Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics - Business Analytics and Quantitative Marketing

Advancing Cyberdefense through BERT: A Natural Language Processing Approach for Vulnerability to Attack Mapping within a Responsible Artificial Intelligence Framework

Jule Koenders (609892)

Supervisor:	P.C Bouman
Second assessor:	O. Kuryatnikova
Company supervisor:	C. Yuen
Date final version:	30-06-2024

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Traditional cybersecurity methods are struggling to keep pace with the rapidly evolving landscape of sophisticated cyber-attacks and the expanding complexity of digital infrastructures. Many conventional cyber security approaches rely on manual processes, which are ineffective against new, unidentified threats and cannot scale with the growing volume of digital data. Static defense mechanisms, such as firewalls and antivirus software, often fail to adapt dynamically to new threats, leading to slower response times and gaps in defense. As a result, more dynamic and adaptive security measures are needed to enhance cybersecurity postures. This thesis investigates the integration of advanced Natural Language Processing (NLP) techniques within a Responsible Artificial Intelligence (AI) framework to enhance the prediction cybersecurity vulnerabilities. It particularly emphasizes the use of Large Language Models (LLMs), to improve the mapping of Common Vulnerabilities and Exposures (CVE) to the MITRE ATT&CK framework. The mapping of CVE to ATT&CK techniques is an important step in the automation for cyber risk assessments, making the mitigation of cyber attacks easier. This research focuses on the mapping step within the automated risk assessment.

This research evaluates two NLP models: the Semantic Mapping of CVE to MITRE ATT&CK techniques (SMET) and MAP-SecureBERT that uses Named Entity Recoginition (NER) combined with a fine-tuned LLM to map CVEs to MITRE ATT&CK techniques. SMET uses Semantic Role Labeling (SRL) to extract semantic structures from CVE descriptions, linking these insights to corresponding ATT&CK techniques through logistic regression (LR). MAP-SecureBERT employs the fine-tuned SecureBERT adapted for cybersecurity contexts, and uses NER to identify critical cybersecurity entities within CVE descriptions to enhance technique mapping performance. These two models are trained and tested on two datasets, one small and one relative larger dataset and compared to a baseline method that is constructed within this thesis. This baseline model uses Bag-of-Words (BoW) and TF-IDF methods combined with a Random Forest (RF) algorithm.

Besides the comparison of these two NLP techniques, a new model is introduced. The Multi-Input Cyber Security (MICS) model, as addition it includes cosine similarities to compare the text from a single CVE entry with all possible MITRE ATT&CK techniques. This approach facilitates a more nuanced understanding of the relationships between CVE entries and cybersecurity techniques, enhancing predictive performance, particularly in larger and more diverse datasets where multiple techniques are harder to distinguish.

The results indicate that the new MICS Model outperforms the baseline model. By incorporating Responsible AI principles, fairness, accountability, and transparency, the implementation of these AI technologies hold on to ethical norms and regulations, like those specified in the EU AI Act. Future efforts will focus on developing a model that can support fully automated risk assessments.

List of Abbreviations

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

BoW Bag of Words

CVE Common Vulnerabilities & Exposures

DAPT Domain Adaptive Pre-training

DT Decision Trees

GPT Generative Pre-trained Transformer

IDS Intrusion Detection System

LR Logistic Regression

LLM Large Language Model

ML Machine Learning

MLM Masked Language Model

NER Named Entity Recognition

NLP Natural Language Processing

NN Neural Network

NSP Next Sentence Prediction

PPFLE Privacy-Preserving Fixed-Legnth Encoding

RF Random Forest

SRL Semantic Role Labeling

SMET Semantic Mapping OF CVE to ATT&CK

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machines

TF-IDF Term Frequency - Inverse Document Frequency

WEAT Word Embedding Association Test

Contents

Chapter 1

Introduction

Digital transformation is reshaping society, integrating technologies into daily life and forming the operational backbone of organizations worldwide. While this transformation encourages innovation and efficiency, it also increases the stakes in cybersecurity. As dependence on digital infrastructure grows, so does vulnerability to cyber attacks. These challenges position cybersecurity not only as a technical issue but as an important protection mechanism (McKinsey, 2023). Traditional cybersecurity strategies, like firewalls, antivirus software, and security audits, are becoming insufficient due to the evolving complexity and frequency of cyber threats. Recent data highlights this insufficiency, with approximately 2,200 cyber attacks daily, adding up to about 800,000 annually (World Economic Forum, 2024). The financial impact is large, with the global average cost of a data breach escalating by 15% over the past three years to reach $4.45 million (IBM, 2023). These figures underscore the need for automated cybersecurity solutions.

Artificial Intelligence (AI) is showing its importance because it can quickly process and analyze huge amounts of data. One important use of AI is machine learning (ML), which helps in spotting signs of cyber attacks. However, using ML in cybersecurity is challenging. One problem is the lack of detailed and well-annotated datasets needed to train powerful algorithms. Cybersecurity problems like viruses, unauthorized access, and service disruptions need specific and thorough data to build strong defenses. However, obtaining sufficient data is often challenging due to its sensitive nature. This limits how well machine learning can predict, and so mitigate, cyber threats (Sarker et al., 2020). Central to cybersecurity data collection are the Common Vulnerabilities and Exposures (CVE) dataset and the MITRE ATT&CK framework dataset (MITRE, 2024b) (MITRE, 2024a). The CVE dataset catalogs documented cybersecurity vulnerabilities, while the MITRE ATT&CK framework offers a matrix of techniques employed by threat actors. A fully integrated dataset linking CVE entries with specific MITRE ATT&CK techniques and mitigation strategies remains not available, while it is necessary for performing automated risk assessments.

Large Language Models (LLMs) present a promising solution by potentially bridging this gap. They can analyze vast amounts of unlabeled data from sources like CVE and MITRE ATT&CK. As they can read large unlabeled data, small annotated dataset can be used only for the prediction layer. This capability is important for training specific classification models, such as Random Forests (RF), and for the broader application of LLMs across both labeled and unlabeled data (Motlagh et al., 2024). Integrating LLMs into cybersecurity can enhance the mapping of CVE entries to corresponding attack techniques in the MITRE ATT&CK framework, overcoming common data limitations. Moreover, the addition of techniques like Named Entity Recognition (NER) and Semantic Role Labeling (SRL) could improve the performance of these models, by extracting the semantic meaning of the text.

The increasing use of AI in cybersecurity emphasizes the need for both technical efficiency and commitment to ethical standards. Responsible AI frameworks, which promote the development of AI systems that are robust, interpretable, and explainable, ensure that these technologies benefit humanity while being transparent, fair, and accountable. Such frameworks align with global regulations like the EU AI Act (European Parliament, 2023). This study leverages a Responsible AI framework to explore how pre-trained LLMs can enhance threat detection and mitigation. This is particularly relevant for organizations like Ernst & Young (EY), which emphasize ethical compliance and data security in their operations (Lu et al., 2023).

This research proposes a new approach to enhance the prediction and mapping of CVE to MITRE ATT&CK techniques through the integration of LLMs with advanced NLP techniques such as SRL and NER. Operating within a Responsible AI framework, this study not only assesses the technical capabilities of pre-trained LLMs in improving mapping strategies but also the development of the Multi-Input Cyber Security (MICS) model. The MICS model represents a significant advancement by incorporating both SRL and NER to analyze and interpret cybersecurity data. This approach not only aims to refine the prediction accuracy of cybersecurity threats but also ensures that these predictions adhere to ethical standards, promoting transparency and accountability in AI applications.

The main research question guiding this study is: “How does the integration of different NLP models enhance the capability of cyber attack technique prediction within a Responsible AI Framework?”

This leads to the following sub-questions:

- Does combining SRL and NER with pre-trained LLMs improve the performance of cybersecurity threat predictions compared to traditional attack models?
- Under what conditions related to the quality and size of datasets does semantic mapping perform well or poorly in predicting cyber attack techniques?
- How can principles of Responsible AI be effectively embedded and operationalized within the architecture of prediction models to enhance ethical practices in cybersecurity prediction?

The structure of this thesis will explore these questions across several key chapters, starting with a background on the evolution of AI in cybersecurity, a literature review of current applications, and a detailed methodology of the proposed solutions, culminating in a discussion of the results and conclusions.

Chapter 2

Background

2.1 Cybersecurity Fundamentals

A cyber attack is any intentional effort to steal, expose, disable, or destroy data, applications, or other assets through unauthorized access to a network, computer system, or digital device (IBM, 2024). In recent years, numerous organizations have struggled to defend effectively against the increasing speed and complexity of cybersecurity attacks (NVIDIA, 2024). Research on cyber threat detection and mitigation has primarily focused on risk identification and framework development. Effective prevention of cyber attacks can be achieved through a phased approach outlined by the National Institute of Standards and Technology (NIST). These phases, Identification, Protection, Detection, Response, and Recovery, are important for a successful cybersecurity program (National Institute of Standards and Technology, 2018).

In the Identification phase, organizations identify their cybersecurity policies, risk management strategies, and resource weaknesses. The Protection phase aims to limit the potential impact of a cybersecurity event, including measures such as identity management and access control, raising awareness throughout the organization, and establishing a secure data environment. Detection is critical for identifying specific attacks through monitoring for anomalies and rare events, although this can be labor-intensive. After an attack is detected, the Response and Recovery phases determine the extent of the aftermath and involve implementing recovery planning processes and making improvements based on lessons learned. These steps help organizations limit damage and enhance their resilience to future attacks.

This thesis specifically focuses on enhancing the Detection phase, utilizing advanced NLP techniques within a Responsible AI framework to improve the accuracy and efficiency of identifying cyber threats. This approach integrates fine-tuned AI models that not only detect but also predict and mitigate potential cybersecurity vulnerabilities, significantly reducing the labor-intensive nature of traditional detection methods. Building on the foundation of these cybersecurity practices, Section 2.2 will delve into the recent research conducted on AI applications in cybersecurity, exploring how these technologies, especially the models developed in this research, can further strengthen defense mechanisms against evolving cyber threats.

2.2 Introduction to Artificial Intelligence in Cybersecurity

In addition to the increased cyber risk, there are several other reasons to use AI in cybersecurity. AI can be particularly effective in the Detect and Protect pillars of the NIST framework (National Institute of Standards and Technology, 2018). AI can learn from past experiences, where it is analyzing and learning from past events or cybersecurity threats, to prevent similar threats in the future (Ansari et al., 2022). AI systems are quick, they can analyze vast amounts of data and identify anomalies to provide quicker detection. Because they can monitor network activity in real-time, AI can quickly identify signs of similar attacks in the future. Although AI brings efficiency and maintenance gains, it also introduces risks. While AI can automatically scan large quantities of data to identify vulnerabilities, this capability also makes it easier for attackers to launch targeted assaults. The algorithms can manipulate other AI systems into making harmful decisions; for instance, introducing bias can lead to harmful decision-making. AI tends to operate on complex algorithms that may be hard for humans to understand, lacking transparency (Tan, 2023). The integration of AI within cybersecurity operations not only promises improved efficiency and real-time thereat detection but also introduces ethical considerations that necessitate a Responsible AI deployment, which will be discussed in Section 3.2.

2.2.1 Evolution of AI in Cybersecurity

Machine Learning, a subset of AI, is a computational process that uses input data to achieve a desired task without being hard coded to produce a particular outcome (El Naqa & Murphy, 2015). ML models are often made up of a set of rules, these rules are capable of detecting data patterns, recognize sequences or anticipate behavior. Techniques like Support Vector Machines (SVMs), Naive Bayes and Decisions Trees (DTs) are commonly used ML techniques with a cyber application (Shaukat et al., 2020). In all the five stages of the NIST Framework, ML techniques can be applied. However it depends on the application which ML technique is most suitable. The study of Ahsan et al. (2022) highlights the importance of choosing the right ML per objective. This paragraph shows different ML applications, all within the cybersecurity sector.

One of the key applications of ML in cybersecurity is in the area of intrusion detection systems (IDS). Handa et al. (2019) demonstrates the enhancement of IDS through automation provided by ML algorithms. Specifically identification of cyber threats by an automated IDS shows importance. This system could detect threats such as the advanced Stuxnet worm, which commandeered the uranium enrichment centrifuges in Iran's Nuclear program (Langner, 2011). This automation is facilitated through techniques such as SVMs, which analyze network traffic to identify malicious activities. Beyond intrusion detection, ML algorithms are instrumental in classifying data, for instance, distinguishing between spam and legitimate messages. Bayesian classifiers, in particular, have proven effective in spam detection, underscoring the versatility of ML not only in recognizing spam but also in detecting malware and phishing attempts (Martínez Torres et al., 2019).

In addition to classic machine learning techniques, research underscores the dominance of NNs in cybersecurity detection processes (Shaukat et al., 2020). Unlike traditional ML methods, NNs offer the advantage of being able to learn and improve automatically from experience without being explicitly programmed with specific rules (Mukhamediev, 2021). This capability enables them to effectively manage increasingly large and complex datasets prevalent in cybersecurity. Particularly adept at processing unstructured data types such as images, audio, and text, NNs are important for detecting malware, identifying network intrusions, and recognizing phishing attempts, as detailed in the comprehensive review by Podder (2020) on artificial NNs for cybersecurity applications. However, while NNs are effective for pattern recognition in structured data, they face challenges with the nuances and context of unstructured textual data, often requiring extensive pre-processing which can be inefficient. This limitation underscores the need for developing more sophisticated models, such as LLMs, which can better understand the nuances of human language in the vast and varied datasets typical in cybersecurity applications, thereby enhancing threat detection and response capabilities.

2.3 Introduction to Large Language Models

According to (NVIDIA, 2024), "Large Language Models are deep learning algorithms (NNs) that can recognize, summarize, translate, predict, and generate content using very large datasets." LLMs process text by encoding it, assigning weights to specific words, and learning to perform various tasks based on these weights. They play a important role in enhancing productivity across different industries. For example, LLMs can decode the language of protein sequences and suggest viable compounds for developing new vaccines (NVIDIA, 2024).

LLMs are primarily used in generating text, summarizing content, translating languages, classifying information, and powering chatbots. Standing at the forefront of AI, these models have helped our ability to process, interpret, and generate human language with remarkable efficiency, thanks to their training on large data collections.

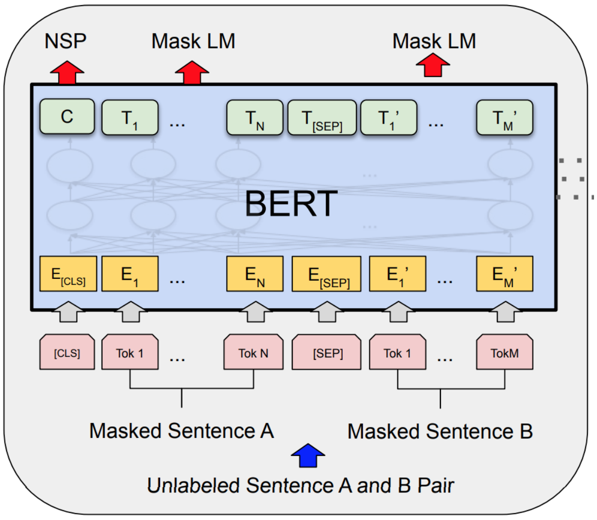


Figure 2.1: BERT Model.

As shown in Figure 2.1, among the most advanced models in NLP is BERT (Bidirectional Encoder Representations from Transformers). Developed by (Devlin et al., 2019), BERT understands context by analyzing text from both directions (left and right) at all layers. This bidirectional processing is part of what sets BERT apart from its predecessors. The model employs a unique pre-training approach known as Masked Language Modeling (MLM), where 15% of the words in each sentence are randomly replaced with a [MASK] token during training. BERT then learns to predict the original words based on the context provided by the other words in the sentence. This technique forces the model to develop a deep understanding of language context and structure, enhancing its capability for tasks such as sentiment analysis, question answering, and language inference. Additionally, BERT uses a Next Sentence Prediction (NSP) task during training, where it predicts whether one sentence logically follows another, further refining its understanding of text structure.

Following BERT, the Generative Pre-trained Transformer series, particularly GPT-3, described by (Brown et al., 2020), has set new benchmarks for text generation. Unlike BERT, which primarily focuses on understanding context, GPT-3 employs an autoregressive model that predicts the likelihood of word sequences, enabling it to generate coherent and contextually appropriate text. GPT-3's capability for few-shot learning, where it performs complex language tasks with minimal specific training data, underscores its versatility and profound comprehension of language. These models are based on the transformative architecture known as transformers, which have accelerated advancements in NLP. As the field evolves, newer versions of GPT and other domain-specific models continue to expand the capabilities of LLMs.

Chapter 3

Literature Review

This literature review investigates the application of LLMs in cybersecurity. It specifically focuses on their role in improving the mapping of CVE to MITRE ATT&CK techniques. The review addresses a gap in current research, the ethical integration of LLMs in cybersecurity, particularly for enhancing threat detection and response strategies. This work aims to highlight recent developments for CVE to MITRE ATT&CK mapping and explore the practical application of Responsible AI principles in this context.

3.1 Large Language Models in Cybersecurity

As highlighted in Section 2.3, BERT and GPT stand out as the most outstanding LLMs currently available. Since their debut, a wide spectrum of applications has been explored, ranging from healthcare (Thirunavukarasu et al., 2023) to supply chain management (Li et al., 2023), and extending into the energy sector (Dong et al., 2024). The origin of cybersecurity challenges, driven by the large volume of data, underscores the impracticalities of manual screening and threat detection. The complexity of cyber attacks surpass human capabilities, necessitating the availability and analysis of extensive datasets.

It is within this context that LLMs can prove to be valuable. However, it is important to note that LLMs were not originally designed for cyber attack detection. Their use in this field is often limited by the lack of specific training on relevant datasets. Despite these limitations, the review by Nourmohammedzadeh Motlagh et al. (Motlagh et al., 2024) reveals the impact of LLMs within the cybersecurity domain, guided by the NIST Framework's five pillars 2.1. The review examines the role of LLMs in identifying potential risks, managing cybersecurity through automated risk assessments (Kereopa-Yorke, 2023), enhancing proactive protections like web content filtration, and contributing to anomaly detection for network logs and software vulnerabilities (Ferrag et al., 2023; Tuor et al., 2018). It delves into how LLMs create response mechanisms, particularly through the development of honeypots (traps) that engage and delay attackers, buying important time for mitigation efforts. Despite their broad training, LLMs show to be adaptable to specialized tasks. This flexibility is important, considering the complex nature of cyber threats. By applying fine-tuning and customization techniques, LLMs can be adapted to detect and mitigate cyber threats, reducing manual effort. These fine-tuning and customization applications will be discussed in Section 3.1.1.

3.1.1 Fine-Tuning LLMs for Cybersecurity Applications

Most applied models are based on architectures like BERT, GPT-3, or GPT-4. Research often involves customizing these architectures for specific tasks or domains rather than using the base models. This customization is important as the effectiveness of general models lacks in highly specialized tasks (Zhang et al., 2023). Fine-tuning LLMs with domain-specific data can increase performance and efficiency while reducing irrelevant data or "noise" in the process (Naveed et al., 2024). However, creating a domain-specific LLM has its own set of challenges, especially in gathering high-quality training data. Examples of such tailored LLMs include SecureBERT, CyBERT, CySecBERT, and SecurityBERT.

SecureBERT distinguishes itself not only through its language understanding derived from BERT but also via the application of fine-tuning techniques focused on cybersecurity texts from diverse sources, including books, blogs, security reports, and academic papers. This model uses a custom tokenizer, building upon RoBERTa's tokenizer to include merge English vocabulary with new cybersecurity-relevant tokens (Liu et al., 2019). This makes SecureBERT applicable for tasks such as phishing detection, code and malware analysis, and intrusion detection (Aghaei et al., 2022). Likewise, CyBERT and CySecBERT represent adaptations of the BERT model fine-tuned for cybersecurity applications. CyBERT aims to increase accuracy and efficiency of cybersecurity tasks such as entity recognition and threat classification; it is specifically trained for industrial control systems device documentation (Ranade et al., 2021). CySecBERT engages in domain-adaptive pre-training (DAPT) to grasp cybersecurity language nuances, it is trained on less data within the cyber domain compared to SecureBERT (Bayer et al., 2022). SecurityBERT employs Privacy-Preserving Fixed-Length Encoding (PPFLE) to transform unstructured network data into a structured format that BERT can efficiently process (Ferrag et al., 2024). This approach, unique to SecurityBERT, integrates privacy into the training data, which is an important feature when using sensitive data.

When deciding among CySecBERT, SecureBERT, CyBERT, and SecurityBERT, the choice depends on the specific requirements of the task. CySecBERT suits tasks that demand a mix of general and cybersecurity text, making it ideal for scenarios requiring domain-specific insights. SecureBERT, with its deep domain knowledge, is preferable for tasks where an understanding of cybersecurity terminologies and contexts is important. Being trained on more data than CySecBERT, it can provide more detailed insights. CyBERT specializes in recognizing and understanding cybersecurity-specific entities in control systems device documentation, making it optimal for specific cybersecurity entity recognition tasks within this field. SecurityBERT, with its innovative PPFLE technique and real-time IoT device deployment capabilities, is the go-to model for high-accuracy, privacy-aware cyber threat detection. Each model offers unique strengths: CySecBERT for versatility, SecureBERT for deep domain knowledge, CyBERT for focused entity recognition, and SecurityBERT for high-accuracy, real-time detection and privacy preservation.

Based on this comparison, the models trained on vulnerabilities, techniques and mitigations for cyber attacks could be a good fit for this research. SecureBERT and CySecBERT fit this description the most, based on the amount of data the models are trained on SecureBERT shows to have the deepest domain knowledge. Besides training BERT models on domain-specific data, it is important to fine-tune the model for the specific task it will be used for, to enhance their performance and efficiency. As demonstrated in the research by Chi Sun et al. (Sun et al., 2019), opportunities to further fine-tune the BERT model for a better understanding of context, such as in text classification tasks, are important. The architecture of BERT consists of multiple layers that capture different levels of semantic and syntactic information. For example, layers closer to the input may capture more generic features, while those closer to the output may capture more task-specific features. Adjusting the learning rate and employing strategies to prevent catastrophic forgetting are important in this step to retain previously learned information while adapting to new data. The architecture of the BERT model will be further elaborated upon within Chapter 5.

3.1.2 Generating Mitigation Strategies by LLMs

The development of automated risk assessments by LLMs is progressing, as these models identify vulnerabilities and threats, suggesting specific mitigation tactics aligned with an organization's unique risk profile. The models are not yet implemented because ensuring legal compliance and ethical integrity is difficult to ensure. The article "Using Large Language Models to Mitigate Ransomware Threats" by Fang Wang (Wang, 2023), explores the potential of LLMs like GPT for the development of cybersecurity policies and strategies to counter ransomware threats. In the article they recommend to perform further research into generating mitigation strategies out of fine-tuned domain-specific LLMs. The article by Jin et al.(Jin et al., 2024) takes a different approach, trying to correlate CVE with MITRE ATT&CK techniques. This paper highlights the importance of the generation of mitigation strategies trough LLMs.

One innovative approach, detailed by Basel Abdeen et al. (Abdeen, 2023) in their work on "Semantic Mapping of CVE to ATT&CK and its Application to Cybersecurity," uses SMET (Semantic Mapping of CVE to ATT&CK) to automatically map CVE entries to ATT&CK techniques based on their textual similarity. This tool leverages SRL within a domain-specific language model, ATT&CK BERT, trained on ATT&CK techniques and mitigations, to understand the semantic meaning of attack descriptions and make accurate mappings. This allows for a more complete understanding of which mitigations can be applied to which vulnerabilities.

Another instance that employs NER instead of SRL to derive useful insights from CVE entries is highlighted in the work by Grigorescu et al. Their study, titled 'CVE2ATT&CK: BERT-Based Mapping of CVEs to MITRE ATT&CK techniques', demonstrates the effective use of a BERT model to associate CVE entries with corresponding MITRE ATT&CK techniques (Grigorescu et al., 2022). This method not only enhances the understanding of cybersecurity threats by enriching the contextual insights for threat mitigation but also showcases how NER can be used to categorize and prioritize threats effectively. Such a model proves important in improving automated risk assessments and formulating precise, mitigation strategies. Another advancement is presented by Ehsan Aghaei et al. (Aghaei & Al-Shaer, 2023) in their article on "Automated CVE Analysis for Threat Prioritization and Impact Prediction." Their tool, CVEDrill, goes beyond simple threat identification to accurately estimate the Common Vulnerability Scoring System (CVSS) vector for threat mitigation and priority ranking, as well as automating the classification of CVEs into the appropriate CWE hierarchy classes. This approach promises to streamline the process of vulnerability analysis and countermeasure implementation, outperforming even sophisticated tools like ChatGPT in terms of efficiency and accuracy.

Research highlights that the effectiveness of strategies generated by LLMs depends on the models' accuracy and the relevance of their training data. Hence, continuous updates with the latest threat intelligence are essential. Ultimately, the incorporation of LLMs into cybersecurity strategy development represents a step forward, providing organisations with the means to respond more effectively and proactively to the myriad threats in the digital domain.

Despite the potential of LLMs in strengthening threat detection and generating mitigation tactics, the introduction of biases remains a important concern. These biases can manifest when models are trained on data that lack diversity, such as datasets predominantly featuring one gender, leading to outputs that may not reliably represent the entire population (Jiang et al., 2023). Such limitations underscore the critical need for Responsible AI practices in cybersecurity. This necessity not only involves scrutinizing the data used for training models but also entails a broader commitment to ensuring that AI technologies promote fairness, transparency, and accountability. The following discussion in Section 3.2 delves into the principles and practices that constitute Responsible AI, aiming to mitigate the adverse effects of biases and uphold the integrity of AI applications in cybersecurity.

3.2 The Need for Responsible AI

As AI systems become integrated into important applications, the threat of these systems acting in ways that are unaligned with ethical guidelines or human values grows, underscoring the need for Responsible AI in modern applications. AI systems have become so advanced that they require minimal human intervention. For most humans, the systems that perform these tasks are hard to grasp. This raises the need for understanding how decisions are made by AI methods. Within the last decade, researchers started implementing this concept into AI systems. It started with the concept of interpretable AI, then came explainable AI and currently they are trying to incorporate Responsible AI into the systems (Barredo Arrieta et al., 2020).

Interpretable AI can improve the implementability of ML techniques by providing robustness and guaranteeing causality in model reasoning. Building on this, Explainable AI aims to create a set of ML techniques that help human understand and manage AI systems better. The concept of Responsible AI goes beyond these two methods, Responsible AI refers to the ethical development of AI systems to benefit the humans, society, and environment (Lu et al., 2023). The pillars that fall within Responsible AI are widely discussed, overall they include explainability, fairness, accountability and privacy (Accenture, 2024).

Because of new European legislation, instances need to assure that their AI systems are safe, transparent, and accountable, while fostering innovation and competitiveness within the EU (European Parliament, 2023). This includes a risk assessment, classifying the system based on the perceived risk level of AI applications from minimal to unacceptable risk. The AI Act explicitly bans certain uses of AI that are deemed to have unacceptable risks. High-risk AI systems require high-quality datasets to operate reliably and without bias. Besides integrating these risk frameworks the Eu will implement market monitoring frameworks that oversee AI products and services available in the market to ensure they adhere to the regulations.

Based on various research Ernst & Young, developed its own framework for Responsible AI to asses their clients AI systems, which will be explained in Section 3.2.1. This paragraph will explore various approaches and interpretations of Responsible AI, delving into the implementation of Responsible AI principles within LLMs.

3.2.1 Use Case: EY's AI Framework

The Responsible AI framework developed by Ernst & Young (EY) assists clients in mitigating AI risks while remaining compliant with emerging AI regulations (Ernst & Young Global, 2024). This framework evaluates AI risks and builds controls around seven trust attributes defined explicitly by EY. These attributes are Accountability, Sustainability, Transparency, Fairness, Reliability, Privacy, and Explainability. These attributes are integral parts of the framework as defined by Ernst & Young and are not altered by the author of this thesis.

- Accountability means that there must be unambiguous ownership over an AI system and its impact throughout the AI development life-cycle.
- Sustainability refers to the design and deployment of AI systems that are compatible with the goals of sustaining physical safety, social well-being, and planetary health.
- Transparency ensures appropriate levels of openness regarding the purpose, design, and impact of AI systems, enabling end users and system designers to understand, evaluate, and correctly use AI outputs.
- Fairness ensures AI systems are designed with consideration for the needs of all impacted stakeholders and to promote inclusiveness and positive societal impacts.
- Explainability provides levels of explanation sufficient for decision criteria of AI systems to be reasonably understood, challenged, and/or validated by human operators.
- Reliability ensures that the outcomes of AI systems align with stakeholder expectations and perform at a desired level of precision and consistency, while being secure from unauthorized access and/or corruption.
- Privacy involves designing AI systems with consideration for data rights regarding how personal information is collected, stored, and used.

For some of the trust attributes like Sustainability, Accountability, and Privacy, documenting the model development process and integrating these attributes within the governance structures of the company might suffice. For other attributes like Transparency, Fairness, Explainability, and Reliability, specific implementations within the model are necessary. These implementations will be discussed in Section 3.2.2.

3.2.2 Responsible AI in LLMs

Fairness within AI refers to the assurance that biases in the data and model inaccuracies do not lead to models that treat individuals unfavorably on the basis of characteristics such as race, gender, disabilities or political orientation (Oneto & Chiappa, 2020). An example within the domain of cybersecurity is less straight forward. For instance, an AI-based cybersecurity algorithm uses for detecting fraudulent activities may disproportionately flag transactions from specific geographic locations, thereby creating a form of geographical discrimination (Kamoun et al., 2020). The effectiveness of mitigation strategies to reduce bias within ML models depends on the context of the data. Fairness metrics like Equalized Odds, Predictive Parity and Demographic Parity ensure that these models do not favor or disadvantage any group on protected characteristics (Agarwal & Mishra, 2021). If these metrics indicate unfavorable bias within the model, reweighing the data could be an option.

The complexity of LLMs, which leverage deep learning algorithms, poses challenges to their explainability. This issue is critical in high-stakes areas like cybersecurity, where the accuracy of model outcomes is important. According to Zhao et al. (2024), both local and global explanations are essential for understanding LLM decisions. Local explanations analyze how specific predictions are made, highlighting influential features. Conversely, global explanations offer an overarching view of the model's decision-making process, using techniques such as probing for learned information, neuron activation analysis, and identifying key concepts understood by the model. Besides fairness and explainability, making a model responsible is equally important. Accountability, in this context, is about ensuring that LLMs operate in a manner that is responsible, traceable, and transparent, contributing to a more trustworthy and ethically sound application of artificial intelligence. The paper written by Huang & Chang (2023) propose incorporating a citation mechanism in LLMs as a solution, these would allow for transparency and verifiability of the information generated by LLMs. However, these implementation are complex because of the combination of data sources within these models, making it hard to figure out the exact source. Ensuring privacy, especially when dealing with sensitive identifiable information (which could be the case within the cybersecurity domain), is important. To address this challenge, researchers use Privacy Protection Language Models, this integrates robust privacy protection methods into the pre-processing stage and the fine-tuning of LLMs (Xiao et al., 2023). By incorporating these methods in AI models we can try to ensure a responsible model.

3.3 Summary and Objectives of This Study

LLMs can play a central role in developing digital defense mechanisms, from enhancing threat detection to the formation of mitigation strategies. By customizing models like BERT and GPT for specific cybersecurity tasks, these technologies offer a promising results for reducing the manual burden and quicken threat response. The lack of complete datasets makes automating risk assessments difficult, using mapping techniques based on LLMs might solve this problem. Developing LLMs for organisations needs to be done in a responsible way. The debate around fairness, explainability, accountability, and privacy within AI models show a broader concern: the importance to develop AI systems that not only perform effectively but are also based on ethical standards and societal values. Further research deepening the technical approach by creating LLMs specifically engineered for particular tasks within cybersecurity, thereby increasing their Precision and adaptability to the threat landscape. Besides achieving higher Precision and adaptability, addressing the ethical challenges, need to ensure responsible deployment.

Chapter 4

Data

This research uses two fundamental databases to navigate the complexities of cybersecurity vulnerabilities and attack techniques: the Common Vulnerabilities & Exposures (CVE) database and the MITRE ATT&CK database (MITRE, 2024a,b). Both repositories are used for standardizing information across various cybersecurity frameworks, tools, and organizations, thus playing a important role in enhancing cybersecurity practices globally.

The CVE database stands as a public repository that catalogs identified cybersecurity vulnerabilities. It facilitates the universal sharing of data concerning security vulnerabilities across diverse security tools and databases through each entry, known as a CVE record. These records are systematically identified by a unique CVE-ID and include:

- CVE ID: Uniquely identifies the vulnerability (e.g., CVE-2015-7007).
- Public Date: The date the vulnerability was made public (e.g., 2015-10-21).
- Affected Products: Lists the products and versions affected by the vulnerability.
- Description: Provides a summary of the vulnerability, including its impact and how it can be exploited.
- Problem Types: Categorizes the nature of the vulnerability.
- References: Lists URLs and other references that provide further information about the vulnerability.
- Mitigations: Offers strategies and advice on how to mitigate the vulnerability.
- Vendor Information: Details about the vendor managing the vulnerability data.
- Update Dates: Indicates when the CVE record was last updated.

This research will focus on analyzing the CVE Description text, as it contains the most information. An example of an CVE Description is shown here:

"Multiple vulnerabilities in Cisco SPA100 Series Analog Telephone Adapters (ATAs) could allow an authenticated, adjacent attacker to execute arbitrary code with elevated privileges. The vulnerabilities are due to improper validation of user-supplied input to the web-based management

interface. An attacker could exploit these vulnerabilities by authenticating to the web-based management interface and sending crafted requests to an affected device. A successful exploit could allow the attacker to execute arbitrary code with elevated privileges. Note: The web-based management interface is enabled by default."

Complementing the CVE database, the MITRE ATT&CK database emerges as an exhaustive collection of hostile tactics and techniques designed for a comprehensive understanding of cyber attack behavior. This knowledge base contains:

- Tactics: The objectives attackers are trying to achieve, representing the "why" of an ATT&CK technique (e.g., Initial Access, Execution, Persistence).
- Techniques: The methods attackers use to achieve tactical goals, detailed with a description, examples, and mitigation advice. Techniques are often mapped to specific tactics to show how they fit into broader attack goals.
- Procedures: Real-world examples of techniques used by cyber threat groups or malware, providing context and illustrating how attackers operate in practice.
- Mitigations: Recommendations and strategies for defending against or reducing the impact of specific techniques.
- Software: Information about software tools, including malware and legitimate software, that attackers use to carry out attacks.

The ATT&CK database is regularly updated to encapsulate the latest insights and evolving threats, providing an valuable resource for cybersecurity professionals engaged in threat modeling, security assessments, and the formulation of defensive mechanisms.

For training the models used in this research, annotated datasets are used. The small dataset consists out of 300 entries and 41 classes, annotated by Abdeen (2023). The larger dataset consist out of 809 entries and 66 classes, from (Center for Threat-Informed Defense, 2024), but annotated with CVE descriptions by me. The first five entries of this large dataset is shown in the Appendix in Table A.

The lack of seamless integration of data from the CVE and MITRE ATT&CK databases underpins the methodology of this research, facilitating a nuanced exploration of the semantic mapping between CVE entries and ATT&CK techniques. This data leverages the CVE entries spanning from 2014 to 2024. Through this structured approach, the study bridges the gap between identified vulnerabilities and the tactics and techniques of cyber attackers, thereby proposing robust strategies for enhancing cybersecurity defenses.

Chapter 5

Methodology

This chapter explains the implemented methodology that helps answer the research questions proposed in this report. The focal point centers on the development of a model that leverage SecureBERT and/or ATT&CK BERT combined with Named Entity Recognition (NER) and/or Semantic Role Labeling (SRL) to identify vulnerabilities related to specific software or hardware versions. Drawing on the methods and performance analytics described in Sections 5.2, 5.3.1 and 5.3.2, this thesis introduces a new model: the Multi-Input Cyber Security (MICS) Model. This model is introduced in Section 5.3.3

5.1 Data Processing and Exploratory Data Analysis (EDA)

The initial step involves cleaning and preparation of the data extracted from JSON files containing CVE and the MITRE ATT&CK dataset. Not all information within the datasets is relevant for training our model and linking the CVE to the ATT&CK mitigation. Variables like, *data_type* and *data_format* can be removed, only the *CVE_ID*, *CVE_Description*, *Technique_ID* and *Technique_Description*. Text normalization is a important step for ensuring data consistency in text processing, where inputs are standardized by converting all text to lowercase and removing special characters. This process is important for more complex operations like NER and SRL, which require tokenization. Tokenization divides sentences into discrete words or "tokens" that are easier for models to process. For instance, the sentence "Jule started writing her thesis at the first of March." is tokenized into a list of words: ['Jule', 'started', 'writing', 'her', 'thesis', 'at', 'the', 'first', 'of', 'March']. BERT, incorporates its own tokenizer that not only splits the text into basic tokens but also into sub-tokens, which is critical for matching words with its pre-defined vocabulary (Devlin et al., 2019). For example, the word "unsuccessfully" might be split into ['un', '##success', '##full']. This approach can complicate tasks like NER, where accurately identifying entities across sub-tokens is challenging. The typical solution is to assign the entity label to the first sub-token of a word and a continuation label to subsequent sub-tokens. Furthermore, BERT necessitates uniform sequence lengths within batches for computational efficiency, achieved by padding shorter sequences with a [PAD] token. It also uses attention masks to ensure the model focuses only on meaningful tokens, not padding.

Besides data cleaning and pre-processing, EDA can be performed by analyzing the frequency of words and bigrams. This uncovers the textual data's core components and linguistic patterns.

This insight into the text's structure lays a foundation for developing robust features that are important for building effective ML models. Such an approach ensures that model training is based on a well-understood dataset, maximizing the potential for achieving accurate and meaningful analytical outcomes.

Data imbalance is common in cybersecurity datasets, where some types of cyber threats or vulnerabilities are less common than others. This imbalance can lead to the model developing a bias towards the majority class, reducing its effectiveness in identifying less frequent but potentially more dangerous threats. To mitigate the effect of an imbalanced dataset, resampling techniques can be used. Oversampling the minority class or under sampling the majority class helps balance the dataset. For instance, the Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples from the minority class to increase its representation in the training set (Kotsiantis et al., 2006). Within this research it is decided to test the performance of SMOTE on the dataset, because of the modest size of the test dataset. SMOTE is advantageous in such scenarios as it generates synthetic samples rather than replicating existing ones, thereby enriching our dataset without the risk of losing valuable information through under-sampling methods.

5.2 Baseline Model

Before assessing the performance of the proposed models, it is essential to establish a baseline model for comparison. This baseline model, constructed using traditional NLP approaches, serves as a reference point to illustrate the enhancements achieved by the proposed methodologies. The baseline methodology employs two widely-used text representation techniques: Bag-of-Words (BoW) and TF-IDF. These methods are instrumental in translating the textual content of CVE descriptions into numerical data that can be processed by ML algorithms.

The Bag-of-Words model represents text by counting the frequency of words within the documents, ignoring the order of words but maintaining a robust approach for capturing the presence of significant terms (Qaiser & Ali, 2018). On the other hand, TF-IDF goes a step further by reducing the weight of words that appear frequently across documents, thus highlighting words that are more unique to each document (HaCohen-Kerner et al., 2020). This method is effective in identifying key terms that are indicative of specific cybersecurity threats. To classify and map vulnerabilities described in the CVEs to corresponding ATT&CK techniques, a RF classifier is used. Known for its efficacy in handling both linear and non-linear data, RF involves constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Breiman, 2001). This model is chosen for its ability to manage the complex nature of multi-label classification inherent in mapping CVE descriptions to multiple possible ATT&CK techniques.

The performance of this baseline model is evaluated using multiple metrics designed to assess the accuracy and effectiveness of multi-label classification systems. These metrics are described in Section 5.4.2. By employing these methods and the classifier, the baseline sets a foundational benchmark against which the more advanced methodologies. These advanced methodologies will be described in Section 5.3.

5.3 Advanced Models for CVE-to-ATT&CK Semantic Mapping

This Section explores modeling techniques designed to enhance the semantic mapping of CVE descriptions to tactics, techniques, and procedures outlined in the MITRE ATT&CK framework. The integration of these advanced models aims to bridge the informational gap between the detailed, often technical descriptions found in CVE entries and the insights provided by the ATT&CK framework. By utilizing NLP and ML strategies, raw text data is converted into structured formats such as NER or SRL to better fit the cyber context. Section 5.3.3 also evaluates the performance of the new MICS model. This model incorporates SRL, embeddings, and cosine similarities between CVE descriptions and all possible MITRE ATT&CK techniques as inputs within a neural network architecture.

5.3.1 Semantic Mapping from CVE to ATT&CK Technique (SMET)

In order to systematically analyze and interpret the vulnerabilities associated within CVE entries, we apply a Natural Language Processing (NLP) technique known as Semantic Role Labels (SRL). SRL is a natural language processing technique designed to identify the basic who-did-what-to-whom structure of sentences (Shi & Lin, 2019). By extracting verbs (actions), subjects (agents), and objects (entities) that compose the semantic structure of the text, SRL facilitates a deeper comprehension of the narrative within CVE descriptions. For example, in a CVE description like "An attacker could send a specially crafted email to exploit a vulnerability in the email client, potentially allowing unauthorized access to user data," SRL would identify "send" and "exploit" as actions, "attacker" as the agent, and "specially crafted email" and "email client" as entities. This level of analysis facilitates a deeper comprehension of the narrative within CVE descriptions and serves as the foundation for mapping these descriptions to the ATT&CK framework.

The steps of our method are shown in 5.1, which is called the Semantic Mapping from CVE to ATT&CK Technique (SMET) process. This figure shows how plain text from CVE reports is turned into a list of ATT&CK techniques, ordered by importance.

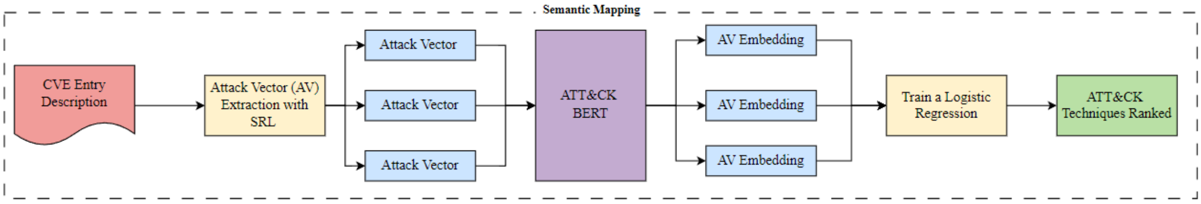


Figure 5.1: SMET - CVE to ATT&CK Technique

SRL is important in enhancing the understanding of CVE entries, as it systematically breaks down the text into manageable, interpretable components. In the context of cybersecurity, SRL is instrumental because it extracts and labels the actions and entities within a sentence, aligning them to the roles they play in cybersecurity incidents. This identification helps in mapping the narrative of CVE reports to actionable intelligence.

The ATT&CK BERT model merges the capabilities of BERT with the domain-specific intelligence of the MITRE ATT&CK framework. Initially, raw text is broken down into manageable pieces called tokens, which can be words, phrases, or parts of words. This text is also cleaned and preprocessed, as described in Section 5.1. Each token is then processed through BERT, which transforms it into a high-dimensional vector. Thanks to its training on a large amount of text, BERT comprehends the contextual use of each word. This allows the same word to yield different vectors based on its surrounding words. After processing through BERT, the vectors produced for each token are aggregated into a single comprehensive vector for the entire text snippet or document. Upon processing, ATT&CK BERT descriptions of discrete attack vectors, such as "attacker sends specially crafted email" and "attacker exploits email client vulnerability", into a numerical vector space. For instance, the vector representation for the first scenario might appear as [0.85, -0.23, 0.45, 0.90, -0.12], and for the second as [0.15, 0.99, -0.30, 0.80, 0.20]. Each dimension of these vectors captures different aspects of the cybersecurity context related to the attack (Abdeen, 2023).

These vectorized representations are then leveraged by the LR classifier, trained to discern the probabilities of linkage between CVE descriptions and specific ATT&CK techniques. The probabilistic modeling, fed with the structured data derived from SRL and interpreted through ATT&CK BERT, provides outputs on the probabilities of various ATT&CK techniques being relevant. For example, there might be an 85% probability that the technique "Spear Phishing" is associated with the action of sending a specially crafted email, and a 90% probability that "Exploitation for Client Execution" corresponds to the exploitation of the email client vulnerability.

Despite the valuable insights garnered from SRL and its application in SMET, this method primarily dissects the structural elements of language within CVE descriptions. While it effectively aligns CVE details with corresponding ATT&CK techniques, it primarily focuses on linguistic structure, which may lead to the oversight of deeper, contextual nuances specific to cybersecurity terminology and interactions.

5.3.2 Named Entity Recognition using SecureBERT

For categorizing entities within CVE descriptions, NER using SecureBERT is implemented, a pre-trained version of the BERT model tailored for cybersecurity contexts. The proposed model is shown in Figure 5.2. NER is important for identifying specific entities within text, such as software names, version numbers, and cybersecurity-related terms within CVE descriptions. This method can provide a more nuanced understanding and categorization of the information in these descriptions, which is essential for the mapping phase. Key methodologies employed for NER are rule-based approaches, supervised learning, NN-based approaches, and transformer-based models. This section focuses on a transformer-based model that is pre-trained and fine-tuned on NER-specific data, this model is called CyNER. CyNER is a Python library designed for cybersecurity NER. CyNER combines transformer-based models for extracting cybersecurity-related entities (Alam et al., 2022). Their ability to handle long-range dependencies and capture nuanced semantic relationships makes them popular for NER. These models show significant improvements in recognizing and classifying named entities with high Precision and Recall (Lothritz et al., 2020).

The pre-processing explained in Section 5.1 is necessary to achieve accurate results in NER. By fine-tuning BERT models on NER specific data they are able to adapt to the task and achieve accurate performance in NER. As mentioned in Section 3.1, there exists several LLMs trained

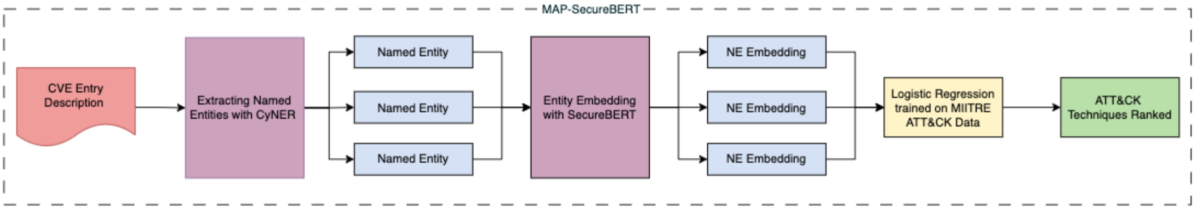


Figure 5.2: MAP-SecureBERT - CVE to ATT&CK Technique

on cyber specific datasets, besides this they are trained for NER purposes as well. Because of the selected cyber specific training data for SecureBERT and it capabilities to perform NER, it is selected as model to apply in this research. By leveraging SecureBERT, we enhance our methodology's ability to detect and categorize important information from CVE descriptions.

The method used for mapping a CVE to the ATT&CK techniques is described in Figure 5.2. After the data pre-processing, CyNER is employed for NER. This involves identifying and extracting cybersecurity-related entities present in the CVE descriptions. For example, consider a CVE description that states, “The vulnerability in Apache Struts version 2.5 allows remote attackers to execute arbitrary code via a crafted URL.” In this scenario, CyNER would identify “Apache Struts” as the software, “2.5” as the version number, “remote attackers” as the threat agent, and “crafted URL” as the attack vector. These identified entities are crucial for mapping the CVE to specific ATT&CK techniques.

Once these entities are recognized, SecureBERT is used again to generate entity embeddings. These high-dimensional vectors capture the essence of the extracted entities and are meticulously designed to encapsulate the relational nuances of these entities, turning textual data into a format ready for machine processing and analysis. For instance, the embedding for “Apache Struts” might encode information about common vulnerabilities and exploits associated with this software, while the embedding for “crafted URL” might highlight methods of delivery and exploitation in cyber attacks. These vectors then serve as input to a LR model, which has been previously trained on a labeled dataset derived from the ATT&CK Matrix. The model evaluates these vectors and calculates a probability score for each ATT&CK technique, effectively ranking them based on these scores. The top-ranking techniques, such as “Remote Code Execution” for exploiting “Apache Struts,” are presumed to be the most relevant to the CVE entry, guiding cybersecurity professionals in prioritizing their defensive strategies.

While the use of CyNER and SecureBERT for NER effectively identifies and categorizes specific elements within descriptions of cybersecurity vulnerabilities, this technique primarily focuses on extracting isolated pieces of information. This consideration highlights the need to evaluate the effectiveness of NER against SRL in cybersecurity contexts, given the importance of understanding relationships and data context. Such relationships can be important for a deeper understanding of the narratives within CVE reports. This comparison prompts us to consider whether the addition of SRL could complement the entity recognition capabilities of SecureBERT by providing a more contextual understanding of the text, which is important for cybersecurity threat analysis. This leads to a detailed view of data points, which could benefit from additional analytical depth provided by another model, further described in section 5.3.3.

5.3.3 Multi-Input Cyber Security Model (MICS)

The Multi-Input Cyber Security Model (MICS), developed as part of this thesis, represents a significant advancement over existing approaches. It integrates a complex set of features including embeddings, SRL data, and cosine similarities. This innovative combination enhances the model's ability to analyze and predict cyber threats with greater performance. The capabilities and performance of MICS are illustrated in Figure 5.3.

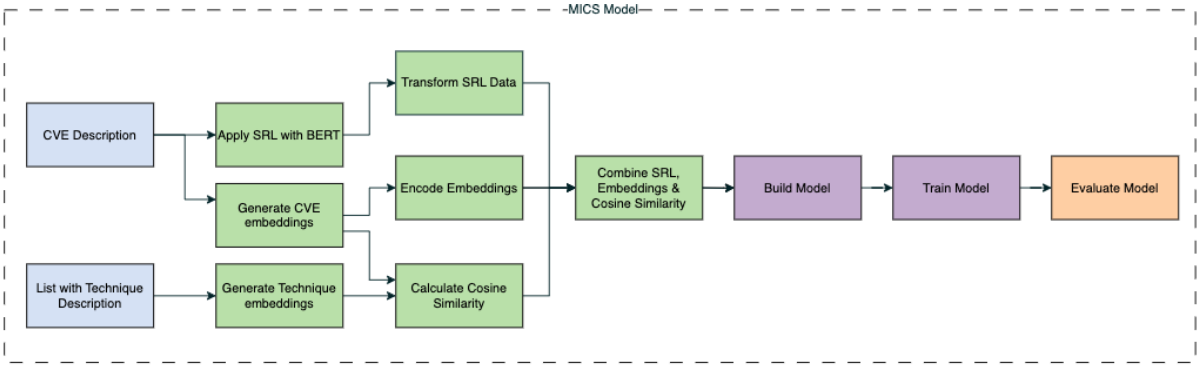


Figure 5.3: MICS Model Development

The MICS model begins by encoding CVE descriptions and technique descriptions using the SentenceTransformer, producing dense vector representations. These embeddings aim to capture the deeper semantic meanings within texts, important for understanding nuanced cybersecurity data. To enhance the analysis further, the model employs cosine similarities between these vector representations. Cosine similarity is calculated using the formula:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where $\mathbf{A}$ and $\mathbf{B}$ are vector representations of text data, $\mathbf{A} \cdot \mathbf{B}$ is the dot product of vectors $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the norms (magnitudes) of these vectors. This measure focuses on vector orientation over magnitude, accurately quantifying the semantic closeness between various CVE entries and cybersecurity techniques. This metric is particularly useful in text analysis within cybersecurity, where the semantic alignment of terms often holds more significance than their frequency or occurrence (Han et al., 2012). Cosine similarity, by assessing how vectors point in relation to each other, effectively captures this alignment, making it superior to measures like Euclidean distance which might emphasize volume or absolute differences in term usage. By utilizing cosine similarities, MICS is able to better match vulnerabilities with corresponding cybersecurity techniques based on their content, overcoming previous models' limitations that might not account for the true closeness between terms described differently.

The NN architecture of MICS, shown in Figure 5.4, adeptly combines these diverse inputs. The model features several dense layers with ReLU activation and dropout layers to prevent overfitting, effectively learning from the rich, combined data inputs. The input layer takes the combined feature set consisting of CVE embeddings, technique embeddings, and SRL features alongside the calculated cosine similarities. This combination allows the model to make informed predictions across multiple labels, reflecting the complex nature of cybersecurity threats where

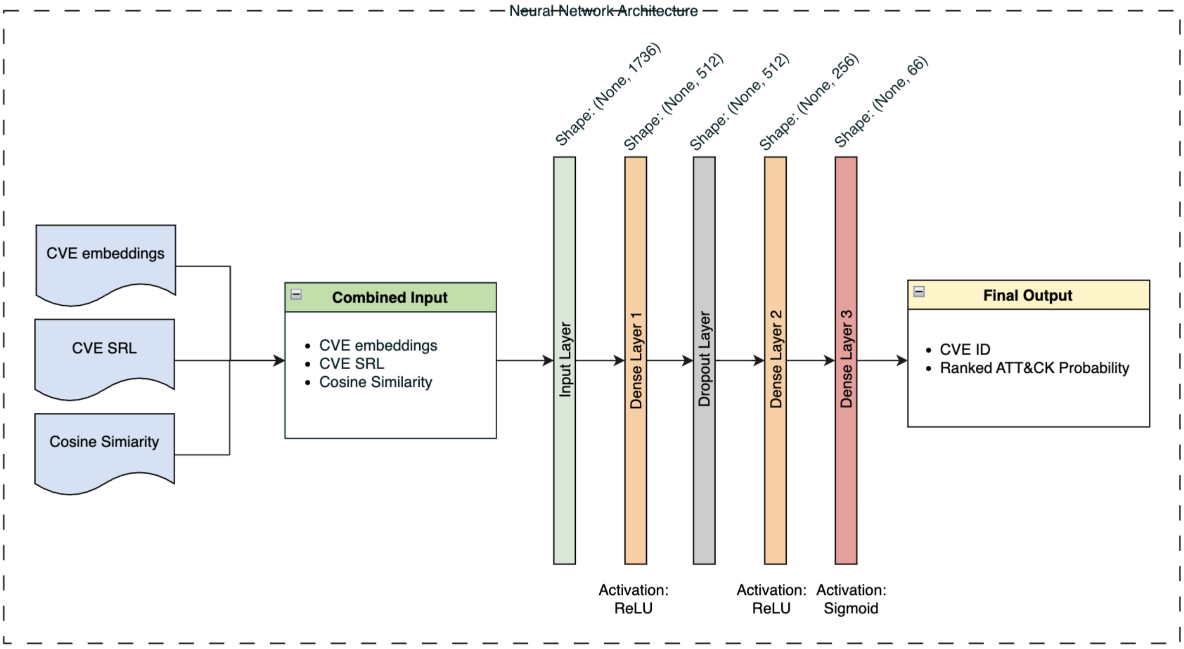


Figure 5.4: Neural Network Architecture

multiple techniques may apply to a single vulnerability. Training involves splitting the data into training and test sets, using cross-validation to ensure the model generalizes well across different data subsets. The model employs the Adam optimizer and binary crossentropy for loss computation, suitable for the multi-label classification tasks at hand.

The introduction of cosine similarities in MICS is innovative, quantifying the semantic distances between CVE entries and predefined ATT&CK techniques. This not only overcomes the limitations of previous models that did not account for the degree of closeness between data points but also enables the model to use technique descriptions effectively.

5.4 Performance Testing and Evaluation

This segment of the chapter describes the metrics used to assess the model's performance, providing an understanding of its predictive power and practical utility.

5.4.1 Training, Testing and Evaluation

The methodology employed in this study aims to bridge the informational divide between two different datasets: the CVE database and the MITRE ATT&CK database. Given the absence of a large, fully annotated dataset linking CVE entries to ATT&CK techniques, we navigate this challenge as explained in this section.

During the training phase, the SecureBERT LLM is employed to identify entities within the CVE database. This extraction process gathers a wide range of named entities essential to cybersecurity from the textual data of CVE descriptions. These entities are then transformed into vector embeddings via SecureBERT. Simultaneously, embeddings for ATT&CK techniques are generated from the ATT&CK database. This multi-input appraoch facilitates the development of a semantically informed training environment for the LR model. By leveraging the semantic properties captured in the embeddings and the contextual cybersecurity knowledge they contain, the model is trained to detect associations between CVE entities and ATT&CK techniques through semantic similarity. Previous research in this domain has often relied on smaller, expert-curated datasets where CVE entries were manually mapped to corresponding ATT&CK techniques. To enhance the data, we incorporate additional CVE entries, which are manually linked to the MITRE ATT&CK database, thereby broadening the scope of our evaluation dataset.

Given the constrained size of our test dataset, the application of cross-validation techniques is important. Cross-validation divides the dataset into k subsets, sequentially engaging k-1 subsets in model training while employing the remaining subset for testing. This iterative process optimizes data usage and enhances the robustness of our evaluation by mitigating overfitting risks. The performance of our models will be evaluated using metrics designed to capture the Precision of mappings and the applicability of the models in real-world scenarios

5.4.2 Evaluation Metrics

The task inherently involves a multi-label classification problem, where each CVE entry might map to multiple ATT&CK techniques based on its description. Traditional metrics such as Precision, Recall, and F1 score, typically used for evaluating binary classification models, are less representative and sometimes misleading for multi-label tasks. This is due to their design for scenarios where each instance is associated with a single label. In multi-label classification, each instance may associate with multiple labels, introducing complexities these metrics do not accommodate. Because of this we use adjusted versions of these metrics.

- **Precision@K**: Measures the proportion of true labels among the top K predictions, where n is the total number of instances, Y_i is the set of true labels for the i-th instance, Y_i' is the set of predicted labels, and K is the number of top predictions considered.

$$\text{Precision@K} = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap \text{top K predictions in } Y_i'|}{K}$$

- **Recall@K**: Evaluates how well the model captures the relevant labels within its top K predictions. This metric reflects the sensitivity of the model by measuring the proportion of true labels Y_i that appear among the top K predictions Y_i' for each instance, averaged over all n instances.

$$\text{Recall@K} = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap \text{top K predictions in } Y_i'|}{|Y_i|}$$

- **F1@K**: The harmonic mean of Precision@K and Recall@K, providing a balanced measure of the model's performance. It combines the assessments of both precision and recall for the top K predictions into a single metric.

$$\text{F1@K} = 2 \cdot \frac{\text{Precision@K} \cdot \text{Recall@K}}{\text{Precision@K} + \text{Recall@K}}$$

To evaluate the overall ranking and error rates of the model's predictions and make sure they are not limited to the top K predictions. The metrics described below provide a broader view of the model's performance across all predictions.

- **Label Ranking Average Precision (LRAP)**: Assesses the average precision with respect to label ranking across all labels. It calculates the average precision by considering the order in which true labels j from Y_i are predicted among all labels, where higher rankings of correct labels indicate better performance (Abdeen, 2023).

$$\text{LRAP} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\sum_{j\in Y_i}\frac{|\{\text{correct labels ranked above } j\}| + 1}{\text{rank of } j}$$

- **Coverage Error**: Calculates the average number of top-ranked predictions needed to cover all true labels for an instance. This metric evaluates the depth of predictions required to ensure that no relevant labels are missed, with lower values indicating better performance (Alvarez & VanBeselaere, 2005).

$$\text{Coverage Error} = \frac{1}{n}\sum_{i=1}^{n}\max(\text{ranks of true labels in } Y_i)$$

- **Ranking Loss**: Measures the average number of incorrectly ordered label pairs per instance, considering the ranks of relevant j and irrelevant k labels. Lower values indicate that the model is more effective at ranking relevant labels higher than irrelevant ones.

$$\text{Ranking Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\{(j,k) : j \in Y_i, k \notin Y_i, \text{rank}(k) < \text{rank}(j)\}|}{|Y_i| \times |\overline{Y_i}|}$$

- **Hamming Loss**: Reflects the fraction of labels that are incorrectly predicted, normalized over the total number of labels L. It provides an overall measure of the error rate across all labels and predictions. (Ganda & Buch, 2018).

$$\text{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \Delta Y_i'|}{L}$$

Examining the confusion matrix for each class is important in multi-label classification because it offers detailed understanding of the model's performance on each specific label. This helps in identifying specific labels that the model struggles with, which can inform targeted improvements. The confusion matrix components for each class can be defined as follows:

- **True Positives ($\mathbf{TP}_c$)**: This counts how many times class c was correctly identified among the top K predictions (TopK_i) for the instances where c is indeed a true label (Y_i). The indicator function $\mathbb{I}$ outputs 1 if c is both a true label and predicted within the top K.

$$\text{TP}_c = \sum_{i=1}^{n}\mathbb{I}(c \in Y_i \cap \text{TopK}_i)$$

- **False Positives ($\mathbf{FP}_c$)**: Measures the occurrences where class c was predicted within the top K predictions, but it was not a true label for that instance. This metric highlights the over-prediction or mislabeling of class c.

$$\mathrm{FP}_c = \sum_{i=1}^{n} \mathbb{I}(c \notin Y_i \cap c \in \mathrm{TopK}_i)$$

- **False Negatives ($\mathbf{FN}_c$)**: Indicates how often class c, while being a true label for an instance, fails to appear in the top K predictions. This metric is crucial for understanding under-predictions or misses.

$$\mathrm{FN}_c = \sum_{i=1}^{n} \mathbb{I}(c \in Y_i \cap c \notin \mathrm{TopK}_i)$$

- **True Negatives ($\mathbf{TN}_c$)**: Counts the instances where class c is correctly identified as not being applicable, both as a true label and within the top K predictions. This helps assess the model's ability to accurately exclude irrelevant classes.

$$\mathrm{TN}_c = \sum_{i=1}^{n} \mathbb{I}(c \notin Y_i \cap c \notin \mathrm{TopK}_i)$$

By using these metrics and confusion matrix components, it is possible to better evaluate the model's ability to prioritize relevant ATT&CK techniques for each CVE entry and minimize errors in label assignments. This evaluation framework helps the understanding and mitigation of cybersecurity threats by providing a comprehensive view of model performance in multi-label classification tasks.

5.5 Integration of Responsible AI in the Model

In this methodology section, we explain how Responsible AI principles are integrated within the model, aligning with EY's AI Framework detailed in subsection 3.2.1.

The integration of Accountability, Sustainability, Transparency, and Privacy will not be coded. Accountability is maintained through rigorous documentation and version control that tracks the development life cycle, ensuring clarity in ownership and responsibility for actions taken by the AI system. Sustainability is promoted through strategic planning that encompasses long-term operational viability, focusing on practices that minimize environmental impact. Transparency is achieved by maintaining open lines of communication about the AI system's capabilities and limitations, supported by detailed user documentation and transparent reporting. Lastly, Privacy is maintained through stringent data governance policies that respect user consent, secure data handling, and compliance with strict privacy laws and regulations. In this case, which involves open-source datasets, privacy concerns are minimal because they do not contain personal data.

The principles of Fairness, Explainability and Reliability can be incorporated within the model by testing them with certain metrics. For Fairness the metrics Equalized Odds and Predictive Parity do not apply, because the dataset does not include subgroups. But even without the existence of subgroups, bias can still exist in the way CVEs are categorized. When using NLP techniques to convert text into embeddings, the Word Embedding Associate Test (WEAT) is a method that can be employed to quantify bias by measuring the cosine similarity between word vectors in the embeddings pace.

$$s(X, Y, A, B) = \sum_{a \in A} (\text{mean}_{x \in X} \cos(\mathbf{a}, \mathbf{x}) - \text{mean}_{y \in Y} \cos(\mathbf{a}, \mathbf{y}))$$

Here, X and Y are sets of word vectors associated with different groups, and A and B are vectors representing the target concepts or attributes. This formula helps identify if certain CVE-related terms are more closely associated with specific attack techniques than others.

Explainability within the model is detailed by how decisions are made through both local and global explanation techniques. On a local level, techniques such as feature importances and Shapley values can be employed. The Shapley value, measures the contribution of each feature to the prediction of a particular instance. The formula for calculating the Shapley value for a feature i is given by:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

where N is the set of all features, S is a subset of features excluding i, $v(S)$ is the prediction model's output with only the features in set S, and $v(S \cup \{i\})$ is the output when feature i is added to S. The difference $v(S \cup \{i\}) - v(S)$ indicates the marginal contribution of feature i when combined with features in S. Globally, techniques such as model probing and neuron activation analysis provide insights into the model's overall functioning and response patterns to various inputs. These methods help to assess the general behavior of the model. By implementing both local and global explanatory techniques, a deeper understanding is gained of how decisions are made within the model, thus enhancing transparency and accountability.

Reliability is ensured by implementing confusion matrices to monitor the model's performance, ensuring it consistently identifies and classifies threats with high accuracy and minimal errors. How a confusion matrix is constructed is described in Section 5.4.2

By integrating these Responsible AI principles, the model not only aligns with EY's Responsible AI Framework but also enhances its capability to deliver precise, understandable, and fair outcomes in real-world applications, ensuring that all actions and decisions made by the AI are well-documented and ethical.

Chapter 6

Experimental Results and Discussion

This chapter presents a comparative analysis of multiple predictive models to assess their effectiveness in identifying and categorizing cybersecurity vulnerabilities from CVE descriptions to MITRE ATT&CK techniques. The models tested include a baseline model employing simpler NLP techniques, the semantic-enhanced SMET model, the entity-focused MAP-SecureBERT model, and the MICS model, which integrates advanced NLP and deep learning techniques. These models were chosen to cover a broad spectrum of approaches from basic to advanced analytics in cybersecurity text processing. The analysis is structured to not only quantify the models' performance in terms of relevant metrics, described in 5.4.2, but also to interpret these results within the context of EY's responsible AI framework, as discussed in Section 3.2.1.

6.1 Exploratory Data Analysis

To understand the characteristics of the dataset used in this thesis, an exploratory data analysis was conducted. An examination of the textual CVE description data was conducted to understand the underlying themes and to take certain information for making modelling decisions. Initial analysis, shown in Figure 6.1, shows a large number of stopwords alongside domain-specific terms such as 'vulnerability', 'attackers' and 'Windows'.

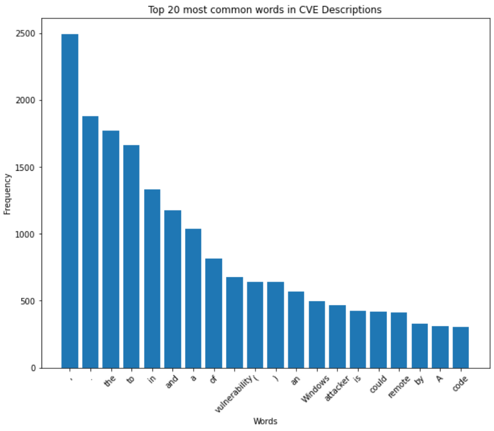


Figure 6.1: Top 20 most common words in CVE Descriptions

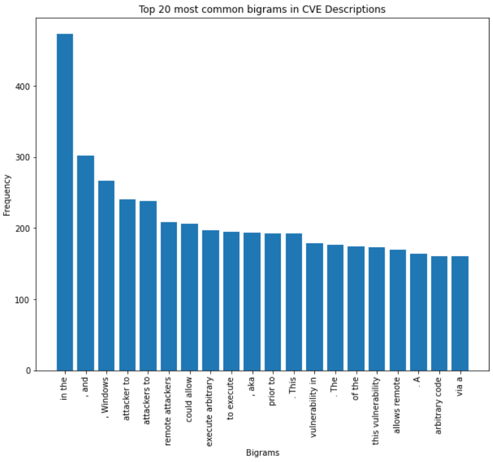


Figure 6.2: Top 20 most common bigrams in CVE Descriptions

Within Figure 6.2, a bi-gram frequency plot is shown. This shows the frequency distribution of every bi-gram, which are strings of words. This graph shows the importance of extracting informative bi-grams like ‘remote attacks’ and ‘execute code’, which underscore typical cyber-security concerns. These observations show the necessity of text pre-processing techniques, including stop-word removal and extracting relevant bi-grams. Especially the relevance of verbs seems important and should be taken into account when designing the final model.

Besides looking into the CVE description, it is important to look at the distribution of techniques in the small and large dataset. Within the figures 6.3 and 6.4 below the technique distribution for both datasets is shown. These figures reveal an imbalance in the distribution of techniques, with ‘Exploitation for Client Execution’ and ‘Exploit Public Facing Application’ being the largest classes in both datasets. To address this imbalance, the application of the SMOTE could be considered to enhance model performance by better representing underrepresented techniques.

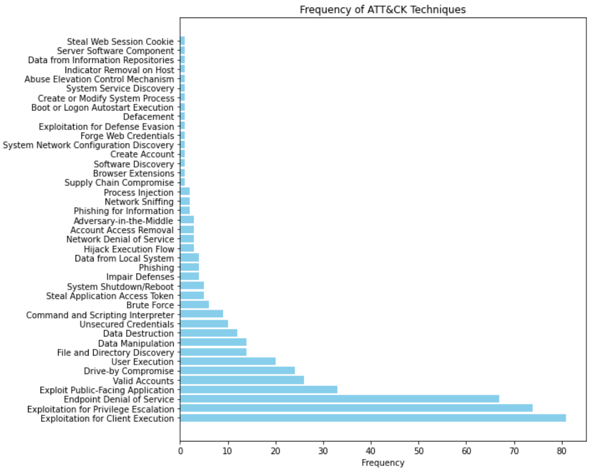


Figure 6.3: ATT&CK Frequency Small Dataset

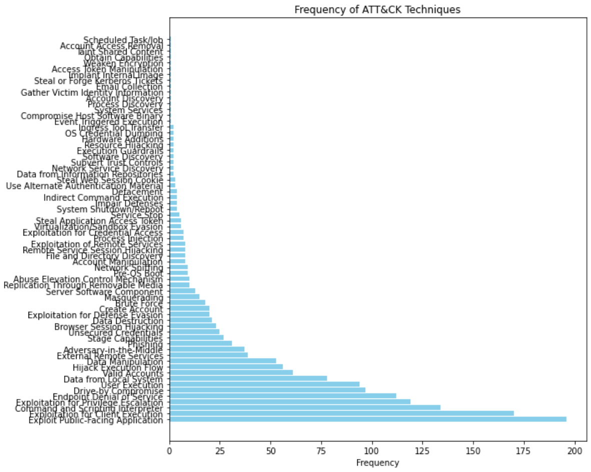


Figure 6.4: ATT&CK Frequency Large Dataset

The models that are being discussed in this chapter will show us the ability to correctly predict the technique classes per CVE description. Because this problem is a multi-label classification problem, multiple techniques can be assigned per CVE entry. The average number of techniques per CVE within the small dataset are 1.47, while the average number of techniques per CVE within the large dataset are 2.00. This is important to have an idea what is more or less the optimal solution for certain metrics, like coverage error.

6.2 Baseline Model Performance

To compare the models, a baseline model was constructed following the methodology outlined in Section 5.2. The baseline model employs simpler NLP methods, including BoW and TF-IDF, along with a RF classifier for mapping vulnerabilities to ATT&CK techniques. Table 6.1 presents the performance metrics for the baseline model using BoW and TF-IDF.

The baseline model exhibits moderate performance across various evaluation metrics. While the F1 Score, Precision, and Recall metrics indicate a balanced performance in terms of false positives and false negatives, they do not reach high performance levels. For evaluation purposes it is chosen to look at the top 5 ranked predictions for the evaluation metrics. The model coverage, averaging around 6 techniques predicted per vulnerability for the small dataset, suggests that while it captures a reasonable amount of relevant information, some details may be overlooked. This coverage increases notably in the larger dataset to approximately 12 techniques, reflecting the broader diversity and increased examples which may enable the model to capture a wider array of techniques.

Metrics	**Small DS**		**Large DS**	
	BoW	**TF-IDF**	**BoW**	**TF-IDF**
Model Coverage	6.049	**6.115**	12.374	**11.859**
LRAP	0.646	**0.659**	0.580	**0.569**
Ranking Loss	0.109	0.101	0.091	0.093
Hamming Loss	**0.122**	0.122	0.067	**0.069**
Precision@5	0.203	**0.203**	0.258	**0.248**
Recall@5	0.710	0.702	0.744	0.711
F1@5	0.316	0.315	0.383	0.368

Table 6.1: Performance Metrics BoW & TF-IDF. This table presents a detailed comparison of multilabel-specific metrics for the baseline model using Bag of Words and TF-IDF methodologies across two different dataset sizes (small and large).

Furthermore, the Ranking Loss and LRAP metrics demonstrate a moderate ability to rank techniques accurately, with the larger dataset showing a slight improvement in Ranking Loss and a decrease in LRAP. This could be due to the larger dataset providing more instances of less frequent techniques, which helps in better modeling the relationships between techniques but also introduces more challenges in consistently ranking highly relevant techniques. Although the Recall@5 metric is relatively high in both datasets, it is a bit higher in the larger dataset, indicating significant relevant information captured within the top predictions and suggesting that the model benefits from the increased data volume in terms of Recall capability.

Despite these insights, both BoW and TF-IDF representations displayed similar trends across the datasets. Nonetheless, the TF-IDF model exhibited a slightly superior performance in the larger dataset, particularly in handling diverse data, leading to its selection as the final baseline model. In conclusion, while the baseline model shows adequate performance, there is room for improvement in accuracy and coverage to better meet the demands of larger datasets. This evaluation not only provides a benchmark for comparing the performance of other models but also highlights areas for potential enhancements to address the complexities introduced by expanding data volumes.

6.3 Advanced Model Performance

This chapter describes the performance of advanced models developed to improve predictive performance. It builds upon the baseline models and explores the capabilities of the SMET, MAP-SecureBert and MICS models. While SMET will show the impact of SRL, Map-SecureBert will show the impact of NER on the performance metrics. Besides to the different implementations of NLP techniques, the MICS model will show the impact of including cosine similarities of a CVE description to all possible MITRE ATT&CK techniques on predicting the correct labels for each entry.

6.3.1 SMET Performance

In comparing the SMET and baseline models, the SMET model shows substantial improvements across various performance metrics. It achieves a lower model coverage within the small dataset, of 5.013 versus the baseline's 6.115, suggesting better capture of relevant information (Table 6.2).

The SMET model, while demonstrating strengths in semantic modeling, shows mixed performance compared to the baseline model across different dataset sizes. On the small dataset, the SMET model exhibits a Precision of 0.234, which is slightly higher than the baseline's Precision that is equal to 0.203. Furthermore, its Recall at 0.826 is higher compared to the baseline's 0.702, leading to a F1 score of 0.372.

Metrics	**Small DS**		**Large DS**	
	SMET	**TF-IDF**	**SMET**	**TF-IDF**
Model Coverage	**5.013**	**6.115**	**19.707**	**11.859**
LRAP	0.623	0.659	0.336	0.569
Ranking Loss	0.068	0.101	0.188	0.093
Hamming Loss	0.035	0.122	0.030	0.069
Precision@5	**0.234**	**0.203**	**0.146**	**0.248**
Recall@5	**0.826**	**0.702**	**0.423**	**0.711**
F1@5	0.372	0.315	0.217	0.368

Table 6.2: Performance Comparison of SMET and TF-IDF. This table provides a comprehensive evaluation of the SMET model's performance relative to the baseline model's TF-IDF technique, across both small and large datasets.

Upon applying the SMET model to the larger dataset, expectations for similar enhancements were not met. Contrary to the small dataset, the Precision drops to 0.146, and the Recall decreases to 0.423. These metrics illustrate a reduction in the model's ability to identify relevant instances accurately as the dataset size increases, which contrasts with the baseline model that maintains relatively stable Recall and Precision metrics across dataset sizes.

These observations suggest that while the SMET model introduces advanced semantic understanding and effective predictive performance in smaller or more controlled environments, its scalability and adaptability to larger, more complex datasets need further refinement. The

Section below explores potential areas for model improvement. This analysis is important to understand where the SMET model may under-perform and how it can be adjusted for better scalability and robustness.

SMET Performance Analysis

To analyze the in which cases the SMET model on the large and small dataset does not perform correct, a confusion matrix per class is extracted. This confusion matrices are shown in the Appendix B.1. The SMET model, which uses LR with simple labels, has several limitations. LR is a linear model that may not capture the complex, non-linear relationships present in the data. This can lead to misclassifications, as evidenced by the confusion matrix. For example, in the confusion matrix, shown in Appendix B.1. For "Process Injection," a high number of false negatives (60) and false positives (0) can be seen, indicating that the model struggles to correctly identify instances of this technique. Similarly, the "Exploit Public-Facing Application" category shows a high number of false positives (101), suggesting that the model incorrectly predicts this technique when it is not present. The confusion matrices reveal misclassification patterns, which are further visualized in the heatmap in Figure B.2. There is confusion between conceptually similar labels such as "Exploitation for Privilege Escalation," "Exploitation for Client Execution," and "Exploitation for Defense Evasion." Additionally, "Endpoint Denial of Service" is often confused with "Exploitation for Privilege Escalation," while "Data Manipulation" is frequently mistaken for "Data from Local System" and "Data from Information Repositories."

The heatmap for the large dataset, shown in Appendix Figure B.4, reveals similar patterns. "Exploitation for Client Execution" is the most frequently misclassified label, often confused with "Exploit Public-Facing Application," "Command and Scripting Interpreter," and "Endpoint Denial of Service." misclassifications among various exploitation techniques suggest insufficiently distinct features due to their conceptual similarities. Labels such as "Compromise Host Software Binary" and "Implant Internal Image" show minimal false positives, indicating these might be underrepresented or more distinct in the dataset.

These trends highlight areas for improvement, such as enhancing feature engineering to better capture the unique characteristics of each label, employing SMOTE to balance the dataset, and exploring more complex models to improve classification accuracy. Given these limitations, we explore more advanced models to better handle the complexities of cybersecurity data. One such approach is the MAP-SecureBERT model, which integrates named entities SecureBERT embeddings to analyze CVE descriptions. The second approach is a multi-input model that calculates the cosine similarity for each CVE entry to all the possible MITRE ATT&CK techniques and uses the technique embeddings and cosine similarities as input besides the CVE SRL and CVE embeddings.

6.3.2 MAP-SecureBERT Performance

As outlined in Section 5.3.2, the Map-SecureBERT model integrates NER and SecureBERT embeddings to analyze CVE descriptions. As a first step, the model employs NER to identify and extract relevant named entities within the text. After that, these entities are processed through SecureBERT to generate embeddings that capture their contextual significance. These embeddings are then used as inputs for a LR model specifically trained to predict the corresponding MITRE ATT&CK techniques. The NER technique that is being used, is CyNER. As described in Section 5.3.2, CyNER is a python library designed enhance cybersecurity efforts through NER specifically tailored to cyber threat intelligence. NER tools like CyNER are designed to detect and categorize entities such as malware names, IP addresses, software vulnerabilities, and other cybersecurity-specific terms.

The performance metrics for the MAP-SecureBERT model, as outlined in Table 6.3, highlight challenges in the model's effectiveness in cybersecurity text analysis. This model is only performed on the small dataset due to the lack of performance. Specifically, the model exhibits a high ranking loss (0.497) compared to SMET and a low LRAP of 0.090, indicating difficulties in ranking relevant entities correctly. The Recall@5 is also low at 0.062, suggesting the model often fails to identify key entities among the top predictions, and a coverage error of 23.363 points to inefficiency, as the model must predict a large number of labels to ensure all relevant entities are captured. The model's Recall is zero, suggesting it fails to detect most relevant entities, leading to an F1 score of zero. This performance indicates a possible over-conservatism in entity prediction.

Metrics	**Small DS**	
	NER	**SMET**
Coverage Error	**23.363**	5.013
LRAP	**0.090**	0.623
Ranking Loss	**0.497**	0.068
Hamming Loss	0.036	0.035
Precision@5	**0.020**	0.234
Recall@5	**0.062**	0.826
F1@5	0.030	0.372

Table 6.3: Performance Metrics Comparison for MAP-SecureBERT using CyNER versus SMET on Small Dataset: This table quantitatively compares the performance of the MAP-SecureBERT model, which integrates NER through the CyNER library, with the SMET model across various evaluation metrics on a small dataset.

NER is effective at pinpointing the pre-decided types of named entities, it does not inherently capture their roles or relationships within the text, which appear to be critical for understanding the narrative or functional context of cybersecurity threats and attacks. Semantic roles seem more important then the named entities for this classification problem. To make sure the performance of MAP-SecureBERT lacks performance due to NER instead of using SRL, a MAP-ATT&CK BERt model was created. This models works the same as MAP-SecureBERT but uses instead of SecureBERT, ATT&CK BERT as LLM which is trained on more specific data. As shown in Table 6.4, the model shows some enhancement compared to SecureBERT. Notably, ATT&CK BERT achieved a Ranking Loss of 0.379, indicating a more accurate prioritization of relevant entities. Additionally, the model showed an improvement in LRAP with a score of 0.210 and managed a Recall@5 of 0.274 compared to MAP-SecureBET. The Coverage Error was reduced to

18.690, further confirming the model's enhanced ability to cover critical entities without excessive over-prediction.

Metrics	**Small DS**	
	ATT&CK BERT	**SecureBERT**
Coverage Error	**18.690**	**23.363**
LRAP	**0.210**	**0.090**
Ranking Loss	0.379	0.497
Hamming Loss	0.036	0.036
Precision@5	0.079	0.020
Recall@5	**0.274**	**0.062**
F1@5	0.122	0.030

Table 6.4: CyNER with ATT&CK BERT & SecureBERT: This table compares the performance of the MAP-ATT&CK BERT model, which uses ATT&CK BERT embeddings, with the MAP-SecureBERT model employing SecureBERT embeddings, evaluated on a small dataset.

Despite these improvements, ATT&CK BERT still performs below the baseline model and the SMET approach, particularly in extracting and leveraging relevant textual relationships as SRL techniques do. Additionally, the impact of a different fine-tuned LLM and a separate NER model (SecBert) was evaluated. However, SecBert did not yield improved performance and has been omitted from this thesis for this reason. This reflects the challenge of adapting NER models to fully grasp and utilize the context in which entities operate within cybersecurity texts.

6.3.3 Multi-Input Cybersecurity Model (MICS) Performance

This Section presents the comparative analysis of the MICS and the SMET across two different dataset sizes: small and large as described in Section 4. In addition to testing different datasets, the study also explored the use of CNN, RNN and LSTM as alternatives to the designed NN, as discussed in Chapter 5. However, these methods were excluded from the results due to their lack of performance compared to the NN structure. The multi-label classification performance of MICS and SMET is summarized in Table 6.5

For the small dataset, MICS exhibits higher model coverage than SMET (32.459 compared to 5.013), suggesting that MICS can capture a broader spectrum of vulnerabilities. SMET outperforms MICS as well in LRAP, indicating better average Precision across the labels. In terms of Ranking loss and Hamming loss, SMET shows superior performance with lower values, implying more accurate label rankings and predictions, respectively. In the large dataset, the roles somewhat reverse. MICS shows lower model coverage (10.265) compared to SMET (19.707) and excels in LRAP (0.498 vs. 0.336), indicating a more precise handling of labels in larger datasets. In this scenario, the MICS model outperforms even the baseline model, described in 6.2

Metrics	**Small DS**		**Large DS**	
	MICS	**SMET**	**MICS**	**SMET**
Model Coverage	**32.459**	5.013	**10.265**	19.707
LRAP	**0.305**	0.623	**0.498**	0.336
Ranking Loss	0.294	0.068	0.084	0.188
Hamming Loss	0.009	0.035	0.031	0.030
Precision@5	0.085	0.234	0.230	0.146
Recall@5	**0.426**	0.826	**0.597**	0.423
F1@5	0.142	0.372	0.332	0.217

Table 6.5: Combined Comparison of MICS and SMET across Multilabel Metrics: This table provides a comparison of the MICS and the SMET model using various multilabel-specific metrics for both small and large datasets.

The performance of MICS and SMET on Precision, Recall, and F1 score at the top 5 predictions is detailed in Table 6.5 as well. For the small dataset, SMET shows a clear advantage in Precision, Recall, and F1 score, demonstrating a more effective top-5 prediction capability when handling smaller datasets. Conversely, in the large dataset, MICS outperforms SMET in both Precision and Recall, resulting in a higher F1 score (0.332 vs. 0.217). This suggests that MICS is better at handling the complexity and volume of larger datasets, particularly in identifying the most relevant top-5 predictions.

Implementing SMOTE in high-dimensional text data, such as Sentence-BERT embeddings, presents challenges. The high-dimensional space results in noisy synthetic samples that do not represent the true data distribution. Severe class imbalance, with many classes having few samples, renders SMOTE ineffective as it requires multiple samples to function. The complexity of semantic roles further complicates synthetic data generation, leading to poor model performance and overfitting. Practical constraints like computational resources and time also limit SMOTE's feasibility. Due to these issues, results using SMOTE are excluded from this thesis.

This research aims to improve the SMET model using a larger dataset, but the MICS model has shown superior performance. This highlights the importance of model selection based on dataset characteristics and underlying algorithms. MICS, utilizing a deep learning framework, integrates multiple data inputs like text embeddings, SRL data, and cosine similarities. Its NN architecture excels in capturing complex patterns in large datasets with non-linear relationships. In contrast, SMET uses LR, which assumes linear relationships and is more interpretable, making it effective for smaller datasets with simpler feature interactions.

6.4 Evaluation of Responsible AI Principles

This Section elaborates on the incorporation and evaluation of Responsible AI principles, fairness, explainability, and reliability, within our cybersecurity threat detection model. The methodology aligns with EY's AI Framework as referenced in subsection 3.2.1, focusing on rigorous assessment techniques and integrative approaches to uphold these principles.

To assess fairness within the model, the Word Embedding Association Test (WEAT) and t-SNE visualizations is used to analyze the word embeddings generated from cybersecurity related texts. By conducting WEAT, we evaluate potential biases in the embeddings that could unfairly associate CVE terms with specific attack techniques.

$$\text{Cross-Validation WEAT Scores} = \begin{bmatrix} -0.04353 \\ -0.05342 \\ -0.01397 \\ 0.00515 \\ 0.05559 \end{bmatrix} \tag{6.1}$$

Cross-validation of WEAT scores indicates minimal bias, as shown in vector 6.1, suggesting that our model does not generate unfair associations, which is essential for AI applications in cybersecurity. The t-SNE visualization further confirms these findings by displaying a distinct clustering of semantically similar terms and effective separation between CVE-related terms and attack techniques, as shown in Figure 6.5.

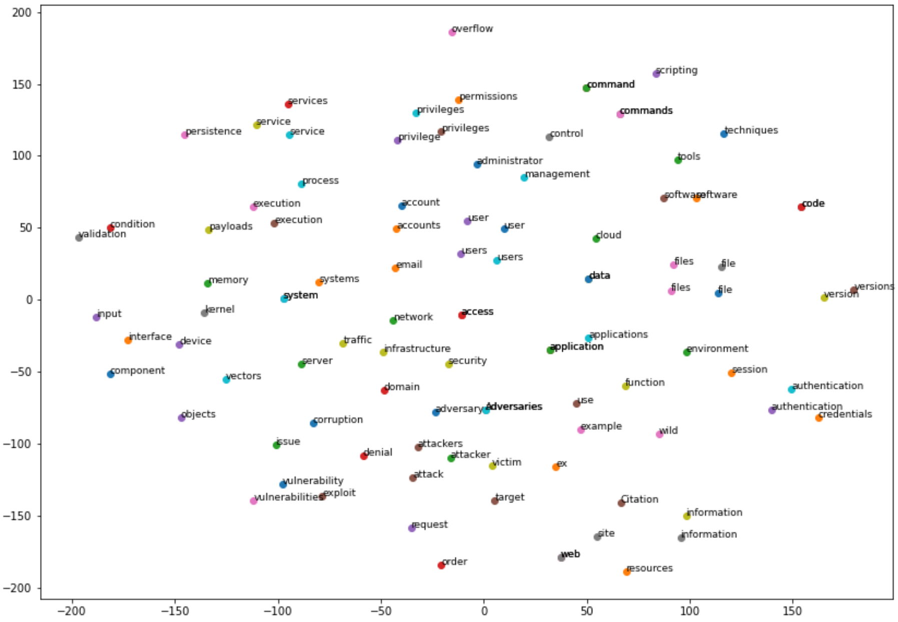


Figure 6.5: t-SNE Visualization of term embeddings. Each point represents a unique term, with clusters indicating related terms based on their proximity in the embedding space.

The SHAP values summary graph, shown in Figure 6.6, underscores the critical role of both SRL features and cosine similarity metrics in the model's predictions. This visualization highlights the individual contributions of features to the model's output, where each point on the plot represents the impact of a feature value on a specific prediction. SRL features like "SRL Feature 90" and "SRL Feature 69" appear at the top of the graph, indicating large influence. Similarly, various cosine similarity metrics, essential for measuring semantic alignments between CVE descriptions and ATT&CK technique descriptions, demonstrate substantial impacts.

This visualization supports the principles of explainability, by making the model's decision-making process transparent. The notable impact of both cosine similarity and SRL features within

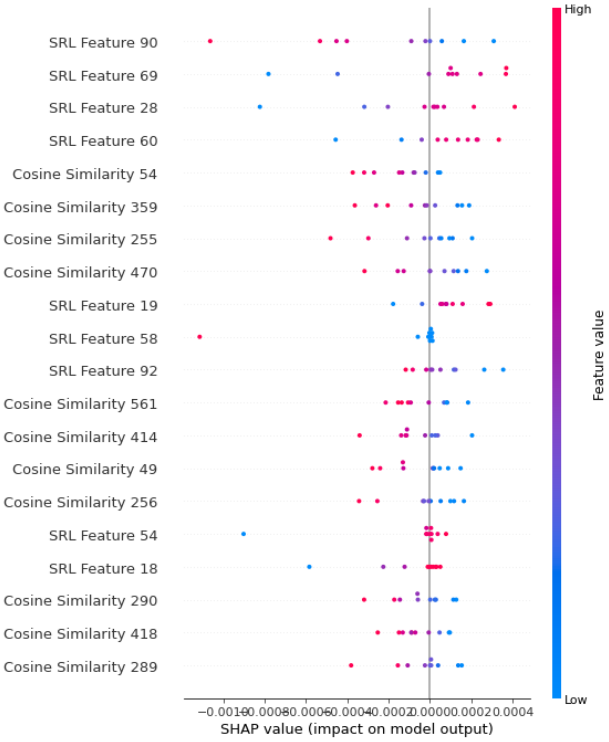


Figure 6.6: SHAP Value Summary for the MICS Model. Each dot on the plot represents the SHAP value for a feature for an individual prediction, where the color indicates the feature's value: red dots signify higher feature values and blue dots signify lower feature values. The position on the x-axis reflects the impact of the feature on the model's output. This plot provides insight into how different feature values influence the model's predictions across the dataset.

the SHAP values indicates that the model's ability to accurately interpret and contextualize cybersecurity threats relies significantly on these features, thereby enhancing its predictive accuracy. However, due to the computational intensity of calculating these values, they are derived only from a small subset of data. While SHAP values provide detailed insights into specific features, Permutation Feature Importance (PFI) offers a more scalable approach for assessing which features are most critical across the entire model.

PFI analysis across metrics such as coverage error, LRAP, precision at k, and rank loss reinforces the importance of features identified by SHAP analysis. "Cosine Similarity 221" and "Cosine Similarity 93", which appear notable in coverage error and rank loss metrics, are pivotal in minimizing predictive errors and ensuring rank integrity. The integration of SRL and CVE features in specific importance metrics like LRAP and precision at k also provides insights into enhancing model robustness and interpretability. For example, the prominence of "SRL Feature 18" and "CVE Embedding 631" in the LRAP metric suggests their potential contributions to model stability and explanatory power.

By integrating WEAT, t-SNE visualization, SHAP values, and PFI metrics, the MICS model robustly adheres to responsible AI principles. These analyses ensure that the model is fair, interpretable, and reliable, supporting its deployment in critical real-world applications and alignment with the Responsible AI framework constructed by EY.

Chapter 7

Conclusion

This thesis explores the integration of four NLP models to enhance cyber vulnerability mapping to attack techniques within a Responsible AI framework. The primary objective was to map CVE to MITRE ATT&CK techniques, thereby improving the mitigation of cyber threats. The study focused on evaluating the performance of three main NLP models: SMET, MAP-SecureBERT, and the MICS Model, comparing their effectiveness against the baseline model.

While the baseline model does not perform optimal, it is showing quite average results already. The baseline model is easy to construct and not computational intensive. The SMET model demonstrates improvements over the baseline model in the small dataset, achieving higher Precision, Model Coverage, and LRAP while maintaining a lower Ranking Loss. However, it struggles with scalability and adaptability to larger datasets, showing a substantial drop in Recall and a high increase in coverage error. In contrast, the MAP-SecureBERT model, which uses SecureBERT and NER techniques, faces challenges in correctly identifying and ranking relevant entities. The instances that SMET predicts incorrectly are often linked to techniques that are closely related to the correct ones. This shows the importance of extracting semantic relations within the text instead of just named entities and the importance of including technique descriptions make it easier to distinguish between techniques.

The newly constructed model, known as the Multi-Input Cyber Security Model (MICS), outperforms the baseline and individual models for the large dataset, by integrating multiple inputs and leveraging the strengths of various NLP techniques. MICS uses SentenceTransformer to generate dense vector representations of CVE descriptions and technique descriptions, allowing for a more nuanced understanding of the relationships between vulnerabilities and attack techniques. Besides the inclusion of structured SRL data, the inclusion of cosine similarities between CVE embeddings and all the MITRE ATT&CK technique embeddings enhances the model's ability accurately map CVE entries to ATT&CK techniques for the large dataset.

The MICS model strives to align with EY's Responsible AI Framework, emphasizing transparency, fairness, and explainability. It employs tools like WEAT and t-SNE visualizations to rigorously assess fairness, ensuring minimal bias in its operations. Furthermore, evaluations using SHAP interaction values and permutation importance metrics highlight the critical role of cosine similarity features in boosting model accuracy and reliability. Although the thesis successfully integrates a Responsible AI framework, its effectiveness largely depends on the sensitivity of the data used. The thesis acknowledges the importance of feature performance in adherence to fairness, transparency, and explainability. However, it provides only a broad overview, making it challenging to trace the specific impact of individual words on performance measures and to establish causality within the model. This limitation underscores the difficulty of fully implementing responsible AI principles in such complex models, given the current state of technology.

The findings of this thesis highlight the importance of integrating advanced NLP models with Responsible AI principles to enhance the capability of cyber threat prediction systems. By combining SRL, vector embeddings and cosine similarities in a model that is tested on an ethical framework, the MICS model not only improve technical performance but also ensure responsible deployment. This dual focus addresses both the immediate need for effective vulnerability mapping and the integration of responsible AI deployment in cybersecurity. The research highlights the need for high-quality, annotated datasets to optimize the performance of semantic mapping models, particularly for automated risk assessments. While traditional models may be sufficient for simpler tasks, advanced NLP models like MICS are important for managing the complex, contextual nature of cybersecurity data. These models provide a more precise method for mapping vulnerabilities to attack techniques, thereby improving the prediction and mitigation of cyber threats.

Chapter 8

Limitations & Future Work

While this research demonstrates the potential of integrating NLP models with Responsible AI principles to enhance cyber vulnerability mapping to attack techniques, several limitations must be acknowledged. These limitations provide a foundation for future research in the field.

One of the key strengths of this study is the utilization of the most relevant and up-to-date database. Additionally, the study employs a sufficiently large dataset, enhancing the robustness and reliability of the models. This allows for results that can better represent real-world scenarios. Furthermore, the integration of the most recent and refined NLP models, along with ethical considerations, ensures the study meets high standards of accuracy, fairness, and transparency. This is important for building trustworthy AI systems.

However, the availability of high-quality, annotated datasets is important for training effective classification models. Acquiring well-annotated datasets is challenging, which limits the model's ability to generalize across various types of cyber vulnerabilities and lowers its overall performance when predicting within larger datasets. Consequently, much of the research focuses solely on predicting the techniques represented in the data, neglecting the broader goal of mapping them to all available techniques within the MITRE ATT&CK matrix. To address this issue, it is advisable to develop more rule-based descriptions of the MITRE ATT&CK techniques, making them easier to distinguish and apply accurately in predictive models.

Besides the availability of high-quality annotated datasets, it is important that the model is trained on a less skewed dataset. Utilizing LLMs to create new descriptions for under-represented MITRE ATT&CK techniques may enhance the prediction performance metrics for these less common techniques.

The SMET model demonstrates good results with smaller datasets but faces challenges in scalability and adaptability when applied to larger, more complex datasets. Conversely, the MICS model excels with large datasets but tends to underperform on smaller ones. Further research is necessary to develop alternative models that offer a more robust framework for mapping CVE descriptions to attack techniques. This study makes improvements in this direction by incorporating cosine similarity vectors, yet there remains room for improvement within large datasets. To address this, research should be conducted to a hybrid model approach that combines the strengths of SMET and MICS. This model could dynamically adjust its approach based on the dataset size, employing different strategies for small and large datasets.

The models are primarily trained and validated on specific datasets that may not be representative for the goal of applying it for a risk assessment. This necessitates further validation to ensure its applicability across different contexts. Additionally, the complexity and size of the models make them time-consuming to train and implement. This makes it difficult to use them in real-time for advice in daily practice, which is a significant limitation for applications requiring quick decisions. Especially when comparing them to the baseline models, this model does make improvements but it does not reach the required precision necessary.

Fully automating the process of risk assessment, including the mapping of vulnerabilities to specific mitigation techniques (such as CWE predictions), remains a significant challenge. Current models have not yet reached the level of precision required for fully autonomous operations in high-stakes environments like cybersecurity. Future research could focus on developing a model that predicts both main techniques and sub-techniques. By incorporating predictions for sub-techniques, mitigation advice could become more accurate and effective.

In conclusion, while this thesis has made improvements in integrating NLP models with Responsible AI for cybersecurity threat prediction, several areas require further research. Addressing these limitations and exploring new ways for improvement will be important for developing more robust, accurate, and ethical AI systems for cybersecurity. Future research should focus on developing more annotated and high-quality datasets to improve model training and generalization, creating hybrid models that can adapt to different dataset sizes and complexities, enhancing the models' scalability and real-time applicability to ensure practical deployment in various cybersecurity scenarios, and fully automating risk assessments with high precision to support autonomous cybersecurity operations.

References

Abdeen, e. a., Basel. (2023). Semantic mapping of cve to att&ck and its application to cybersecurity. *Provided by User*. (file-UlAgGnJu46TWSmRdQoDxEdos)

Accenture. (2024, 29th Mar). *Responsible ai principles to practice.* https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice. (Accessed: 2024-05-23)

Agarwal, S. & Mishra, S. (2021). Fairness and proxy features. In *Responsible ai: Implementing ethical and unbiased algorithms* (chap. 2). Springer. Retrieved from https://doi.org/10.1007/978-3-030-76860-7 doi: 10.1007/978-3-030-76860-7

Aghaei, E. & Al-Shaer, E. (2023). Automated cve analysis for threat prioritization and impact prediction. *Preprint*. (file-sHsqBKXHnB9AOy5ZXCKGoJ10)

Aghaei, E., Niu, X., Shadid, W. & Al-Shaer, E. (2022). *Securebert: A domain-specific language model for cybersecurity.* (Preprint arXiv:2204.02685. Available at ht-tps://arxiv.org/abs/2204.02685)

Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N. & Connolly, J. F. (2022). Cybersecurity threats and their mitigation approaches using machine learning—a review. *Journal of Cybersecurity and Privacy*, *2*(3), 527–555. doi: 10.3390/jcp2030027

Alam, M. T., Bhusal, D., Park, Y. & Rastogi, N. (2022). *Cyner: A python library for cybersecurity named entity recognition.*

Alvarez, R. M. & VanBeselaere, C. (2005). Web-based survey. In *Encyclopedia of social measurement* (p. 955-962). Elsevier. doi: 10.1016/B0-12-369398-5/00390-X

Ansari, M. F., Dash, B., Sharma, P. & Yathiraju, N. (2022, 10). The impact and limitations of artificial intelligence in cybersecurity: A literature review. *International Journal of Advanced Research in Computer and Communication Engineering*, *11*, 81-90. doi: 10.17148/IJARCCE.2022.11912

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. Retrieved from www.elsevier.com/locate/inffus

Bayer, M., Kuehn, P., Shanehsaz, R. & Reuter, C. (2022). *Cysecbert: A domain-adapted language model for the cybersecurity domain.*

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. Retrieved from `https://doi.org/10.1023/A:1010933404324` doi: 10.1023/A:1010933404324

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Center for Threat-Informed Defense. (2024). *Mapping of att&ck to cve.* GitHub repository. Retrieved from `https://github.com/center-for-threat-informed-defense/attack_to_cve/blob/master/README.md`

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* (Preprint arXiv:1810.04805. Available at https://arxiv.org/abs/1810.04805)

Dong, L., Majumder, S., Doudi, F., Cai, Y., Tian, C., Kalathi, D., ... Xie, L. (2024). *Exploring the capabilities and limitations of large language models in the electric energy sector.*

El Naqa, I. & Murphy, M. J. (2015). What is machine learning? In I. El Naqa, R. Li & M. Murphy (Eds.), *Machine learning in radiation oncology* (p. 3-11). Cham: Springer. Retrieved from `https://doi.org/10.1007/978-3-319-18305-3_1` doi: 10.1007/978-3-319-18305-3_1

Ernst & Young Global. (2024). *Responsible ai.* Ernst & Young Global Limited. Retrieved from `https://www.ey.com/en_ch/ai/responsible-ai`

European Parliament. (2023, 8th Jun). *Eu ai act: First regulation on artificial intelligence.* European Parliament News. Retrieved from `https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence` (Last updated on 2023-12-19 at 11:45)

Ferrag, M. A., Battah, A., Tihanyi, N., Debbah, M., Lestable, T. & Cordeiro, L. C. (2023). Securefalcon: The next cyber reasoning system for cyber security. *arXiv preprint arXiv:2307.06616*. Retrieved from `https://arxiv.org/abs/2307.06616`

Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., Lestable, T. & Thandi, N. S. (2024). Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *Technology Innovation Institute*.

Ganda, D. & Buch, R. (2018). A survey on multi label classification. *Recent Trends in Programming Languages*, *5*(1), 19–23. Retrieved from `http://www.stmjournals.com/`

Grigorescu, O., Nica, A., Dascalu, M. & Rughinis, R. (2022). Cve2att&ck: Bert-based mapping of cves to mitre att&ck techniques. *Algorithms*, *15*, 314. Retrieved from `https://www.mdpi.com/journal/algorithms` doi: 10.3390/a15090314

HaCohen-Kerner, Y., Miller, D. & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, *15*(5), e0232525. Retrieved from `https://doi.org/10.1371/journal.pone.0232525` doi: 10.1371/journal.pone.0232525

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Handa, A., Sharma, A. & Shukla, S. K. (2019). Machine learning in cybersecurity: A review. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1306. doi: 10.1002/widm.1306

Huang, J. & Chang, K. C.-C. (2023). A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

IBM. (2023). *Cost of a data breach report 2023.* https://www.ibm.com/security/digital-assets/cost-data-breach-report/. (Accessed: 2024-03-05)

IBM. (2024). *Ibm x-force threat intelligence index 2024.* (Available at https://www.ibm.com/downloads/cas/L0GKXDWJ. Accessed 2024-03-04)

Jiang, F., Xu, Z., Niu, L., Wang, B., Jia, J., Li, B. & Poovendran, R. (2023). *Identifying and mitigating vulnerabilities in llm-integrated applications.*

Jin, J., Tang, B., Ma, M., Liu, X., Wang, Y., Lai, Q., ... Zhou, C. (2024). Crimson: Empowering strategic reasoning in cybersecurity through large language models. *Peking University, Beijing, China and National University of Defense Technology, Changsha, China*. (*Corresponding author: Changling Zhou)

Kamoun, F., Iqbal, F., Esseghir, M. A. & Baker, T. (2020). Ai and machine learning: A mixed blessing for cybersecurity. In *2020 international symposium on networks, computers and communications (isncc)* (pp. 1–7).

Kereopa-Yorke, B. (2023). Building resilient smes: Harnessing large language models for cyber security in australia. *arXiv preprint arXiv:2306.02612*. Retrieved from https://arxiv.org/abs/2306.02612

Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*, 955–962.

Langner, R. (2011, Feb). The real story of stuxnet. *IEEE Spectrum*. Retrieved from https://spectrum.ieee.org/the-real-story-of-stuxnet

Li, B., Mellou, K., Zhang, B., Pathuri, J. & Menache, I. (2023). *Large language models for supply chain optimization.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.*

Lothritz, C., Allix, K., Veiber, L., Bissyandé, T. F. & Klein, J. (2020, December). Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3750–3760). Barcelona, Spain (Online). (December 8-13, 2020)

Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D. & Jacquet, A. (2023). *Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering.* (Preprint arXiv:2209.04963. Available at https://arxiv.org/abs/2209.04963)

Martínez Torres, J., Iglesias Comesaña, C. & García-Nieto, P. (2019). Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*, 2823–2836. doi: 10.1007/s13042-018-00906-1

McKinsey. (2023). *Cybersecurity in a digital era.* https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/cybersecurity-in-a-digital-era. (Accessed: 2023-10-10)

MITRE. (2024a). *Cve - common vulnerabilities and exposures.* https://cve.mitre.org/. (Accessed: 2024-04-11)

MITRE. (2024b). *Mitre att&ck matrix for enterprise.* https://attack.mitre.org/matrices/enterprise/. (Accessed: 2024-04-11)

Motlagh, F. N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F. & Meinel, C. (2024). *Large language models in cybersecurity: State-of-the-art.* (Preprint arXiv:2402.00891. Available at https://arxiv.org/abs/2402.00891)

Mukhamediev, e. a., R.I. (2021). From classical machine learning to deep neural networks: A simplified scientometric review. *Applied Sciences*, *11*(12), 5541. Retrieved from https://doi.org/10.3390/app11125541 doi: 10.3390/app11125541

National Institute of Standards and Technology. (2018). *The csf 1.1 five functions.* Retrieved from https://www.nist.gov/cyberframework/getting-started/online-learning/five-functions (Updated February 26, 2024)

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... Mian, A. (2024). *A comprehensive overview of large language models.*

NVIDIA. (2024, Mar). *How generative ai is transforming cybersecurity.* https://blogs.nvidia.com/blog/generative-ai-cybersecurity/. (Accessed: 2024-03-05)

NVIDIA. (2024). *What are large language models?* https://www.nvidia.com/en-us/glossary/large-language-models/. (Accessed: 2024-03-05)

Oneto, L. & Chiappa, S. (2020). Fairness in machine learning. In *Recent trends in learning from data* (Vol. 896, pp. 155–196). Springer. doi: 10.1007/978-3-030-76977-2_2

Podder, P. e. a. (2020). Artificial neural network for cybersecurity: A comprehensive review. *Not specified*, *Not specified*(Not specified), Not specified.

Qaiser, S. & Ali, R. (2018, July). Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25.

Ranade, P., Piplai, A., Joshi, A. & Finin, T. (2021). Cybert: Contextualized embeddings for the cybersecurity domain. In *2021 ieee international conference on big data (big data)* (p. 3334-3342). doi: 10.1109/BigData52589.2021.9671824

Sarker, I. H., Kayes, A. & Badsha, S. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, *7*(1), 41. Retrieved from `https://doi.org/10.1186/s40537-020-00318-5` doi: 10.1186/s40537-020-00318-5

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A. & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*. doi: 10.1109/ACCESS.2020.3041951

Shi, P. & Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019). How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*. Retrieved from `https://arxiv.org/abs/1905.05583`

Tan, J. (2023). *The danger of artificial intelligence in cybersecurity: A theoretical analysis* (Doctoral dissertation, New York University). doi: 10.13140/RG.2.2.34877.64485

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K. & et al. (2023). Large language models in medicine. *Nature Medicine*, *29*, 1930–1940. Retrieved from `https://doi.org/10.1038/s41591-023-02448-8` doi: 10.1038/s41591-023-02448-8

Tuor, A. R., Baerwolf, R., Knowles, N., Hutchinson, B., Nichols, N. & Jasper, R. (2018). Recurrent neural network language models for open vocabulary event-level cyber anomaly detection. In *Workshops at the thirty-second aaai conference on artificial intelligence.*

Wang, F. (2023). *Using large language models to mitigate ransomware threats.* Preprints. Retrieved from `https://doi.org/10.20944/preprints202311.0676.v1` doi: 10.20944/preprints202311.0676.v1

World Economic Forum. (2024, 10th January). *2023 was a big year for cybercrime – here's how we can make our systems safer.* `https://www.weforum.org/agenda/2024/01/cybercrime-2023-review-and-prevention/`. (Accessed: 2024-04-15)

Xiao, Y., Jin, Y., Bai, Y., Wu, Y., Yang, X., Luo, X., … Cheng, W. (2023). Large language models can be good privacy protection learners. *arXiv preprint arXiv:2310.02469*.

Zhang, Z., Zheng, C., Tang, D., Sun, K., Ma, Y., Bu, Y., … Zhao, L. (2023). *Balancing specialized and general skills in llms: The impact of modern tuning and data strategy.*

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., … Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), Article 20. doi: 10.1145/3639372

Appendix A

Dataset Sample

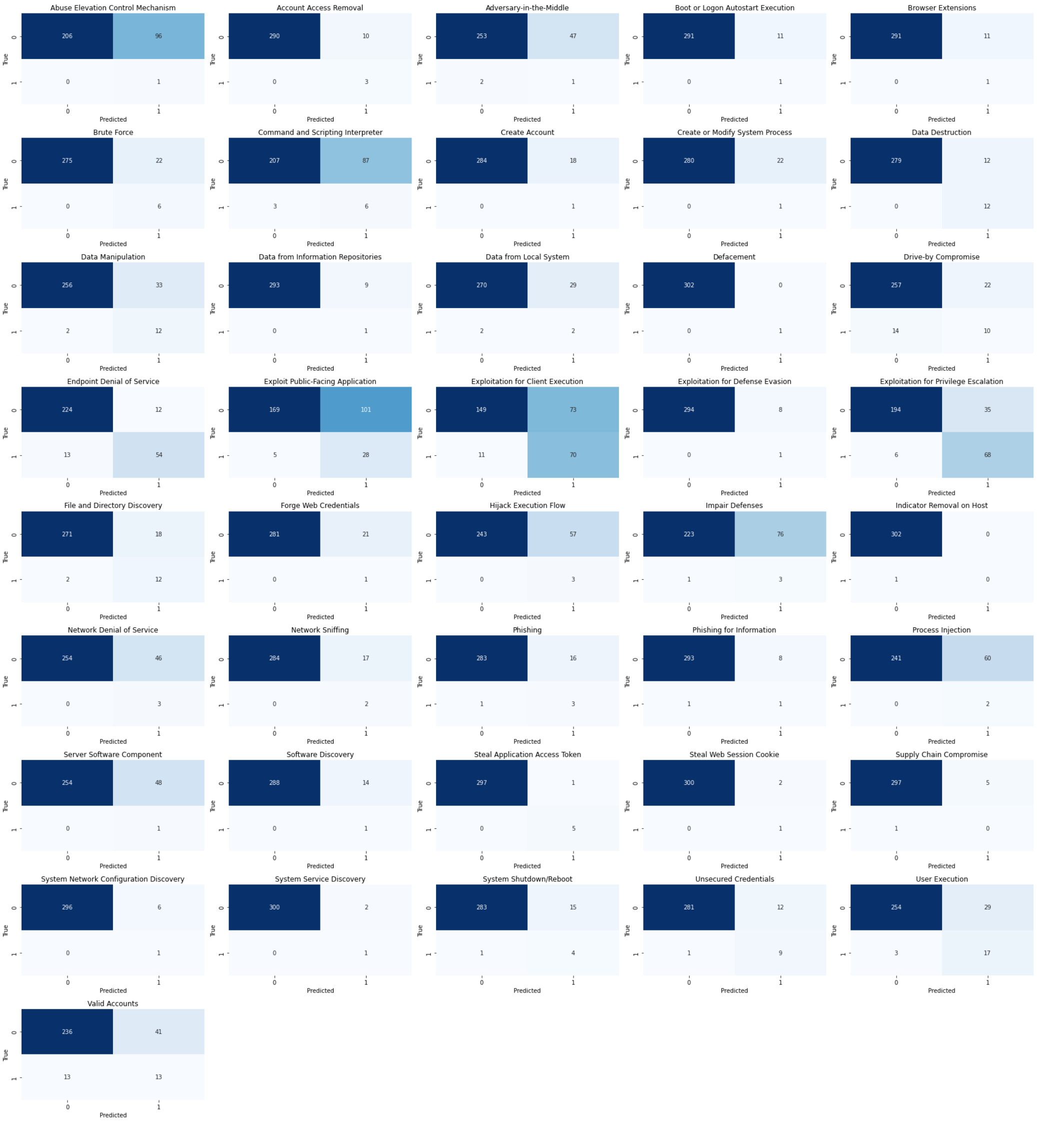
Index	CVE ID	Main Techniques	Description
0	CVE-2019-15243	Command and Scripting Interpreter, Valid Accounts, Exploit Public-Facing Application	Multiple vulnerabilities in Cisco SPA100 Series ATAs could allow an authenticated attacker to execute arbitrary code with elevated privileges.
1	CVE-2019-15976	Command and Scripting Interpreter, Exploitation for Privilege Escalation, Exploit Public-Facing Application	Multiple vulnerabilities in Cisco DCNM authentication mechanisms could allow an attacker to bypass authentication and execute with administrative privileges.
2	CVE-2019-15956	Account Manipulation, Endpoint Denial of Service, Exploit Public-Facing Application, Valid Accounts	A vulnerability in Cisco AsyncOS could allow an authenticated attacker to perform an unauthorized system reset.
3	CVE-2019-15958	Command and Scripting Interpreter, Exploit Public-Facing Application	A vulnerability in Cisco PI and EPNM REST API could allow an attacker to execute arbitrary code with root privileges.
4	CVE-2019-12660	Valid Accounts, Impair Defenses, Hijack Execution Flow	A vulnerability in the CLI of Cisco IOS XE Software could allow an authenticated attacker to write values to device memory.
5	CVE-2019-1753	Command and Scripting Interpreter, Valid Accounts, Exploitation for Privilege Escalation, Exploit Public-Facing Application	A vulnerability in Cisco IOS XE Software web UI could allow an attacker to run privileged commands via the web UI.

Table A.1: Sample entries from annotated CVE Dataset

Appendix B

Extended Results

B.1 Confusion Matrix Small Dataset



Abuse Elevation Control Mechanism

Account Access Removal

Adversary-in-the-Middle

Boot or Logon Autostart Execution

Browser Extensions

Brute Force

Command and Scripting Interpreter

Create Account

Create or Modify System Process

Data Destruction

Data Manipulation

Data from Information Repositories

Data from Local System

Defacement

Drive-by Compromise

Endpoint Denial of Service

Exploit Public-Facing Application

Exploitation for Client Execution

Exploitation for Defense Evasion

Exploitation for Privilege Escalation

File and Directory Discovery

Forge Web Credentials

Hijack Execution Flow

Impair Defenses

Indicator Removal on Host

Network Denial of Service

Network Sniffing

Phishing

Phishing for Information

Process Injection

Server Software Component

Software Discovery

Steal Application Access Token

Steal Web Session Cookie

Supply Chain Compromise

System Network Configuration Discovery

System Service Discovery

System Shutdown/Reboot

Unsecured Credentials

User Execution

Valid Accounts

Figure B.1: Confusion Matrix Small SMET.

B.2 Heatmap SMET Small Dataset

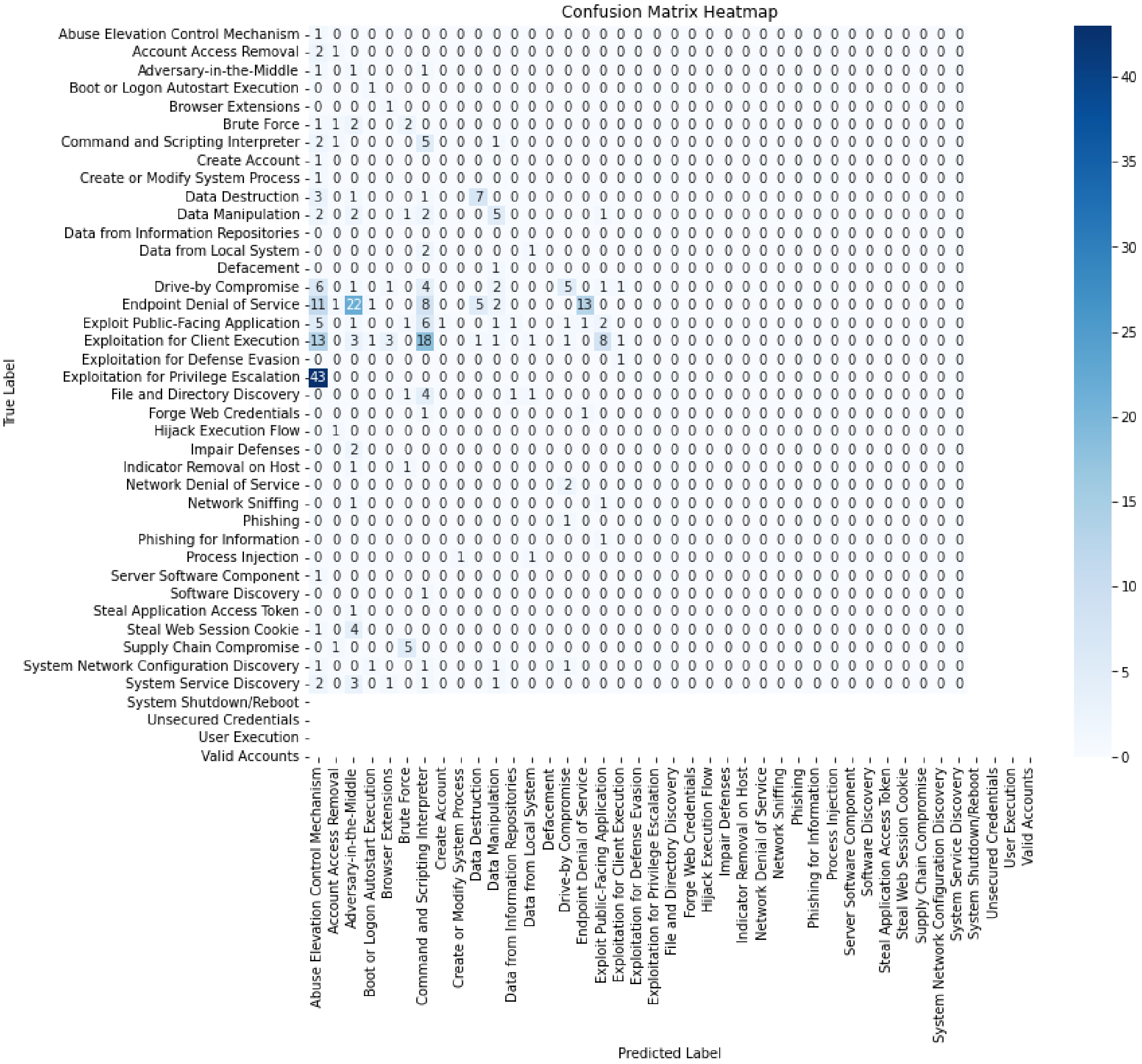


Figure B.2: Heatmap Small Dataset

B.3 Confusion Matrix SMET Large Dataset

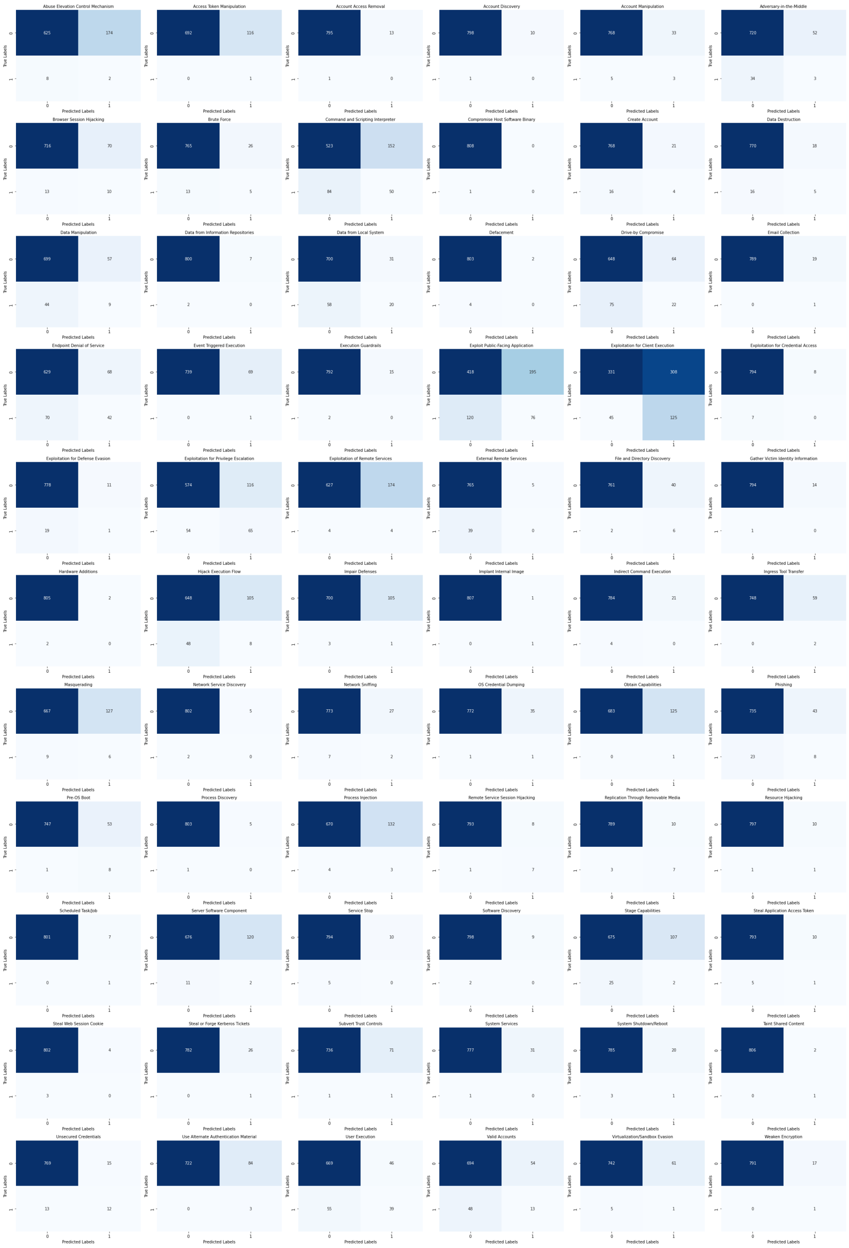


Figure B.3: Confusion Matrix Large SMET.

B.4 Heatmap SMET Large Dataset

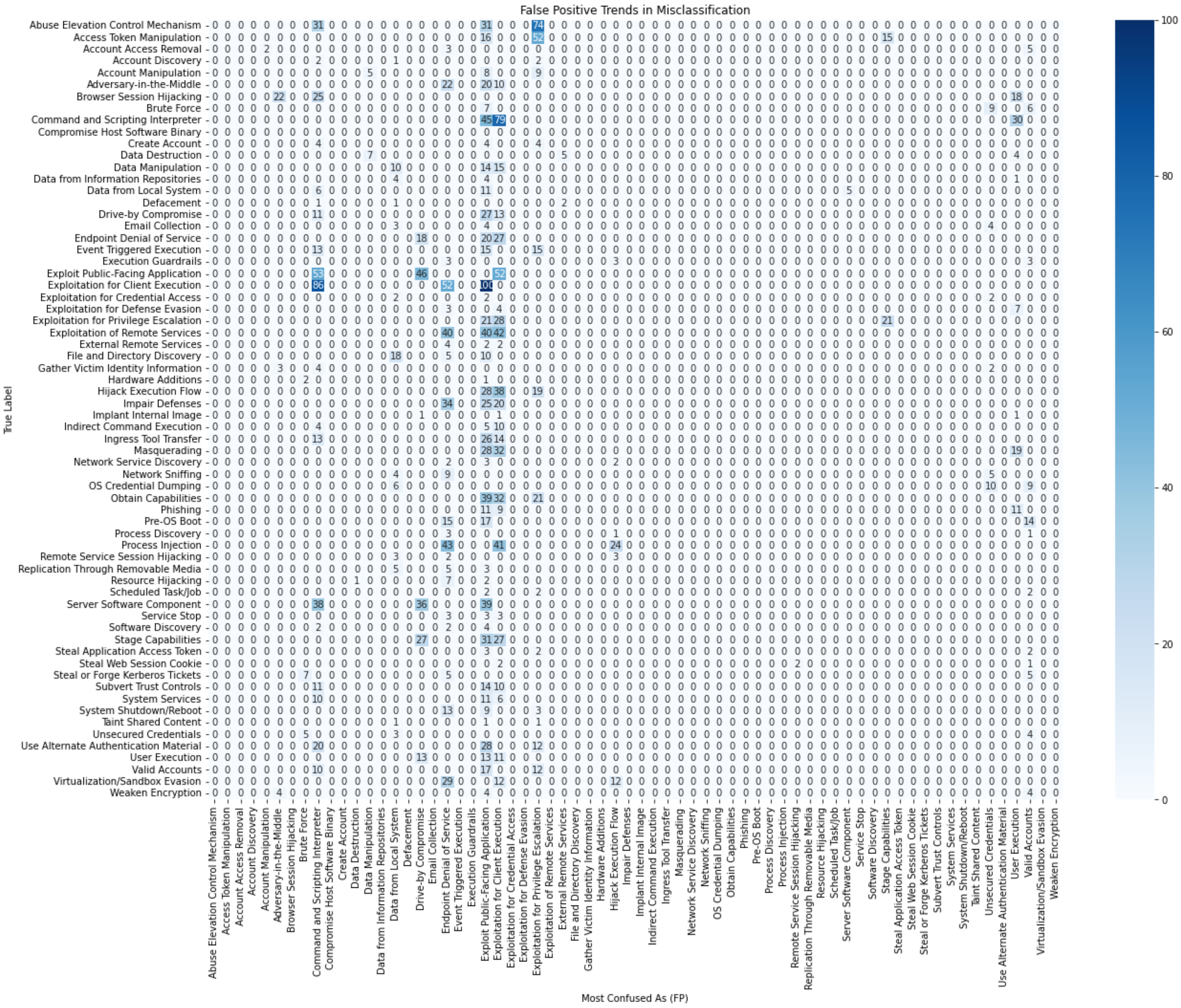


Figure B.4: Heatmap Large Dataset.

B.5 Confusion Matrix MICS

B.6 Permutation Feature Importances MICS

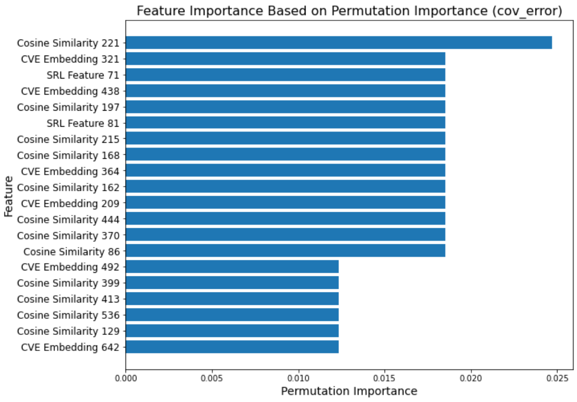


Figure B.6: Feature Importance Coverage Error

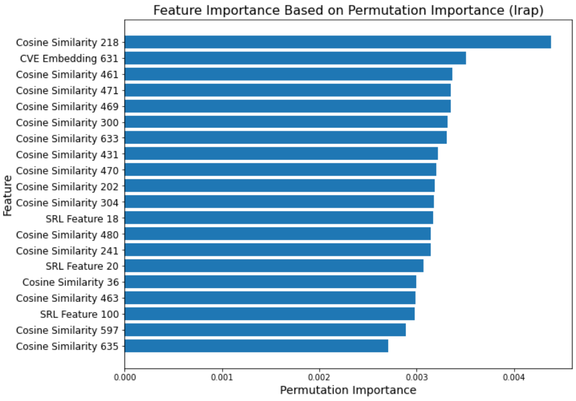


Figure B.7: Feature Importance LRAP

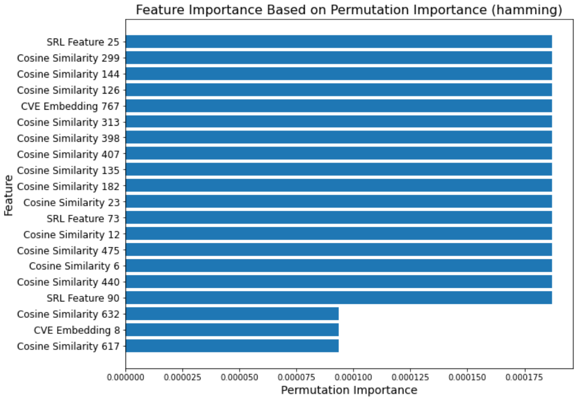


Figure B.8: Feature Importance Hamming Loss

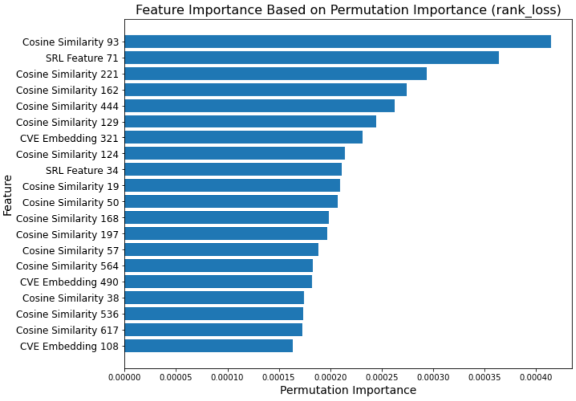


Figure B.9: Feature Importance Rank Loss

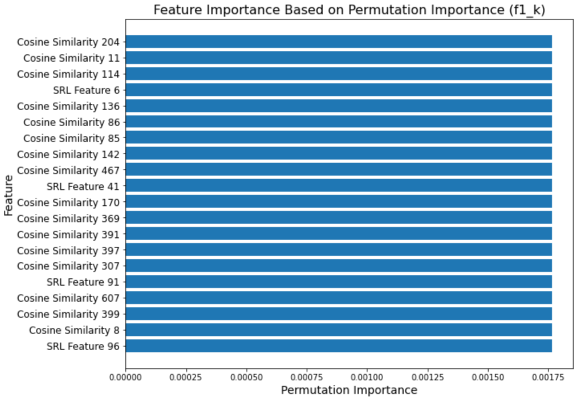


Figure B.10: Feature Importance F1 Score

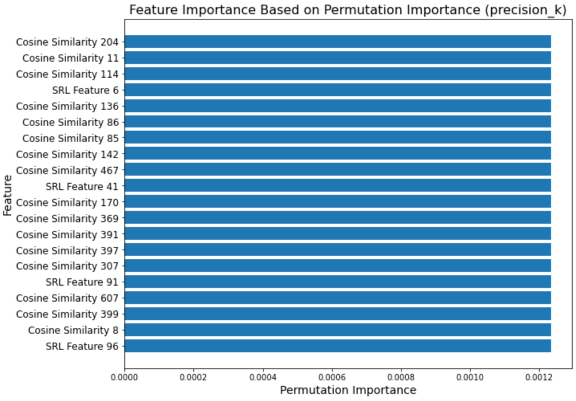


Figure B.11: Feature Importance Precision

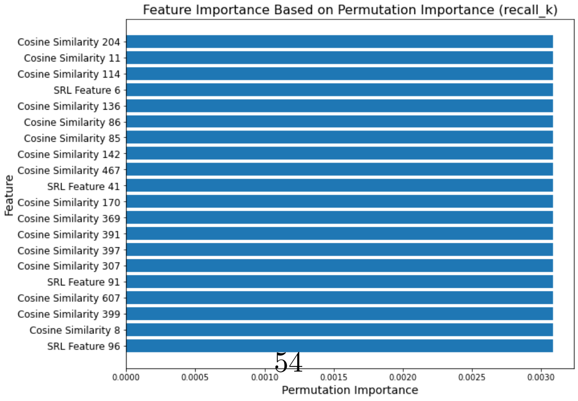


Figure B.12: Feature Importance Recall

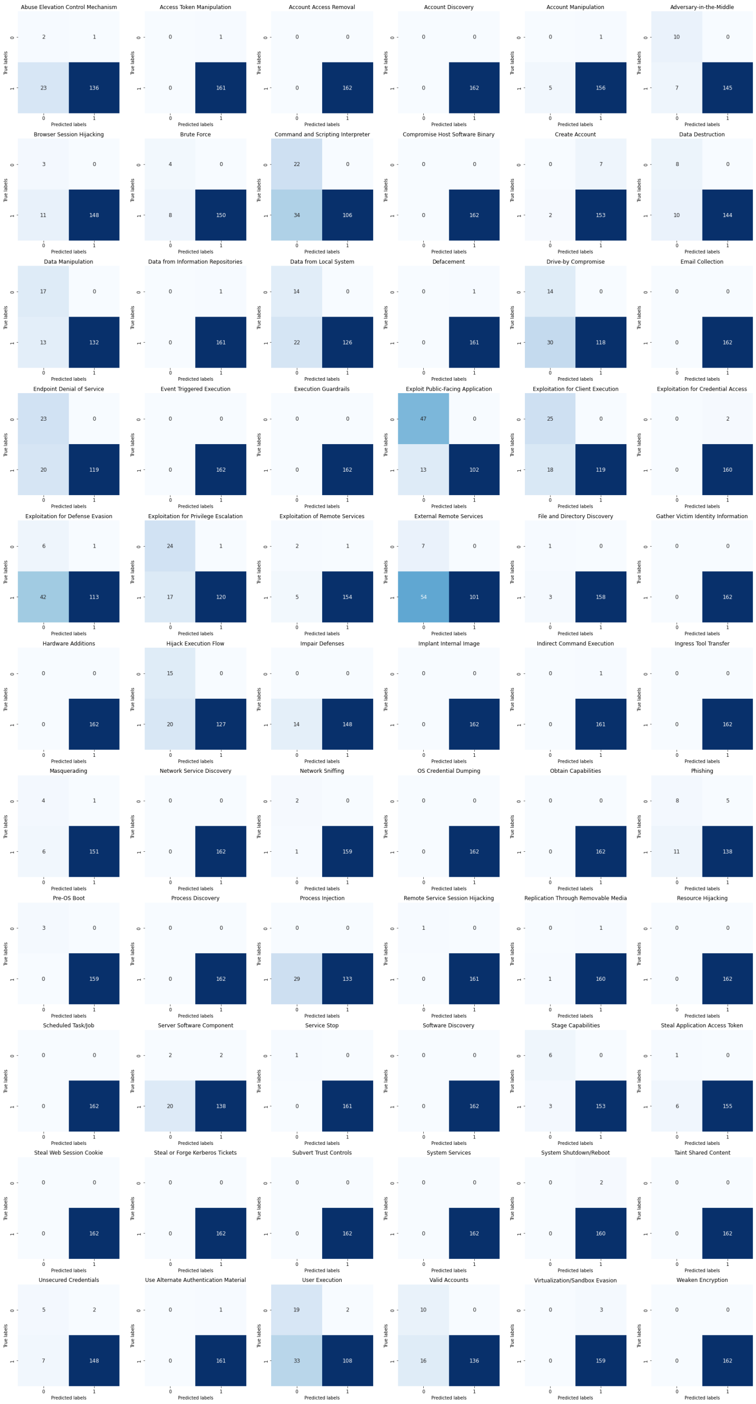


Figure B.5: Confusion Matrix MICS.