ensemble of two techniques is highly effective in detecting phishing URLs compared to other methods. The study compares different structures, including CNN, LSTM, CNN-CNN, and CNN-LSTM, with the proposed ensemble of CNN and multi-head self-attention.

The proposed ensemble of two techniques is highly effective in detecting phishing URLs compared to other methods. The model was compared with five commonly used methods and it achieved the lowest false positive rate (FPR) of 0.26%, the highest accuracy of 99.84%, and the highest F1 score of 99.84%. Moreover, the method outperformed all the previous methods in terms of Recall, which is 99.95%. Although the FPR of the method is higher than that of CNN–LSTM, which is 0.82%, it is still lower than that of all the other compared methods. Therefore, it can be concluded that the proposed ensemble of two techniques is highly effective in detecting phishing URLs compared to other methods.

The below information shows that the combination of two networks helps increase the performance of the model. The training time of the proposed method is also lower than CNN-LSTM and other methods, which indicates that the proposed ensemble is more efficient.

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

Phishing attacks continue to be a major threat to online security, and various methods have been developed to detect and prevent them. Machine learning algorithms have emerged as a promising approach to detecting phishing URLs due to their ability to learn from data and adapt to new and evolving threats. In this study, we examined the types of machine learning algorithms used for detecting phishing URLs, the datasets used to train them, and the accuracy measures used to evaluate their performance.

The findings indicate that different machine learning algorithms have different strengths and weaknesses in detecting phishing URLs. Supervised learning can be effective for detecting known phishing attacks, while unsupervised learning can be effective for detecting new and unknown attacks. Semi-supervised learning can provide a balance between these two approaches, while deep learning can learn features directly from raw data and can be effective for detecting new and unknown attacks but requires a large amount of labeled data and computing resources.

In terms of accuracy measures, it has been found that the Naive-Bayes and Random Forest methods had high accuracy rates of 97.18% and 97%, respectively. The CNN method had a slightly lower accuracy rate of 96.61%, but a relatively high false positive rate of 3.50%, indicating that it classified some legitimate URLs as phishing URLs. The LSTM method had the highest accuracy rate of 97.20% with a low false positive rate of 1.80%, indicating that it correctly identified the majority of phishing URLs while misclassifying only a few legitimate URLs.

The most efficient model of detection phishing URLs, based on the research, is a proposed ensemble of CNN and Multi-head self-attention since it achieved the best performance on the task with an accuracy of 98.34%, the lowest false positive rate of 1.76%, a recall of 98.44%, precision of 98.16%, and an F1 score of 98.30. This model combines a CNN with MHSA, allowing it to capture long-term dependencies and attend to multiple parts of the sequence simultaneously. Overall, it outperformed the other models, including LSTM, CNN+RNN, and Random Forest.

Overall, the study highlights the potential of machine learning algorithms for detecting phishing URLs and suggests that a combination of methods, including blacklists, DNS filters, user awareness training, and machine learning algorithms, may be needed to provide effective protection against phishing attacks. Additionally, the choice of algorithm may depend on the specific needs of the organization and the nature of the phishing attacks they are trying to detect. Further research is needed to explore the effectiveness of different machine learning algorithms and their potential applications in real-world settings.

6.2 Future work

There are several possible future works that can be done in the sphere of phishing URLs detection based on machine learning, including:

a. Developing more sophisticated models: Researchers can develop more sophisticated machine learning models that can detect more complex phishing URLs, such as those that use obfuscation techniques to evade detection.

b. Incorporating more features: Researchers can incorporate more features into their models, such as website content analysis, network traffic analysis, and user behavior analysis, to improve the accuracy of phishing URL detection.

c. Enhancing model explainability: Researchers can develop methods to enhance the explainability of their machine learning models, which can help to build trust and increase their adoption in real-world settings.

d. Conducting large-scale evaluations: Researchers can conduct large-scale evaluations of their models on real-world datasets, which can help to identify the strengths and weaknesses of different approaches and facilitate the development of more effective models.

e. Adapting to new threats: As phishing techniques evolve, researchers must constantly adapt their machine learning models to detect new and emerging threats.

Overall, there is significant potential for future research in this area, and the development of more accurate and effective machine learning models could have a significant impact on improving online security.

7 REFERENCES

[1] James, L. (2006). Banking on phishing. In James, L. (Ed.), Phishing Exposed (pp. 1-35). Syngress. ISBN 9781597490306

[2] Sundara Pandiyan, S., Selvaraj, P., Burugari, V. K., Benadit P, J., & Kanmani, P. (2022). Phishing attack detection using Machine Learning. Measurement: Sensors, 24, 100476. ISSN 2665-9174

[3] Ahammad, S. K. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, M. D. K. J. (2022). Phishing URL detection using machine learning methods. Advances in Engineering Software, 173, 103288. ISSN 0965-9978

[4] Berners-Lee, T., Masinter, L., & McCahill, M. (Eds.). (1994). Uniform Resource Locators (URL). Request for Comments: 1738. Network Working Group. CERN. Standards Track. Updated by: 1808, 2368, 2396, 3986, 6196, 6270, 8089. Obsoleted by: 4248, 4266. Errata Exist

[5] L. Wenyin, G. Liu, B. Qiu and X. Quan, "Antiphishing through Phishing Target Discovery," in IEEE Internet Computing, vol. 16, no. 2, pp. 52-61, March-April 2012, doi: 10.1109/MIC.2011.103

[6] Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. Journal of King Saud University - Computer and Information Sciences, 35(2), 590-611. ISSN 1319-1578

[7] Vrbančič, G., Fister, I., & Podgorelec, V. (2020). Datasets for phishing websites detection. Data in Brief, 33, 106438. ISSN 2352-3409

[8] Zheng, F., Yan, Q., Leung, V. C. M., Yu, F. R., & Ming, Z. (2022). HDP-CNN: Highway deep pyramid convolution neural network combining word-level and character-level representations for phishing website detection. Computers & Security, 114, 102584. ISSN 0167-4048

[9] Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., & Woźniak, M. (2020). Accurate and fast URL phishing detector: A convolutional neural network approach. Computer Networks, 178, 107275. ISSN 1389-1286

[10] Sahingoz, O. K., Baykal, S. I., & Bulut, D. (2018). Phishing detection from urls by using neural networks. Computer Science & Information Technology (CS & IT), 41-54.

[11] Remmide, M. A., Boumahdi, F., Boustia, N., Feknous, C. L., & Della, R. (2022). Detection of Phishing URLs Using Temporal Convolutional Network. Procedia Computer Science, 212, 74-82. ISSN 1877-0509.

[12] Marwa M. Emam, Nagwan Abdel Samee, Mona M. Jamjoom, Essam H. Houssein, Optimized deep learning architecture for brain tumor classification using improved Hunger Games Search Algorithm, Computers in Biology and Medicine, Volume 160, 2023, 106966, ISSN 0010-4825

[13] Sundara Pandiyan S, Prabha Selvaraj, Vijay Kumar Burugari, Julian Benadit P, Kanmani P, Phishing attack detection using Machine Learning, Measurement: Sensors, Volume 24, 2022, 100476, ISSN 2665-9174,

[14] Kai Florian Tschakert, Sudsanguan Ngamsuriyaroj, Effectiveness of and user preferences for security awareness training methodologies, Heliyon, Volume 5, Issue 6, 2019, e02010, ISSN 2405-8440

[15] Mohsen Soori, Behrooz Arezoo, Roza Dastres, Machine learning and artificial intelligence in CNC machine tools, A review, Sustainable Manufacturing and Service Economics, 2023, 100009, ISSN 2667-3444,

[16] Tianyuan Liu, Hangbin Zheng, Pai Zheng, Jinsong Bao, Junliang Wang, Xiaojia Liu, Changqi Yang, An expert knowledge-empowered CNN approach for welding radiographic image recognition, Advanced Engineering Informatics, Volume 56, 2023, 101963, ISSN 1474-0346,

[17] Jun Ma, Guolin Yu, Weizhi Xiong, Xiaolong Zhu, Safe semi-supervised learning for pattern classification, Engineering Applications of Artificial Intelligence, Volume 121, 2023, 106021, ISSN 0952-1976

[18] Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Rodriguez-Galan, G., Martínez-Cepeda, V., Nuñez-Agurto, D. (2023). Comparative Study of Deep Learning Algorithms in the Detection of Phishing Attacks Based on HTML and Text Obtained from Web Pages. In: Botto-Tobar, M., Zambrano Vizuete, M., Montes León, S., Torres-Carrión, P., Durakovic, B. (eds) Applied Technologies. ICAT 2022. Communications in Computer and Information Science, vol 1755. Springer, Cham. https://doi.org/10.1007/978-3-031-24985-3_28

[19] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104161.

[20] Do, Q.N.; Selamat, A.; Krejcar, O.; Yokoi, T.; Fujita, H. Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical Study. Appl. Sci. 2021, 11, 9210. https://doi.org/10.3390/ app11199210

[21] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1173-1179, doi: 10.1109/ICSSIT48917.2020.9214225.

[22] Y. Huang, Q. Yang, J. Qin and W. Wen, "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN," 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 2019, pp. 112-119, doi: 10.1109/TrustCom/BigDataSE.2019.00024.

[23] M. A. Adebowale, K. T. Lwin and M. A. Hossain, "Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection," 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Island of Ulkulhas, Maldives, 2019, pp. 1-8, doi: 10.1109/SKIMA47702.2019.8982427.

[24] Bahnsen, A. C., Bohorquez, C. E., Villegas, S., Vargas, J., & González, F. A. (2017). Classifying phishing URLs using recurrent neural networks. In *2017 APWG symposium on electronic crime research (eCrime)* (pp. 1–8). Scottsdale, AZ, USA.

[25] Bahnsen, A. C., Bohorquez, C. E., Villegas, S., Vargas, J., & González, F. A. (2017). Classifying phishing URLs using recurrent neural networks. In *2017 APWG symposium on electronic crime research (eCrime)* (pp. 1–8). Scottsdale, AZ, USA.

[26] Zhang J., Li X. Phishing detection method based on borderline-smote deep belief network security, privacy, and anonymity in computation, communication, and storage. SpaCCS 2017, Lecture notes in computer science, vol. 10658, Springer, Cham (2017), pp. 45-53

[27] Yang P., Zhao G., Zeng P. Phishing website detection based on multidimensional features driven by deep learning IEEE Access, 7 (2019), pp. 15196-15209

[28] Abdelnabi et al., 2020, S. Abdelnabi, K. Krombholz, M. Fritz, VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity, Proceedings of the ACM Conference on Computer and Communications Security (2020), pp. 1681-1698, 10.1145/3372297.3417233

[29] Adebowale et al., 2019, M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain, Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text

[30] Al-Ahmadi and Alharbi, 2020, S. Al-Ahmadi, Y. Alharbi, A deep learning technique for web phishing detection combined Url features and visual similarity, Int. J. Comput. Netw. Commun., 12 (5) (2020), pp. 41-54, 10.5121/ijcnc.2020.12503

[31] AlEroud and Karabatis, 2020, A. AlEroud, G. Karabatis, Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks, IWSPA 2020 - Proceedings of the 6th International Workshop on Security and Privacy Analytics (2020), pp. 53-60, 10.1145/3375708.3380315

[32] Almeida and Westphall, 2020, Almeida, R., Westphall, C., 2020. Heuristic Phishing Detection and URL Checking Methodology Based on Scraping and Web Crawling. In: Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020, doi: 10.1109/ISI49825.2020.9280549.

[33] N.A. Azeez, S. Misra, I.A. Margaret, L. Fernandez-Sanz, S.M. Abdulhamid, Adopting automated whitelist approach for detecting phishing

attacks, Comput. Security, 108 (2021), Article 102328, 10.1016/j.cose.2021.102328

[34] G.G. Geng, Z.W. Yan, Y. Zeng, X.B. Jin, RRPhish: Anti-phishing via mining brand resources request, 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, 2018-Janua (2018), pp. 1-2, 10.1109/ICCE.2018.8326085

[35] Lakshmi et al., 2021, L. Lakshmi, M.P. Reddy, C. Santhaiah, U.J. Reddy, Smart phishing detection in web pages using supervised deep learning classification and optimization technique ADAM, Wireless Pers. Commun., 118 (4) (2021), pp. 3549-3564, 10.1007/s11277-021-08196-7

[36] Opara et al., 2020, C. Opara, B. Wei, Y. Chen, HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis, Proceedings of the International Joint Conference on Neural Networks (2020), 10.1109/IJCNN48605.2020.9207707

[37] Ozker and Sahingoz, 2020, U. Ozker, O.K. Sahingoz, Content Based Phishing Detection with Machine Learning, 2020 International Conference on Electrical Engineering, ICEE 2020 (2020), 10.1109/ICEE49691.2020.9249892

[38] Qabajeh et al., 2018, I. Qabajeh, F. Thabtah, F. Chiclana, A recent review of conventional vs. automated cybersecurity anti-phishing techniques, Computer Sci. Rev., 29 (2018), pp. 44-55, 10.1016/j.cosrev.2018.05.003

[39] S. Sindhu, S.P. Patil, A. Sreevalsan, F. Rahman, A.N. Saritha, Phishing detection using random forest, SVM and neural network with backpropagation, Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020 (2020), pp. 391-394, 10.1109/ICSTCEE49637.2020.9277256

[40] Singh et al., 2020, S. Singh, M.P. Singh, R. Pandey, Phishing detection from URLs using deep learning approach, Proceedings of the 2020 International Conference on Computing, Communication and Security, ICCCS 2020 (2020), pp. 16-19, 10.1109/ICCCS49678.2020.9277459

[41] I. Saha, D. Sarma, R.J. Chakma, M.N. Alam, A. Sultana, S. Hossain, Phishing attacks detection using deep learning approach, Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Icssit (2020), pp. 1180-1185, 10.1109/ICSSIT48917.2020.9214132