TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

[Hua Zhong] [221910IVCM]

A DEEP LEARNING SOLUTION FOR DETECTING IMAGE-BASED PHISHING/SPAM EMAILS

Master's Thesis

Supervisor: Sven Nõmm
Ph.D.

Co-supervisor: Adrian Nicholas Venables
Ph.D.

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

[Hua Zhong] [221910IVCM]

SÜVAÕPPE LAHENDUS PILDIPÕHISTE ÕNGITSUS- JA RÄMPSKIRJADE TUVASTAMISEKS

Magistritöö

Juhendaja: Sven Nõmm
Ph.D.

Kaasjuhendaja: Adrian Nicholas Venables
Ph.D.

Tallinn 2024

Author’s Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: [Hua Zhong]

17.05.2024

Acknowledge

I would like to extend my heartfelt thanks to my family: Julia, Leo, and David, for your trust, love, and support.

Additionally, I am grateful to Professor Sven Nõmm, Adrian Venables, and Raimundas Matulevičius for their guidance and steadfast support throughout this journey.

I also want to express my appreciation to my classmates Mihhail, Marco, and Semjon, whose insightful reviews and valuable suggestions were absolutely necessary.

Abstract

According to the annual report from Estonia's Information Authority, criminals have lists with the email addresses of tens of thousands of people in Estonia. 1722 phishing incidents were recorded by CERT-EE and 546 incidents of fraud in 2023 - an increase of 250% from the year before. [1] This trend has underscored the need for more diverse phishing and spam filters capable of detecting different types of emails, including those based on images. To reduce the workload on human operators, more artificial intelligence solutions should be employed. Phishing, a specialized type of social engineering attack, aims to deceive victims into revealing sensitive information, such as personal and financial data. Common forms of phishing attacks include emails with malicious attachments or redirecting URLs, presenting significant security challenges. Spam emails are typically known for spreading unwanted advertising, scams, or malware. Malware spreading through spam emails is one potential method of phishing attacks. Both phishing and spam emails are considered cybersecurity threats as they can lead to unauthorized access, financial loss, and distribution of malware.

Finding patterns, especially unusual ones, is becoming difficult with conventional methods. These methods typically analyze text content, sender information, or metadata. Deep learning provides sophisticated tools for managing complex and large-scale data, similar to those encountered in network settings. It is particularly effective at noticing small patterns and unusual changes that traditional methods might overlook. Furthermore, deep learning can continuously improve and adapt, making it more suitable for dynamic settings where data patterns are always changing.

This research study's contribution encompasses exploring and implementing innovative image-based deep-learning methodologies aimed at enhancing the detection of phishing and spam emails. The primary goal of this research is to collect a comprehensive dataset, utilize data augmentation, and achieve the functionality of a novel deep learning architecture, particularly the Vision Transformer (ViT). Key outcomes of this study include the development of deep learning models; the creation of datasets that include email-associated images; a data augmentation solution; and a brief validation of the effectiveness of these models in identifying phishing and spam emails. The classification metrics were used for validation.

This thesis is written in English and is 49 pages long, including 5 chapters, 4 figures, and 9 tables.

Annotatsioon
Süvaõppe lahendus pildipõhiste õngitsus- ja rämpskirjade tuvastamiseks

Vastavalt Riigi Infosüsteemi Ameti aastaaruandele on kurjategijatel nimekirjad kümnete tuhandete Eesti inimeste e-posti aadressidega. CERT-EE registreeris 2023. aastal 1722 õngitsusjuhtumit ja 546 pettusejuhtumit, mis on 250% rohkem kui eelmisel aastal. See trend rõhutab vajadust mitmekesisemate õngitsus- ja rämpspostifiltrite järele, mis suudaksid tuvastada erinevat tüüpi e-kirju, sealhulgas piltidel põhinevaid. Inimoperaatorite koormuse vähendamiseks tuleks kasutada rohkem tehisintellekti lahendusi. Õngitsus, mis on spetsialiseerunud tüüpi sotsiaalse manipulatsiooni rünnak, on suunatud ohvrite petmisele, et nad avaldaksid tundlikku teavet, nagu isiku- ja finantsandmed. Levinud õngitsusrünnakute vormid hõlmavad pahatahtlike manuste või ümbersuunamise URL-idega e-kirju, mis kujutavad endast olulisi turvariske. Rämpspost e-kirjad on tavaliselt tuntud soovimatu reklaami, pettuste või pahavara levitamise poolest. Pahavara levitamine rämpspostide kaudu on üks võimalik õngitsusrünnakute meetod. Nii õngitsus- kui ka rämpspost e-kirju peetakse küberjulgeolekuohtudeks, kuna need võivad põhjustada volitamata juurdepääsu, rahalisi kaotusi ja pahavara levikut.

Mustrite, eriti ebatavaliste mustrite leidmine muutub tavapäraste meetoditega raskemaks. Need meetodid analüüsivad tavaliselt tekstisisu, saatja teavet või metaandmeid. Süvõpe pakub keerukaid tööriistu suurte ja keerukate andmekogumite haldamiseks, nagu neid kohatakse võrgukeskkondades. See on eriti tõhus väikeste mustrite ja ebatavaliste muutuste märkamisel, mida traditsioonilised meetodid võivad tähelepanuta jätta. Peale selle saab süvõpe pidevalt paremaks ja kohanevamaks, muutes selle sobivamaks dünaamilisteks olukordadeks, kus andmemustrid pidevalt muutuvad.

Selle uurimistöö panus hõlmab õngitsus- ja rämpsposti e-kirjade tuvastamise parandamisele suunatud uuenduslike põhinevate süvaõppemeetodite uurimist ja rakendamist. Uurimistöö peamine eesmärk on koguda kõikehõlmav andmekogum, kasutada andmete suurendamist ja saavutada uue süvaõppe arhitektuuri, eriti Vision Transformer (ViT), funktsionaalsus. Uuringu peamised tulemused hõlmavad süvaõppemudelite arendamist; e-kirjadega seotud piltide andmekogumite loomist; andmete suurendamise lahendus; ja nende mudelite tõhususe lühike valideerimine õngitsus- ja rämpsposti e-kirjade tuvastamisel. Valideerimiseks kasutati standardseid klassifikatsioonimeetrikaid.

See lõputöö on kirjutatud inglise keeles ja selle pikkus on 49 lehekülge, sealhulgas 5 peatükki, 4 joonist ja 9 tabelit.

List of Abbreviations and Terms

AI	Artificial Intelligence
CEN	Computational Engineering and Networking
CERT	Computer Emergency Response Team
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
DL	Deep Learning
GAN	Generative adversarial networks
HSV	Hue Saturation and Value
ML	Machine Learning
POC	Proof Of Concept
RNN	Recurrent Neural Network
SLR	Systematic Literature Review
SVM	Support Vector Machines
ViT	Vision Transformer

Table of Contents

List of Figures

List of Tables

1. Introduction

Phishing is the combination of social engineering and technical exploits designed to convince a victim to provide personal information, usually for the monetary gain of the attacker. Phishing emails contain messages to lure victims into performing certain actions, such as clicking on a URL where a phishing website is hosted, or executing a malware code. Phishing has become the most popular practice among criminals of the Web.[2] Email spam is the most prevalent form of spam. In e-mail spam, messages are sent to a large number of e-mail addresses. Such spam messages can include product advertisements, links to phishing websites, or links to malware installers.[3] In contrast, ham emails are non-spam or legitimate emails. [4] In this research, phishing and spam emails are analyzed together due to their shared characteristics as cybersecurity threats. Both types of attacks can lead to unauthorized access, financial damage, and the distribution of malware.

1.1 Background

The landscape of cybersecurity threats in 2023, particularly in Estonia, vividly illustrates the prevalent and escalating challenge posed by phishing attacks, which accounted for 1,722 incidents out of the total recorded. This makes phishing the most significant cybersecurity threat among others.[1]Moreover, the impact of spam emails should not be overlooked, as these can also be tied to the incidents categorized under fraud and malware. Such statistics not only underscore the persistent and evolving nature of phishing and spam emails but also spotlight the necessity for innovative and effective countermeasures. This insidious strategy is crafted to exploit and inflict damage on unsuspecting victims and organizations.

Table 1. Incidents with an Impact in 2023 Estonia

Incident Type	Number of Incidents
Phishing	1,722
Fraud	546
Service Interruption	312
Account Takeover	207
Compromising	165
Denial-of-service Attack	139
Malicious Redirect	113
Malware	47
Data Leak	26
Defacement	24
Ransomware	13

1.2 Motivation

The motivation for the proposed research comes from the urgent need to address the high incidence of phishing, which conventional text-based detection systems may overlook, especially as attackers continually refine their strategies to bypass traditional security measures.[5] Considering the complexity and adaptability of phishing techniques, which often include sophisticated visual elements to deceive victims, an image-based approach using deep learning could provide a more robust and accurate detection system.

This thesis will explore the potential of leveraging cutting-edge AI technologies to enhance email security and reduce the susceptibility of individuals and organizations to these prevalent cyber threats. Through this approach, the research aims to contribute significantly to the field of cybersecurity by introducing an innovative solution to combat phishing and spam emails.

1.3 Main Research Questions

This thesis will attempt to answer one main research question: **[MRQ]How can a new deep learning solution be employed to detect image-based phishing and spam emails?** To address this question, three sub-research questions have been formulated, corresponding to different sections of the thesis:

1. For the **literature review**, the following research question was addressed: **[RQ1] What is the research gap between current phishing and spam email detection technology and the novel methodology?**
2. In the **contributions** section, the question to be answered is: **[RQ2] What imaged-based deep learning technology should be employed for phishing/spam emails detection?**
3. In the **research results presentation** section, the focus will be on: **[RQ3] What are the research results?**

Each research question is designed to ensure a thorough investigation of the innovative image-based deep learning approach within the context of existing methods and its implementation effectiveness.

1.4 Scope and Goal

The scope of this research study encompasses the exploration and implementation of innovative image-based deep-learning methodologies aimed at enhancing the detection of phishing and spam emails. The primary goal of this research is to build a proof-of-concept (POC) prototype to validate the functionality of a novel deep learning architecture, which aims to fill the gap in the field of image-based phishing email detection. Key outcomes of this study include the development of deep learning models; the creation of datasets that include email-associated images; advanced data augmentation; and an evaluation of the effectiveness of these models in accurately identifying phishing/spam or Ham emails.

2. Literature Review

In this section, a systematic literature review (SLR) is conducted to investigate image-based deep-learning solutions for detecting phishing and spam emails. The goal of this SLR is to evaluate current deep-learning methods that use image analysis to enhance the detection of phishing and spam in emails. The literature review is approached from a deep learning perspective, focusing on solutions adept at image processing and pattern recognition within emails.

2.1 Literature review research question

The primary research question addressed is: **[RQ1] What is the research gap between current phishing and spam email detection technology and the novel methodology?** To effectively explore this question, it is divided into three sub-research questions:

[RQ1.1] What is the traditional approach to detect phishing and spam emails? This sub-research question seeks to identify and describe conventional email security methods, particularly those not involving image-based techniques, to establish a comparison baseline for the effectiveness of newer, image-based methods. **[RQ1.2] What is the current status of deep learning technology used to detect phishing and spam emails?**This sub-research question aims to pinpoint advancements in deep learning that utilize image recognition and classification technologies to identify fraudulent content in emails. **[RQ1.3] Is data augmentation necessary to develop a robust AI model for detecting phishing and spam emails?** This sub-research question investigates whether increasing the diversity of training datasets with synthetic or altered images can enhance the reliability of the model in practical settings. Through this structured analysis, the review aims to provide a detailed overview of the potential and limitations of employing image-based deep learning to combat email threats.

2.2 Literature Sources

The initial search for relevant papers was conducted using Google Scholar, the IEEE Digital Library, ScienceDirect, and SpringerOpen. Additional relevant papers were identified by examining the related work sections and citations of the papers found in the initial search.

2.3 Search Terms

The search terms used in the study: *"Deep Learning"* AND *"phishing emails"* AND *"Image"*. The terms *"Machine learning"* and *"Data augmentation"* were included to capture a broader range of potential methodologies relevant to the detection of phishing emails: *"Machine learning"* AND *"Image"* AND *"phishing emails"*;*"Deep Learning"* AND *"Data augmentation"*. Other search terms were considered but ultimately excluded from the final search strategy.

2.4 Inclusion and Exclusion Criteria

Inclusion Criteria

- Papers that apply email filtering techniques to detect spam emails, including deep learning methods.
- Papers that include Image-based deep learning methods.
- Papers that include machine learning, and data augmentation strategies for spam email detection.

Exclusion Criteria

- Papers published before 2005.
- Papers not written in English.
- Papers that are shorter than 3 pages.
- Papers that are not freely accessible.
- Modeling solutions that are not intuitive or understandable by individuals who are not deep learning experts, such as methods heavily reliant on pure mathematical models or text-based approaches.

2.5 Papers Selection

The first step of the selection was to look at digital libraries using the previously mentioned.

Table 2. Summary of Search Results and Selection Process

Source	**Initial**	**Inclusion**	**Manual Inspection**	**Snowballing**
IEEE	110	6	4	–
Google Scholar	1403	293	4	4
Springer Open	23	22	2	–
Others*	–	–	2	–

*Other sources include Taltech Digital Library and the Estonian Riigi Infosüsteemi Amet's Studies and analyses.

2.6 Summary of Selected Articles

In this section, we summarize each publication that was chosen during the paper selection phase of the literature review.

"Learning Fast Classifiers for Image Spam" by Dredze et al. [6] This paper introduces a novel Just in Time feature extraction method for image spam classification. This approach dynamically extracts features on a per-image basis during the classification phase, rather than using the traditional two-stage process of first extracting all potential features and then performing classification. By extracting only the necessary features for making a prediction on each specific image, this method significantly cuts down on processing time.

Building email datasets is challenging because emails are private. Another important part of Dredze et al.'s research is that they made their own image spam dataset. This is crucial because there aren't many public datasets specifically for image spam, which makes it hard to develop and test effective spam detection methods. By creating their own dataset, Dredze et al. were able to design it to suit their study's needs and improve the usefulness of their results for real-world situations. This dataset is valuable not only for testing their Just in Time feature extraction method but also for helping other researchers looking to improve image spam detection techniques.

"Image Spam Hunter" by Yan Gao et al. [4] The authors address the evolving challenge of image-based spam. The paper highlights the adaptability of spammers who employ varied image processing techniques to avoid detection, such as altering image colors, backgrounds, and fonts, and introducing rotations and artifacts. A notable contribution of this work is the development of a robust image spam dataset, which could greatly benefit further research and refinement of anti-spam technologies. This dataset is particularly valuable for training and testing purposes, given the diverse and challenging nature of the image variations created by spammers.

"Using Visual Features for Anti-Spam Filtering" by Ching-Tung Wu et al.[7] This study explores the integration of visual features with traditional text-based approaches to enhance the effectiveness of anti-spam filters. Recognizing the limitations of solely text-based filters, the authors introduce a system that utilizes visual cues from images within emails, such as embedded text and graphic banners. A novel aspect of their approach is the use of a one-class Support Vector Machine (SVM) classifier, which focuses on identifying spam based on these visual characteristics without needing a comparative set of non-spam emails. Experimental results indicate that this method substantially improves spam detection rates when integrated with existing text-based systems, demonstrating an increase from 47.7% to 84.6% in detection accuracy. This study highlights the potential of incorporating visual information to address the evolving complexity of spam emails.

"Efficient Modeling of Spam Images" by Qiao Liu et al.[8]The paper presents a novel statistical approach to classify spam images without relying on embedded text, enhancing resilience against obfuscation techniques used by spammers. The study proposes a model that utilizes color and shape features extracted from images, which are shown to robustly differentiate spam from legitimate content. The effectiveness of this model was validated through experiments on two open datasets, where it demonstrated superior performance compared to previous methods, achieving high accuracy with minimal false positives. This approach offers a promising advancement for spam filtering technologies, particularly in handling image-based spam.

"Detecting Image Spam using Visual Features and Near Duplicate Detection" by Bhaskar Mehta et al.[9] The paper addresses the challenge of image-based spam, which bypasses traditional text-based filters by embedding spam messages within images. The authors propose two innovative solutions to enhance the detection of image-based spam. One of the approaches is to utilize visual features such as color, texture, and shape to classify emails, achieving a notable accuracy improvement of at least 6% over existing methods, reaching about 98% effectiveness. This method employs Support Vector Machines (SVMs) for classification.

"Detecting Image Spam Using Image Texture Features" by Basheer Al-Duwairi, Ismail Khater, and Omar Al-Jarrah[10] This paper presents a technique for filtering image-based email spam by utilizing low-level image texture features for more effective characterization and identification. The authors propose the Image Texture Analysis-Based Image Spam Filtering (ITA-ISF) method, which incorporates a variety of machine learning classifiers including C4.5 Decision Trees, Support Vector Machines, Multilayer Perceptions, Naive

Bayes, Bayesian Networks, and Random Forests to analyze and classify images based on extracted texture features. The effectiveness of these classifiers is evaluated using publicly available datasets, with the Random Forest classifier showing superior performance, achieving an average precision, recall, accuracy, and F-measure of 98.6%. This study contributes to the ongoing efforts to combat spam by enhancing the capability of spam filters to recognize and block image-based spam, which continues to evolve and present significant challenges in content-based email filtering systems.

"Deep Learning Based Phishing E-mail Detection" by Hiransha M et al.[11] This paper introduces a deep learning model, "CEN-Deepspam," designed to identify phishing emails which are a significant threat, especially to the financial sector. The model leverages Keras Word Embedding and Convolutional Neural Network (CNN) technologies to differentiate phishing emails from legitimate ones. The researchers used a dataset comprising emails with and without headers to train and validate their model. They utilized word embeddings to transform text data into numerical form, allowing the CNN to process and classify the emails effectively. This combination of word embedding and CNN, followed by pooling and fully connected layers, achieves a notable classification accuracy.

"Enhancement and Augmentation of Drawing Tests for Deep Learning Based Diagnostics of Neurological Disorders" by Nõmm et al.[12] In this thesis at Tallinn University of Technology, Nõmm et al. investigates the application of convolutional neural networks (CNNs) to diagnose Parkinson's Disease using digital drawing tests. The study significantly augmented the dataset by using OpenCV for better model training. The research systematically evaluates various CNN architectures, including AlexNet, LeNet5, and Xception.

"Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism" by Yong Fang et al.[13] This paper presents an advanced phishing email detection system called THEMIS, leveraging an improved Recurrent Convolutional Neural Network (RCNN) integrated with an attention mechanism. The proposed model uniquely analyzes emails by segmenting them into multi-level vectors—character and word levels of both email headers and bodies. This method allows for a nuanced understanding of the contextual and syntactical nuances of phishing emails, aiding in more accurate detection.

"Phishing Email Detection Model Using Deep Learning" by Samer Atawneh and Hamzah Aljehani[14] This paper explores the use of advanced deep learning techniques to enhance the detection of phishing emails, a significant cybersecurity challenge. The authors employ a combination of convolutional neural networks (CNNs) recurrent neural networks (RNNs), and other architectures to analyze and classify emails. The model benefits from natural language processing to extract a comprehensive set of features from both phishing and ham emails.

"Convolutional Neural Network Optimization for Phishing Email Classification" by Cameron McGinley and Sergio A. Salinas Monroy[15]The paper addresses the challenge of detecting phishing emails using optimized Convolutional Neural Network (CNN) architectures, focusing on text analysis without relying on URLs or metadata. The authors implement various CNN configurations to analyze the semantic content of email bodies, aiming to identify phishing attempts through text patterns alone.

"Efficient Spam and Phishing Emails Filtering Based on Deep Learning" by Safaa Magdy et al.[16] The paper presents a deep learning-based framework for effectively filtering spam and phishing emails, which poses a significant threat in digital communication. The study introduces an innovative approach utilizing a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to analyze and classify email content based on textual features. By employing deep learning models, the framework aims to capture complex patterns and anomalies that distinguish malicious emails from legitimate ones. The methodology demonstrates high effectiveness in detecting varied types of email threats, ensuring robust security measures. The results indicate substantial improvements over traditional spam filtering techniques, suggesting that deep learning can significantly enhance email security systems against sophisticated cyber threats.

"A survey on Image Data Augmentation by Connor Shorten, Taghi M. Khoshgoftaar[17] This survey by Shorten and Khoshgoftaar thoroughly reviews image data augmentation techniques crucial for enhancing the performance of deep learning models, especially when there is a lack of training data. It details basic image manipulations:

- Flipping: This involves creating flipped versions of the images, both horizontally and vertically, thereby doubling the dataset size.
- Cropping: Extracting random crops from the original image simulates partial visibility, a common occurrence in real-world scenarios
- Rotation: Rotating the image by specific degrees ensures model reliability against orientation changes.
- Translation: Shifting the image by a certain number of pixels in any direction aids in developing a model resilient to positional variations.
- Color space: digital image data, typically structured as a tensor with dimensions representing height, width, and color channels. It outlines practical color space augmentations, including isolating individual RGB color channels and modifying

them through simple matrix operations. These operations can adjust image brightness or alter color histograms to change the overall lighting, similar to techniques used in photo editing software.

- Scaling: Resizing images to a percentage of their original size helps in training models to recognize phishing emails irrespective of image size.
- Noise Injection: Adding random noise mimics the effect of high ISO camera settings, adding a layer of complexity to the dataset.

Advanced methods like generative adversarial networks (GANs), neural style transfer, and meta-learning are also discussed, showing their effectiveness in complex vision tasks like medical image analysis. The authors highlight the widely accepted idea that larger and more varied datasets can significantly improve deep learning models. This detailed overview helps researchers and practitioners with practical strategies to expand datasets systematically, leading to more accurate and robust models.

"Cyber security in Estonia 2024" by Martin Mileiko et al.[1] The cybersecurity landscape in Estonia during 2023 was marked by a surge in sophisticated and targeted cyberattacks, with phishing and spam being identified as primary threats. These threats have not only become more complex but also more frequent, causing a significant impact on both the public and private sectors. The year witnessed a notable increase in phishing incidents, making up over half of all cyber incidents, signaling a troubling trend in the cybersecurity domain.

"An Intelligent Classification Model for Phishing Email Detection" by Adwan Yasin and Abdelmunem Abuhasan[18] In their study, they develop a sophisticated classification model utilizing knowledge discovery, data mining, and text processing techniques to identify phishing emails effectively. This model is particularly notable for its use of textual-based feature extraction, popular datasets like Nazario and Nigerian Fraud datasets exclusively contain phishing emails were used. These text-based enhancements significantly improve the model's ability to classify emails with high precision. The paper also provides a comparative analysis with other classification techniques, highlighting its advanced capability in detecting phishing emails through enriched textual analysis.

"Image spam analysis and detection" by Annadatha and Stamp[19] In their study, the researchers analyze two primary image-based spam detection strategies. The first employs Principal Component Analysis to identify eigenvectors from a set of spam images and uses these vectors to project images onto an eigenspace for spam identification. The second method involves extracting a comprehensive set of image features and selecting an optimal subset through Support Vector Machines (SVM) for effective spam detection. The paper also emphasizes the creation of their so-called "improved dataset", which is designed to test and improve the effectiveness of spam detection technologies.

2.7 Presentation of Results

The presentation of results in this systematic literature review highlights the various feature sets, classifications, and deep-learning architectures applied in detecting phishing and spam emails, as detailed in the selected articles. Here are the summarized findings presented in table formats:

Table 3. Comparison of Feature Sets and Classifications

Source	**Feature sets**	**Classification(s)**
Ching-Tung Wu et al 2005 [7]	Graphic banners, Embedded text	Spam detection
Dredze et al. 2007 [6]	Metadata, Image Properties (e.g., size, color saturation), Simple Image Analysis	Spam detection, Ham detection
Mehta et al. 2008 [9]	Color, Texture, Shape	Spam detection, Ham detection
Liu et al 2010 [8]	Color, Shape	Spam detection
Basheer Al-Duwairi et al. 2013 [10]	Image histogram, Run-length matrix, Co-occurrence matrix, Image gradient, Autoregressive model, Wavelet transform	Spam detection, Ham detection

Table 3 summarizes traditional machine learning techniques used in spam and phishing detection. It outlines various studies that have employed different feature sets, such as graphic banners, embedded text, color, texture, and shape, to classify emails. The techniques range from simple image analysis to more complex pattern recognition, indicating the diversity in approaches to detect email threats.

Table 4. Deep-Learning Architectures and Their Applications

Source	**Deep-learning architectures**	**Purpose**	**Classification(s)**
Hiransha et al 2016 [11]	CNN	Spam detection	Spam emails Texture, Ham emails Texture
Fang et al 2019[20]	RCNN	Phishing email detection	Phishing emails, Ham emails
McGinley and Monroy 2021 [15]	CNN	Classifying phishing emails	Phishing emails URL, Ham emails URL
Nõmm et al 2021 [12]	CNN	Diagnostics of neurological disorders	Health control images, Parkinson’s disease images
Gogoi and Ahmed 2022 [16]	Pre-trained transformer	Identifying phishing emails	Phishing emails Texture, Ham emails Texture
Samer Atawneh and Hamzah Aljehani 2023[21]	CNN, RNN	Detection of phishing emails	Phishing emails Texture, Ham emails Texture

Table 4 specifically focuses on deep learning solutions employed in the detection of phishing and spam emails. It details the types of deep learning architectures used, such as CNNs, RNNs, and RCNNs, and their specific applications in email security.

Table 5. Comparison of Ham Corpus and Spam/Phishing Corpus

Source	**Ham corpus**	**Personal spam/phishing corpus**	**Spam/Phishing Archive**
Ching-Tung Wu et al 2005 [7]	428	N/A	8500
Dredze et al. 2007 [6]	2550	3239	9503
Gao Yang et al.2008 [4]	810	928	N/A
Mehta et al 2008 [9]	5373	N/A	N/A
Liu et al 2010 [8]	3784	3112	8719
Basheer Al-Duwairi et al. 2013[10]	2580 (1770 + 810)	N/A	4135 (3209 + 926)
Hiransha et al. 2016 task1 [11]	4583	N/A	501
Hiransha et al. 2016 task2 [11]	5088	N/A	612
Annadatha and Stamp 2018[19]	810	1000	920
Fang et al. 2019 [20]	7781	N/A	999
McGinley and Monroy 2021[15]	1870	N/A	1934
Samer Atawneh et al. 2023[21]	3081	N/A	2331

Table 5 presents a comprehensive overview of the training datasets used in various studies over the last two decades. The table categorizes the number of samples in each dataset into ham (legitimate emails) and spam/phishing emails, and identifies whether these datasets are publicly accessible or proprietary. Notably, the summary emphasizes that spam and phishing dataset sizes are often limited, a constraint primarily due to ethical considerations and privacy concerns. This limitation is critical, as it affects the robustness and comprehensiveness of the training data used across different studies, crucial for developing effective spam and phishing detection systems.

2.8 Answers to Research Question [RQ1]

The findings from the systematic literature review in Section 3.6 allow us to address and interpret the primary research question:**[RQ1] What is the research gap between current phishing and spam email detection technology and the novel methodology?**

2.8.1 [RQ1.1] What is the traditional approach to detect phishing and spam emails?

Traditional methods primarily rely on text-based analysis, using keyword filtering and heuristics to detect suspicious content. These methods, while effective against simpler forms of phishing/spam, often fail to catch more sophisticated attempts that use images and complex layouts to mimic legitimate sources.

2.8.2 [RQ1.2] What is the current status of deep learning technology used to detect phishing and spam emails?

Current deep learning approaches leverage image recognition and classification techniques to identify fraudulent content. Advancements include feature extraction methods that dynamically process images on a per-instance basis, and the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to analyze visual and textual content together, improving detection rates significantly.

2.8.3 [RQ1.3] Is data augmentation necessary to develop a robust AI model for detecting phishing and spam emails?

Data augmentation is crucial in building effective models, the literature review result emphasizes that spam and phishing dataset sizes are often limited due to ethical considerations and privacy concerns, larger and more varied datasets can significantly improve deep learning models.[14] Techniques such as image manipulation (flipping, cropping, rotating), noise injection, and synthetic image generation expand the diversity of the training set, enabling the model to perform well across various real-world scenarios. Advanced methods like GANs for generating new training samples are particularly promising for enhancing model robustness against evolving phishing tactics. On-the-fly data augmentation is also commonly employed in numerous deep-learning workflows.[3]

The systematic analysis reveals that integrating deep learning with traditional methods and expanding datasets through data augmentation can substantially enhance the effectiveness of spam and phishing detection systems. These findings support the potential of image-based deep learning as a powerful tool in combating email-based threats.

Answers to [RQ1] After my literature review, it's clear that deep learning has transformed the way we detect phishing and spam emails through image analysis. However, there's still a notable gap in the application of novel deep-learning architecture like Vision Transformer models. Data augmentation, which introduces a greater variety of training images, is underutilized in this domain. Similarly, Vision Transformers, which analyze images in segments to better understand the overall picture, have not been widely applied to email images. Exploring these areas could provide a new method of detecting email threats based on images, offering fresh perspectives and approaches.

3. Contribution

3.1 Contribution research question

This chapter sets the stage for a detailed discussion of how to build the Proof of Concept (POC) prototype to validate the functionality of a novel deep-learning architecture.

The central focus of this section is the primary research question:

[RQ2] What image-based deep learning technology should be employed for phishing/spam email detection?

To effectively explore this question, it is divided into three sub-research questions:

1. **[RQ2.1] How to create a Comprehensive Dataset:** This sub-question addresses the challenges and strategies involved in assembling a dataset that is both diverse and representative of real-world scenarios, which is crucial for the effective training of deep learning models.
2. **[RQ2.2] How to Implement Data Augmentation:** This explores the techniques and methodologies for enhancing the dataset through advanced data augmentation techniques. Such techniques are essential for improving the model's robustness and ability to generalize from training data to unseen data.
3. **[RQ2.3] How to train the Recognition Model:** This sub-question discusses the approaches for training the recognition model, focusing on the architectural choices, optimization algorithms, and tuning parameters that contribute to a successful detection system.

3.2 Collect original dataset

The planned utilization of datasets from Dredze et al[6], Image Spam Hunter (ISH), and a personal collection aims to compile approximately 1,000 to 3,000 phishing/spam email images. To effectively harness these datasets, specific steps including dataset regrouping and similarity analysis are essential.

The selection of datasets from Dredze et al., the Image Spam Hunter (ISH), and a personal collection for the study of phishing/spam email images is strategically justified based on several criteria crucial for research in this domain.

Firstly, the dataset from Dredze et al. is widely recognized and utilized in phishing email research. Its frequent application in similar studies underscores its importance and suitability for this field. By incorporating this dataset, the current research aligns with established methodologies in phishing detection, facilitating meaningful comparisons and benchmarks against other studies. The proven utility of the Dredze et al. dataset in previous research highlights its reliability and the comprehensive nature of its data, making it a critical component for any robust phishing detection study.

Secondly, the Image Spam Hunter (ISH) dataset offers a unique perspective by focusing on image-based spam. This specialization is particularly pertinent given the evolving nature of phishing attacks, which increasingly incorporate visual elements. The inclusion of the ISH dataset enables the study to cover a wider spectrum of phishing/spam tactics, enhancing the comprehensiveness of the analysis. This dataset is instrumental in developing detection models that can effectively recognize and differentiate various types of phishing content, especially those relying heavily on images.

The ISH dataset's inclusion of 'ham' images, which are non-spam or legitimate emails, plays a crucial role in providing a comprehensive understanding of the email landscape. These 'ham' images serve as a counterbalance to the phishing and spam images, enabling the development of more nuanced and accurate detection models. The presence of 'ham' images aids in training machine learning models to distinguish between legitimate and malicious content more effectively. This aspect is essential in reducing false positives, a common challenge in phishing detection systems.

Lastly, a personal collection of phishing and spam emails, identified as such by individuals, was used. The inclusion of such real-world data ensures that the research is grounded in current trends. Due to access and time limitations, only 10 images of phishing emails were collected during this study, which were sourced from the author's personal email address.

3.3 Dataset augmentation by using “OpenCV library”

In their 2021 study, S. Nõmm and E. Valla employed OpenCV technology to advance and enrich drawing tests, utilizing them for deep learning-based diagnostics of neurological disorders.[12] Therefore, in the context of data augmentation for the phishing email image dataset, a thorough exploration of techniques using the OpenCV library in Python has been conducted. These augmentation techniques are designed to artificially expand the dataset, thereby enhancing the robustness and generalizability of the recognition models. Based on preliminary estimations, it is anticipated that the dataset size could expand to between 12,000 and 20,000 images after augmentation. This range aligns with the recommendation by Professor Sven Nõmm and E. Vfalla's study.[12]

In this project, given the data-intensive nature of the transformer model and the limited size of the initial dataset, data augmentation was deemed necessary. To address this, the dataset was expected to expand by twenty times, and a random shuffling procedure was incorporated to enhance the diversity of the training data. On-the-fly data augmentation is another practice in machine learning. Originally, the plan was to utilize Generative Adversarial Networks (GANs) for the purpose of data augmentation. Due to constraints in time and expertise, this approach was replaced with the use of OpenCV. Moreover, the decision to pre-generate the augmented data was considered a significant contribution to the work that could potentially benefit future research.

3.4 Shuffling the image dataset

Shuffling the image dataset is a crucial step in preparing the data for machine learning models, particularly in the context of image recognition tasks like phishing email detection. This process ensures that the data fed into the training and validation phases is randomized, which helps in mitigating any order bias that may exist due to the way the images were originally collected or stored. By randomizing the order of the images, we promote a more robust learning environment where the model is less likely to overfit to sequences or patterns that might be present in an unshuffled dataset. This approach not only enhances the generalizability of the model but also ensures that each batch seen by the model during training is statistically independent, providing a comprehensive exposure to the variations within the data across different epochs. Therefore, shuffling is integral to achieving a balanced model that performs well on unseen data, maintaining consistency and reliability in its predictions.

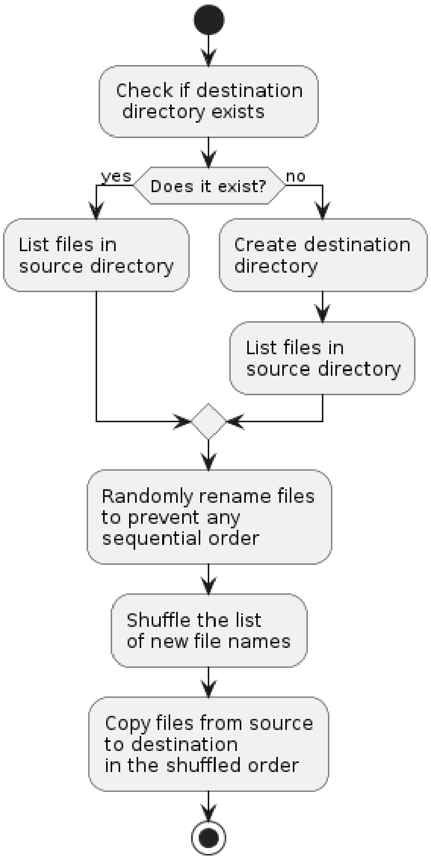
Check if destination directory exists
Does it exist?
yes
no
List files in source directory
Create destination directory
List files in source directory
Randomly rename files to prevent any sequential order
Shuffle the list of new file names
Copy files from source to destination in the shuffled order

Figure 1. Flowchart of the shuffling process

3.5 Recognition model training

The decision to employ Vision Transformers (ViT) as the deep learning architecture for this research over traditional Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) is founded on several strategic advantages that ViTs offer, particularly for the task of image-based email classification.

The primary advantage is Vision Transformers are inherently adept at handling sequential data, which allows them to consider the entire image context, capturing global dependencies within the image.[22] This holistic approach to processing images is contrasted with CNNs that process images locally and might miss the larger patterns necessary for distinguishing sophisticated phishing attempts.

Secondly, ViTs excel at transfer learning.[22] The capability to transfer knowledge from extensive pre-trained models enable ViTs to outperform CNNs in situations where annotated data is scarce, which is often the case in phishing detection due to the rapid evolution of phishing techniques.

The research starts with building a training model prototype.

3.5.1 Building model training prototype

The Python script employs a pre-trained Vision Transformer (ViT) model, specifically the "ViT_B_16" from PyTorch's torchvision library. Key steps in this function include:

- Model Initialization and Modification: The script modifies the pre-trained ViT model by freezing its existing parameters and altering the classifier head with a new linear layer. This modification tailors the model for a binary classification task targeting the 'Phishing' and 'Ham' categories.
- Data Loader Setup: This sets up data loaders for both the training and testing phases. These loaders handle image data from specified directories, applying transformations suitable for the pre-trained ViT model.
- Training Process: The training employs an Adam optimizer and CrossEntropyLoss. It is executed over a defined number of epochs. Functions from an engine module facilitate this process, along with "set_seeds" for ensuring reproducibility.
- Model Saving: Post-training, the script saves the model using a utility function from the utils module.

3.5.2 Training in Taltech AI lab environment

A Bash shell script was designed to set up an environment for running a machine learning model training project, specifically using a Vision Transformer (ViT), on a remote AI lab server. The script is structured to configure necessary environment variables, activate a Python virtual environment, execute the training script, and then deactivate the environment. Here's a detailed breakdown of its components:

Environment Variable Configuration:

- First setting up the required environment variables:
- LD_LIBRARY_PATH: prepends the CUDA 11.4 library path. It is used by the Linux dynamic linker to find shared libraries. This is essential for ensuring that the CUDA libraries are correctly located by the system.
- PATH: sets CUDA 11.4's binary directory, which allows the system to find and execute CUDA binaries, like the NVIDIA compiler nvcc.
- CUDA_HOME: sets the CUDA 11.4 installation directory. This is often required by various tools and scripts to locate the CUDA installation.

- PYTHON_PATH: sets CUDA's Python libraries. This is crucial for Python to find CUDA-related modules.

Activation of Python Virtual Environment:

- *source huaenv/bin/activate*: This command activates a Python virtual environment named huaenv. Virtual environments in Python are used to create isolated spaces with specific packages and versions, which is crucial for maintaining project dependencies and avoiding conflicts with system-wide Python packages.

Model Training Execution:

- */usr/bin/time -v python vit/main.py*: This command runs the main Python script (main.py) located in the vit directory. The script is executed with Python, and its execution time and resources are logged verbosely (-v) by the time command. This is useful for performance analysis, as it provides detailed information about the script's resource usage, including CPU time and memory consumption.

Deactivation of the Virtual Environment:

- *deactivate*: After the training script finishes, the script deactivates the Python virtual environment. This is a clean-up step to ensure that any subsequent commands run in the default system environment.

3.5.3 Training model with augmented dataset

The script is designed for training a Vision Transformer (ViT) model on a classification task with images stored in zip files. Here's a summary of the key components in the "data loader setup" step:

- ZipDataset Class:
  A custom dataset class ZipDataset is defined, which inherits from PyTorch's Dataset class. It's designed to handle image datasets stored in zip files, with the ability to apply transformations to the images. This class extracts image names and labels from the zip file, using the directory names in the zip file as labels.
- create_dataloaders Function:
  This function sets up the data loaders for training and testing. It uses the ZipDataset class to create datasets from zip files containing training and test data, and returns DataLoader objects for each.

- Data Preparation:
  The script sets up paths to the training and testing data (stored in zip files). It applies transformations to the data (derived from the pre-trained ViT model's default transformations).

3.5.4 Prediction and Displaying the Result

Finally, the script iterates through images in a specified directory and applies a function pred_and_plot_image, which makes a prediction using the loaded model on each image and then plots and saves the result. The function takes the model, image name, image path, and class names as arguments.

3.6 Answer Research question [RQ2]

A critical component of this research is the collection and utilization of diverse datasets, including contributions from Dredze et al.[6], the Image Spam Hunter (ISH) [7], and a personal archive. These datasets collectively form a robust foundation for the study, providing a wide spectrum of phishing and spam email examples for analysis. A key aspect of this research involves the use of OpenCV for data augmentation. The primary prototype workflow was developed to utilize a Vision Transformer model, training it on an augmented dataset within a specialized AI lab environment:

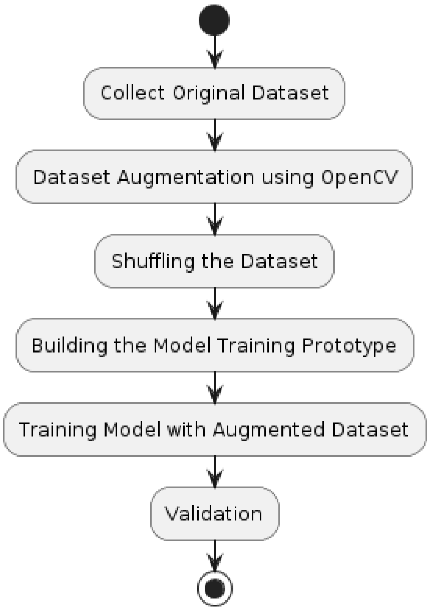
Collect Original Dataset
Dataset Augmentation using OpenCV
Shuffling the Dataset
Building the Model Training Prototype
Training Model with Augmented Dataset
Validation

Figure 2. Main prototype workflow process

3.6.1 [RQ2.1]How to create a Compensative Dataset

The development of a compensative dataset is imperative due to the limitations in size and potential imbalances in the existing dataset. Recognizing the necessity for diverse and representative data collection, this project has focused on aggregating a wide array of phishing and spam email datasets. These datasets are sourced from various platforms, including Dredze et al[6], Image Spam Hunter (ISH), and a personal archive.

3.6.2 [RQ2.2]How to implement Data Augmentation

The primary objective of utilizing such sophisticated augmentation strategies is to elevate the diversity and quality of the dataset, a critical factor in the success of using the OpenCV libraries for data augmentation. While on-the-fly data augmentation is employed in numerous machine learning workflows due to its efficiency and effectiveness, I have opted for the approach of utilizing pre-generated image data augmentation. This decision arises from my recognition of the augmented dataset as a substantial scholarly contribution, with potential benefits for other researchers, particularly in light of the limited availability of publicly accessible image phishing/spam datasets.

3.6.3 [RQ2.3]How to train the Recognition Model

The training process involves several key steps:

- **Model Prototype Development:** Start by building a basic model prototype to establish a structural foundation.
- **Fine-tuning:** Optimize the model by adjusting key parameters like batch size and epochs.
- **Advanced Training:** Upload the complete project to the TalTech AI Lab for advanced training.
- **Advanced Training with augmented dataset:** Enhance the training process by incorporating the pre-generated augmented dataset.

4. Research Results Presentation

4.1 Research Results Presentation research question

In this section, the research question **[RQ3] What are the research results** is answered.

4.2 Dataset collection and Distribution result

The dataset is derived from diverse sources, including academic contributions from Dredze et al.[6], datasets from the Image Spam Hunter (ISH)[7], and a small portion of the personal archive. This diversity ensures a wide spectrum of email types and scenarios, enhancing the robustness and applicability of the model.

Additional datasets were also explored during the selection process, including one referred to as a "challenging dataset" by its creators, Annapurna Annadatha and Mark Stamp. In the image-based phishing/spam detection research domain, datasets such as ISH and Dredze's are often regarded as **Standard Datasets** due to their comprehensive and realistic phishing scenarios. [19] Conversely, the dataset by Annadatha and Stamp primarily consisted of a simplistic combination of text and randomly selected images, which does not adequately reflect real-world phishing scenarios. Due to its limited applicability and the artificial nature of its content, this "challenging dataset" was deemed unsuitable for this study.

During the literature review and result presentation section, nine other datasets were also explored and studied. Unfortunately, they were either text-based datasets or not publicly accessible. Therefore, they were also unsuitable for this study.

The dataset for this study is partitioned into an **80%/20% training and testing distribution** to minimize the risk of data leakage. Data leakage occurs when information from the testing data is inadvertently used to train the model, which can lead to overfitting where the model performs well on the training data but poorly on unseen data. By carefully splitting the dataset into separate training and testing sets and ensuring that no information is shared between them, the integrity of the evaluation process is maintained. Below, we detail the dataset breakdown, the augmentation process, and the validation set. The training and testing datasets are composed as follows:

- **Total Ham Dataset**: The total number of ham images used in this study is the sum of the images in the test and training sets. Specifically, Ham_test = 400 and Ham_train = 1552, totaling 1952 ham images.

- **Total Phishing/Spam Dataset**: The total number of phishing/spam images is Phishing_test = 184 and Phishing_train = 700, totaling 884 phishing/spam images.

Table 6. Training, Testing, and Validation Dataset Distribution

Category	**Test Set**	**Train Set**	**Validation Set**
Ham	400	1552	50
Phishing/Spam	184	700	50

It is also important to mention that not all images from the ISH and Dredze datasets were suitable for training deep learning models. The most common reason is that these images were either recognized as broken or had errors during the original editing process, which was intended to remove sensitive personal data. Consequently, the aggregate count of ham and phishing/spam emails is less than the combined totals of these two datasets.

4.3 Data Augmentation result for Model Training

To demonstrate the data Augmentation result, taking an image, designated as "aaas.jpg,"(as named so in the oringal datasets) representative of a phishing email, was subjected to these augmentation procedures. The process generated 20 distinct images, each resulting from a unique transformation. These images, systematically named from "augmented_image_0.jpg" to "augmented_image_19.jpg," expanded the initial dataset from one to twenty images. These images were subsequently saved to disk, providing a richer dataset for further analysis or reuse.

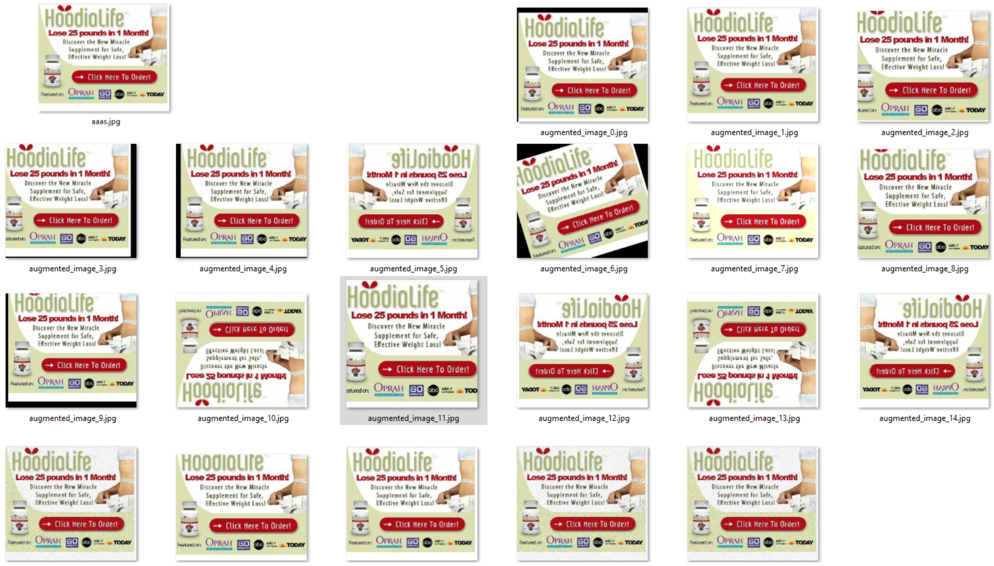

Figure 3. Data augmentation by OpenCV

The following data augmentation techniques were applied to enhance the diversity and size of the training dataset, which also aligns with the recommendation by Professor Sven Nõmm and E. Valla's study.[9]:

Table 7. Summary of Image Augmentation Techniques

Method	**Description**	**Parameter Range**
Random Brightness	Modifies the value channel in HSV	0.5 to 1.5
Random Contrast	Scales pixel intensities	0.5 to 1.5
Random Saturation	Modifies the saturation channel in HSV	0.5 to 1.5
Random Scaling	Resizes the image by a scale factor	0.7 to 1.3
Random Translation	Shifts the image in X and Y directions	Up to ±20 pixels
Random Rotation	Rotates the image around its center	-15 to +15 degrees
Random Noise	Adds Gaussian noise to the image	Mean = 0, SD = 25

After applying advanced data augmentation techniques, the training dataset was significantly enlarged, expanding nearly 20 times its original size. This substantial increase not only enriches the diversity of the dataset but also enhances the robustness and generalizability of the model by allowing it to learn from a broader spectrum of variations within the data. Such augmentation is crucial for improving the model's performance, particularly in accurately detecting nuanced and sophisticated phishing and spam emails under varied conditions.

Table 8. Augmented Training Dataset

Category	**Train Set**	**Train Augmented**
Ham	1552	31040
Phishing/Spam	700	13860

4.4 Model Training Result

The primary parameters configured for training are outlined below, reflecting the specific requirements aimed at optimizing the learning process:

- **Batch Size**: The model was trained using a batch size of 32. This size was chosen to balance the computational load and training dynamics, ensuring efficient utilization of system resources while maintaining adequate gradient estimation.
- **Epochs**: Training was conducted 30 epochs to allow the model sufficient time to converge on the optimal weights and biases.

The trained model was meticulously saved to a designated directory, utilizing the `utils.save_model` function to ensure the integrity and reusability of the model state.

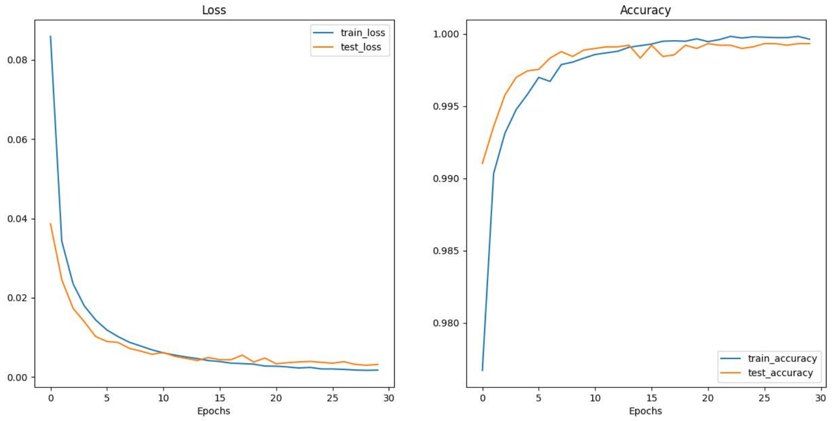

Figure 4. Training model in AI lab with augmented dataset

Loss Chart: The training loss (blue line) starts high and decreases sharply within the first few epochs, then continues to decrease gradually, flattening out as it approaches zero. The test loss (orange line) also decreases sharply initially and then exhibits some fluctuation, indicating some variance between epochs but generally maintaining a downward trend. Accuracy Chart: The training accuracy (blue line) begins just below 1.0 and remains stable throughout the training process, showing the model performs consistently well on the training data. The test accuracy (orange line) starts around 0.97, and shows more variability, peaking around 0.99 but with some dips. This suggests the model may be experiencing some overfitting.

4.5 Classification Metrics

The performance of the developed model was rigorously evaluated using classification metrics. Upon detailed examination, the model displayed a flawless prediction rate for phishing emails, with all 50 phishing emails accurately identified. Furthermore, the model successfully recognized 49 out of 50 ham emails, with a singular instance of a ham email being erroneously classified as phishing. These results validate the model's high precision and reliability in a real-world application scenario.

The following table presents the key metrics obtained from the evaluation of the model:

Metric	Value (%)
Accuracy	99
Precision	98
Recall	100
Specificity	98
F1-Score	99

Interpretation of Metrics:

1. **Accuracy (99%):** The model achieved an exceptionally high accuracy rate, correctly identifying 99% of the emails in the test set, which indicates strong overall performance in distinguishing between phishing/spam emails and legitimate ones.
2. **Precision (98%):** The precision metric signifies that 98% of the emails identified by the model as phishing/spam were indeed correct. This high precision rate is crucial in minimizing false positives, and ensuring that legitimate emails are not incorrectly flagged.
3. **Recall (100%):** With a recall rate of 100%, the model efficiently identifies a significant majority of actual phishing and spam emails. This is critical for a cybersecurity application where missing harmful emails could have serious implications.
4. **Specificity (98%):** The specificity of the model stands at 94%, indicating a strong ability to correctly identify genuine emails. High specificity is important to avoid the inconvenience and potential operational disruptions caused by misclassifying legitimate emails as threats.
5. **F1-Score (99%):** The F1-score, which balances precision and recall, is 99%. This demonstrates that the model maintains a good balance between accurately identifying phishing/spam emails and minimizing false positives.

The combined analysis of these metrics underscores the robustness of the Vision Transformer model, enhanced by the data processing capabilities of OpenCV, in the context of phishing and spam email detection. However, it is important to note that the validation dataset constitutes a small portion of the original datasets. This limited scope might introduce potential validation bias, which should be considered when interpreting the results.

4.6 Answer research question[RQ3]

This study has presented a comprehensive exploration of image-based deep learning techniques for the detection of phishing and spam emails, structured around a series of meticulously designed experiments and evaluations. The development and testing of the model were anchored by an extensively augmented dataset, derived from multiple sources, ensuring both variety and volume to simulate real-world conditions as closely as possible.

Key Findings

- The training process utilized advanced data augmentation techniques, which expanded the dataset nearly twenty times its original size. This substantial increase significantly enhanced the model's ability to generalize across different types of email data, effectively reducing overfitting and improving overall model robustness.
- The configured model training parameters, such as a batch size of 32 and 30 epochs, were optimized to ensure efficient learning without compromising the system's performance. The use of the `utils.save_model` function guaranteed the preservation of model integrity for future use and validation.
- Upon evaluation, the model demonstrated exceptional performance metrics, achieving an accuracy of 99%, a precision of 98%, a recall of 100%, specificity of 98%, and an F1-score of 99%. These results highlight the model's precision in identifying phishing and spam emails while minimizing false positives, ensuring that legitimate emails are seldom misclassified.

5. Discussion

This research successfully developed a proof-of-concept prototype leveraging a transformer architecture combined with an augmented dataset to identify ham or phishing/spam emails. The augmentation of the dataset served to enhance the model's exposure to diverse email types, potentially increasing its robustness in real-world scenarios. This prototype signifies an innovative step in employing advanced deep-learning techniques for cybersecurity purposes.

5.1 Limitations

The study faced several limitations:

1. Dataset Comprehensiveness: In this research, standard datasets were selected that are over 16 years old[19], while new datasets were also explored; however, the new ones were either poorly collected or not publicly accessible. Due to time and resource constraints, the study primarily included mainly older images and content from U.S. sources targeting English-speaking individuals. The reliance on older datasets may lead to concept drift. This limitation potentially affects the model's generalizability and effectiveness in identifying phishing emails in languages other than English or from non-U.S. sources.

2. Size of Personal Collection: The dataset was further limited by the small size of the personal collection used, which may not have provided a sufficiently varied representation of phishing email characteristics.

3. Lack of Comparative Analysis: The study did not include a comparative analysis of the transformer model against other machine learning architectures, such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs). Such a comparison could provide valuable insights into the relative effectiveness of these different approaches for phishing email detection.

4. Bias in Model Training: During the model training process, biases may have been unintentionally introduced if the data was not selected randomly or representatively. Moreover, the dataset is still relatively small, which can exacerbate these issues. This can result in skewed performance, especially when the model is presented with data that were underrepresented in the training set, leading to noticeable overfitting.

5. Evaluation Metrics Limitations: The metrics used to evaluate the model might not fully capture its effectiveness in a real-world scenario. For example, the validation dataset constitutes a small portion of the original datasets. This limited scope might introduce potential validation bias.

6. Dependency on Feature Engineering: The effectiveness of traditional machine learning models, including some types of neural networks, often heavily depends on the quality and selection of feature engineering. This study may not have explored the impact of different feature sets comprehensively, which could affect the model's performance.

7. Adaptability to New Phishing Techniques: Phishing attacks evolve rapidly, and models trained on current techniques may not adapt well to new or emerging tactics. The study might not have addressed the model's ability to update or retrain in response to new threats.

8. Model Vulnerability to Adversarial Attacks: The robustness of the model against adversarial attacks, where inputs are deliberately manipulated by attackers to cause the model to make errors, was not thoroughly examined. This vulnerability is critical in cybersecurity applications, as adversaries continuously seek new ways to circumvent detection systems. The absence of testing against adversarial examples might limit the model's reliability and effectiveness in a real-world environment where phishing tactics are constantly evolving.

9. Legal and Ethical Considerations: The study may not have fully explored the legal and ethical implications of deploying such a model, including privacy concerns and the potential consequences of false positives and false negatives.

5.2 Answer Main Research Question [MRQ]

This thesis advances the cybersecurity field by introducing an advanced, image-based deep learning approach to detect phishing and spam emails. It highlights the efficacy of the Vision Transformer in conjunction with OpenCV for data augmentation, laying a foundational path for future research and development in cybersecurity measures against email-based threats.

MRQ: How can a new image-based deep learning solution be employed to detect phishing and spam emails?

The study's primary objective was to explore and implement the Vision Transformer architecture, a novel approach in the cybersecurity realm. The development and testing of a prototype model using ViT, complemented by the augmented dataset via OpenCV, marked a significant advancement in email security.

A critical component of this research is the collection and utilization of diverse datasets, including contributions from Dredze et al., the Image Spam Hunter (ISH), and a personal archive. These datasets collectively form a robust foundation for the study, providing a wide spectrum of phishing and spam email examples for analysis. The dataset collection is shared publicly.

The findings of this research underline the effectiveness of the Vision Transformer model, augmented by OpenCV data processing capabilities, in the domain of email security. During the initial research plan, two research approaches were considered:

- Developing different image datasets and refining deep learning algorithms to target optimal detection performance.
- Exploring the latest novel deep learning architectures for image-based phishing and spam email detection.

The second approach was selected due to several reasons:

- **Innovation and Advancement:** The latest deep learning architectures, particularly those based on transformers, represent the cutting edge in machine learning technology. Opting for the newest models could potentially uncover new insights and methodologies not achievable with older technologies. Additionally, deep learning architectures like Convolutional Neural Networks (CNNs) have been extensively researched for their efficacy in detecting image-based phishing/spam emails.[3]
- **Scalability and Future-proofing:** Newer architectures often come with improved scalability and efficiency, making them more adaptable to evolving threats and larger datasets.
- **Community and Support:** Engaging with the latest technologies also taps into active research communities, ensuring access to ongoing improvements and collaborative opportunities.

Due to the constraints of time and knowledge, OpenCV was chosen for more refined data augmentation. On-the-fly data augmentation is also a common method used in many machine-learning workflows due to its efficiency and effectiveness. However, there are compelling reasons to employ pre-generated data augmentation in certain scenarios. One primary reason for adopting pre-generated augmentation in this study is the potential for the augmented dataset to serve as an independent contribution to the field.

Due to the selected research approach, the comparison with other solutions was not sufficiently addressed. This oversight represents a significant area for further investigation. While this study is not without limitations, it paves the way for further research into the application of advanced machine learning models to improve email security protocols and combat cyber threats.

5.3 Conclusion

This study represents a pioneering effort in applying deep learning with transformer architectures for the detection of phishing/spam emails. The successful implementation of this prototype underscores the potential of transformers in areas beyond their traditional applications, such as text-based processing.

The dataset created in this study is of significant importance. It comprises verified image-based phishing/spam data, which is suitable for training deep learning models, and incorporates a novel shuffling process.

The current model primarily serves as a prototype, it has undergone solid validation with robust classification metrics results. With a precision rate of 98.04%, coupled with the limitation that the validation dataset is a relatively small subset of the original dataset, there is a potential bias that must be considered. The primary contribution of this research lies in the exploration and implementation of innovative image-based deep-learning methodologies, aimed at enhancing the detection of phishing and spam emails. This includes a detailed discussion of the model's limitations and recommendations for future research. Successful refinement and validation of the model could eventually lead to its integration into corporate email systems, webmail services, or other applications where security is crucial.

5.4 Recommendations for Future Research

Future research could significantly extend and enhance this initial work:

1. Dataset Enhancement: The dataset could be enriched with newer images and content from diverse sources[3], including but not limited to different countries and linguistic backgrounds. Incorporating various forms of email content, such as QR codes and other graphical elements, would potentially increase the model's accuracy and applicability. An expanded classification system could also be implemented to distinguish more effectively between phishing and spam emails, thereby refining the dataset's utility in practical

scenarios.[10]

2. Use of GANs for Dataset Augmentation: Employing Generative Adversarial Networks (GANs) could be explored for further dataset augmentation.[23] GANs can generate synthetic, yet realistic, email content, which might help in training the model to recognize more sophisticated phishing attempts.[24]

3. Comparative Model Analysis: Future studies should include a comparative analysis of the transformer model with other machine learning architectures like RNNs and CNNs.[3] This comparison would provide valuable insights into the strengths and weaknesses of different approaches in phishing email detection and help refine the choice of models for this task.[2]

4. Robustness Testing Against Adversarial Attacks: It is critical to conduct comprehensive evaluations of the model's robustness against adversarial attacks. Future research could develop and test new defensive strategies to strengthen the model against such tactics, ensuring reliability in adversarial environments.[14]

Through these recommendations, subsequent research can build upon the foundational work presented in this thesis, potentially leading to more robust and effective solutions in the field of email security and phishing/spam detection.

References

[1] Estonian information system authority. *cyber security in Estonia 2024*. pp. 10-11. 2024.

[2] Jayshree Hajgude. *Phishing mail detection technique by using textual and URL analysis*. IEEE pp. 1-2. 2012.

[3] Sharmin Tazmina et al. *Convolutional neural networks for image spam detection*. Information Security Journal: A Global Perspective 29.3 pp. 103-117. 2020.

[4] Gao Yang et al. *Image Spam Hunter*. IEEE pp. 2-4. 2008.

[5] Basheer Al-Duwairi, Ismail Khater, and Omar Al-Jarrah. *Detecting Image Spam Using Image Texture Features*. IJISR, Volume 3, Issue 4, pp.334- 352. 2013.

[6] Dredze et al. *Learning Fast Classifiers for Image Spam*. CEAS pp. 3-7. 2007.

[7] Wu et al. *Using visual features for anti-spam filtering*. IEEE pp. 2-4. 2005.

[8] Qiao Liu et al. *Efficient Modeling of Spam Images*. IEEE pp. 664-666. 2010.

[9] B. Mehta et al. *Detecting Image Spam Using Visual Features and Near Duplicate Detection*. 17th WWW pp. 501-504. 2008.

[10] Basheer Al-Duwairi et al. *Detecting Image Spam Using Image Texture Features*. IJISR Volume 3, Issue 4 pp. 664-666. 2013.

[11] Hiransha M et al. *Deep Learning Based Phishing E-mail Detection*. ceur-ws pp. 2-4. 2016.

[12] S Nõmm and E Valla. *Enhancement and augmentation of drawing tests for deep learning-based diagnostics of neurological disorders*. pp.5-7. 2021.

[13] Dhruv Rathee and Suman Mann. *Detection of E-Mail Phishing Attacks – using Machine Learning and Deep Learning*. 2022.

[14] H. Zhang et al. *Self-Attention Generative Adversarial Networks*. Proceedings of the 36th International Conference on Machine Learning pp. 2-7. 2019.

[15] Cameron McGinley and Sergio A. Salinas Monroy. *Convolutional Neural Network Optimization for Phishing Email Classification*. IEEE pp. 5611-5612. 2021.

[16] Safaa Magdy et al. *Efficient Spam and Phishing Emails Filtering Based on Deep Learning*. ScienceDirect pp. 10-11. 2022.

[17] Taghi M. Khoshgoftaar Connor Shorten. *A survey on Image Data Augmentation*. SpringerOpen pp. 4-21. 2019.

[18] Adwan Yasin and Abdelmunem Abuhasan[. *An Intelligent Classification Model for Phishing Email Detection*. ScienceDirect pp. 59-62. 2016.

[19] Annadatha and Stamp. *Image spam analysis and detection*. Journal of Computer Virology and Hacking Techniques Volume 14, pp. 39–52. 2018.

[20] YONG FANG et al. *Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism*. IEEE volumn7 pp. 56331-56335. 2019.

[21] Samer Atawneh and Hamzah Aljehani. *Phishing Email Detection Model Using Deep Learning*. Electronics pp. 8-13. 2023.

[22] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Cornell University, pp. 1–9. 2020.

[23] Erik Dzotsenidze Elli Valla Sven Nõmm et al. *Generative Adversarial Networks as a Data Augmentation Tool for CNN-Based Parkinson's Disease Diagnostics*. ScienceDirect pp. 110-111. 2022.

[24] C. Ledig et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4684*. 2017.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I [Hua Zhong]

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “ A deep Learning Solution for Detecting Image-Based Phishing/spam Emails”, supervised by Sven Nõmm and Adrian Nicholas Venables
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

17.05.2024

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 - Repositories

A table with some datasets and main source code repositories (*The full datasets' list is committed to the main code repository; not all images are used for this study*):

Table 9. A table with Repositories

Nr	**URL**	**Name**
1	https://livettu-my.sharepoint.com/:f:/r/personal/huzhon_taltech_ee/Documents/augmented/Spam?csf=1&web=1&e=nxMWez (access right is required)	Augmented Phishing (total 17,680)email image dataset
2	https://gitlab.cs.ttu.ee/huzhon/vit/-/blob/main/test_data.zip	Ham (total 2002) email image dataset
3	https://gitlab.cs.ttu.ee/huzhon/vit/-/blob/main/test_data.zip	Original phishing(total 929) email image dataset
4	https://gitlab.cs.ttu.ee/huzhon/vit	Main source code repository
5	https://gitlab.cs.ttu.ee/huzhon/vit/-/tree/main/data/Prediction_set	Test sets for prediction validation(total 100)
6	https://gitlab.cs.ttu.ee/huzhon/vit/-/tree/main/data/Single%20error?ref_type=heads	Singular prediction Error(total 1)