Forecasting cyber threats & pertinent alleviation technologies

Birkbeck
UNIVERSITY OF LONDON

Forecasting Cyber Threats and Pertinent Alleviation Technologies

Student Name

Zaid Almahmoud

Degree Programme

PhD in Computer Science

Submission Date

Wednesday 3$^{\text{rd}}$ July, 2024

School of Computing and Mathematical Sciences
Birkbeck, University of London

Abstract

Traditionally, cyber-attack detection relies on reactive techniques, where pattern-matching algorithms help human experts to scan system logs and network traffic for known virus signatures. Recent research has introduced effective Machine Learning (ML) models for cyber-attack detection. However, approaches that can forecast attacks likely to happen in the long term are also desirable, as this gives defenders more time to develop defensive actions and tools. Today, long-term predictions of attack waves are based on the subjective perceptiveness of human experts, susceptible to bias. This work introduces a novel ML-based approach that leverages unstructured big data to forecast the trend of cyber-attacks, years in advance. To this end, we develop a framework that utilises a monthly dataset of major cyber incidents in 36 countries over the past 11 years, with new features extracted from big data sources, namely news, government advisories, research literature, and tweets. Our framework not only forecasts attack trends automatically, but also generates a threat cycle that drills down into five key phases that constitute the life cycle of 42 known cyber threats.

Our research advances to the next level, by predicting the disparity between cyber-attack trends and the trend of the relevant alleviation technologies. These predictive analyses inform investment decisions in cyber security technologies and provide a fundamental basis for strategic choices by national defence agencies. Here, we expand our dataset with records for the trend of 98 alleviation technologies. Using our expanded dataset, we construct a graph that elucidates the interplay between cyber threats and pertinent alleviation technologies. To forecast the graph, we propose a Bayesian adaptation of a Graph Neural Network (GNN) model. Furthermore, we generate future data projections for the next three years, including the gap between the trend of cyber-attacks and the associated technologies. Consequently, we introduce the concept of "alleviation technologies cycle", delineating the key phases in the life cycle of 98 technologies. To bolster the transparency of our model, we incorporate explainability features, fostering a clear and informed decision-making process.

Publications and Conferences

Publications

Title	Journal/Conference	Status	Date
A Holistic and Proactive Approach to Forecasting Cyber Threats	*Nature - Scientific Reports*	*Published*	*May 17, 23*
Forecasting Cyber Threats and Pertinent Alleviation Technologies	*Technological Forecasting and Social Change*	*Minor revision*	*Jun 24, 24*

Conferences

Title	Conference name	Type	Date
Forecasting Cyber Events and Trend Analysis	*School of Business, Economics and Informatics PhD conference*	*Poster*	*May 17, 22*
Forecasting Cyber Events and Trend Analysis	*School of Business, Economics and Informatics PhD conference*	*Presentation*	*Jun 21, 23*

Notes

The first row in the above tables corresponds to our contributions in chapter 3, while the second row corresponds to our contributions in chapter 4.

Acknowledgements

First, and foremost, I would like to use this opportunity to thank my professor and advisor, Dr. Paul D. Yoo for his hard effort, and outstanding guidance and assistance to me in the pursuit of my PhD degree. I sincerely appreciate the knowledge and exemplary conduct I have learned from him, and I am deeply grateful for his unwavering support and firm belief in me. I am very thankful to Dr. Omar Alhussein for his support and supervision. I appreciate the time he spent and the suggestions he provided me with, in the writing of my thesis. I am truly thankful to Birkbeck College and the University of London for giving me the opportunity to pursue my PhD degree in the area I love, and for providing me with a scholarship to conduct the research I have always aspired to undertake. It is an honour to be part of a university dedicated to advancing knowledge and innovation in various fields. Also, special thanks to my friends and colleagues in Birkbeck College including Alberto Matuozzo, Angelina Mcdonald, Dilara Uysal, and Ibrahim Alsaggaf for their help and support.

Lastly, and most importantly, I would like to thank my father Jamal, my mother Wafa, my brothers, my sister Safia, and my beloved and supportive wife Rana, for their continuous support, care, and love, and whose encouragement has been my pillar of strength. To them, and to my beloved children Laya and Amir (Algo), I dedicate this thesis, with deep love and appreciation.

Wednesday 3rd July, 2024
Zaid Almahmoud

"Our thoughts are with those facing hardships."

Table of contents

List of Tables

List of Figures

1 | Introduction

1.1 Background

In the rapidly evolving domain of cyber security, the past decade has witnessed a disturbing surge in both the frequency and sophistication of cyber-attacks. Governments, organisations, and enterprises find themselves increasingly targeted, grappling not only with the immediate digital ramifications but also contending with repercussions that extend to impact their overall financial stability [1]. Faced with these escalating threats, the conventional approach to cyber defence has predominantly adhered to reactive methodologies.

Reactive methods, characterised by their responsive nature, involve responding to cyber threats after they have been identified or manifested. One prevalent technique within the field of reactive cyber security is the use of pattern-matching algorithms, designed to identify known virus or malware signatures within system logs and network traffic [2]. This method involves creating a database of previously identified cyber threats and their distinctive signatures, which are then compared against the incoming data for a match. While such reactive measures are effective in identifying and mitigating known threats, their limitation lies in their dependence on the availability of pre-existing signatures [3]. Consequently, they may fall short in addressing novel or previously unseen cyber threats, leaving a critical vulnerability in the cyber security infrastructure.

Amidst the challenges posed by the reactive nature of cyber defences, the cyber security landscape has undergone further transformation with the emergence of highly sophisticated threats. The complexity of modern cyber-attacks has reached unprecedented levels, resulting in waves of incidents that inflict substantial damage on various entities, significantly affecting their bottom lines [1].

Acknowledging the inadequacies of traditional cyber security measures, recent efforts have pivoted towards the integration of ML models as a potential solution [3]. ML, equipped with the capability to learn from data and adapt to evolving patterns, presents a promising avenue for addressing the shortcomings of the traditional defences. A significant advancement in this direction involves the utilisation of anomaly detection algorithms within ML-based models, offering a more intelligent approach to threat detection.

However, the adoption of anomaly-based models introduces its own set of challenges. Notably, the potential for false alarms arises, as these models may classify benign behaviour as abnormal [4]. This phenomenon heightens alert levels for security teams, leading to unnecessary investigations and resource allocation. Striking the right balance between sensitivity to potential threats and specificity to normal behaviour becomes crucial for the success of ML-based anomaly detection in cyber security.

As cyber threats continue to evolve and diversify, there is a growing recognition within the cyber security community of the need to complement these reactive approaches with more proactive and anticipatory strategies [5, 6]. This shift is essential to stay ahead of cyber adversaries who are continually innovating and devising new methods to exploit vulnerabilities in digital systems.

1.2 Proposal

We firmly advocate that leveraging the abundance of available data can usher in a paradigm shift toward a *proactive* defence strategy, enabling actions to be taken before a potential cyber threat escalates into an actual incident. Drawing parallels with non-cyber threats, such as terrorism and military attacks, it is evident that proactive approaches have been instrumental in alleviating, delaying, and even preventing incidents from occurring. Notably, sophisticated software programs have been developed for assessing the intention, potential damages, attack methods, and alternative options in the context of traditional threats like terrorism [7]. In this vein, we posit that the field of cyber security should be no exception, and the contemporary landscape affords us the capabilities to deploy proactive, low-latency cyber defences, particularly through the application of ML techniques [5].

ML, with its capacity to analyse vast datasets and discern patterns, stands as a formidable tool in constructing proactive cyber defence mechanisms. This assertion is supported by the success of ML models in various domains, showcasing their ability to provide accurate and reliable forecasts. Noteworthy examples include ML models like AlphaFold2 [8] and RoseTTAFold [9], which have demonstrated the capability to predict a protein's intricate three-dimensional structure from its linear sequence with remarkable precision.

However, the application of ML in the field of cyber security encounters distinctive challenges. Unlike other fields where ML models thrive, cyber security data is often shrouded in confidentiality due to the sensitive nature of cyber incidents, impacting the reputation of the involved organisations. Compounding this challenge is the inherent difficulty in tracking cyber incidents, as they can frequently go unnoticed even by the victimised entities. Additionally, the pre-processing of cyber security data presents unique hurdles, characterised by a lack of structure, diversity in format, and high rates of missing values that can potentially distort analytical findings [10, 11].

Navigating these challenges requires a nuanced approach that considers the intricacies of cyber security data. It involves developing ML models that not only account for the sensitivity and confidentiality of cyber incidents but also address the complexities in data pre-processing. By doing so, we can harness the full potential of ML to construct proactive cyber defence systems that not only predict potential threats but also adapt to the dynamic and evolving landscape of cyber security. This research endeavour aims to contribute novel insights and methodologies to overcome these challenges, thereby advancing the frontier of proactive cyber defence.

In this work, we aim to conduct the first study on the long-term forecast of cyber-attacks, employing a fully automated approach to predict cyber-attack trends years in advance [6]. Nonetheless, achieving an informed decision-making process in the context of technology investment for the mitigation of these forthcoming threats necessitates a more forward-looking perspective [12]. This perspective includes the discernment of the divergence between the trajectory of cyber threats and the Pertinent Alleviation Technologies (PATs). For instance, being able to predict a substantial future disparity between the trajectory of an attack trend and the trend of the corresponding technical solution empowers us to make judicious choices regarding our investments. Consequently, we can prioritise our defence strategies based on these predictive insights, bridging the gap between threats and alleviation technologies. Therefore, in our extended version of the problem, we study the forecast of the cyber-attack trends and the trend of the PATs. Ultimately, we aim to recommend future investment decisions and strategic defences based on our forecast.

The main challenge we face is the scarcity of data since we are looking at the scale of years, and neural networks are hungry for data. We also note that there are many attacks and technologies, each demanding thorough analysis and forecasting. An intriguing avenue of inquiry emerges: can features such as those associated with geopolitical conflicts be harnessed to enhance our predictive capabilities? Additionally, our research seeks to determine the most effective ML model for addressing this complex problem.

Furthermore, we aim to incorporate explainability as a crucial aspect of our contribution, given the promising applications of Explainable AI in cyber trend forecasting. Explainability in cyber-attack forecasting and PATs prediction presents a pivotal opportunity to enhance the transparency of our predictive models. Also, by employing models with inherent interpretability, analysts can gain valuable insights into the underlying factors influencing cyber trends. In fact, we will show that explainability not only fosters trust and confidence in the model, but also facilitates the identification of hidden associations and dependencies between cyber trends. This, in turn, enables a more effective prioritisation of defences, fostering a proactive approach to cyber security.

1.3 Rationale

The persistent reliance on reactive measures, characterised by identifying and responding to threats after their occurrence, has exacerbated vulnerabilities. One significant challenge lies in the overhead incurred by the time-consuming pattern-matching operations essential for discerning the signatures of polymorphic malware [13]. Polymorphic malware, with its ability to exhibit different behaviours with each execution, poses a formidable challenge for traditional signature-based detection methods.

The constant adaptation and mutation of attack patterns demand continuous updates to signature databases, resulting in a perpetual cat-and-mouse game. Defenders find themselves struggling to keep pace with the evolving tactics of malicious actors, leading to a noticeable time lag between the emergence of a new threat and the implementation of corresponding defence measures.

The current paradigm in cyber threat prediction predominantly focuses on short and midterm forecasting, typically extending only to a few hours, days, or months in advance [14, 15, 16, 17]. While these predictions are valuable for immediate response measures, the broader landscape of cyber security demands a shift towards long-term prognostic capabilities [6]. The temporal limitations of existing approaches hinder the ability to foresee cyber threats on a larger scale, spanning years into the future. As cyber threats evolve and become more sophisticated over extended periods, there is an imperative to develop forecasting methodologies that extend beyond the confines of short or midterm predictions, providing defenders with sufficient time for strategic insights and response planning.

It follows that the primary rationale behind this work lies in the critical need for accurate and objective forecasting of cyber threats over the long term. In today's digital landscape, cyber-attacks are becoming increasingly sophisticated and prevalent. Hence, the ability to anticipate and prepare for future threats ahead of time is paramount for effective cyber security management. As previously mentioned, the long-term forecast of cyber threats provides cyber security agencies sufficient time to assess existing defence measures and identify areas where preventive solutions can be developed proactively. However, the existing approaches to cyber threat forecasting in the long term remains, to a considerable extent, dependent on the subjective interpretations of human security experts [18, 19]. In contrast to fully automated procedures grounded in quantitative metrics, the human-based approach introduces an element of subjectivity susceptible to bias, often influenced by scientific or technical inclinations [20]. This reliance on human judgment may inadvertently introduce variations in threat assessments, potentially leading to inconsistencies in long-term predictions.

It is imperative to acknowledge that while human expertise is invaluable, the inherent subjectivity in their judgments can hinder the attainment of scientific objectivity in long-term cyber threat predictions. The integration of quantitative metrics provides a crucial avenue for introducing a more objective and standardised approach to threat assessment. This objectivity is vital for ensuring the reliability and credibility of long-term predictions in the ever-evolving landscape of cyber security.

Quantitative predictions, grounded in empirical data and measurable metrics, offer a systematic and transparent framework for evaluating and forecasting cyber threats. By relying on quantitative indicators, the risk of subjective biases is minimised, fostering a more objective and consistent analytical process [21]. This scientific objectivity not only enhances the reliability of long-term predictions but also facilitates a more comprehensive understanding of the evolving threat landscape.

A notable research gap also exists in forecasting the disparity between the trajectory of cyber threats and the development of relevant alleviation technologies. To make informed decisions regarding technology investment and strategic defence measures, it is crucial to predict the future misalignments between emerging threats and the technologies designed to counteract them. A systematic approach to forecasting this gap is paramount for allocating resources effectively and prioritising the development of technologies that will be most instrumental in mitigating the imminent threats on the cyber security horizon.

In addition to forecasting, ensuring the transparency and explainability of ML models employed in cyber threat prediction is fundamental. Transparent models not only enhance the interpretability of predictions but also foster informed decision-making processes. Establishing a clear understanding of the relationships between cyber-attacks and the associated technologies is pivotal for developing robust defence strategies. Therefore, integrating explainability features into predictive models contributes to a more comprehensive and interpretable framework, instilling confidence in stakeholders and facilitating proactive defence measures.

1.4 Motivation

The proposed research aligns seamlessly with the initiatives undertaken within the School of Computing and Mathematical Sciences, owing to the school's notable focus on applied ML in the field of cyber security. Furthermore, it is crucial to acknowledge the pressing need for cyber threat forecasting, particularly in light of the recent ransomware attack on the NHS, which garnered significant attention from the UK government [1]. By addressing this imperative issue, the proposed work contributes directly to the growing demand for effective measures in safeguarding digital security.

Academic curiosity and a keen interest in literature gaps have been the driving forces behind the exploration of ML algorithms and their potential applications. Exploring the complexities of algorithmic design and the transformative power these technologies hold over real-world challenges has always intrigued researchers at the intersection of theory and application. The field of cyber security, with its dynamic landscape of threats and ongoing demand for innovative solutions, presents a compelling opportunity to harness the potential of ML. This research endeavour serves as a valuable platform to channel academic curiosity and contribute meaningfully to the evolving field.

Looking beyond the confines of academia, motivation is derived from recognising the practical implications of research endeavours. Recent incidents, such as the ransomware attack on the NHS, have starkly highlighted the tangible consequences of cyber threats on critical infrastructure. These events underscore the urgency of effective cyber threat forecasting and fortifying digital security measures. This work extends beyond theoretical exploration to address broader societal imperatives, offering proactive strategies to enhance the resilience of digital infrastructure.

1.5 Research Questions

This is the first work answering the following research questions:

1. How can we create a comprehensive dataset using big unstructured data for

predicting cyber-attacks across various regions and attack types?

2. What methodologies can we employ to achieve long-term forecasting of cyber threats, up to three years in advance?
3. How can we adopt a holistic approach to forecasting that accounts for multiple regions and attack types?
4. How can we develop an automated and quantitative approach to forecasting that minimises human bias?
5. How can we classify and forecast a wide range of cyber threats using historical and predictive data?
6. What are the distinct phases in the life cycle of different cyber threats, and how can we model them?
7. How can we construct a graph linking cyber threats to PATs?
8. How can we expand our dataset to include comprehensive trend data for both cyber threats and PATs?
9. How can we develop a Bayesian variation of the GNN model to forecast cyber threat trends and address model uncertainty?
10. What strategies can we propose for future investment and defence based on the forecasted trends in cyber threats and technologies?
11. What are the key phases in the life cycle of PATs, and how can we predict their future states?
12. How effective is the proposed model compared to traditional models in predicting cyber threat trends?
13. How can we enhance the explainability of the model to reveal hidden relationships and improve decision-making?

1.6 Research Contribution

1.6.1 Cyber Threat Forecasting

In the first problem of cyber threat forecasting studied in chapter 3, our objective is to fill the research gaps by a proactive, long-term, and holistic approach to attack prediction. The proposed framework gives cyber security agencies sufficient time to evaluate existing defence measures while also providing objective and accurate representation of the forecast. Our study is aimed at predicting the trend of cyber-attacks up to three years in advance, utilising big data sources and ML techniques. Our ML models are learned from heterogeneous features extracted from massive, unstructured data sources, namely, news, blogs, government advisories [10], Elsevier [22], Twitter [23], and Python APIs [24]. The news, blogs, and government advisories provide more than $15,000$ records of global cyber incidents since the year 2011, while Elsevier API offers access to the Scopus database, the largest abstract and citation database of peer-reviewed literature with over 27,000,000 documents [25]. The number of relevant tweets we collected is around 9 million. Our study covers 36 countries and 42 major attack types. The proposed framework not only provides the forecast and categorisation of the threats, but also generates a threat life-cycle model, whose the five key phases underlie the life cycle of all 42 known cyber threats. The key contribution of the first part of this work consists of the following points, each aligned with a corresponding research question number:

- We constructed a novel dataset using big unstructured data including news and government advisories, in addition to Elsevier, Twitter, and Python API. The dataset comprises monthly counts of cyber-attacks and other unique features, covering 42 attack types across 36 countries. (1)
- Our proactive approach offers long-term forecasting by predicting threats up to 3 years in advance. (2)
- Our approach is holistic in nature, as it does not limit itself to specific entities or regions. Instead, it provides projections of attacks across 36 countries situated

in diverse parts of the world. (3)

- Our approach is completely automated and quantitative, effectively addressing the issue of bias in human predictions and providing a precise forecast. (4)
- By analysing past and predicted future data, we classified threats into four main groups and provided a forecast of 42 attacks until 2025. (5)
- We proposed the first threat cycle, which delineates the distinct phases in the life cycle of 42 cyber-attack types. (6)

1.6.2 Forecasting Threats and Pertinent Alleviation Technologies

In the extended version of the problem studied in chapters 4 and 5, we construct a comprehensive graphical representation known as the Threats and Pertinent Technologies graph (TPT). This graph links cyber threats with their respective PATs. The connections between threats and PATs are established through edges, with the weight of each edge quantifying the gap between the trend of these interconnected nodes. To accomplish the construction of this graph, we employ a semi-automated methodology, utilising the capabilities of the Generative Pre-trained Transformer (GPT) model [26], in conjunction with Elsevier API [22]. This approach facilitates the extraction of PATs associated with each threat. Furthermore, we acquire the monthly trend data for each threat node by leveraging news, blogs, and government advisories data, allowing us to tally the number of monthly incidents. Also, for each PAT node, we use Elsevier platform to retrieve the monthly mentions (trend) of that PAT, thereby augmenting the dataset proposed in our prior work [6]. Our methodology extends to the development of a new Bayesian variation of the GNN model, building upon the framework introduced in the study by Wu *et al.* [27]. This enhanced model is deployed for the purpose of forecasting the TPT graph over a forthcoming 3-year period, while addressing inherent model uncertainties. The ultimate goal of this endeavour is to provide insightful recommendations for future investments in the cyber threat landscape. In addition to the aforementioned contributions, our analysis extends to the introduction of a novel concept called the Alleviation Technologies Cycle (ATC), which delineates the principal phases within the life cycle of 98 PATs. Finally, we incorporate explainability in our model to reveal hidden relationships and facilitate trust and confidence in the prediction. The contributions of the second part of this work are highlighted below along with the corresponding research questions:

- We constructed the graph of 26 emerging and rapidly increasing threats and their PATs, through a semi-automated approach using GPT-3 model and Elsevier API. A novel algorithm called Extractive GPT (E-GPT) which prompts GPT-3 to extract PATs from Elsevier research documents is presented. (7)
- We used big data sources, such as news, blogs, government advisories [10], Elsevier research documents [25], Twitter tweets [23], and the Python Holidays package [24], to expand upon our dataset. This expansion includes incorporating monthly trends of 98 PATs from Elsevier, covering the years 2011 to 2022. Additionally, we included recent trends in cyber threats from news articles and blogs, as well as other features related to attack mentions from Elsevier, wars and conflicts from Twitter, and public holidays from Python API, up to the end of 2022. (8)
- We built a novel Bayesian variation of the Multivariate Time-series Graph Neural Network model (B-MTGNN) proposed in [27] to forecast the graph while addressing the epistemic uncertainty. (9)
- We provided 3 years forecast for the TPT graph, followed by an analysis and categorisation of future gaps, along with recommendations for future investment and defence strategies. (10)
- We proposed the first ATC, illustrating the state of 98 PATs in the coming 3 years, and identifying the key phases in the life cycle of these PATs. (11)
- We provided comparative analysis to show the effectiveness of the proposed model over traditional models and the importance of the features in our dataset. (12)

- We provided explainability analysis, including analysis for saliency maps and attention scores. This meticulous approach not only enhances transparency in the model's decision-making process but also instils confidence in the accuracy of its predictions. Additionally, our focus on explainability unveils concealed relationships within the data, providing invaluable insights that are particularly beneficial for decision makers. (13)

1.7 Thesis Structure

The rest of the thesis is structured as follows. Chapter 2 provides a comprehensive literature review while identifying and highlighting the research gaps. In chapter 3, we describe the first problem of cyber threat forecasting and present our framework and results. In chapter 4, we describe the extended version of the problem, which covers the forecast of both the threats and the PATs using GNN with the objective of identifying the future gaps and recommend strategic defence decisions. Next, chapter 5 presents our preliminary endeavour to integrate explainability into our GNN model. This enhancement aims to endow the model with transparency, fostering confidence in its predictions. This is followed by a discussion for the thesis provided in chapter 6. In chapter 7, we conclude our thesis, and suggest directions for future work.

2 | Literature Review

2.1 Background

2.1.1 Emergence and Complexity of Cyber Threats

The digital landscape has evolved rapidly, leading to a significant increase in the volume and complexity of cyber threats. Cyber-attacks range from data breaches and ransomware to advanced persistent threats that exploit vulnerabilities in critical infrastructure. The impact of these attacks is profound, affecting individuals, corporations, and governments globally [10]. Understanding the nature and dynamics of these threats is crucial for developing effective defence mechanisms.

Cyber threats are increasingly sophisticated and diverse. Attack vectors include malware, phishing, adversarial attacks, and deepfakes. The Center for Strategic and International Studies (CSIS) highlights that cybercrime costs the global economy approximately $600 billion annually [28]. Additionally, nation-state actors have been implicated in cyber-espionage and cyber-warfare, further complicating the security landscape [29].

Traditional methods for cyber-attack detection are primarily reactive, relying on signature-based systems that compare current activity against known threat patterns [30]. These methods are limited in their ability to detect novel or evolving threats. Recent advances in ML offer promising alternatives by enabling predictive models that can identify potential threats before they manifest [31]. Despite these advances, existing forecasting approaches often lack the capacity to predict long-term trends in cyber threats effectively [6].

One major challenge in cyber threat forecasting is the dynamic nature of the threat landscape. Attack techniques and patterns evolve rapidly, making it difficult for static models to remain effective over time. Additionally, the reliance on historical data can introduce biases, leading to inaccurate predictions [32]. Another significant challenge is the integration of diverse data sources, such as news articles, research papers, and social media, which contain valuable but unstructured information that is often difficult to process and analyse [6].

To address these challenges, there is a growing need for advanced predictive models that can incorporate big data analytics and leverage the power of ML to forecast long-term trends in cyber threats. Such models can help in anticipating future attack vectors and improving the proactive defence strategies of organisations and nations. However, it is also essential to first consider the human element in cyber security responses. Theoretical frameworks can offer valuable insights into how individuals perceive and react to these evolving threats [33], complementing the analytical approaches.

2.1.2 Theoretical Framework

The Protection Motivation Theory (PMT) serves as a robust theoretical framework for understanding individuals' responses to perceived threats and their adoption of protective measures [34]. Originating from the field of health psychology, PMT has been widely applied in various domains, including cyber security [33], to elucidate the cognitive processes underlying risk perception and risk management strategies.

Proposed by Rogers in 1975, PMT was initially developed to explain how individuals respond to health-related threats, such as illness or disease [35]. Building upon earlier theories of fear appeals and cognitive appraisal, PMT posits that individuals engage in protective behaviours when they perceive a threat to be sufficiently severe and when they believe that recommended actions are effective in reducing that threat. Over the years, PMT has evolved to encompass a broader range of threats, including those posed by cyber-attacks and online security breaches [36].

At the core of PMT lies the concept of threat appraisal, wherein individuals assess the severity and vulnerability associated with a threat. In the context of cyber security, this involves analysing historical data and current trends to evaluate the evolving landscape of cyber threats. Unfortunately, experience has shown that human experts tend to show poor inter-rater agreements when exposed to raw data [20]. On the other hand, leveraging big data analytics enables the quantification of the severity and likelihood of various cyber-attacks, ranging from malware infections to sophisticated phishing campaigns [6, 14]. Through this comprehensive threat assessment, emerging patterns can be identified, facilitating anticipation of future cyber threat trends.

PMT emphasises individuals' evaluation of the efficacy of available coping strategies in mitigating perceived threats. In the realm of cyber security, coping strategies encompass a wide array of technological measures and behavioural interventions. Existing literature explores the effectiveness of cyber security technologies, such as Intrusion Detection Systems (IDS) and encryption protocols, in combating identified cyber threats [37, 38]. Additionally, research delves into the role of user education and training programmes in enhancing cyber security awareness and promoting safe online behaviours [39]. By examining the perceived effectiveness of these coping strategies, insights can be gained to inform the development of targeted interventions to bolster cyber-defence mechanisms.

Adopting a data-driven version of PMT as the basis for our research framework provides a comprehensive understanding of individuals' responses to cyber threats. By assessing the intensity of future threats and bridging gaps with relevant mitigation technologies through continuous evaluation and informed investments, individuals will have confidence in the effectiveness of security measures and will engage in proactive actions to bolster cyber security preparedness and resilience.

Having explored the theoretical foundations of PMT and its application in understanding responses to cyber threats, it becomes evident that analysing historical data is crucial for effective threat assessment. Transitioning to practical methodologies, time series analysis emerges as a pivotal tool in this domain [14]. By studying temporal patterns and behaviours within data, such as cyber incident logs or attack frequencies, analysts can uncover underlying trends and anticipate future threats.

2.1.3 Time Series Analysis

A time series is a sequence of data points, typically measured or recorded at successive points in time, and is represented in chronological order. Time series data is collected or recorded over a continuous period, and each observation in the series corresponds to a specific time point [40].

In the context of various fields such as statistics, econometrics, signal processing, and ML, time series analysis involves studying and modelling the patterns, trends, and behaviours within the data to make predictions or gain insights into its underlying structure. Time series data is commonly encountered in diverse domains, including finance, economics, weather forecasting, stock market analysis, signal processing, and cyber security.

The fundamental components of time series analysis include:

- Trend: The long-term movement or tendency in the data, indicating whether it is increasing, decreasing, or remaining stable over time.
- Seasonality: The repeating patterns or fluctuations that occur at regular intervals, often related to specific seasons, months, days of the week, or other calendar-based factors.
- Cyclic Patterns: Longer-term fluctuations or cycles that are not strictly tied to specific calendar intervals.
- Irregular or Random Components: The unpredictable and random variations in the data that are not explained by the trend, seasonality, or cycles.

2.1.4 Univariate and Multivariate Time Series

In the context of time series analysis, a univariate time series involves a single variable or data stream observed or measured over a period of time [41]. The primary focus is on understanding and forecasting the behaviour of this single variable. Common examples include stock prices, temperature readings, or daily sales figures. Analysis of univariate time series often includes identifying trends, seasonality, and irregular fluctuations.

On the other hand, a multivariate time series involves multiple variables or data streams that are observed simultaneously over time [41]. Each variable in the series is interrelated, and the goal is to model the complex dependencies and interactions between them. Multivariate time series are prevalent in various fields, such as economics, environmental science, and finance. Examples include studying the simultaneous movements of stock prices, exchange rates, and interest rates.

In summary, while univariate time series focuses on a single variable's temporal behaviour, multivariate time series extends the analysis to encompass the interconnected dynamics of multiple variables, providing a more comprehensive understanding of the underlying system.

2.1.5 Time Series Forecasting

Given a time series of data, various prediction algorithms can be employed to forecast the future trends and patterns of the time series. The first type of prediction algorithms are statistical algorithms such as Autoregressive Integrated Moving Average (ARIMA), which has been well studied in the literature [14, 15]. The more recent approaches apply ML and deep learning algorithms to forecast the future trend, such as Long Short-Term Memory (LSTM). LSTM is a recurrent neural network that has feedback connections and can process an entire sequence of data [6, 21]. Compared to statistical methods, ML can capture complex relationships between different time series data to provide more accurate predictions. It is also possible to utilise a hybrid-approach to achieve the prediction with the lowest error [42].

2.1.6 Time Series as a Feature

In the context of ML, a multivariate time series can be conceptualised as a collection of features, with each variable in the time series representing a distinct feature [27]. In a multivariate time series, different variables (features) are observed simultaneously over successive time points [43]. Each of these variables encapsulates specific information about the system being studied. For instance, in the context of cyber security, a multivariate time series could consist of two distinct variables or time series: one representing the number of cyber-attacks observed over time, and the other reflecting rates of wars and conflicts [6]. Each of these time series serves as a feature contributing to the overall understanding of the cyber security landscape in conjunction with geopolitical events. For instance, the "Number of Attacks" time series might capture the frequency and intensity of cyber-attacks on a network, while the "Wars and Conflicts Rates" time series could provide information about geopolitical instability. Analysing these multivariate time series collectively allows for exploring potential correlations or dependencies between cyber threats and global events. ML models leveraging these features could uncover patterns and associations that enhance the ability to predict or respond to cyber security incidents in the context of broader geopolitical dynamics.

When devising an ML-based method, one can rely on manual feature identification and engineering, or try and learn the features from raw data. In the context of cyber incidents, there are many factors (*i.e.*, potential features) that could lead to the occurrence of an attack. Wars and political conflicts between countries often lead to cyberwarfare [10, 44]. The number of mentions of a certain attack appearing in scientific articles may correlate well with the actual incident rate. Also, cyber-attacks often take place on holidays, anniversaries and other politically significant dates [5]. Finding the right features out of unstructured big data is one of the key objectives of our research.

2.1.7 Multivariate Time Series as a Graph

In scenarios involving intricate datasets, particularly when the intricate relationships among time series data can be explicitly modelled in a graph structure, a highly effective strategy is the utilisation of GNNs for forecasting multivariate time series data. The application of GNN models, as demonstrated in diverse domains [27], proves to be a promising avenue for significantly improving prediction performance.

The inherent capacity of GNNs to capture complex dependencies and patterns within graph-structured data aligns with the nuanced nature of cyber security trends. Leveraging this approach not only enhances the accuracy of predictions but also allows for a more comprehensive exploration of the underlying relationships between nodes in the cyber trend graph. This exploration, facilitated by GNNs, contributes to the uncovering of hidden connections, enabling decision makers to gain valuable insights into the dynamics of cyber trends.

2.1.8 Bayesian Modelling

Bayesian models are a powerful framework for reasoning under uncertainty, rooted in Bayes' theorem, which provides a systematic way to update beliefs based on new evidence [45]. At its core, Bayesian inference revolves around estimating the probability of hypotheses given observed data. In this paradigm, probabilities are not just measures of uncertainty, but also express degrees of belief. This makes Bayesian models particularly suited for tasks where uncertainty plays a critical role, such as decision-making under ambiguity or prediction in noisy environments.

In the context of deep learning, researchers often turn to Bayesian models to tackle several challenges. One key motivation is the ability to incorporate prior knowledge into the learning process [46]. By encoding prior beliefs about the parameters of a model, Bayesian approaches can help regularise learning, especially in scenarios with limited data. Additionally, Bayesian models offer a principled framework for uncertainty quantification, which is crucial for making reliable predictions in safety-critical applications.

Unfortunately, exact Bayesian inference in complex models is often computationally intractable due to the high-dimensional parameter space and non-convexity of the likelihood function. As a result, researchers resort to approximations, such as variational inference or Markov chain Monte Carlo methods, to make Bayesian modelling feasible in deep learning settings [46]. These approximations strike a balance between computational efficiency and maintaining the benefits of Bayesian reasoning, enabling scalable and effective probabilistic modelling in modern ML applications.

Within the field of deep learning for cyber security, the Monte Carlo dropout method offers a practical means to approximate Bayesian models [47]. Monte Carlo dropout leverages dropout regularisation, a technique commonly used in deep neural networks to prevent overfitting, in order to estimate model's uncertainty. By performing multiple forward passes with dropout enabled during inference, Monte Carlo dropout generates a distribution of predictions, allowing for the quantification of uncertainty in the model's output. This epistemic uncertainty estimation is crucial in cyber security applications, where decisions must be made under data scarcity, incomplete information, and evolving threats. By leveraging Bayesian modelling techniques like Monte Carlo dropout, deep learning systems can enhance their robustness and reliability in detecting and defending against cyber-attacks.

2.2 Cyber Threat Forecasting

We categorise cyber threat forecasting based on the prediction timeframe into three main categories. These are long-term (years ahead) [6], midterm (months ahead) [17, 48], and short-term (hours ahead) [49, 50]. The practicality and usefulness of each category depend on the specific objectives and the context of cyber security efforts. Each category has its own set of advantages and challenges as outlined below:

- Long-term Predictions (Years Ahead): Long-term predictions are crucial for strategic planning, policy formulation, and setting cyber security standards. They help organisations and governments anticipate major trends in cyber threats, such as the rise of new types of malware or attack vectors, enabling proactive development of defence mechanisms [6]. By understanding potential future threats, organisations can allocate resources more effectively, investing in the development of new technologies, training, and infrastructure improvements that will be most relevant in the face of anticipated threats. Yet, a primary challenge with long-term predictions is the high level of uncertainty. The cyber threat landscape evolves rapidly due to technological advancements, changes in attacker tactics, and geopolitical developments [17]. Long-term predictions may become outdated quickly, requiring continuous monitoring and adjustment.
- Midterm Predictions (Months Ahead): Midterm predictions are valuable for operational planning, including the deployment of specific security measures, conducting targeted training sessions, and performing security drills or simulations based on anticipated attack scenarios. Organisations can adjust their security postures based on midterm predictions, fine-tuning firewalls, intrusion detection systems, and response protocols to guard against expected threats

[51]. However, the accuracy of midterm predictions can be affected by sudden changes in attacker behaviour or the emergence of unforeseen vulnerabilities [6]. These predictions require a balance between specificity and adaptability.

- Short-term Predictions (Hours Ahead): Short-term predictions are critical for immediate threat detection and response. They can enable real-time security measures, such as blocking an imminent attack or isolating affected systems to prevent the spread of malware. Predictions over shorter timeframes can be more precise and actionable, leveraging real-time data analytics and ML models to identify and respond to threats as they emerge [49]. Nevertheless, short-term predictions require extensive monitoring and data analysis capabilities. The high volume of false positives and the need for rapid, automated decision-making systems can be challenging to manage [6].

While we are relatively successful in detecting and classifying cyber-attacks when they occur [52, 53, 54], there has been a much more limited success in predicting them. Most of the existing studies on cyber threat forecasting focus on predicting the attacks in the short or midterm [14, 15, 16, 17, 48, 49, 50, 51, 55, 56, 57], such as predicting the number or source of attacks within a few hours, days, or months. The majority of this work performs the prediction in restricted settings (*e.g.*, against a specific entity or organisation) where historical data are available [17, 51, 55]. Forecasting attack occurrences has been attempted by using statistical methods, especially when parametric data distributions could be assumed [14, 16], as well as by using ML models [15]. Other methods adopt a Bayesian setting and build *event graphs* suitable for estimating the conditional probability of an attack following a given chain of events [56]. Such techniques rely on libraries of predefined attack graphs: they can identify the known attack most likely to happen, but are helpless against never-experienced-before, *zero-day* attacks.

Other approaches aim to identify potential attackers by using network entity reputation and scoring [50]. A small but growing body of research explores the fusion of heterogeneous features (warning signals) to forecast cyber threats using ML. Warning signs may include the number of mentions of a victim organisation on Twitter [17], mentions in news articles about the victim entity [55], and digital traces from dark web hacker forums [15]. Table 2.1 provides a summary for the related work on cyber threat forecasting and highlights our contribution.

Forecasting the cyber threats that will most likely turn into attacks in the long term is of significant importance. It not only gives to cyber security agencies the time to evaluate the existing defence measures, but also assists them in identifying areas where to develop preventive solutions. Long-term prediction of cyber threats, however, still relies on the subjective perceptions of human security experts [18, 19]. Unlike a fully automated procedure based on quantitative metrics, the human-based approach is prone to bias based on scientific or technical interests [20]. Also, quantitative predictions are crucial to scientific objectivity [21]. In summary, we highlight the following research gaps, which will be addressed in chapter 3:

- Current research primarily focuses on detecting (*i.e.*, reactive) rather than predicting cyber-attacks (*i.e.*, proactive).
- Available predictive methods for cyber-attacks are mostly limited to short or midterm predictions.
- Current predictive methods for cyber-attacks are limited to restricted settings (*e.g.*, a particular network or system).
- Long-term prediction of cyber-attacks is currently performed by human experts, whose judgement is subjective and prone to bias and disagreement.

2.3 Technology Forecasting

Many of the existing work on technology forecasting rely on the judgement of human experts or adopt semi-automated approach [58, 59, 60, 61]. The early work forecasted future generations of tools and technologies based on human imagination and creativity [58]. This is possible by exploring the idea that many of the high technology products we use today were once conceptualised in science fiction before becoming a reality through technological advancements. In [59], a technology ecosystem model

Table 2.1: Cyber Threat Forecasting - Literature review summary and our contribution

Ref.	**Problem**	**Detection/ Prediction**	**Forecast Period**	**Forecast Coverage**	**Methods**
[53]	Detect different types of attacks	Detection	N/A	N/A	Feature extraction, deep reinforcement learning
[54]	Malicious traffic detection	Detection	N/A	N/A	Deep neural network with attention mechanism
[14, 16]	Forecast attack count	Prediction	1-7 days	Multiple targets	ARIMA model
[17]	Forecast attack count	Prediction	Months	Organisation	Unconventional signals, lagged feature selection, concept drift training
[55]	Forecast attack motivation and opportunity	Prediction	1 week	1 target	Social media analysis, SVM, CNN
[15]	Forecast attack count	Prediction	1 week or month	Organisation	Digital traces, ARIMA, ARIMAX, LSTM
[56]	Predict next attack in the chain	Prediction	N/A	1 target	Bayesian network
[49]	Predict intrusion detection alerts	Prediction	Minutes or hours	Organisation	Stream processing, sequential rule mining
[48]	Forecast if a data breach will occur	Prediction	Months	Organisation	Externally measurable features, Random Forest
[57]	Reconnaissance detection	Detection	N/A	N/A	LSTM, CNN
[51]	Forecast if a machine will be infected	Prediction	Months	Machine	Binary file analysis, semi-supervised learning
[50]	Forecast if an IP address will attack	Prediction	24 hours	N/A	Entity reputation and scoring, decision trees
Ours	**Forecast cyber-attack trends**	**Prediction**	**3 years**	**36 countries**	**Big data, multivariate time series analysis, LSTM, GNN**

was introduced, which offers analysts a tool to navigate the intricate relationships among technologies. The model aids in dissecting the interplay of various factors influencing technological change, enhancing technology forecasts, investments, and development decisions. More recently, Li *et al.* [60] proposed a framework that utilises scientific papers and patents as data sources, while incorporating text mining and expert judgement techniques to predict technology trends.

The Gartner Hype Cycle (GHC) is a graphical representation and methodology that helps organisations understand the maturity and adoption of technologies over time [61]. It was developed by the research and advisory firm Gartner, Inc [62]. The GHC is based on the premise that technologies go through predictable stages of enthusiasm, disillusionment, and eventual adoption. It was derived from observing and analysing the patterns of technology adoption and understanding how people perceive and adopt new technologies. Gartner's analysts study the life cycle of various technologies, their visibility, and their market expectations to position them on the GHC. Table 2.2 summarises the existing work on technological forecasting compared to our approach. In summary, we highlight the following research gaps, which will be addressed in chapter 4:

- Most of the existing approaches to technology forecasting are not fully automated.
- There is a lack of research specific to the forecast of cyber threat related technologies.
- There is a lack of employment for the big data and GNNs to forecast cyber threat related technologies.

2.4 Time Series Forecasting with Graph Neural Networks

Time series forecasting with GNNs has been heavily applied in the domain of traffic prediction [43, 63]. In [63], a deep learning framework called Spatio-Temporal Graph

Table 2.2: A comparison between existing approaches to technology forecasting and our approach

Ref.	Problem Domain	Approach	Methods
[58]	Forecast future generations of tools and technologies	Human-based	Human imagination and creativity
[59]	Understand evolution in technology ecosystems	Human-based	Navigating the complex relationships among technologies
[60]	Forecast technology trends	Semi-automated	text mining and expert judgement
[61]	Understand the maturity and adoption of technologies over time	Human-based	Observation and analysis by human expert
Ours	**Forecast the trend of cyber threat related technologies**	**Machine-based**	**Big data, multivariate time series analysis, GNN**

Convolutional Networks (STGCN) was developed for learning spatio-temporal correlations by modelling multi-scale traffic networks. In [43], dynamic spatial temporal correlations were studied through the use of spatial-temporal attention mechanism. Other methods jointly learn inter-series correlations and temporal dependencies in the spectral domain, by combining Graph Fourier Transform (GFT) and Discrete Fourier Transform (DFT) [64].

A recent study introduced a generic GNN model for forecasting multivariate time series data [27], applicable across various domains. The model includes a graph learning layer capable of learning the hidden adjacency matrix in the graph using latent representation of nodes. In addition, the model includes temporal convolution modules and graph convolution modules interleaved with each other for learning both the temporal and the spatial dependencies in the graph. The model was evaluated on multiple datasets and was shown to be effective compared to the state-of-the-art baselines. In our work, we propose the Bayesian variation of this model which expresses the model uncertainty, and apply the model in the cyber security domain. Table 2.3 summarises the existing GNN models for time series forecasting and highlights our contribution. In summary, we highlight the following research gaps, which will be addressed in chapter 4:

- There is a shortage of research on employing GNNs in the field of cyber security trend forecasting.

- Most of the current GNN models do not express the epistemic uncertainty, which can be informative for decision makers.

Table 2.3: A comparison between existing GNN models for time series forecasting and our model

Ref.	Model	Description	Domain
[63]	Spatio-Temporal Graph Convolutional Network (STGCN)	Learns spatio-temporal correlations by modelling multi-scale traffic networks.	Traffic
[43]	Attention based Spatial-Temporal Graph Convolutional Network (ASTGCN)	Learns dynamic spatial temporal correlations using spatial-temporal attention mechanism.	Traffic
[64]	Spectral Temporal Graph Neural Network (StemGNN)	Learns inter-series and temporal dependencies in the spectral domain using GFT and DFT.	Traffic, energy, ECG
[27]	Multivariate Time-series Graph Neural Network (MTGNN)	Jointly learns the adjacency matrix and the spatial and temporal dependencies.	Traffic, energy, exchange
Ours	**Bayesian Multivariate Time-series Graph Neural Network (B-MTGNN)**	**Jointly learns the adjacency matrix and the spatial and temporal dependencies, and expresses model uncertainty.**	**Cyber security**

2.5 Explainability

2.5.1 Categorisation

Explainability in ML models has garnered significant attention due to its critical role in understanding model decisions and building trust in AI systems. In this section, we review two prominent categories of explainability, namely correlational and causal explainability [65, 66].

Correlational explainability focuses on identifying statistical correlations between input features and model predictions. These approaches often employ techniques such as feature importance scores [67], Shapley Additive Explanations (SHAP) [68], and partial dependence plots [69] to quantify the influence of each feature on the model output. While correlational explainability provides valuable insights into how features are associated with predictions, it falls short in elucidating the underlying causal relationships between variables [65].

On the other hand, causal explainability aims to uncover the causal mechanisms driving model predictions [65]. Unlike correlational approaches, causal methods attempt to discern cause-and-effect relationships within the data, enabling more nuanced interpretations of model behaviour. Within causal explainability, two main subcategories emerge: adhoc causal models and inherently explainable models.

Adhoc causal models involve retrofitting existing ML models with causal inference techniques to infer causal relationships from observational data [70]. Techniques such as causal mediation analysis, propensity score matching, and instrumental variable analysis are commonly used to estimate causal effects from observational datasets. While adhoc causal models offer a practical approach to inferring causality from observational data, they often require strong assumptions and careful consideration of confounding factors, limiting their applicability in complex real-world scenarios.

Inherently explainable models, on the other hand, are designed from the ground up to incorporate causal reasoning into the model architecture [71]. These models leverage causal inference principles, such as structural causal models and do-calculus, to explicitly model causal relationships between variables. Examples of inherently explainable models include causal Bayesian networks, structural equation models, and counterfactual-based models. By explicitly modelling causal mechanisms, inherently explainable models offer greater transparency and interpretability compared to adhoc causal models.

Despite the promise of causal explainability, the field is still in its infancy, and several challenges remain. First, causal inference from observational data is inherently fraught with methodological challenges, such as confounding bias, selection bias, and measurement error, which can undermine the validity of causal conclusions [72]. Second, developing inherently explainable models requires a deep understanding of causal inference theory and may entail computational complexity and scalability issues [73]. Finally, the adoption of causal explainability in real-world applications necessitates interdisciplinary collaboration between ML researchers, statisticians, and domain experts to ensure the robustness and validity of causal inferences.

Overall, while causal explainability holds great promise for enhancing the transparency and interpretability of ML models, its adoption in practice is still premature. Further research is needed to address methodological challenges, develop scalable computational techniques, and validate causal conclusions in real-world settings. By investing in the exploration of causal explainability, researchers can advance the state-of-the-art in ML explainability and pave the way for more trustworthy and accountable AI systems.

2.5.2 Graph Neural Networks

Explainability in GNNs involves elucidating how the model utilises input graph structures and node features to arrive at specific predictions or classifications. Interpretability techniques aim to reveal the importance and contributions of individual nodes, edges, or graph substructures in influencing the model's output. This understanding is essential for users, stakeholders, or domain experts who seek insights into why a GNN makes particular predictions, as well as for addressing concerns related to bias, fairness, or ethical considerations in the decision-making process.

Various correlational explainability methods are adapted or developed specifically for GNNs, such as graph attention mechanisms [74], influence analysis [75], and saliency maps [76]. These techniques shed light on the graph elements that have the most impact on the model's predictions, providing transparency and accountability. The explainability of GNNs is particularly crucial in applications such as drug discovery, fraud detection, or social network analysis, where comprehensible and trustworthy decisions are paramount. As GNNs continue to be employed in diverse domains, ongoing research focuses on advancing and standardising techniques for enhancing the interpretability of these powerful graph-based models.

The most straightforward and widely accepted method for the explainability of time series forecasting with GNNs is *contrastive gradient-based saliency maps* [77]. This technique involves differentiating the model output with respect to the input, forming a heat map where the norm of the gradient highlights the relative importance of input variables. The resultant gradient in the input space aligns with the direction of the maximum positive rate of change in the model output. Consequently, negative values in the gradient are disregarded to selectively preserve input components that make a positive contribution to the solution.

The attention mechanism has gained significant interest lately, demonstrating its effectiveness across a multitude of fields [43, 78, 79]. In [43], dynamic spatial temporal correlations were studied while using GNN through the use of spatial-temporal attention mechanism, applied in the field of traffic forecasting.

To explore the role of attention mechanism in such model and enhance its interpretability, it is possible to select a sub-graph comprising 10 nodes and obtain the average spatial attention matrix among these nodes in the training set [43]. As depicted on the right side of Figure 2.1, each row in the spatial attention matrix reflects the correlation strength between a given node and the corresponding node in that row. For instance, examining the last row reveals a close relationship between data flows on the 9th node and those on the 3rd and 8th nodes. This observation aligns with spatial proximity in the real network, illustrated on the left side of Figure 2.1. Consequently, the model exhibits a noteworthy advantage in terms of interpretability.

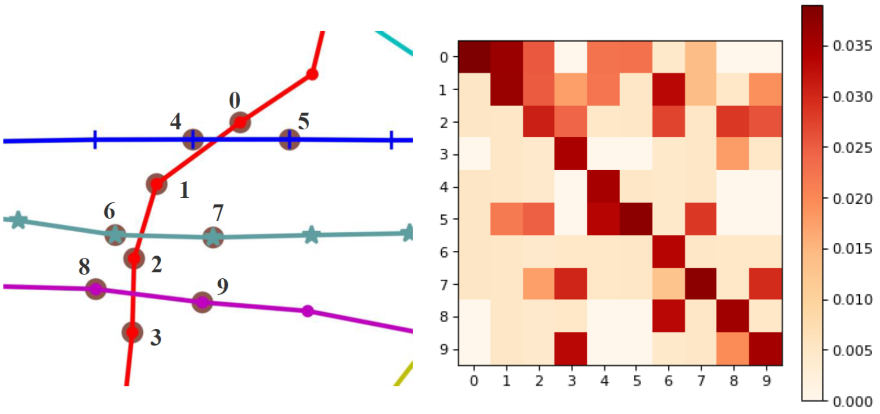


Figure 2.1: The attention matrix obtained from the spatial attention mechanism [43].

2.5.3 Cyber Security Trend Forecasting

In the context of cyber security trend forecasting, the explainability provides decision makers with a transparent view into the model's learning process. By revealing the knowledge acquired by the model, decision makers can discern the intricate web of relationships among cyber trend nodes. This transparency empowers decision makers to make more informed and strategic decisions, as they gain a deeper understanding of the factors influencing cyber security trends.

It is also essential for an effective automation approach to incorporate such explainability techniques into the forecasting process. This involves interpreting the decisions made by ML models, which is crucial for bridging the gap between human-based and machine-based forecasting methodologies [80]. Achieving this transparency is crucial for creating a completely automated approach that can effectively replace human judgement. However, while human experts can explain and justify their decisions, ML models often operate as black boxes, making it challenging to understand their decision-making process unless it is unravelled. Thus, addressing the issue of explainability is vital for filling the gaps in cyber security trend forecasting.

It is also imperative to acknowledge the advantages of machine-based explainability over human-based explainability in the context of our research. While human-based explanations rely on subjective judgments and expert opinions, which can vary widely and introduce bias, machine-based approaches offer objectivity and quantitative analysis [81]. By leveraging data-driven techniques, machine-based explainability provides transparent insights into the decision-making process of predictive models. This objectivity ensures consistency and reliability in forecasting, as it is based on empirical evidence rather than individual perspectives. Additionally, machine-based explainability allows for scalability and automation, enabling the analysis of large datasets and the rapid adaptation to evolving cyber threats. Thus, embracing machine-based explainability enhances the effectiveness and robustness of cyber security trend forecasting methodologies. In chapter 5, we will explore explainability features that we incorporate into our GNN model.

2.6 Summary and the Road Ahead

In this chapter, our literature review uncovered significant gaps in current research within the field of cyber security trend forecasting. We emphasised the prevalent focus on reactive rather than proactive threat detection strategies, signalling a critical need for automated long-term forecasting approaches. Furthermore, the lack of research on forecasting the disparity between threats and PATs and the employment of GNNs in this context have become apparent. Moreover, the demand for explainable models alongside effective uncertainty quantification mechanisms has emerged as a crucial area for further investigation.

In the next chapter, we will address part of these gaps by introducing a proactive and long-term approach to cyber threat forecasting. Our objective will be to predict cyber-attack trends up to three years in advance, utilising diverse big data sources and Bayesian ML models. We will also present further contributions including threat categorisation and the development of the first threat cycle model.

3 | Cyber Threat Forecasting

3.1 Background

Cyber threats have become a pervasive and persistent concern in today's interconnected digital landscape [82]. As technology continues to advance, so does the sophistication and diversity of cyber threats. Organisations, governments, and individuals face an ever-evolving landscape of malicious actors, ranging from individual hackers to well-funded and organised cybercriminal groups [10, 83].

The need for effective cyber threat forecasting has never been more critical, especially when considering long-term planning on a scale of years in advance [6]. Forecasting cyber threats involves the analysis and prediction of potential malicious activities in the digital world. It serves as a proactive measure to anticipate and mitigate potential risks before they materialise into actual attacks. The dynamic nature of cyber threats, coupled with the speed at which technology evolves, necessitates advanced methodologies and frameworks to stay ahead of potential risks.

Traditional approaches to cyber security often involve reactive measures, responding to incidents after they occur [84]. However, with the increasing frequency and complexity of cyber-attacks, organisations are recognising the importance of adopting proactive strategies [5]. Cyber threat forecasting emerges as a key component of proactive cyber security, enabling stakeholders to anticipate, prepare for, and counteract potential threats.

One of the fundamental challenges in cyber threat forecasting is the sheer volume and diversity of data generated in the digital domain. Threat intelligence sources, including incident reports, social media, and open-source intelligence, contribute to a vast and unstructured dataset [10]. Effectively harnessing this data requires advanced analytical frameworks that can extract meaningful insights, identify patterns, and correlate information across disparate sources.

In recent years, ML and artificial intelligence have played a pivotal role in enhancing cyber threat forecasting capabilities [85]. These technologies enable the development of models that can analyse historical data, detect emerging patterns, and predict potential future threats. From anomaly detection to natural language processing for extracting insights from textual data, ML techniques have become indispensable in the arsenal of cyber security professionals.

Moreover, the interconnectedness of cyberspace with scientific discoveries, geopolitical events, and seasons further complicates the landscape. Cyber threat forecasting must consider not only the incidents but also scientific trends and emerging technologies, geopolitical developments, social unrest, and the time of the year. Any of these factors may influence the motivations and tactics of threat actors [6, 86].

The scientific literature serves as a vital source for discerning patterns and insights into cyber threat occurrences. Academic research papers and articles often contain valuable information regarding emerging cyber threats, attack methodologies, and vulnerabilities [87]. The frequency and prominence of cyber-attacks discussed in scientific literature can serve as an indicator of the evolving threat landscape. Researchers and cyber security professionals contribute to this repository of knowledge by documenting their findings, detailing case studies, and proposing countermeasures. Monitoring the mentions of cyber-attacks in scholarly publications allows for a proactive approach to cyber threat forecasting.

The prominence of an attack in scientific discourse can reflect its severity and potential impact. High-profile attacks are more likely to garner attention from the research community, resulting in increased mentions in academic papers [88]. Conversely, the absence of literature on certain types of cyber threats may indicate their novelty or underrecognition, emphasising the need for heightened vigilance in forecasting these emerging risks. Leveraging data from scientific literature, particularly through advanced techniques like natural language processing and sentiment analysis, enables cyber threat forecasters to gain nuanced insights into the evolving tactics, techniques, and procedures employed by threat actors. This multidimensional understanding enhances the accuracy and relevance of forecasting models, contributing to a more resilient cyber security posture.

Incorporating scientific literature into the cyber threat forecasting framework broadens the scope of available data sources, fostering a holistic and informed approach to risk assessment. Researchers can analyse trends, correlations, and anomalies in the frequency of cyber-attack mentions across diverse academic disciplines, providing a comprehensive view of the cyber threat landscape. This integration of scientific knowledge enhances the adaptability of forecasting models and empowers cyber security professionals to proactively address emerging threats based on the collective intelligence derived from scholarly insights.

Next, wars and conflicts, particularly in the contemporary digital age, play a crucial role in shaping the landscape of cyber threats [89]. The importance of wars and related discourse on social media platforms, such as Twitter, as factors influencing the occurrence of cyber-attacks cannot be overstated. During times of geopolitical tension or armed conflicts, nations and threat actors often leverage cyberspace as a domain for both strategic advantage and retaliation.

The intensity of cyber-attacks tends to escalate in the wake of geopolitical events, as state-sponsored or politically motivated threat actors exploit the chaos or uncertainty to advance their objectives. Tweets and public discourse on platforms like Twitter provide valuable insights into the prevailing sentiments, geopolitical tensions, and potential flashpoints [90]. Monitoring these social media channels can offer early indicators of the likelihood of cyber-attacks, as hostile entities may use online rhetoric as a precursor to their malicious activities [15, 55].

Moreover, the interconnected nature of cyberspace allows for asymmetric warfare, where even smaller entities can inflict significant damage through cyber-attacks. Wars and conflicts serve as catalysts for the development and deployment of advanced cyber capabilities, as nations invest heavily in cyberwarfare strategies to gain a competitive edge. This dynamic interplay between physical conflicts and cyber activities underscores the importance of considering geopolitical events and their digital footprints as integral components of cyber threat forecasting.

Then, understanding the significance of public holidays in the context of cyber threat forecasting is also imperative for comprehensive risk assessment [5, 91]. Holidays, whether national or cultural, often present a unique set of circumstances that can be exploited by malicious actors. During holiday periods, there is a notable surge in online activities, ranging from increased online shopping to heightened social media engagement. This uptick in digital interactions creates an opportune environment for cybercriminals seeking to capitalise on distracted users and overwhelmed cyber security defences. Additionally, holidays may coincide with strategic timing for threat actors, aligning with political events or significant anniversaries. Recognising holidays as potential factors influencing cyber threats allows forecasters to tailor their analyses, anticipate heightened risks during specific periods, and implement targeted cyber security measures to safeguard individuals and organisations during these vulnerable times.

In this context, the integration of diverse data sources, including cyber security incident reports, scientific literature, geopolitical events and social media feeds, and public holidays becomes paramount. A holistic approach to cyber threat forecasting involves not only technical indicators but also the broader context in which these threats unfold [6].

This chapter presents a comprehensive framework for forecasting cyber threats, leveraging a diverse set of data sources and employing advanced ML techniques such as LSTM neural networks. By combining technical indicators with contextual information, the proposed framework aims to provide a more nuanced and accurate prediction of cyber threats, empowering organisations to bolster their cyber security defences in an increasingly complex digital landscape, with a specific emphasis on long-term prediction spanning years into the future.

3.2 Methods

3.2.1 The Framework of Forecasting Cyber Threats

The architecture of our framework for forecasting cyber threats is illustrated in Figure 3.1. As seen in the Data Sources component (l.h.s), our framework utilises various sources of unstructured data. One of our main sources is Hackmageddon dataset [10], which includes massive textual data on major cyber-attacks (approx. 15,334 incidents) dating back to July 2011. We refer to the monthly number of attacks as the *Number of Incidents* (NoI). Also, Elsevier's API gives access to the large corpus of scientific articles and data sets from thousands of sources. Utilising this API, we obtained the *Number of Mentions* (NoM) (*e.g.*, monthly) of each attack that appeared in the scientific publications. During the preliminary research phase, we examined all the potentially relevant features and noticed that wars/political conflicts are highly correlated to the number of cyber-events. These data were then extracted via Twitter API as Armed Conflict Areas/Wars (ACA). Lastly, as attacks often take place around holidays, Python's holidays package was used to obtain the number of public holidays per month for each country, which is referred to as *Public Holidays* (PH).

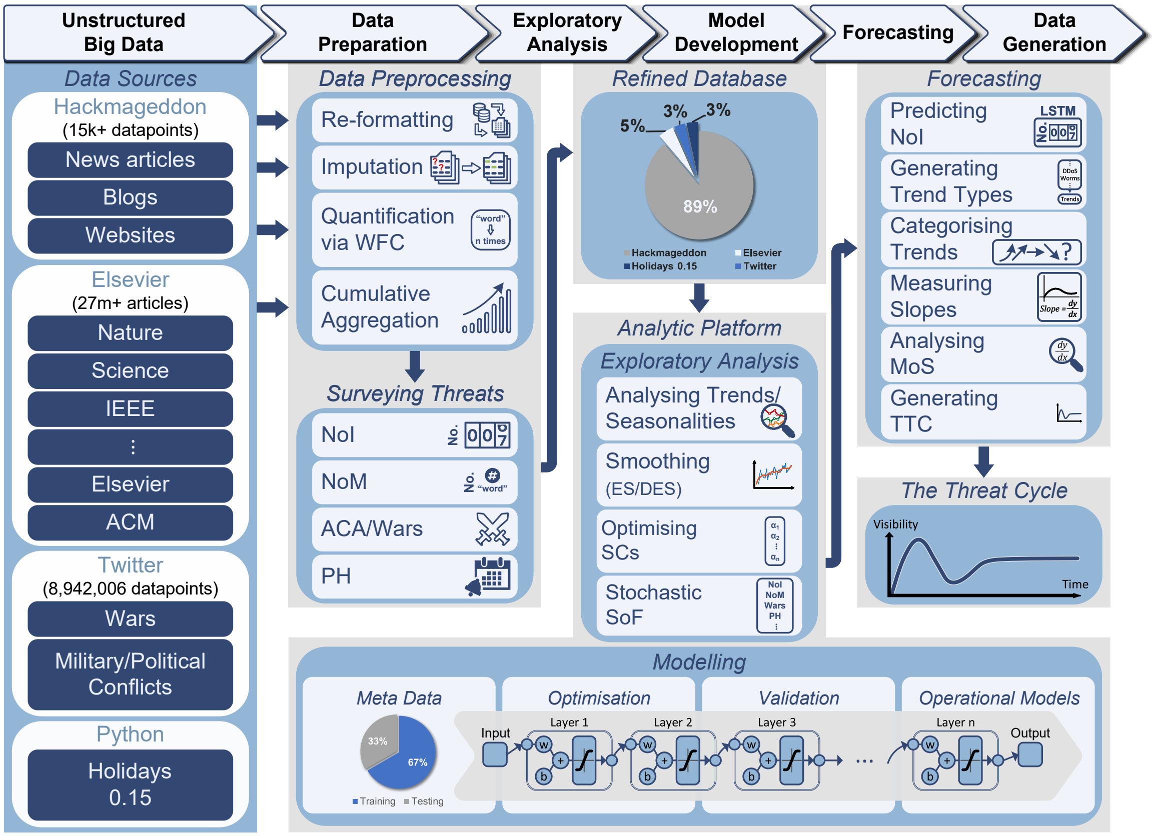


Figure 3.1: The workflow and architecture of forecasting cyber threats. WFC: Word Frequency Counter, NoI: Number of Incidents, NoM: Number of mentions, ACA: Armed Conflict Areas/Wars, PH: Public Holidays, ES: Exponential Smoothing, DES: Double Exponential Smoothing, SCs: Smoothing Constants, SoF: Selection of Features, MoS: Magnitude of Slope, TTC: The Threat Cycle.

In our work, we selected Hackmageddon as the data source for the ground truth (NoI), due to its unique characteristics and advantages. Based on our research, Hackmageddon is the only accessible platform that aggregates major cyber incidents data comprehensively without confidentiality restrictions, addressing a critical challenge in the field where much of the relevant data remains proprietary or classified. This openness allows for transparent and reproducible research. Additionally, Hackmageddon stands out by providing data on a global scale, a feature that is scarce among available datasets. This global coverage ensures a broad representation of cyber threats, enhancing the robustness and generalisability of our analysis. Furthermore, Hackmageddon's extensive daily reporting of incidents offers a rich dataset with a high frequency of data points, enabling us to closely approximate the ground truth of cyber-attacks and enhancing the statistical reliability of our findings. Importantly, Hackmageddon has been extensively studied in the literature and utilised by numerous cyber security researchers [14, 92, 93], further validating its credibility and reliability as a data source for our research.

To ensure the accuracy and quality of Hackmageddon data, we validated it using the statistics from official sources across government, academia, research institutes and technology organisations. For a ransomware example, the Cybersecurity and Infrastructure Security Agency stated in their 2021 trend report that cyber security authorities in the United States, Australia, and the United Kingdom observed an increase in sophisticated, high-impact ransomware incidents against critical infrastructure organisations globally [94]. The WannaCry attack in the dataset was also validated with Ghafur *et al*'s [1] statement in their article: "WannaCry ransomware attack was a global epidemic that took place in May 2017".

For identifying research trends (NoM), we used the Elsevier API, specifically Scopus indexing [22], which grants access to over 27 million research abstracts [25]. This extensive database is crucial for capturing the comprehensive landscape of academic discourse on cyber-attack types, providing a robust dataset that reflects the evolution of research trends over time. This vast and diverse collection of articles makes it an ideal big data source for ML and predictive modelling, allowing for accurate trend analysis and forecasting. The breadth and depth of data available through Scopus enable us to track the frequency and context of mentions related to various cyber threats, ensuring that our research trends are grounded in the most current and extensive scholarly work available.

As for political tensions and conflicts (ACA), Twitter is an optimal choice for data collection due to its widespread use as a platform for public discourse and real-time discussion on global events, including politics [95]. Unlike other social media platforms such as Facebook, which are often used for private, personal conversations, Twitter is recognised for its role in public discussions where users frequently share and debate political events, including wars and international conflicts. This makes it a valuable resource for capturing real-time information and sentiments related to geopolitical issues. Additionally, Twitter's well-documented API provides researchers with the tools to write specific queries and collect large volumes of tweets efficiently [23], making it an accessible and powerful source for extracting data on political tensions and conflicts. This capability allows us to monitor and analyse the dynamics of political discussions and their potential impact on cyber security trends.

An example of an entry in the Hackmageddon dataset is shown in Table 3.1. Each entry includes the incident date, the description of the attack, the attack type, and the target country. Data pre-processing (Figure 3.1) focused on noise reduction through imputing missing values (*e.g.*, countries), which were often observed in the earlier years. We were able to impute these values from the description column or occasionally, by looking up the entity location using Google.

The textual data were quantified via our *Word Frequency Counter*, which counted the number of each attack type per month as in Table 3.2. Cumulative aggregation obtained the number of attacks for all countries combined and an example of a data entry after transformation includes the month, and the number of attacks against each country (and all countries combined) for each attack type. By adding features such as NoM, ACA, and PH, we ended up having additional features that we appended to the dataset as shown in Table 3.3. Our final dataset covers 42 common types of attacks in 36 countries. The full list of attacks is provided in Figure 3.5. The list of countries is given in Table 3.4.

Table 3.1: Hackmageddon data - entry example

Date	**Description**	**Attack**	**Target country**
31/07/2019	Gadsden Independent School District is hit by a ransomware.	Malware	US

Table 3.2: Transformed data - entry example with dummy values

Month	**Malware US**	**Malware UK**	..	**Malware Total**	**DDoS US**	**DDoS UK**	..	**DDoS Total**
July 2019	50	40	..	1000	20	15	..	100

Table 3.3: Additional data features with dummy values

Month	**NoM Malware**	**NoM Ransomware**	**ACA US**	**ACA Total**	**PH UK**	**PH Total**
July 2019	500	400	12,000	100,000	8	400

To analyse and investigate the main characteristics of our data, an exploratory analysis was conducted focusing on the visualisation and identification of key patterns such as trend and seasonality, correlated features, missing data, and outliers. For seasonal data, we smoothed out the seasonality so that we could identify the trend while removing the noise in the time series [96]. The smoothing type and constants were optimised along with the ML model (see Optimisation for details). We applied Stochastic selection of Features (SoF) to find the subset of features that minimises the prediction error, and compared the univariate against the multivariate approach.

For the modelling, we built a Bayesian encoder-decoder LSTM (B-LSTM) network. B-LSTM models have been proposed to predict perfect wave events like the onset of stock market bear periods on the basis of multiple warning signs, each having different time dynamics [97]. Encoder-decoder architectures can manage inputs and outputs that both consist of variable-length sequences. The encoder stage encodes a sequence into a fixed-length vector representation (known as the latent representation). The decoder prompts the latent representation to predict a sequence. By applying an efficient latent representation, we train the model to consider all the useful warning information from the input sequence - regardless of its position - and disregard the noise.

Our Bayesian variation of the encoder-decoder LSTM network considers the weights of the model as random variables. This way, we extract epistemic uncertainty via (approximate) Bayesian inference, which quantifies the prediction error due to insufficient information [98]. This is an important parameter, as epistemic uncertainty can be reduced by better intelligence, *i.e.*, by acquiring more samples and new informative features. Details are provided in the section Bayesian Long Short-Term Memory.

Our overall analytical platform learns an operational model for each attack type. Here, we evaluated the model's performance in predicting the threat trend 36 months in advance. A newly modified symmetric Mean Absolute Percentage Error (M-SMAPE) was devised as the evaluation metric, where we added a penalty term that accounts for the trend direction. More details are provided in the section Evaluation Metrics.

3.2.2 Feature Extraction

Below, we provide the details of the process that transforms raw data into numerical features, obtaining the ground truth NoI and the additional features NoM, ACA and PH.

- NoI: The number of daily incidents in Hackmageddon was transformed from the purely unstructured daily description of attacks along with the attack and country columns, to the monthly count of incidents for each attack in each country. Within the description, multiple related attacks may appear, which are not necessarily in the attack column. Let E_{x_i} denote the set of entries during the month x_i in Hackmageddon dataset. Let a_j and c_k denote the j^{th} attack and k^{th} country. Then NoI can be expressed as follows:

$$NoI(x_i, a_j, c_k) = \sum_{e \in E_{x_i}} Z(a_j, c_k, e) \tag{3.1}$$

where $Z(a_j, c_k, e)$ is a function that evaluates to 1 if a_j appears either in the description or in the attack column of entry e and c_k appears in the country column of e. Otherwise, the function evaluates to 0. Next, we performed cumulative aggregation to obtain the monthly count of attacks in all countries combined for each attack type as follows:

$$NoI(x_i, a_j) = \sum_{k=1}^{K} NoI(x_i, a_j, c_k) \tag{3.2}$$

- NoM: We wrote a Python script to query Elsevier API for the number of mentions of each attack during each month [22]. The search covers the title, abstract and keywords of published research papers that are stored in Scopus database [99]. Let P_{x_i} denote the set of research papers in Scopus published during the month x_i. Also, let W_p denote the set of words in the title, abstract and keywords of research paper p. Then NoM can be expressed as follows:

$$NoM(x_i, a_j) = \sum_{p \in P_{x_i}} \sum_{w \in W_p} U(w, a_j) \tag{3.3}$$

where $U(w, a_j)$ evaluates to 1 if $w = a_j$, and to 0 otherwise.

- ACA: Using Twitter API in Python [23], we wrote a query to obtain the number of tweets with keywords related to political conflicts or military attacks associated with each country during each month. The keywords used for each country are provided in Table 3.5, representing our query. Formally, let T_{x_i} denote the set of all tweets during the month x_i. Then ACA in the country c_k during the month x_i can be expressed as follows:

$$ACA(x_i, c_k) = \sum_{t \in T_{x_i}} Q(t, c_k) \tag{3.4}$$

where $Q(t, c_k)$ evaluates to 1 if the query in Table 3.5 evaluates to 1 given the tweet t and the country c_k. Otherwise, it evaluates to 0.

Table 3.5: Keywords and phrases in the query for obtaining the feature Armed Conflict Areas/Wars (ACA).

	Keyword 1	**Keyword 2**	**Keyword 3**	**Keyword 4**
Phrase 1	c_k	WAR	MILITARY	
Phrase 2	c_k	WAR	ARMED	FORCE
Phrase 3	c_k	CONFLICT	POLITIC	
Phrase 4	c_k	MILITARY	ATTACK	
Phrase 5	c_k	ARMED	FORCE	ATTACK

The keywords are joined by logical AND operator and the phrases are joined by logical OR operator. The variable c_k stands for the k^{th} country. The query was repeated for the 36 countries in the study.

- PH: We used the Python holidays library [24] to count the number of days that are considered public holidays in each country during each month. More formally, this can be expressed as follows:

$$PH(x_i, c_k) = \sum_{d \in x_i} H(d, c_k) \tag{3.5}$$

where $H(d, c_k)$ evaluates to 1 if the day d in the country c_k is a public holiday, and to 0 otherwise. In (3.4) and (3.5), cumulative aggregation was used to obtain the count for all countries combined as in (3.2).

3.2.3 Data Integration

Based on equations (3.1)-(3.5), we obtain the following columns for each month:

- NoI_C: The number of incidents for each attack type in each country (42 × 36 columns) [Hackmageddon].

- NoI: The total number of incidents for each attack type (42 columns) [Hackmageddon].
- NoM: The number of mentions of each attack type in research articles (42 columns) [Elsevier].
- ACA_C: The number of tweets about wars and conflicts related to each country (36 columns) [Twitter].
- ACA: The total number of tweets about wars and conflicts (1 column) [Twitter].
- PH_C: The number of public holidays in each country (36 columns) [Python].
- PH: The total number of public holidays (1 column) [Python].

In the aforementioned list of columns, the name enclosed within square brackets denotes the source of data. By matching and combining these columns, we derive our monthly dataset, wherein each row represents a distinct month. A concrete example can be found in Tables 3.2 and 3.3, which, taken together, constitute a single observation in our dataset. Overall, the process of transforming the unstructured big data into the above numerical format (*i.e.,* monthly trends spanning 11 years) resulted in a total of 129 observations. The dataset can be expanded through the inclusion of other monthly features as supplementary columns. Additionally, the dataset may be augmented with further samples as additional monthly records become available.

3.2.4 Data Smoothing

We tested multiple smoothing methods and selected the one that resulted in the model with the lowest M-SMAPE during the hyper-parameter optimisation process. The methods we tested include exponential smoothing (ES), double exponential smoothing (DES) and no smoothing (NS). Let α be the smoothing constant. Then the ES formula is:

$$S(x_i) = \begin{cases} \alpha D(x_i) + (1-\alpha)S(x_{i-1}), & \text{if } i \geq 1 \\ D(x_0), & \text{otherwise} \end{cases} \tag{3.6}$$

where $D(x_i)$ denotes the original data at month x_i. For the DES formula, let α and β be the smoothing constants. We first define the level $l(x_i)$ and the trend $\tau(x_i)$ as follows:

$$l(x_i) = \begin{cases} \alpha D(x_i) + (1-\alpha)(l(x_{i-1}) + \tau(x_{i-1})), & \text{if } i \geq 1 \\ D(x_0), & \text{otherwise} \end{cases} \tag{3.7}$$

$$\tau(x_i) = \begin{cases} \beta(l(x_i) - l(x_{i-1})) + (1-\beta)\tau(x_{i-1}), & \text{if } i \geq 1 \\ D(x_1) - D(x_0), & \text{otherwise} \end{cases} \tag{3.8}$$

then, DES is expressed as follows:

$$DS(x_i) = \begin{cases} l(x_i) + \tau(x_i), & \text{if } i \geq 1 \\ D(x_0), & \text{otherwise} \end{cases} \tag{3.9}$$

The smoothing constants (α and β) in the aforementioned methods are chosen as the predictive results of the ML model that gives the lowest M-SMAPE during the hyper-parameter optimisation process. Figure 3.2 depicts an example for the DES result.

3.2.5 Bayesian Long Short-Term Memory

LSTM is a type of recurrent neural network (RNN) that uses lagged observations to forecast the future time steps [21]. It was introduced as a solution to the vanishing gradient problem of traditional RNNs [100]. In LSTM, the input is passed to the network cell, which combines it with the hidden state and cell state values from previous time steps to produce the next states. The hidden state can be thought of as a short-term memory since it stores information from recent periods in a weighted manner. On the other hand, the cell state is meant to remember all the past information from

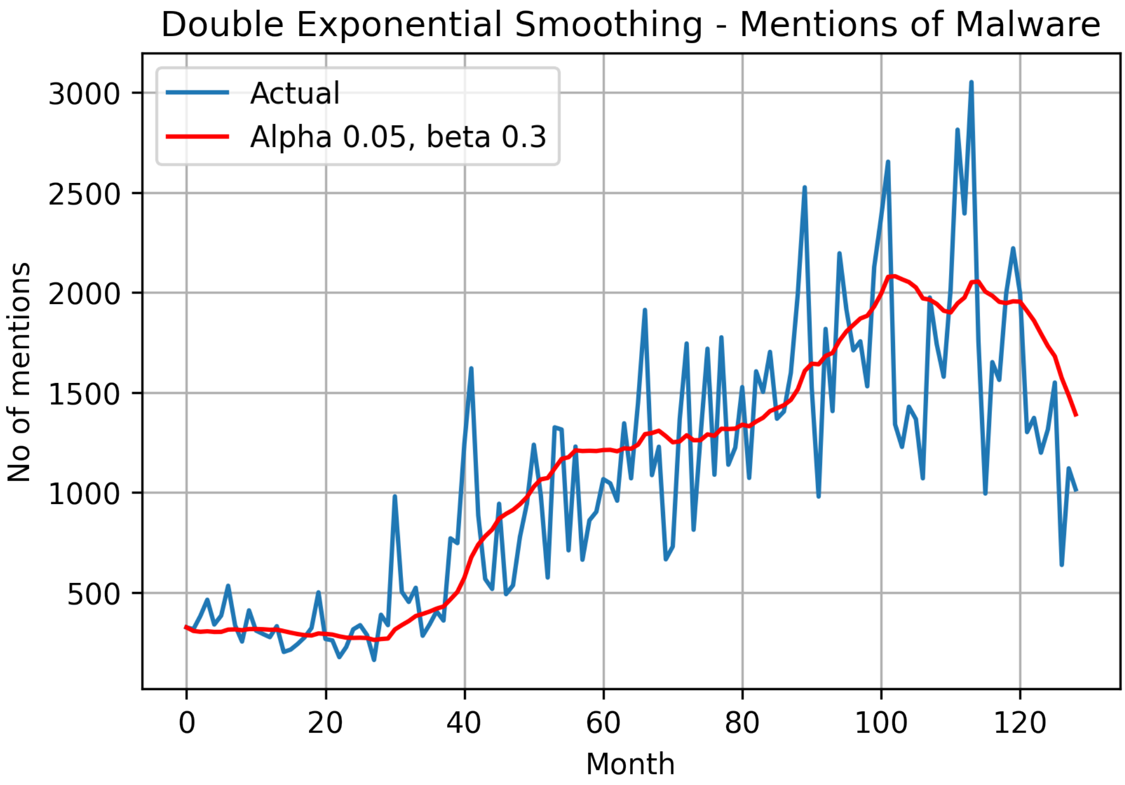


Figure 3.2: Double exponential smoothing (DES) for the number of mentions (NoM) of the malware attack in the scientific literature. The smoothing captures the trend to improve the prediction. The values of alpha and beta are chosen in a such way that the prediction error is minimised.

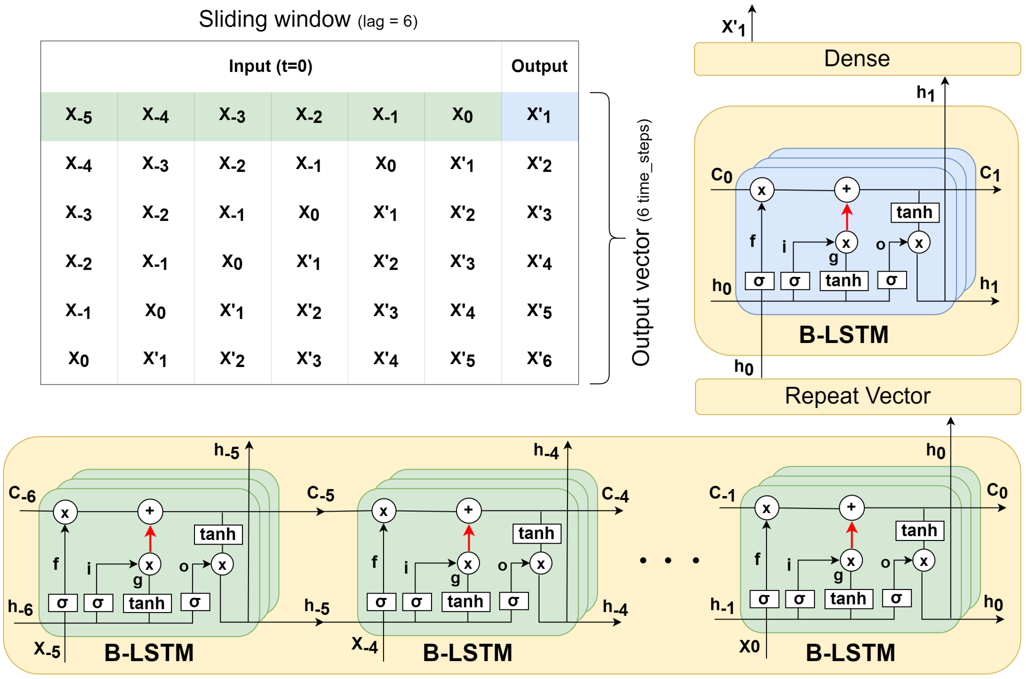
previous intervals and store them in the LSTM cell. The cell state thus represents the long-term memory.

LSTM networks are well-suited for time series forecasting, due to their proficiency in retaining both long-term and short-term temporal dependencies [101, 102]. By leveraging their ability to capture these dependencies within cyber-attack data, LSTM networks can effectively recognise recurring patterns in the attack time series. Moreover, the LSTM model is capable of learning intricate temporal patterns in the data and can uncover inter-correlations between various variables, making it a compelling option for multivariate time series analysis [103].

Given a sequence of LSTM cells, each processing a single time step from the past, the final hidden state is encoded into a fixed-length vector. Then, a decoder uses this vector to forecast future values. Using such architecture, we can map a sequence of time steps to another sequence of time steps, where the number of steps in each sequence can be set as needed. This technique is referred to as *encoder-decoder* architecture.

Because we have relatively short sequences within our refined data (*e.g.*, 129 monthly data points over the period from July 2011 to March 2022), it is crucial to extract the source of uncertainty, known as *epistemic* uncertainty [104], which is caused by lack of knowledge. In principle, epistemic uncertainty can be reduced with more knowledge either in the form of new features or more samples. Deterministic (non-stochastic) neural network models are not adequate to this task as they provide point estimates of model parameters. Rather, we utilise a Bayesian framework to capture epistemic uncertainty. Namely, we adopt the Monte Carlo dropout method proposed by Gal *et al.* [47], who showed that the use of non-random dropout neurons during ML training (and inference) provides a Bayesian approximation of the deep Gaussian processes. Specifically, during the training of our LSTM encoder-decoder network, we applied the same dropout mask at every time step (rather than applying a dropout mask randomly from time step to time step). This technique, known as *recurrent dropout* is readily available in Keras [105]. During the inference phase, we run trained model multiple times with recurrent dropout to produce a distribution of predictive results. Such prediction is shown in Figure 3.5.

Figure 3.3 shows our encoder-decoder B-LSTM architecture. The hidden state and cell state are denoted respectively by h_i and C_i, while the input is denoted by X_i. Here, the length of the input sequence (lag) is a hyper-parameter tuned to produce the optimal model, where the output is a single time step. The number of cells (*i.e.*, the depth of each layer) is tuned as a hyper-parameter in the range between 25 and 200 cells. Moreover, we used one or two layers, tuning the number of layers to each attack type. For the univariate model we used a standard Rectified Linear Unit (ReLU) activation function, while for the multivariate model we used a Leaky ReLU. Additionally, we used recurrent dropout (*i.e.*, arrows in red as shown in Figure 3.3), where the probability of dropping out is another hyper-parameter that we tune as described above, following Gal's method [106]. The tuned dropout value is maintained during inference as previously mentioned. Once the final hidden vector h_0 is produced by the encoder, the Repeat Vector layer is used as an adapter to reshape it from the bi-dimensional output of the encoder (*e.g.*, h_0) to the three-dimensional input expected by the decoder. The decoder processes the input and produces the hidden state, which is then passed to a dense layer to produce the final output.



Sliding window (lag = 6)

Input (t=0)						Output
X_{-5}	X_{-4}	X_{-3}	X_{-2}	X_{-1}	X_0	X'_1
X_{-4}	X_{-3}	X_{-2}	X_{-1}	X_0	X'_1	X'_2
X_{-3}	X_{-2}	X_{-1}	X_0	X'_1	X'_2	X'_3
X_{-2}	X_{-1}	X_0	X'_1	X'_2	X'_3	X'_4
X_{-1}	X_0	X'_1	X'_2	X'_3	X'_4	X'_5
X_0	X'_1	X'_2	X'_3	X'_4	X'_5	X'_6

Figure 3.3: The encoder-decoder architecture of Bayesian Long Short-Term Memory (B-LSTM). X_i stands for the input at time step i, h_i stands for the hidden state, and C_i stands for the cell state. The red arrows indicate a recurrent dropout maintained during inference. The figure shows an example for an input with time lag=6 and a single layer. The table illustrates the concept of sliding window method used to forecast multiple time steps during inference (*i.e.*, using the output at a time step as an input to forecast the next time step).

Each time step corresponds to a month in our model. Since the model is learnt to predict a single time step (single month), we use a sliding window during the prediction phase to forecast 36 (monthly) data points. In other words, we predict a single month at each step, and the predicted value is fed back for the prediction of the following month. This concept is illustrated in the table shown in Figure 3.3. Utilising a single time step in the model's output minimises the size of the sliding window, which in turn allows for training with as many observations as possible with such limited data.

The difference between the univariate and multivariate B-LSTMs is that the latter carries additional features in each time step. Thus, instead of passing a scalar input value to the network, we pass a vector of features including the ground truth at each time step. The model predicts a vector of features as an output, from which we retrieve the ground truth, and use it along with the other predicted features as an input to predict the next time step.

3.2.6 Model Convergence

In the context of big data, the challenge often lies in translating vast amounts of raw, unstructured information into actionable, structured insights. Our dataset originally encompassed thousands of data points spanning various sources, including records of cyber incidents, literature mentions, and tweets and geopolitical events. However, after rigorous data cleaning and transformation processes, we focused on aggregating this information into a more manageable and relevant dataset consisting of 129 monthly observations from July 2011 to March 2022. Each observation encapsulated the number of incidents (NoI) for 42 different attack types and included associated features such as literature mentions, tweets about conflicts, and public holidays for that month.

In the context of ML and statistical modelling, the Law of Large Numbers (LLN) suggests that as the sample size increases, the estimates and predictions become more reliable and closer to the true values [107]. While a commonly used heuristic is to have a sample size of at least 30 observations for reliable estimation, the actual required sample size depends on various factors such as the complexity of the model, the nature of the data, and the desired accuracy of predictions. Given our dataset's scope, which spans 129 months but encompasses multiple variables across 42 attack types, the concern is whether this is sufficient for robust predictive modelling. This challenge is accentuated by the diversity and complexity of cyber-attack types, each requiring tailored analysis and prediction.

Model convergence is crucial in ensuring that the predictions made are stable and reliable. With our data, despite the seemingly limited sample size of 129 observations, we employed several strategies to enhance model convergence. First, we decided to build 42 separate models, one for each attack type, rather than a single comprehensive model. This strategy ensures that each model is tailored to the specific characteristics and patterns associated with each attack type, thereby enhancing prediction accuracy. This approach not only simplifies the modelling process by reducing the dimensionality of the problem but also allows each model to focus on the unique dynamics of its respective attack type. By doing so, we effectively increase the sample size per model in a relative sense, as each model only needs to handle a subset of the overall data, thus improving the statistical power and convergence properties. Each model benefits from a focused set of data points that are directly relevant to the variable it aims to predict, which aligns well with the principles of LLN in terms of ensuring sufficient observations per model.

To further address the challenge of limited data, we used a Bayesian model, approximated using the Monte Carlo dropout method. This technique is particularly well-suited for dealing with smaller datasets as it helps to approximate a Bayesian inference process, effectively quantifying uncertainty in model parameters without explicitly incorporating prior knowledge [47]. Bayesian models, even when approximated, are advantageous because they provide a probabilistic framework that allows for the quantification of uncertainty in predictions [108]. The Monte Carlo dropout method helps achieve this by applying dropout at both training and inference stages, enabling the model to generate multiple predictions that represent a distribution rather than a single deterministic outcome. This helps mitigate the risks associated with smaller sample sizes by capturing a range of possible outcomes, which improves the robustness of predictions.

By approximating a Bayesian model with Monte Carlo dropout, we can better manage overfitting [109], a common issue with smaller datasets, ensuring that our model generalises well to unseen data. This approach does not directly incorporate prior knowledge but leverages the probabilistic nature of Bayesian methods to handle uncertainty effectively, enhancing the model's stability and reliability even with a limited number of samples. This probabilistic handling of uncertainty provides a critical layer of robustness, making our predictions more resilient and less sensitive to the inherent variability in smaller datasets.

The convergence of Bayesian models, including those approximated through Monte Carlo dropout, with small sample sizes is supported by scientific theories that emphasise the value of probabilistic reasoning and the benefits of capturing uncertainty [47, 109]. Studies have shown that these methods can achieve convergence efficiently by providing a probabilistic estimate of the model parameters, thus handling data scarcity more effectively than traditional frequentist approaches. This theoretical backing supports our confidence that, despite the limited sample size, our models are capable of providing reliable and robust predictions by leveraging the strengths of Bayesian approximation techniques.

Overall, the use of Bayesian model, along with our strategy of separating models by attack type, positions us well to overcome the limitations imposed by a small dataset, thus enabling us to make accurate and reliable predictions for future cyber incidents. To ensure that our models are robust and reliable, we continuously monitored their performance using an appropriate evaluation metric, which we describe in detail in the next section. By adopting a rigorous evaluation technique, we ensure that our models are not only theoretically sound but also practically effective in predicting cyber-attack trends.

3.2.7 Evaluation Metrics

The evaluation metric SMAPE is a percentage (or relative) error based accuracy measure that judges the prediction performance purely on how far the predicted value is from the actual value [110]. It is expressed by the following formula:

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{|F_t| + |A_t|} \tag{3.10}$$

where F_t and A_t denote the predicted and actual values at time t. This metric returns a value between 0% and 100%. Given that our data has zero values in some months (*e.g.*, emerging threats), the issue of division by zero may arise, a problem that often emerges when using standard MAPE (Mean Absolute Percentage Error). We find SMAPE to be resilient to this problem, since it has both the actual and predicted values in the denominator.

Recall that our model aims to predict a curve (corresponding to multiple time steps). Using plain SMAPE as the evaluation metric, the best model may turn out to be simply a straight line passing through the same points of the fluctuating actual curve. However, this is undesired in our case since our priority is to predict the trend direction (or slope) over its intensity or value at a certain point. We hence add a penalty term to SMAPE that we apply when the height of the predicted curve is relatively smaller than that of the actual curve. This yields the modified SMAPE (M-SMAPE). More formally, let $I(V)$ be the height of the curve V, calculated as follows:

$$I(V) = \text{max}_{t\in[n]} V_t - \text{min}_{t\in[n]} V_t \tag{3.11}$$

where n is the curve width or the number of data points. Let A and F denote the actual and predicted curves. We define M-SMAPE as follows:

$$MSMAPE = \begin{cases} SMAPE + 100\%\gamma, & \text{if } I(F) < I(A)/d \\ SMAPE, & \text{otherwise} \end{cases} \tag{3.12}$$

where γ is a penalty constant between 0 and 1, and d is another constant ≥ 1. In our experiment, we set γ to 0.3, and d to 3, as we found these to be reasonable values by trial and error. We note that the range of possible values of M-SMAPE is between 0% and $(100 + 100\ \gamma)\%$ after this modification.

3.2.8 Optimisation

On average, our model was trained on around 67% of the refined data, which is equivalent to approximately 7.2 years. We kept the rest, approximately 33% (3 years + lag period), for validation. These percentages may slightly differ for different attack types depending on the optimal lag period selected.

For hyper-parameter optimisation, we performed a random search with 60 iterations, to obtain the set of features, smoothing methods and constants, and model's hyper-parameters that results in the model with the lowest M-SMAPE. Random search is a simple and efficient technique for hyper-parameter optimisation, with advantages including efficiency, flexibility, robustness, and scalability. The technique has been studied extensively in the literature and was found to be superior to grid search in many cases [111]. For each set of hyper-parameters, the model was trained using the mean squared error (MSE) as the loss function, and while using ADAM as the optimisation algorithm [112]. Then, the model was validated by forecasting 3 years while using M-SMAPE as the evaluation metric, and the average performance was recorded over 3 different seeds. Once the set of hyper-parameters with the minimum M-SMAPE was obtained, we used it to train the model on the full data, after which we predicted the trend for the next 3 years (until March, 2025).

The first group of hyper-parameters is the subset of features in the case of the multivariate model. Here, we experimented with each of the 3 features separately (NoM, ACA or PH) along with the ground truth (NoI), in addition to the combination of all features. The second group is the smoothing methods and constants. The set of methods includes ES, DES and NS, as previously discussed. The set of values for the smoothing constant α ranges from 0.05 to 0.7 while the set of values for the smoothing constant β (for DES) ranges from 0.3 to 0.7. Next is the optimisation of the lag period with values that range from 1 to 12 months. This is followed by the model's hyper-parameters which include the learning rate with values that range from 6×10^{-4} to 1×10^{-2}, the number of epochs with values between 30 and 200, the number of layers in the range 1 to 2, the number of units in the range 25 to 200, and the recurrent dropout value between 0.2 and 0.5. The range of these values was obtained from the literature and the online code repositories [113].

3.3 Results

3.3.1 Validation

The results of our model's validation are provided in Figure 3.4 and Table 3.6. As shown in Figure 3.4, the predicted data points are well aligned with the ground truth. Our models successfully predicted the next 36 months of all attacks' trends with an average M-SMAPE of 0.25. Table 3.6 summarises the validation results of univariate and multivariate approaches using B-LSTM. The results show that with approximately 69% of all attack types, the multivariate approach outperformed the univariate approach. As seen in Figure 3.4, the threats that have a consistent increasing or emerging trend seemed to be more suitable for the univariate approach, while threats that have a fluctuating or decreasing trend showed less validation error when using the multivariate approach. According to Table 3.6, the feature of ACA resulted in the best model for 33% of all attack types, which makes it among the three most informative features that can boost the prediction performance. The PH accounts for 17% of all attacks followed by NoM that accounts for 12%.

We additionally compared the performance of the proposed model B-LSTM with other models namely LSTM and ARIMA. The comparison covers the univariate and multivariate approaches of LSTM and B-LSTM, with two features in the case of multivariate approach namely NoI and NoM. The comparison is in terms of the Mean Absolute Percentage Error (MAPE) when predicting four common attack types, namely the Distributed Denial-of-Service attack (DDoS), password attack, malware, and ransomware. The comparison results are provided in Table 3.7. The results illustrate the superiority of the B-LSTM model for most of these attack types.

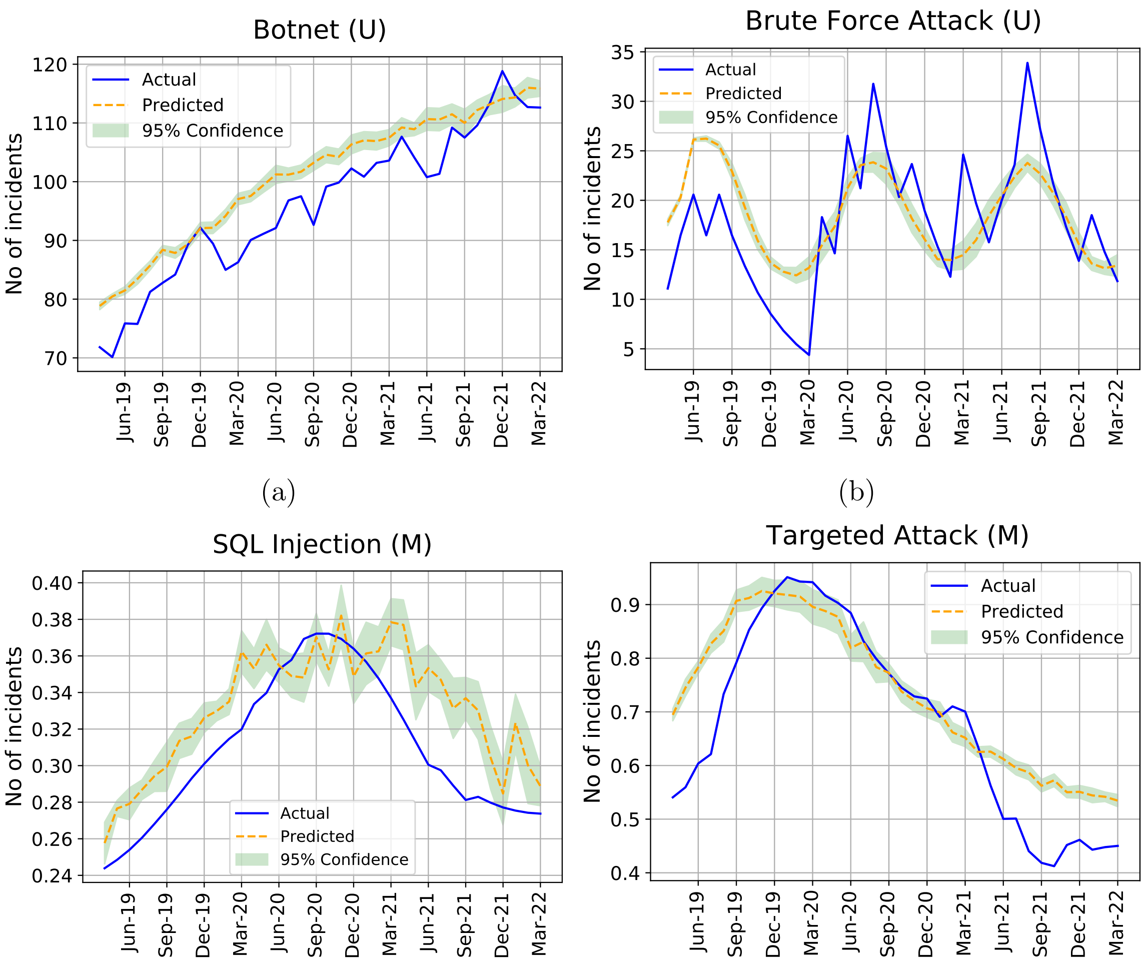
Table 3.7: Mean Absolute Percentage Error (MAPE) for 4 attacks and 5 models

Attack/Model	**ARIMA**	**LSTM (U)**	**LSTM (M)**	**B-LSTM (U)**	**B-LSTM (M)**
DDoS	1.43	0.63	0.93	0.65	**0.53**
Password	0.60	0.69	0.70	**0.56**	0.72
Malware	**0.61**	2.88	1.96	5.47	1.11
Ransomware	0.85	0.37	1.04	**0.36**	0.59

(U) indicates a univariate model while (M) indicates a multivariate model.

3.3.2 Trends Analysis

The forecast of each attack trend until the end of the first quarter of 2025 is given in Figure 3.5. By visualising the historical data of each attack as well as the prediction for the next three years, we were able to categorise the overall trend of each attack. The attacks generally follow 4 types of trends: 1) rapidly increasing, 2) overall increasing,



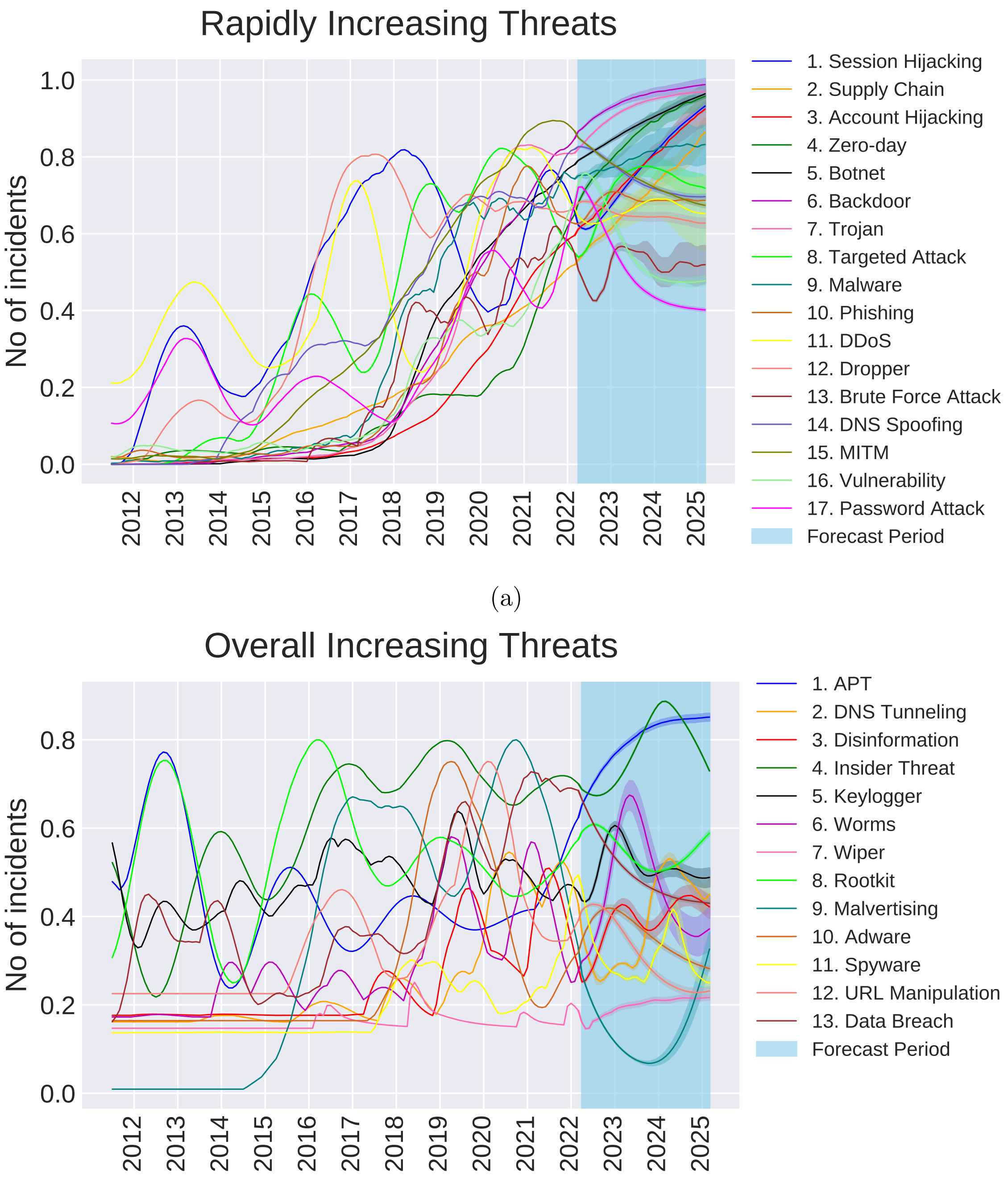
(a) (b)

(c) (d)

Figure 3.4: The B-LSTM validation results of predicting the number of attacks from April, 2019 to March, 2022. (U) indicates an univariate model while (M) indicates a multivariate model. (a) Botnet attack with M-SMAPE=0.03. (b) Brute force attack with M-SMAPE=0.13. (c) SQL injection attack with M-SMAPE=0.04 using the feature of NoM. (d) Targeted attack with M-SMAPE=0.06 using the feature of NoM. Y axis is normalised in the case of multivariate models to account for the different ranges of feature values.

3) emerging and 4) decreasing. The names of attacks for each category are provided in Figure 3.5.

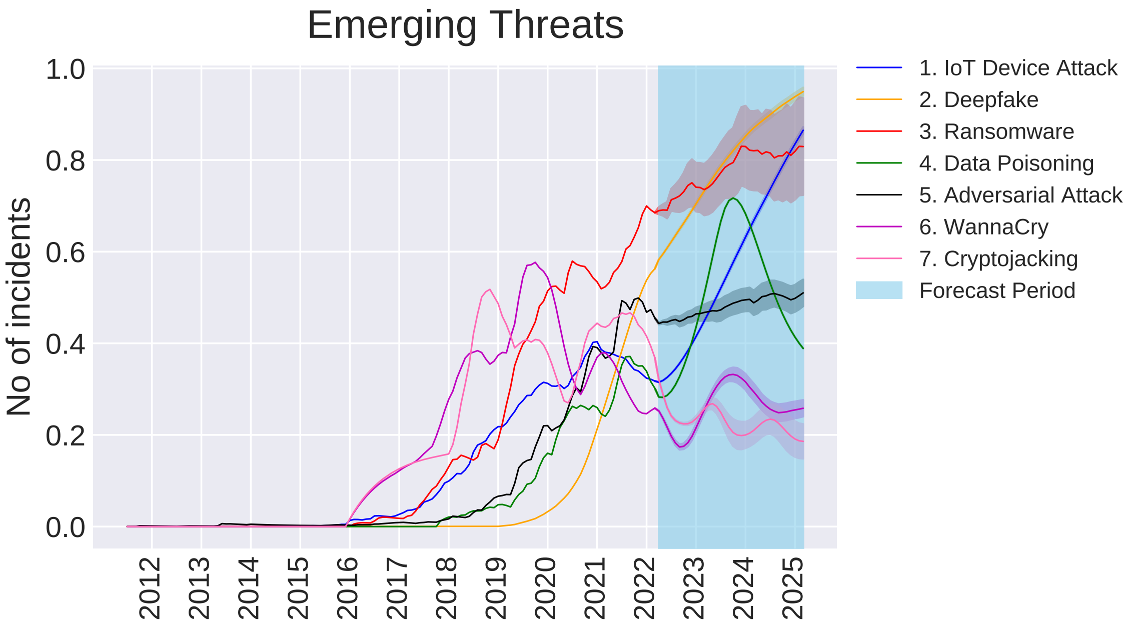
The first trend category is the rapidly increasing trend (Figure 3.5a) - approximately 40% of the attacks belong to this trend. We can see that the attacks belonging to this category have increased dramatically over the past 11 years. Based on the model's prediction, some of these attacks will exhibit a steep growth until 2025. Examples include session hijacking, supply chain, account hijacking, zero-day and botnet. Some



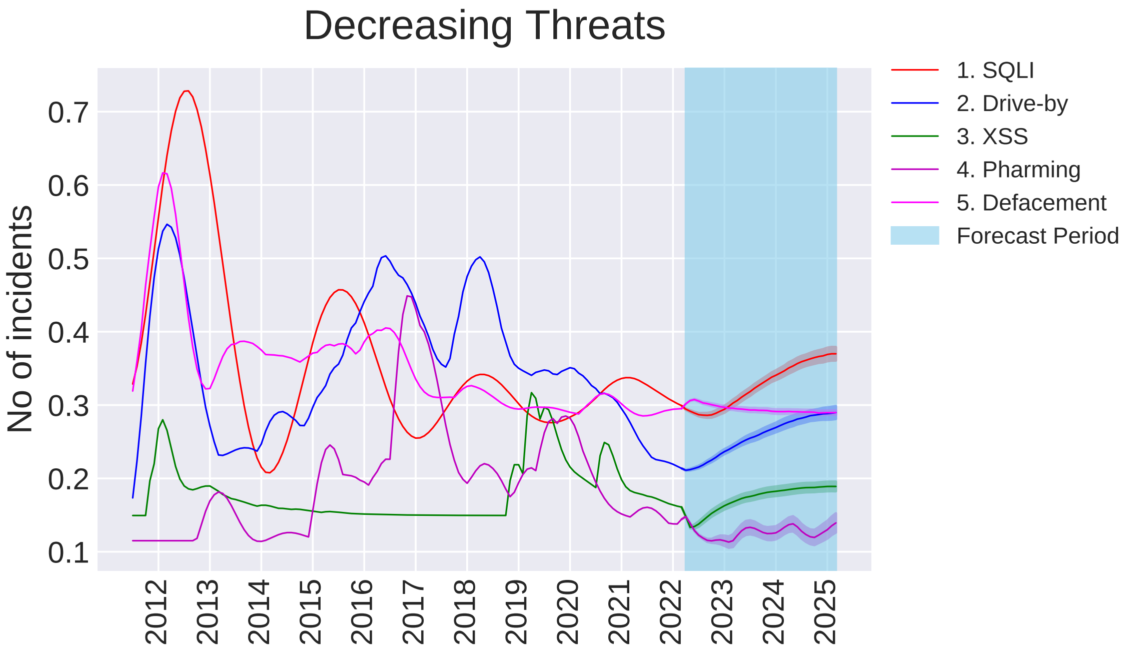
(a)

(b)

Figure 3.5: A bird's eye view of two threat trend categories (rapidly increasing and overall increasing threats). The period of the trend plots is between July, 2011 and March, 2025, with the period between April, 2022 and March, 2025 forecasted using B-LSTM. (a) Rapidly Increasing Threats. (b) Overall Increasing Threats.



(c)



(d)

Figure 3.5: Continued...A bird's eye view of two threat trend categories (emerging and decreasing threats). The period of the trend plots is between July, 2011 and March, 2025, with the period between April, 2022 and March, 2025 forecasted using B-LSTM. (c) Emerging Threats. (d) Decreasing Threats.

of the attacks under this category have reached their peak, have recently started stabilising, and will probably remain steady over the next 3 years. Examples include malware, targeted attack, dropper and brute force attack. Some attacks in this category, after a recent increase, are likely to level off in the next coming years. These are password attack, DNS spoofing and vulnerability-related attacks.

The second trend category is the overall increasing trend as seen in Figure 3.5b. Approximately 31% of the attacks seem to follow this trend. The attacks under this category have a slower rate of increase over the years compared to the attacks in the first category, with occasional fluctuations as can be observed in the figure. Although some of the attacks show a slight recent decline (*e.g.*, malvertising, keylogger and URL manipulation), malvertising and keylogger are likely to recover and return to a steady state while URL manipulation is projected to continue a smooth decline. Other attacks typical of cold cyberwarfare like Advanced Persistent Threats (APT) and rootkits are already recovering from a small drop and will likely to rise to a steady state by 2025. Spyware and data breach have already reached their peak and are predicted to decline in the near future.

Next is the emerging trend as shown in Figure 3.5c. These are the attacks that started to grow significantly after the year 2016, although many of them existed much earlier. In our study, around 17% of the attacks follow this trend. Some attacks under this category have been growing steeply and are predicted to continue this trend until 2025. These are Internet of Things (IoT) device attack and deepfake. Other attacks have also been increasing rapidly since 2016, however, are likely to slow down after 2022. These include ransomware and adversarial attacks. Interestingly, some attacks that emerged after 2016 have already reached the peak and recently started a slight decline (*e.g.*, cryptojacking and WannaCry ransomware attack). It is likely that WannaCry will become relatively steady in the coming years, however, cryptojacking will probably continue to decline until 2025 thanks to the rise of proof-of-stake consensus mechanisms [114].

The fourth and last trend category is the decreasing trend (Figure 3.5d) - only 12% of the attacks follow this trend. Some attacks in this category peaked around 2012, and have been slowly decreasing since then (*e.g.*, SQL Injection and defacement). The drive-by attack also peaked in 2012, however, had other local peaks in 2016 and 2018, after which it declined noticeably. Cross-site scripting (XSS) and pharming had their peak more recently compared to the other attacks, however, have been smoothly declining since then. All attacks under this category are predicted to become relatively stable from 2023 onward, however, they are unlikely to disappear in the next 3 years.

3.3.3 The Threat Cycle

This large-scale analysis involving the historical data and the predictions for the next three years enables us to come up with a generalisable model that traces the evolution and adoption of the threats as they pass through successive stages. These stages are named by the launch, growth, maturity, trough and stability/decline. We refer to this model as The Threat Cycle (or TTC), which is depicted in Figure 3.6. In the launch phase, few incidents start appearing for a short period. This is followed by a sharp increase in terms of the number of incidents, growth and visibility as more and more cyber actors learn and adopt this new attack. Usually, the attacks in the launch phase are likely to have many variants as observed in the case of the WannaCry attack in 2017. At some point, the number of incidents reaches a peak where the attack enters the maturity phase, and the curve becomes steady for a while. Via the trough (when the attack experiences a slight decline as new security measures seem to be very effective), some attacks recover and adapt to the security defences, entering the slope of plateau, while others continue to smoothly decline although they do not completely disappear (*i.e.*, slope of decline). It is worth noting that the speed of transition between the different phases may vary significantly between the attacks.

As seen in Figure 3.6, the attacks are placed on the cycle based on the slope of their current trend, while considering their historical trend and prediction. In the trough phase, we can see that the attacks will either follow the slope of plateau or the slope of decline. Based on the predicted trend in the blue zone in Figure 3.5, we were able to indicate the future direction for some of the attacks close to the split point of the trough using different colours (blue or red). Brute force, malvertising, DDoS, insider threat, WannaCry and phishing are denoted in blue meaning that these are likely on their way to the slope of plateau. In the first three phases, it is usually unclear and difficult to predict whether a particular attack will reach the plateau or decline, thus, denoted in grey.

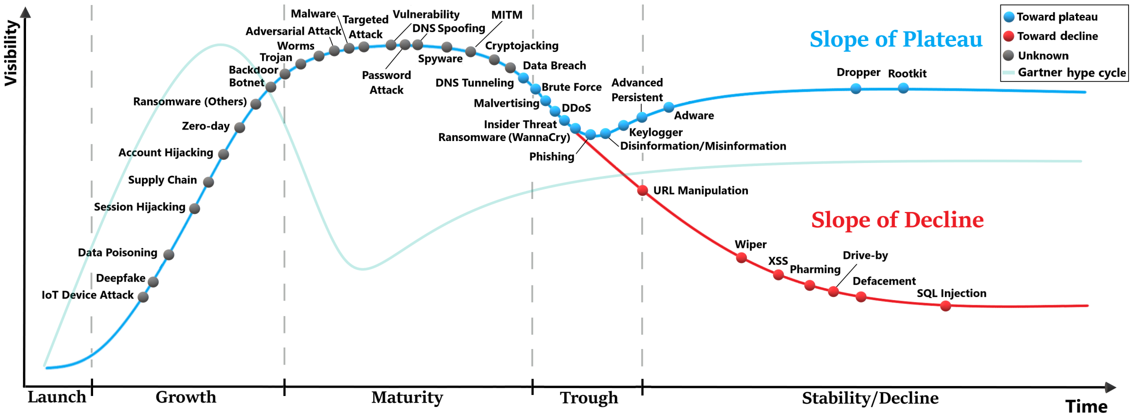


Figure 3.6: The threat cycle (TTC). The attacks go through 5 stages, namely, launch, growth, maturity, trough, and stability/decline. A standard Gartner hype cycle (GHC) is shown with a vanishing green colour for a comparison to TTC. TTC captures the state of each attack in 2022, where the colour of each attack indicates which slope it would follow based on the model prediction until 2025 (*e.g.*, blue: plateau or red: decline). The attacks with unknown final destination are coloured in grey.

There are some similarities and differences between TTC and the well-known GHC [115]. A standard GHC is shown in a vanishing green colour in Figure 3.6. As TTC is specific to cyber threats, it has a much wider peak compared to GHC. Although both GHC and TTC have a trough phase, the threats decline slightly (while significant drop in GHC) as they exit their maturity phase, after which they recover and move to stability (slope of plateau) or decline.

Many of the attacks in the emerging category are observed in the growth phase. These include IoT device attack, deepfake and data poisoning. While ransomwares (except WannaCry) are in the growth phase, WannaCry already reached the trough, and is predicted to follow the slope of plateau. Adversarial attack has just entered the maturity stage, and cryptojacking is about to enter the trough. Although adversarial attack is generally regarded as a growing threat, interestingly, this machine-based prediction and introspection shows that it is maturing. The majority of the rapidly increasing threats are either in the growth or in the maturity phase. The attacks in the growth phase include session hijacking, supply chain, account hijacking, zero-day and botnet. The attacks in the maturity phase include malware, targeted attack, vulnerability-related attacks and Man-In-The-Middle attack (MITM). Some rapidly increasing attacks such as phishing, brute force, and DDoS are in the trough and are predicted to enter the stability. We also observe that most of the attacks in the category of overall increasing threats have passed the growth phase and are mostly branching to the slope of plateau or the slope of decline, while few are still in the maturity phase (*e.g.*, spyware). All of the decreasing threats are on the slope of decline. These include XSS, pharming, drive-by, defacement and SQL injection.

3.4 Discussion

3.4.1 Findings

3.4.1.1 Threat Trends - Analysis

The comprehensive analysis of cyber threat trends and their forecasted trajectories provides valuable insights into the evolving landscape of cyber security. By categorising threat trends into distinct patterns, we can better understand their dynamics, anticipate future developments, and tailor mitigation strategies accordingly.

The identification of four main trend categories, namely rapidly increasing, overall increasing, emerging, and decreasing threats, offers a nuanced perspective on the diverse nature of cyber threats. Notably, the prevalence of rapidly increasing threats underscores the urgency of proactive measures to address their escalating impact. These threats, exemplified by session hijacking, supply chain attacks, and zero-day exploits, demonstrate a trajectory of sustained growth, posing significant challenges to cyber security efforts. Consequently, we believe that policymakers should allocate funds to technologies aimed at mitigating these threats, and should prioritise their defences and allocate their resources accordingly. For example, zero-day attacks exploit a previously unknown vulnerability before the developer has had a chance to release a patch or fix for the problem [116]. Zero-day attacks are particularly dangerous because they can be used to target even the most secure systems and go undetected for extended periods of time. As a result, these attacks can cause significant damage to an organisation's reputation, financial well-being, and customer trust. Our results in Figure 3.5a suggest that zero-day attack is likely to continue a steep growth until 2025. If we know this information, we can proactively invest on solutions to prevent it or slow down its rise in the future, since after all, the ML detection approaches may not be alone sufficient to reduce its effect.

Conversely, the analysis reveals a subset of threats exhibiting an overall increasing trend, characterised by a slower rate of growth and occasional fluctuations. While these threats may not manifest as immediate concerns, their persistent nature necessitates ongoing monitoring and adaptation of defensive measures. For instance, APTs and insider threats, although experiencing temporary declines, are anticipated to regain momentum, highlighting the importance of continuous vigilance. Moreover, addressing overall increasing threats with occasional fluctuations, demands approaches that can accommodate variability. For instance, deploying tools capable of dynamically adjusting security measures in response to evolving threat landscapes can provide an effective means of mitigating such fluctuations.

Emerging threats, marked by their significant growth since 2016, present unique challenges and opportunities in cyber security. While attacks such as IoT device exploits and deepfakes show exponential growth, others like ransomware are projected to plateau in the near future. This divergence underscores the need for adaptive strategies capable of addressing evolving threat landscapes. For instance, the ability to discern emerging threats likely to grow rapidly in the near future, such as IoT device attacks and deepfakes, enables proactive measures before these attacks reach critical levels. On the other hand, we find that other emerging threats such as the WannaCry ransomware which has recently posed a significant threat, are likely to stabilise or decrease in the near future. This projection of declining trends boosts policymakers' confidence in the effectiveness of the current security measures to address these risks, enabling them to uphold the existing defence mechanisms against such emerging attacks.

The analysis also sheds light on decreasing threats, which, although relatively fewer in number, continue to pose risks despite declining incidence rates. Understanding the trajectory of threats like SQL injection and drive-by attacks is crucial for allocating resources effectively and mitigating residual risks. By closely monitoring these declining threats, policymakers can ensure that adequate measures are in place to address any lingering vulnerabilities. Moreover, considering that these threats are projected to continue declining in the near future, policymakers could consider shifting their attention to other emerging or rapidly increasing threats. By giving these declining threats less priority and reallocating resources to address newer challenges, policymakers can stay ahead of evolving cyber threats and better prioritise their defences against potential vulnerabilities. This adaptive approach ensures that cyber security strategies remain dynamic and responsive to the changing threat landscape, ultimately enhancing overall resilience against cyber-attacks.

The multifaceted nature of the identified trend categories underscores the need for holistic cyber security strategies. Organisations can no longer rely on one-size-fits-all approaches but must tailor their defences to the specific characteristics of each threat category. This holistic perspective encourages a comprehensive understanding of the cyber threat landscape, fostering a more resilient cyber security posture that can adapt to the ever-changing nature of cyber threats. In essence, the discovery and analysis of these four trend categories contribute not only to a richer understanding of cyber threats but also pave the way for more targeted, adaptive, and proactive approaches in both research and practical cyber security applications.

3.4.1.2 The Threat Cycle - Analysis

The proposed TTC model provides a structured framework for contextualising threat evolution, encompassing stages from launch to stability/decline. This model not only captures the life cycle of cyber threats but also offers insights into their current states and future destinations. By delineating the phases of threat evolution, organisations can better anticipate shifts in threat landscapes and allocate resources strategically.

The model in Figure 3.6 indicates that there is a common life cycle for cyber-attacks. However, according to Figure 3.5, the attacks can vary in their speed of transition and peak level based on our response strategies and the effectiveness of our security measures. Our analysis reveals that despite the fact that some attacks emerged only recently, they already progressed to the maturity phase (*e.g.,* cryptojacking) while some other older attack types are still in the growth phase (*e.g.,* account hijacking). The model also indicates that while some attacks eventually decline, others may stabilise on the slope of plateau depending on the nature of the attack type and the viability and effectiveness of our security measures in countering the attack.

We believe that policymakers should aim to transition the emerging and rapidly increasing threats found in the launch and growth phases (Figure 3.6) to the stability phase as soon as possible, while minimising the threat's peak. An example for such successful effort can be found in the case of WannaCry ransomware in Figure 3.5c, which despite being an emerging attack is likely to level off in the near future. Moreover, the aim should be to progress threats that did not reach stability to the slope of decline, in order to minimise future losses by preventing their progress to the slope of plateau. Therefore, we believe that in the coming years, the focus should be on investing in technologies and security measures to counter threats observed in the launch and growth phases. For example, to counter the IoT device attack observed in the growth phase, agencies should invest in network segmentation, device authentication, and endpoint security solutions [117, 118]. Similarly, deepfake should be also prioritised and mitigated using deepfake detection tools such as those utilising ML algorithms to analyse media content for signs of manipulation [119], in addition to user education and awareness [120]. Data poisoning attack is also growing with relevant technologies worth investing in including adversarial training, anomaly detection, and data sanitisation [121, 122].

Moreover, attacks currently in the maturity and trough phases should receive high attention, in order to prevent them from plateauing and increase their chances of declining in the future. For instance, it is recommended that policymakers invest in IDS and threat intelligence platforms [123], which are key technologies for countering the *targeted attacks* currently in the maturity phase. Similarly, DDoS is currently in the trough phase and therefore, it is important to invest in relevant mitigating technologies including content delivery networks (CDNs) [124], rate limiting, and traffic shaping [125]. Overall, investing in these technologies helps to minimise the risk of many attack types and has the potential to make a positive change by altering the destination of these attacks.

Regarding prioritisation, organisations should allocate resources based on the severity and prevalence of threats in each phase. While emerging threats in the launch and growth phases require immediate attention, declining threats in the slope of decline phase may warrant less prioritisation. However, maintaining vigilance and preparedness across all phases is essential to effectively manage evolving cyber risks.

It is crucial to recognise that the threat landscape is dynamic. While TTC provides valuable insights, factors such as changes in attacker tactics and advancements in security technologies can influence the trajectory of attacks. Therefore, organisations must remain adaptable, continuously reassessing their security strategies to effectively mitigate evolving risks.

3.4.1.3 Feature Importance

It is worth-noting that the multivariate approach has proven to be more effective than the univariate approach based on our results (Table 3.6), for approximately 70% of all attack types. For the multivariate approach, our results in Table 3.6 show that using the feature ACA only (in addition to NoI) resulted in the best model for 33% of all attack types. This result suggests a pivotal finding: tweets pertaining to wars and conflicts stand out as crucial informative features capable of enhancing the model's overall performance. These tweets emerge as significant predictors for a wide array of cyber-attacks, underscoring their importance in forecasting potential cyber threats. Moreover, the evident correlation between the frequency or content of these tweets and the incidence of cyber-attacks underscores a robust relationship. This linkage not only highlights the relevance of monitoring social media discourse on conflicts but also accentuates the potential for leveraging such data as a key factor in bolstering cyber threat prediction capabilities.

According to the results, the feature PH when used along with the feature NoI resulted in the best model for 17% of all cyber-attacks. This feature reveals a nuanced yet discernible pattern in the cyber threat landscape. The frequency of attacks correlated with public holidays underscores the importance of temporal factors in understanding cyber threats. Public holidays may present opportune moments for threat actors to exploit potential vulnerabilities, making them a notable consideration in cyber security analysis.

Similarly, the Number of Mentions (NoM) in scientific literature resulted in the best model for 12% of cyber-attack types. This metric reflects the prevalence and attention given to specific attack types within the scholarly discourse. A higher number of mentions suggests increased scrutiny or recognition of certain cyber threats in academic circles, potentially signalling their prominence in real-world cyber incidents. Therefore, monitoring the number of mentions can provide valuable insights into the evolving landscape of cyber threats and guide proactive cyber security measures. While not as dominant as the feature ACA, both PH and NoM contribute valuable dimensions to the multifaceted understanding of cyber incidents, warranting careful consideration in comprehensive threat assessment strategies.

3.4.2 Highlights and Contributions

This study presents the development of an ML-based proactive approach for long-term prediction of cyber-attacks offering the ability to communicate effectively with the potential attacks and the relevant security measures in an early stage to plan for the future. This approach can contribute to the prevention of an incident by allowing more time to develop optimal defensive actions/tools in a contested cyberspace. Proactive approaches can also effectively reduce uncertainty when prioritising existing security measures or initiating new security solutions. We argue that cyber security agencies should prioritise their resources to provide the best possible support in preventing fastest-growing attacks that appear in the launch phase of TTC or the attacks in the categories of the rapidly increasing or emerging trend as in Figure 3.5a and 3.5c based on the predictions in the coming years.

In addition, our fully automated approach is promising to overcome the well-known issues of human-based analysis, above all expertise scarcity. Given the absence of the possibility of analysing with human's subjective bias while following a purely quantitative procedure and data, the resulting predictions are expected to have lower degree of subjectivity, leading to consistencies within the subject. By fully automating this analytic process, the results are reproducible and can potentially be explainable with help of the recent advancements in Explainable Artificial Intelligence (XAI).

The identification and exploration of four distinct categories in cyber threat trends (rapidly increasing, overall increasing, emerging, and decreasing) carry profound implications for both research and practical cyber security strategies. Understanding these diverse trend categories provides a roadmap for future research endeavours. Researchers can delve into the unique characteristics of each category, investigating the underlying factors contributing to the observed trajectories. This nuanced exploration can lead to a deeper comprehension of the evolving cyber threat landscape, informing the development of more targeted and effective mitigation strategies.

The predictive nature of the model empowers both researchers and practitioners to adopt a proactive stance. Armed with insights into the potential trajectories of various attack types, cyber security professionals can develop pre-emptive strategies to counteract threats before they escalate. This proactive approach represents a paradigm shift from reactive cyber security practices, enabling organisations to stay ahead of evolving cyber threats and fortify their defences accordingly.

Thanks to the massive data volume and wide geographic coverage of the data sources we utilised, this study covers every facet of today's cyber-attack scenario. Our holistic approach performs the long-term prediction on the scale of 36 countries, and is not confined to a specific region. Indeed, cyberspace is limitless, and a cyber-attack on critical infrastructure in one country can affect the continent as a whole or even globally. We argue that our TTC provides a sound basis to awareness of and investment in new security measures that could prevent attacks from taking place. We believe that our tool can enable a collective defence effort by sharing the long-term predictions and trend analysis generated via quantitative processes and data and furthering the analysis of its regional and global impacts.

Moreover, the study demonstrates the superiority of a multivariate approach in predicting cyber-attacks, showcasing the importance of incorporating diverse data sources such as social media discourse, temporal factors like public holidays, and scholarly attention. By leveraging features like tweets related to wars and conflicts, the research highlights the potential of social media data in enhancing threat prediction capabilities. Additionally, insights from public holidays and academic discourse shed light on the evolving cyber threat landscape, emphasising the need for a holistic understanding to effectively mitigate cyber risks.

3.4.3 Limitations

A limitation of our approach is its reliance on a restricted dataset that encompasses data since 2011 only. This is due to the challenges we encountered in accessing confidential and sensitive information. Extending the prediction phase requires the model to make predictions further into the future, where there may be more variability and uncertainty. This could lead to a decrease in prediction accuracy, especially if the underlying data patterns change over time or if there are unforeseen external factors that affect the data. While not always the case, this uncertainty is highlighted by the results of the Bayesian model itself as it expresses this uncertainty through the increase of the confidence interval over time (Figure 3.4a and Figure 3.4b). Despite incorporating the Bayesian model to tackle the epistemic uncertainty, our model could benefit substantially from additional data to acquire a comprehensive understanding of past patterns, ultimately improving its capacity to forecast long-term trends. Moreover, an augmented dataset would allow ample opportunity for testing, providing greater confidence in the model's resilience and capability to generalise.

Our study leverages long-term features such as the trend of cyber-attacks in news, literature, and social media to forecast threats over extended periods. However, it is important to acknowledge a limitation of this approach: its inability to predict the onset of cyber-attacks or anticipate novel attack types. To address this limitation, integrating features that capture digital traces reflecting attackers' intentions and motivations would be valuable for predicting attack onset. Additionally, incorporating short-term features like network traffic patterns and reconnaissance activities could enhance the ability to forecast imminent attacks in the short term. By integrating these additional features, our predictive models could offer more comprehensive and timely insights into evolving cyber threats.

When considering the potential biases introduced by the selection of our current data sources, such as the Elsevier API (Scopus) for academic research trends and Twitter for political tensions, it is crucial to acknowledge and address these biases in comparison to other sources like industry reports. One primary concern is the difference in focus and timeliness between academic literature and industry reports. Academic research often explores emerging threats and theoretical models, which may lead to a delay or lag in addressing immediate, practical cyber security concerns compared to industry reports that typically highlight current trends and immediate threat landscapes. To mitigate this, we conducted a comparative analysis of the NoM of cyber-attack types in academic literature versus industry reports.

As shown in Figure 3.7, our analysis revealed a general trend where academic discussions often precede industry implementations by a certain lag period, reflecting the time required for theoretical research to be applied practically [94]. By visualising these trends, we observed that while there is a correlation between the NoMs in academia and industry, with significant overlap in the types of threats discussed, industry reports tend to emphasise more urgent, actionable information on cyber threats. For instance, an attack type may be discussed extensively in academic literature for its theoretical implications [126], whereas industry reports might prioritise attacks that have immediate relevance to ongoing cyber security defences [127]. This correlation and the observed lag highlight the dynamic interplay between academic research and practical implementation, suggesting that while our selected sources provide a comprehensive view of emerging trends, they may not capture the immediacy of threats as effectively as industry reports. However, this also suggests that academic attention to an attack type typically anticipates its subsequent manifestation in real-world scenarios. Consequently, features such as NoM could serve as valuable predictors of imminent cyber incidents, potentially enhancing the performance of ML models and justifying their inclusion in predictive frameworks [6].

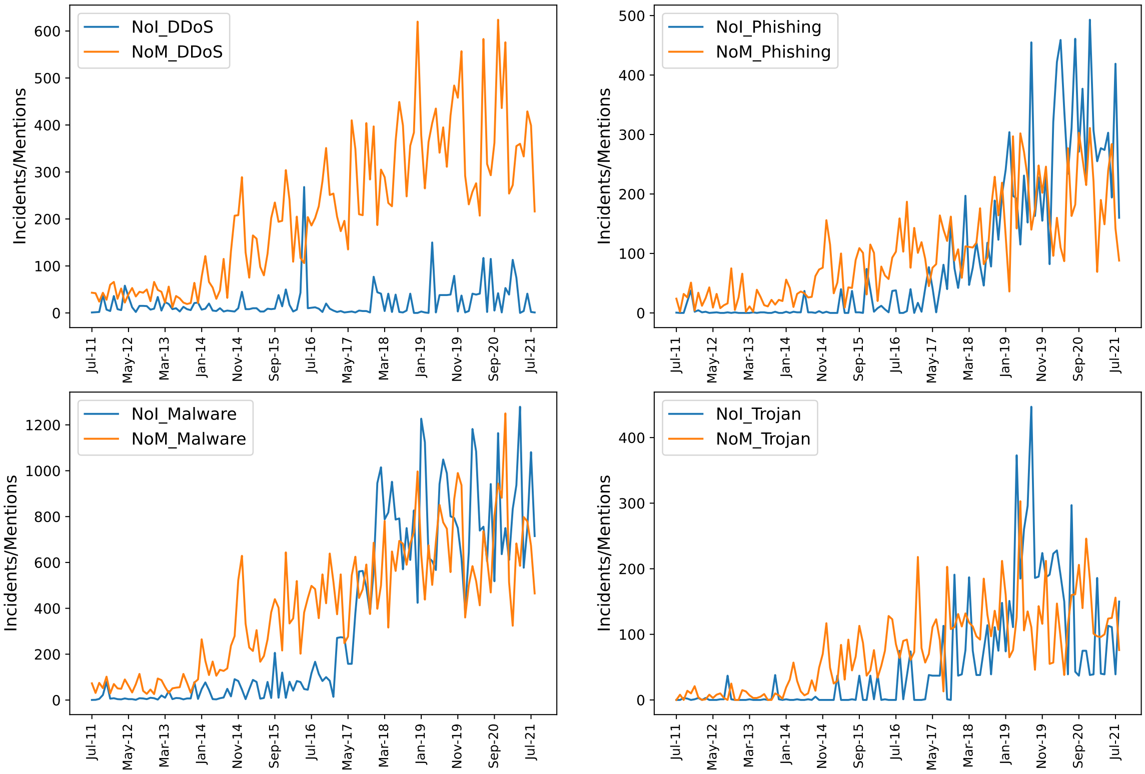


Figure 3.7: Trend visualisation for four attack types: Number of Mentions (NoM) in literature vs. Number of Incidents (NoI). The figure illustrates the trends of four attack types in academic literature compared to their trends in industry reports. It reveals that the mentions in academic discussions for a given attack type often precedes the incidents reported in industry by a certain lag period. This suggests that academic attention to an attack type typically anticipates its subsequent manifestation in real-world scenarios.

Furthermore, our analysis extended to the comparison of political tensions data from Twitter against more formal sources of geopolitical information, such as governmental reports or news archives. We found that while Twitter provides real-time, crowd-sourced insights into political events, which are invaluable for capturing public sentiment and immediate reactions, it may lack the structured and validated perspectives found in more formal reports [128]. This can lead to biases where public discourse may overemphasise or underplay certain issues based on the nature of social media dynamics.

Overall, while the selected data sources offer significant advantages in terms of breadth and real-time information, it is important to acknowledge their limitations and potential biases. The correlations between academic literature, industry reports, and public discourse, as well as the validation of observed trends, suggest that our approach provides a robust foundation for understanding cyber threats. However, it also underscores the importance of integrating multiple data sources to capture a holistic and balanced view of the cyber security landscape, ensuring that our research is both timely and relevant.

3.5 Summary and the Road Ahead

In this chapter, we introduced the development of an ML-based proactive approach for long-term prediction of cyber-attacks. The goal is to effectively communicate with potential attacks early on, allowing for strategic planning and the development of optimal defensive actions. The proactive approach aims to contribute to incident prevention by providing more time for the implementation of robust security measures and facilitating the prioritisation of security measures based on the machine forecast. The chapter compared the ML-based automated approach with traditional human-based analysis, emphasising the potential for overcoming issues such as expertise scarcity and bias. The automated approach was presented as less subjective, providing consistent and reproducible results.

The role of big data in the study was highlighted, emphasising its contribution to covering every facet of the cyber-attack scenario on a global scale. Leveraging this data, our analysis introduced four distinct categories of cyber threat trends (rapidly increasing, overall increasing, emerging, and decreasing), suggesting that understanding these categories is crucial for research and practical cyber security strategies. Tailoring defence mechanisms to the specific characteristics of each threat category was highlighted as essential for effective mitigation strategies. Moreover, the TTC model was presented as a tool that can enable collective defence efforts by sharing long-term predictions and trend analyses globally, emphasising its potential impact on regional and global cyber security.

In the next chapter, we introduce a new area of research that focuses on prognosticating the disparity between the trend of cyber-attacks and the associated mitigation technologies, with the aim of guiding research investment and strategic defence decisions. Subsequently, TTC could be improved by adopting another curve model that can visualise the current development of relevant security measures. The threat trend categories (Figure 3.5) and TTC (Figure 3.6) show how attacks will be visible in the next three years and more, however, we do not know where the relevant security measures will be. For example, data poisoning is an AI-targeted adversarial attack that attempts to manipulate the training dataset to control the prediction behaviour of a machine-learned model. From the scientific literature data (*e.g.*, Scopus), we could analyse the published articles studying the data poisoning problem and identify the relevant keywords of these articles (*e.g.*, Reject on Negative Impact (RONI) and Probability of Sufficiency (PS)). RONI and PS are typical methods used for detecting poisonous data by evaluating the effect of individual data points on the performance of the trained model. Likewise, the features that are informative, discriminating or uncertainty-reducing for knowing how the relevant security measures evolve exist within such online sources in the form of author's keywords, number of citations, research funding, number of publications, *etc.*

Code and Data Availability

The code and dataset used in this chapter are available at the following link: `https://github.com/zaidalmahmoud/Cyber-threat-forecast`.

Table 3.4: The list of 36 countries included in the study

No.	Country Name	Country Code
1	United States of America	US
2	United Kingdom	GB
3	Canada	CA
4	Australia	AU
5	Ukraine	UA
6	Russia	RU
7	France	FR
8	Germany	DE
9	Brazil	BR
10	China	CN
11	Japan	JP
12	Pakistan	PK
13	North Korea	KP
14	South Korea	KR
15	India	IN
16	Taiwan	TW
17	Netherlands	NL
18	Spain	ES
19	Sweden	SE
20	Mexico	MX
21	Iran	IR
22	Israel	IL
23	Saudi Arabia	SA
24	Syria	SY
25	Finland	FI
26	Ireland	IE
27	Austria	AT
28	Norway	NO
29	Switzerland	CH
30	Italy	IT
31	Malaysia	MY
32	Egypt	EG
33	Turkey	TR
34	Portugal	PT
35	Palestine	PS
36	United Arab Emirates	AE

The country code is used in our dataset in the naming of the columns.

Table 3.6: The validation results of univariate and multivariate approach using B-LSTM to forecast 42 attacks next 36 months

Attack	M-SMAPE (univariate)	M-SMAPE (multivariate)	Best features (multivariate)
Adware	0.35	0.29	ACA
Backdoor	0.10	0.03	ACA
Cryptojacking	0.40	0.34	ACA
Data Poisoning	0.47	0.46	ACA
Defacement	0.36	0.06	ACA
DNS Tunneling	0.48	0.42	ACA
Keylogger	0.17	0.14	ACA
Pharming	0.59	0.27	ACA
Trojan	0.31	0.30	ACA
Vulnerability	0.33	0.25	ACA
WannaCry	0.58	0.57	ACA
Wiper	0.43	0.14	ACA
Worms	0.50	0.37	ACA
XSS	0.47	0.17	ACA
Advanced Persistent	0.84	0.32	PH
DNS Spoofing	0.48	0.36	PH
Drive-by	0.46	0.27	PH
Insider Threat	0.17	0.07	PH
Malvertising	0.38	0.25	PH
Session Hijacking	0.39	0.34	PH
URL manipulation	0.47	0.36	PH
Data Breach	0.27	0.24	NoM
Disinformation	0.45	0.36	NoM
Phishing	0.22	0.21	NoM
SQL Injection	0.53	0.06	NoM
Targeted Attack	0.25	0.22	NoM
Password Attack	0.59	0.52	NoM, ACA, PH
Rootkit	0.19	0.15	NoM, ACA, PH
Spyware	0.63	0.48	NoM, ACA, PH
Account Hijacking	0.09	0.49	ACA
Adversarial Attack	0.37	0.63	NoM, ACA, PH
Botnet	0.03	0.17	PH
Brute Force Attack	0.13	0.28	ACA
DDoS	0.22	0.23	PH
Deepfake	0.17	0.52	PH
Dropper	0.12	0.37	PH
IoT Device Attack	0.16	0.21	PH
Malware	0.12	0.27	PH
MITM	0.14	0.32	PH
Ransomware	0.26	0.53	NoM
Supply Chain	0.15	0.33	PH
Zero-day	0.30	0.63	NoM

For each attack, the M-SMAPE value of the model with the better performance is highlighted in purple and the best feature(s) when using the multivariate model are displayed in the last column. NoM stands for the number of attack mentions in the scientific literature. ACA stands for the number of tweets related to armed conflict areas/wars. PH stands for the number of public holidays.

4 | Forecasting Threats and Pertinent Alleviation Technologies

4.1 Background

Forecasting the trend of cyber threats is valuable for enhancing defence strategies and cyber security resilience [5]. By analysing historical data and identifying patterns, organisations can anticipate potential threats and allocate resources accordingly. However, solely focusing on forecasting cyber threats may not provide a comprehensive understanding of the cyber security landscape. To effectively prioritise defence strategies and allocate resources, it is essential to consider not only the trends of cyber threats but also the trends of the *pertinent alleviation technologies* (PATs).

Predicting the trend of cyber threats allows organisations to anticipate the types and frequency of potential attacks [129]. This information enables proactive measures such as implementing security controls, patching vulnerabilities, and training personnel to mitigate risks. However, without considering the corresponding trends in alleviation technologies, organisations may overlook critical aspects of their defence strategies. For instance, if the trajectory of a specific cyber threat is anticipated to escalate significantly in the future, but the corresponding PATs lag behind in development, organisations may encounter challenges in effectively mitigating the risks posed by such threats.

By forecasting the trend of alleviation technologies alongside cyber threats, organisations can gain insights into the effectiveness of their defence mechanisms and identify potential gaps between threats and defences. Predicting the gap between the trend of threats and the trend of PATs enables policymakers to prioritise research investment and make strategic defence decisions [6]. For instance, if the trend of a cyber threat is expected to outpace the development of corresponding PATs, organisations may need to allocate additional resources to accelerate the research and development of defence technologies or explore alternative defence strategies.

Furthermore, predicting the gap between threats and alleviation technologies allows organisations to adopt a proactive approach to cyber security, rather than a reactive one [6]. By anticipating future challenges and addressing potential gaps in defence capabilities, organisations can enhance their resilience to cyber threats and minimise the impact of cyber-attacks. Additionally, considering both threat and PAT trends enables organisations to develop comprehensive defence strategies that take into account the evolving nature of the cyber security landscape [130].

Modelling the data as a graph is both possible and highly useful in the context of cyber trend forecasting, such as forecasting the trend of threats and PATs. In this approach, nodes in the graph represent entities such as cyber threats, PATs, or other relevant factors such as wars and political conflicts. Here, nodes can hold the trend value, and edges can represent relationships between these nodes. Edges can hold multiple values such as gap values or attention scores to reflect different types of relationships. By representing the problem as a graph, we can capture the intricate relationships and dependencies between different elements of the cyber security landscape.

This graph-based representation enables us to analyse the interactions between cyber threats and PATs, identify patterns and trends, and predict future developments. For example, we can use graph algorithms to identify clusters of related threats or technologies, detect anomalies or outliers [131], and uncover hidden relationships within the data [27]. Additionally, by incorporating temporal information into the graph, we can model how cyber threats and PATs evolve over time, allowing us to make predictions about future trends and anticipate potential risks.

GNNs hold significant promise for addressing complex problems [132], particularly those involving graph-structured data such as cyber trend forecasting. GNNs are specifically designed to operate on graph data, allowing them to capture intricate relationships and dependencies between interconnected entities [27]. In the context of cyber security, where threats and defence mechanisms can be represented by nodes and edges in a graph, GNNs offer a powerful framework for analysing and predicting the evolving landscape of cyber threats and PATs.

By leveraging graph convolutional layers, GNNs can effectively aggregate information from neighbouring nodes, enabling them to capture spatial dependencies within the graph [27]. This capability is crucial for identifying patterns and trends in cyber threat data and predicting future attack vectors. Additionally, GNNs can incorporate temporal convolutional layers to model temporal dependencies [43], allowing them to capture how cyber threats and PATs evolve over time. By combining spatial and temporal information, GNNs can provide actionable insights for stakeholders, enabling them to anticipate future threats, prioritise defence strategies, and allocate resources effectively. Overall, GNNs offer a versatile and powerful framework for cyber security forecasting, with the potential to significantly enhance defence capabilities and mitigate cyber risks.

4.2 Problem Extension

In this chapter, we study the problem of forecasting the gap between the trend of cyber threats and PATs by leveraging a graph-based approach. The graph enables us to visualise and analyse the evolving relationship between threats and PATs over time, and the possible gaps between them, providing valuable insights for defence strategies. To construct the graph, we collect data from various big data sources including news and government advisories, scientific articles, tweets, and Python APIs. We extract key features such as the number of incidents for threats and the number of mentions for PATs, which serve as indicators of their respective trends. Through exploratory analysis, we identify patterns and characteristics in the data, facilitating the development of a robust forecasting model. Additionally, we introduce a Bayesian variation of a GNN model, capable of capturing both temporal and spatial dependencies in the graph while quantifying epistemic uncertainty. By forecasting the graph and analysing past and future trends, we categorise these trends and provide actionable recommendations for research investment and strategic defence decisions. Finally, we develop a generalisable model where we identify the key phases that constitute the life cycle of 98 alleviation technologies.

In our pursuit of forecasting the gap between threats and PATs, our approach entails constructing a graph called Threats and Pertinent Technologies (TPT) (Figure 4.1) comprising nodes that represent cyber threats, which are linked to nodes that represent the PATs. Other nodes exist in the graph which represent our proposed external features including the mentions of threats in research documents, tweets about wars and conflicts, and public holidays. The node value indicates the trend of the corresponding entity. The gap between a threat node and its linked PAT node is quantified as the difference in their trends, serving as the edge weight. The existence of an edge between a threat node and a PAT node can be identified through semi-automated means using the GPT model or potentially through automated graph learning [27]. The construction of the graph will be followed by training the GNN model to forecast the graph, predicting the trends (hence the gaps) three years ahead and generating our recommendations. Our built GNN model will be evaluated extensively and compared to other models. The future forecast data will be generated and visualised while highlighting the gap between threats and PATs.

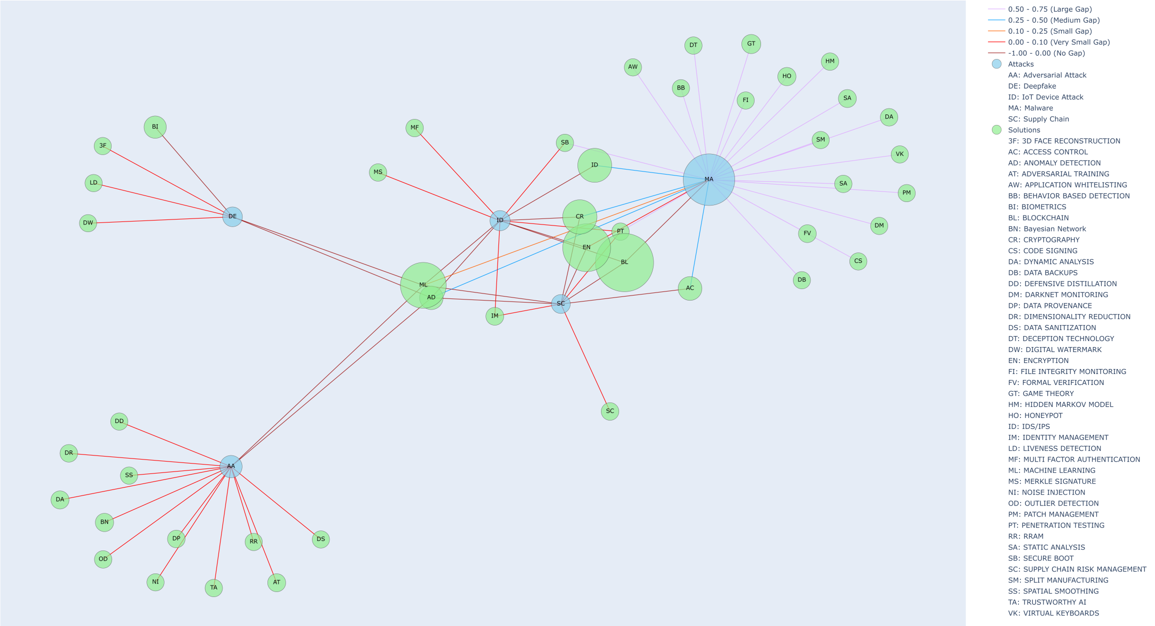


Figure 4.1: Threats and Pertinent Technologies (TPT) graph during December, 2022. The attacks are shown in blue and the PATs are shown in green. The node size signifies the trend. The darkness of the edge colour signifies the gap, where a lighter colour indicates a larger gap between the attack's trend and the trend of the PAT.

In chapter 1, we classified 26 cyber-attacks as emerging or rapidly increasing, indicating the urgency and significance of these threats in the cyber security landscape. In this chapter, we have selected these 26 threats for detailed examination due to their heightened importance and potential impact on cyber security. Emerging and rapidly increasing threats pose the highest risk as they have the propensity to escalate quickly and cause significant harm to individuals, organisations, and critical infrastructure. Compared to other categories, such as declining threats, these emerging and rapidly increasing threats are more critical because they represent ongoing challenges that demand immediate attention. By focusing our analysis on these critical threats, we aim to provide actionable insights that enable policymakers to prioritise their cyber security efforts effectively and timely.

For instance, emerging threats include adversarial attacks and deepfakes, which exploit vulnerabilities in ML algorithms to manipulate data or deceive systems for malicious purposes. Adversarial attacks involve the deliberate manipulation of input data to fool ML models, leading to misclassification or incorrect decisions [133]. Deepfakes, on the other hand, utilise artificial intelligence to generate highly realistic but fabricated images, videos, or audio recordings, often for spreading disinformation or conducting fraud [134]. Another emerging threat is ransomware attack, which encrypts critical data or systems and demand payment for their release, causing substantial financial losses and operational disruptions to targeted organisations [1]. Addressing such emerging threats is crucial as failure to do so could precipitate their proliferation, which leads to severe consequences such as compromised data integrity, reputational damage, financial losses, and disruptions to essential services.

Rapidly increasing threats include a wide array of attacks, such as DDoS and insider threats, which have demonstrated a notable escalation in frequency or severity [6, 135]. DDoS attacks flood targeted systems or networks with an overwhelming volume of traffic, rendering them inaccessible to legitimate users and disrupting services [135]. Insider threats involve individuals within an organisation exploiting their access privileges or knowledge to compromise security, steal sensitive information, or sabotage operations [136]. Being prepared to counter these rapidly increasing threats before they escalate is imperative as neglecting to do so may exacerbate their prevalence, potentially resulting in significant harm such as prolonged service disruptions, compromised data confidentiality, and loss of trust in organisational security measures.

The PATs identified for each of these threats are essential components of comprehensive cyber security defence strategies. For instance, technologies such as Anomaly Detection, ML/DL, and Intrusion Detection/Prevention Systems (IDS/IPS) are instrumental in detecting and mitigating adversarial attacks and deepfakes by identifying anomalous patterns or behaviours indicative of malicious activity [137, 138]. Similarly, measures such as Access Control, Data Loss Prevention, and User Behaviour Analytics are crucial for addressing insider threats by monitoring and controlling access to sensitive resources and detecting aberrant behaviours or unauthorised activities [139].

By forecasting the trend of the emerging and rapidly increasing threats, our study aims to provide actionable insights into evolving cyber threats and inform proactive risk management strategies. By adopting this approach, organisations can effectively prioritise their cyber security efforts, judiciously allocate resources, and implement targeted measures to mitigate emerging risks before they escalate into significant threats. Furthermore, by predicting the trend of PATs tailored to specific attack vectors, organisations can anticipate future gaps between each threat and its corresponding PATs. This foresight enables them to make informed investment and strategic defence decisions, ultimately bolstering their resilience against evolving cyber threats and more effectively safeguarding their assets, data, and operations.

4.3 Methods

4.3.1 Framework Extension

The framework's architecture for forecasting cyber threats and PATs is shown in Figure 4.2. As illustrated in the figure, our framework leverages a variety of unstructured data sources to gather all relevant information and extract valuable insights. Among these sources, the news, blogs, and government advisories' websites play a crucial role, providing an extensive collection of textual data on major cyber-attacks (approximately 18,000 incidents) since July 2011. The monthly count of attacks represents the ground truth of the attacks' trend, and is denoted as the *Number of Incidents*

(NoI). Furthermore, by utilising Elsevier API, we gained access to a vast repository of scientific articles from numerous sources. Through this API, we acquired the *Number of Mentions* (NoM) for each attack type and each PAT, which indicates their frequency in scientific publications, typically on a monthly basis. This NoM feature is particularly significant as it serves as a reliable reference for attack types that may not be present in other sources and also represents the ground truth of the PATs' trend. During the initial research phase, we thoroughly examined all potential features and identified a strong correlation between wars and political conflicts and the occurrence of cyber-events. To capture this information, we extracted relevant tweets using the Twitter API, specifically focusing on the number of tweets about *Armed Conflict Areas/Wars* (ACA). Finally, considering that cyber-attacks often coincide with holidays, we employed Python's Holidays package to obtain the count of public holidays per month for each country, denoted as *Public Holidays* (PH).

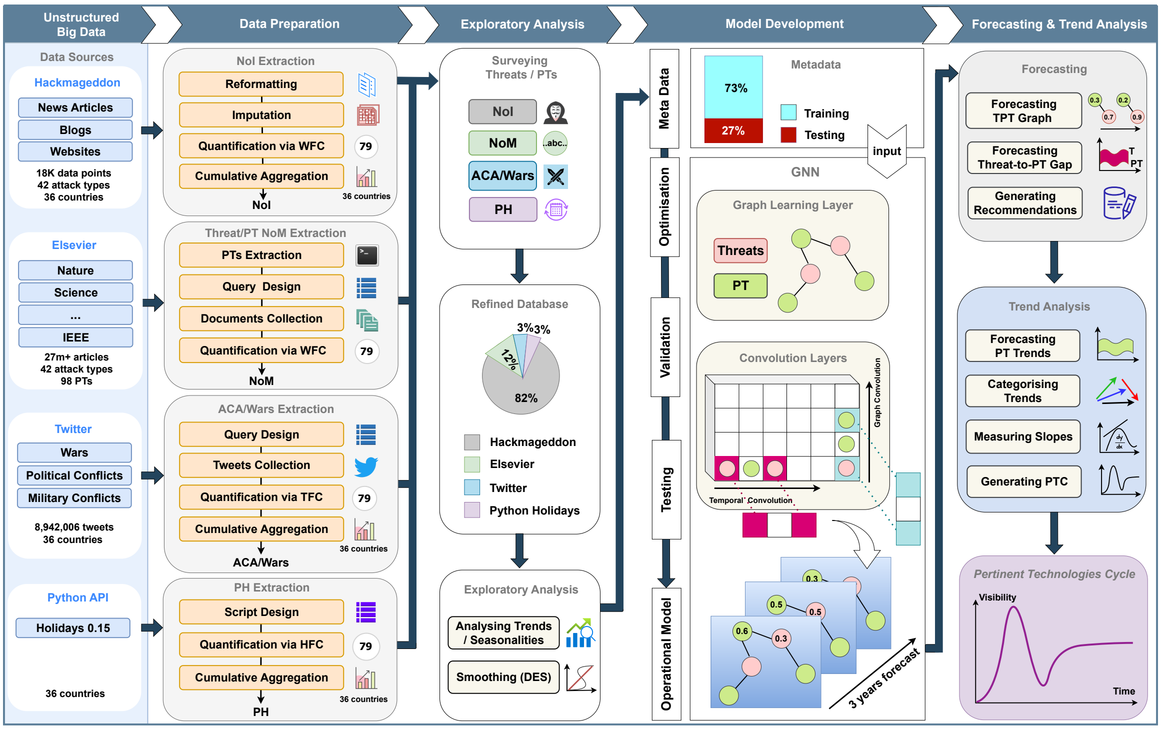


Figure 4.2: The workflow and architecture of forecasting cyber threats and pertinent alleviation technologies. NoI: Number of Incidents, NoM: Number of mentions, ACA: Armed Conflict Areas, PH: Public Holidays, PT: Pertinent Technology, WFC: Word Frequency Counter, TFC: Tweet Frequency Counter, HFC: Holiday Frequency Counter, DES: Double Exponential Smoothing, TPT: Threats and Pertinent Technologies, PTC: Pertinent Technologies Cycle.

For the extraction of NoI, the data preparation phase, as illustrated in Figure 4.2, starts by collecting and arranging all incidents in a tabular format including the date and attack type in addition to the description and country. This is followed by noise reduction through the handling of missing values, particularly in the earlier years. Here, imputation techniques were employed, utilising information from the description column or external sources such as reliable articles found through Google searches to supplement missing country data. Next, quantification of the textual data involved the implementation of a *Word Frequency Counter*, tallying the occurrences of each attack type per month for each country. Finally, cumulative aggregation facilitated the calculation of attack counts per month for all countries collectively (36 countries).

Prior to extracting NoM, we extracted the PATs by prompting GPT to extract relevant technologies to each attack type from Elsevier abstracts and also through a direct prompt to GPT. We then queried Elsevier API to collect the research documents relevant to each attack type and each extracted PAT. This was followed by running a Python script to obtain NoM for each attack type and each PAT per month within the collected documents.

The extraction of ACA from Twitter involved designing a script that included a query for collecting all tweets about wars and political conflicts relevant to each of the 36 countries in the study and during each month. A *Tweet Frequency Counter* was then used to count the number of such tweets for each individual country per month, followed by a cumulative aggregation to obtain the total number of tweets per month. The extraction of PH was done by writing a Python script including a *Holiday Frequency Counter* to obtain the number of public holidays per month for each country followed by a cumulative aggregation to obtain the total number of holidays per month.

While we focus in this study on 26 emerging and rapidly increasing threats, our monthly dataset includes the trend of 42 attack types in 36 countries, in addition to 98 PATs. Based on the above, we obtain the following columns for each month:

- NoI_C: The number of incidents for each attack type in each country (42 × 36 columns) [News, blogs, government advisories].

- NoI: The total number of incidents for each attack type (42 columns) [News, blogs, government advisories].
- NoM_A: The number of mentions of each attack type in research articles (42 columns) [Elsevier].
- NoM_P: The number of mentions of each alleviation technology in research articles (98 columns) [Elsevier].
- ACA_C: The number of tweets about wars and conflicts related to each country (36 columns) [Twitter].
- ACA: The total number of tweets about wars and conflicts (1 column) [Twitter].
- PH_C: The number of public holidays in each country (36 columns) [Python].
- PH: The total number of public holidays (1 column) [Python].

In the aforementioned list of columns, the name enclosed within square brackets denotes the source of data. By matching and combining these columns, we derive our monthly dataset, wherein each row represents a distinct month.

To gain insights into the dataset's main characteristics, an exploratory analysis was conducted. This analysis involved visualisations to identify key patterns such as trends, seasonality, correlated features, missing data, and outliers. Seasonal data was smoothed to unveil underlying trends while mitigating noise, employing double exponential smoothing [96].

In terms of modelling, B-MTGNN was constructed. The MTGNN model has been successfully applied to traffic prediction among other problems [27]. The model captures both temporal and spatial dependencies in the graph through temporal convolution and graph convolution layers and can additionally learn hidden relationships between nodes using a graph learning layer. Learning such relationships is useful for improving prediction performance. This is in contrast to relying on fixed, pre-assumed relationships between the nodes. We demonstrate this improvement experimentally in later sections.

The proposed Bayesian variation of the MTGNN model treats the model weights as random variables, allowing for the quantification of epistemic uncertainty through approximate Bayesian inference. Epistemic uncertainty quantifies the prediction error resulting from insufficient information [98] and can be reduced by acquiring more samples and informative features. The overall model development phase produced an operational model that can be readily used for forecasting the TPT graph. The performance of this model in predicting the trends up to 36 months in advance was evaluated. The model was ultimately used to forecast future trends and provide investment and strategic defence recommendations based on the predicted disparities between threats and PATs. Moreover, the analysis of past and future trends facilitated the development of the ATC model, identifying key phases in the life cycle of 98 technologies. This was achieved through the categorisation of the trends and the analysis of their slope and direction.

4.3.2 Graph Construction

The TPT graph consists of nodes representing the threats and PATs, supplemented by other feature nodes during the modelling step. The value of the node represents the trend level (NoI for threats and NoM for PATs). The edges link each threat to its PATs. The edge weight represents the gap between the threat trend and the connected PAT's trend. Formally, we define T as the total number of rows or months in the dataset, N as the total number of columns or features, and D as the feature dimension, which is set to 1 in our case. Let t denote the threat and p denote the PAT. The gap between t and p in a given month m is given by the following formula,

$$\mathrm{G}_{t,p}(m) = \frac{NoI_{m,t}}{\max_{i \in \mathcal{M}, u \in \mathcal{T}} NoI_{i,u}} - \frac{NoM_{m,p}}{\max_{i \in \mathcal{M}, v \in \mathcal{P}} NoM_{i,v}} \tag{4.1}$$

where $\mathbf{NoI} \in \mathbb{R}^{T \times N_{\text{threats}}}$ and $NoI_{m,t}$ represents the trend of threat t in month m. Similarly, $\mathbf{NoM} \in \mathbb{R}^{T \times N_{\text{pats}}}$ and $NoM_{m,p}$ represents the trend of PAT p in month m. Also, $\mathcal{T}$ is the set of all threats, $\mathcal{P}$ is the set of all PATs, and $\mathcal{M}$ is the set of all months. The formula normalises the node value over the maximum NoI in the dataset in the case of threats, and over the maximum NoM in the dataset in the case of PATs. This normalisation approach is crucial as it effectively bridges the significant scale disparities between NoI and NoM. The resulting gap value falls within the range -1 to 1, with a positive gap denoting a relatively lower research effort compared to the number of incidents, while a negative gap signifies a higher research effort. Ideally, the gap value should approach zero to indicate a balanced alignment between research efforts and incident occurrences.

In our study, we focus on the emerging and rapidly increasing threats identified in chapter 1, since these threats require the highest attention when investing in related technologies, compared to the other declining threats. The threats in our study are shown in Table 4.1.

To extract the PATs of each threat, we propose the E-GPT algorithm shown in Algorithm 1. Given a threat t, the algorithm starts by collecting relevant abstracts from Elsevier database. These abstracts include technology related keywords along with t. The second step is to iteratively prompt the GPT model to extract PATs from each abstract. The prompt to GPT contains an example for an expected answer in order to improve the performance. Given that there are many abstracts and many keywords that could be returned, the ranking of the PATs is then performed to obtain the top n PATs. In our study, we set n to 10. The ranking is done by considering the frequency defined as the number of times the PAT was returned by GPT. Intuitively, we give higher priority to the PATs with higher frequency. Within the same frequency groups, we perform a secondary ranking that prioritises the PATs that appear closely to technology-related keywords in the abstract, such as the word “solution”. This is done by computing the minimum distance within the abstract between the PAT and any of the keywords that belong to a predefined list of keywords S. The average distance is kept track of since a PAT can be returned multiple times by GPT. This secondary ranking is motivated by the fact that technology terms are frequently mentioned in close proximity to other keywords in the text. When they appear further in the text, they are more likely to be irrelevant.

Algorithm 1 Extractive GPT

```
input  : Threat $t$, number of PATs $n$, number of abstracts $b$, list of technology related
         keywords $S$
output: top $n$ PATs to $t$
$\mathcal{P} = \{\}_{\text{set}}$, $frequency = \{\}_{\text{dict}}$, $m\_distance = \{\}_{\text{dict}}$
$\mathcal{A} = \text{Query\_Elsevier\_for\_PATs\_Abstracts}(t, b)$
for $a$ in $\mathcal{A}$ do
    $\mathcal{U} = \text{Prompt\_GPT\_to\_Extract\_PATs}(t, a)$
    for $p$ in $\mathcal{U}$ do
        $frequency[p]{+}{+}$
        $m\_distance[p] = \text{avg\_min\_distance}(p, S, a)$
    $\mathcal{P} = \mathcal{P} \cup \mathcal{U}$
sort $\mathcal{P}$ by $frequency$ in descending order
sort by $m\_distance$ in ascending order within the same frequency groups
return $\{p_1, p_2, \ldots, p_n\}$ where $p_i \in \mathcal{P}$ for $i = 1$ to $n$
```

One important benefit of the extractive method is to ensure that the returned PATs reflect the state-of-the-art, since GPT can be outdated. Another benefit is controlling the source of information to ensure data reliability. However, to obtain more general answers and improve the accuracy, the list of PATs for each threat is further appended with an additional list that we obtain by asking GPT a direct question (*e.g.*, What are the PATs to t?). Finally, manual adjustment by human experts is performed to filter out irrelevant terms or add missing PATs. The final list of threats and PATs in the graph is shown in Table 4.1. The PATs abbreviations table can be found in Figure 4.6.

4.3.3 Bayesian Multivariate Time Series Graph Neural Network

To forecast the TPT graph, we developed a Bayesian variation of the MTGNN model proposed by Wu *et al.* [27]. This model was originally introduced as a general framework for forecasting multivariate time series, while leveraging state-of-the-art GNN components. The model's efficacy was extensively examined and validated across various datasets from different domains [27].

Within the context of our research, we apply the aforementioned model to the task of TPT graph forecasting. Furthermore, we enhance its capabilities by addressing the epistemic uncertainty inherent in the model's forecasts. This augmentation allows the model to articulate and quantify its uncertainty during the prediction process, a valuable asset when confronted with limited data or when seeking a measure of the model's confidence in its predictions [47].

The developed model is depicted in Figure 4.3. The first component in the model is the graph learning layer, which aims to adaptively learn the adjacency matrix in the graph. The learning process is designed in such a way that the resulting adjacency matrix leads to more accurate predictions in terms of node values. This approach is more effective than assuming predefined relationships since these can be hidden, unclear, or difficult to quantify. In our scenario, there are additional feature nodes beyond the threats and PATs, including NoM of the threats, ACA, and PH. The connections between these nodes and the threat/PAT nodes are not predefined. Therefore, we opt to let the model learn these hidden links and their weights within the graph.

Given randomly initialised node embeddings $\mathbf{E}_1$, $\mathbf{E}_2 \in \mathbb{R}^{N\times V}$, where V is a hyper-parameter denoting the node dimension, the graph learning layer extracts uni-directional relationships by computing the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ as follows,

$$\mathbf{M}_1 = \tanh(\alpha \mathbf{E}_1 \boldsymbol{\Theta}_1) \tag{4.2}$$

$$\mathbf{M}_2 = \tanh(\alpha \mathbf{E}_2 \boldsymbol{\Theta}_2) \tag{4.3}$$

$$\mathbf{A} = \mathrm{ReLU}(\tanh(\alpha(\mathbf{M}_1\mathbf{M}_2^T - \mathbf{M}_2\mathbf{M}_1^T))) \tag{4.4}$$

$$\mathbf{A}[i, -\mathrm{argtopk}(\mathbf{A}[i,:])] = 0, \forall i \in [N] \tag{4.5}$$

where $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2 \in \mathbb{R}^{V\times V}$ are model parameters, $\mathbf{M}_1$, $\mathbf{M}_2 \in \mathbb{R}^{N\times V}$, α is a hyper-parameter for controlling the saturation rate of the activation function, and argtopk(.)

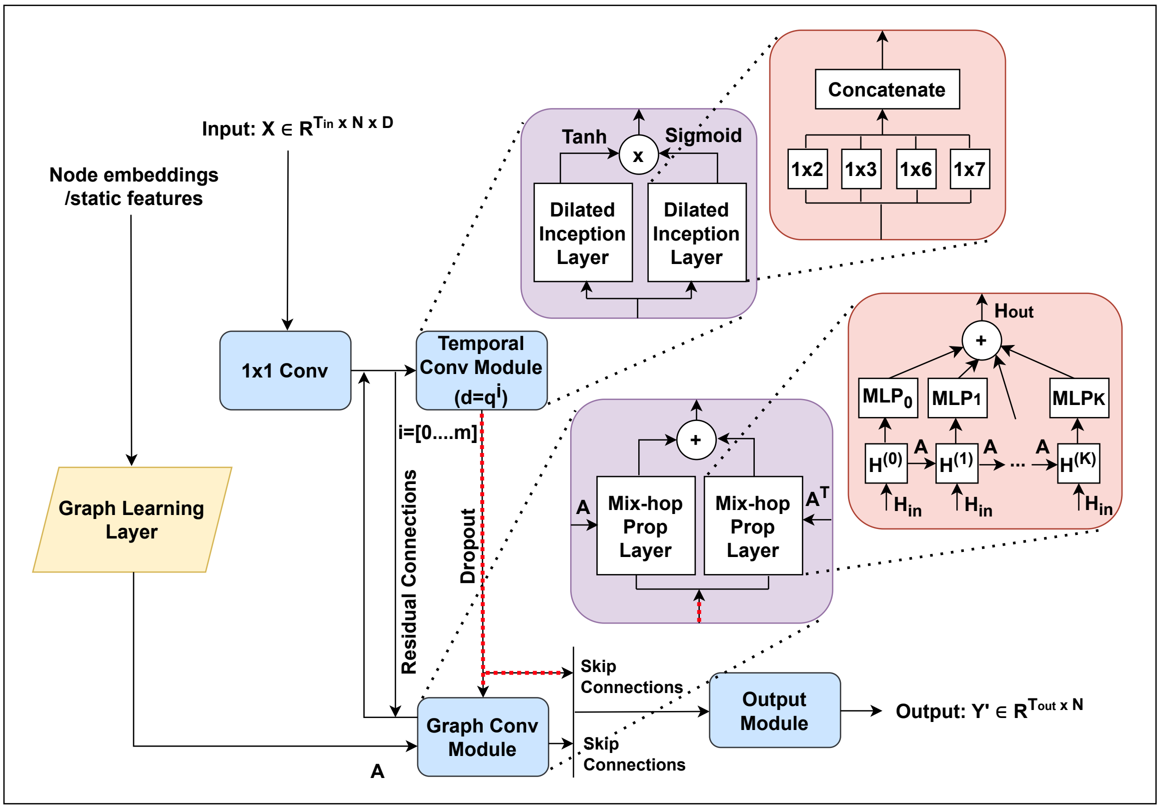


Figure 4.3: The Bayesian Multivariate Time Series Graph Neural Network model (B-MTGNN). The model learns the adjacency matrix of the graph through the graph learning layer, while capturing temporal and spatial dependencies using temporal and graph convolution modules. The dilation factor d increases exponentially with the increase in the number of layers m at the rate of q. The red arrows indicate the use of dropout during inference to approximate a Bayesian model.

returns the index of the top k closest nodes to be selected as neighbours. This selection strategy makes the adjacency matrix sparse while reducing the computation cost of the graph convolution [27].

The graph convolution module aims to fuse a node's information with its neighbours' information to capture the spatial dependencies. As shown in Figure 4.3, it consists of two mix-hop propagation layers for processing inflow and outflow information for each node. The mix-hop propagation layer mainly consists of two steps. The first step is the information propagation step defined as follows,

$$\mathbf{H}^{(k)} = \beta\mathbf{H}_{in} + (1-\beta)\tilde{\mathbf{A}}\mathbf{H}^{(k-1)} \tag{4.6}$$

where $\mathbf{H}^{(k)} \in \mathbb{R}^{B \times C \times N \times O}$. Here, B is the batch size, C is the number of convolution channels, and O is the last dimension of the output from the previous layer. β is a hyper-parameter for controlling the amount of information to be retained from the root node's original states, and $\mathbf{H}_{in} \in \mathbb{R}^{B \times C \times N \times O}$ denotes the input hidden states from the previous layer. The second step is the information selection step given by the following formula,

$$\mathbf{H}_{out} = \sum_{k=0}^{K} \mathbf{H}^{(k)} \mathbf{W}^{(k)} \tag{4.7}$$

where $\mathbf{H}_{out} \in \mathbb{R}^{B \times I \times N \times O}$ denotes the output hidden states of the current layer, where I is a hyper-parameter that denotes the number of residual channels. K is the propagation depth, and $\mathbf{W}^{(k)} \in \mathbb{R}^{I \times C}$ is a feature selector for controlling what to be retained from the original node's information. Further details about these steps can be found in [27].

As shown in Figure 4.3, the temporal convolution module captures the temporal dependencies by utilising dilated inception layers. Given that the receptive field increases exponentially with the increase in the number of layers, the dilation strategy is employed to handle large sequences while reducing the model complexity [140]. The inception strategy is used to handle temporal patterns with different ranges by using filters with multiple sizes [141]. Formally, given a sequence input $\mathbf{z} \in \mathbb{R}^{T_{\text{in}}}$ and four filters of the form $\mathbf{f}_{1\times 2} \in \mathbb{R}^2$, $\mathbf{f}_{1\times 3} \in \mathbb{R}^3$, $\mathbf{f}_{1\times 6} \in \mathbb{R}^6$, and $\mathbf{f}_{1\times 7} \in \mathbb{R}^7$, the dilated inception layer takes the following form,

$$\mathbf{z} = \text{concat}(\mathbf{z} \star \mathbf{f}_{1\times 2}, \mathbf{z} \star \mathbf{f}_{1\times 3}, \mathbf{z} \star \mathbf{f}_{1\times 6}, \mathbf{z} \star \mathbf{f}_{1\times 7}) \tag{4.8}$$

Let d denote the dilation factor. The dilated convolution denoted by $\mathbf{z} \star \mathbf{f}_{1\times k}$ is defined as follows,

$$\mathbf{z} \star \mathbf{f}_{1\times k}(t) = \sum_{s=0}^{k-1} \mathbf{f}_{1\times k}(s) \mathbf{z}(t - d \times s) \tag{4.9}$$

Deep neural networks often suffer from the vanishing gradient problem, where gradients become increasingly small as they propagate backward through many layers during training. This can hinder the training process, especially in very deep networks. Residual and skip connections are techniques commonly used in neural networks, particularly in deep learning architectures like convolutional neural networks, to mitigate the vanishing gradient problem. As shown in Figure 4.3, they involve adding values from previous layers to the output, which helps in preserving gradient flow during backpropagation, thereby addressing the vanishing gradient issue. To obtain the final outputs, the output module maps the hidden features to the desired output dimension.

Because we have relatively short time series within our refined data (*i.e.*, 138 monthly data points between July 2011 and December 2022), it is vital to extract the model's uncertainty. Deterministic neural network models that do not involve randomness are insufficient for this task, since they offer single-point predictions of model parameters. Instead, we employ a Bayesian approach to capture epistemic uncertainty. Specifically, we employ the Monte Carlo dropout method proposed by Gal *et al.* [47], who showed that the use of dropout neurons during inference provides a Bayesian approximation of the deep Gaussian processes. The use of dropout mask in our model during inference is highlighted in red arrows (Figure 4.3). Therefore, during the prediction phase, the trained model runs multiple times, which results in a distribution of prediction (representing the uncertainty) rather than a single point (Figure 4.4).

4.3.4 Experimental Settings

In our experimental setup, we partitioned the dataset into three distinct subsets: 43% for training, 30% for validation, and 27% for testing. This allocation was carefully chosen to ensure that ample data was available for rigorous testing of the model's performance. Specifically, our model's input comprises 10 months of historical data, corresponding to 10 time steps, while the output encompasses forecasts for the subsequent 36 months. This forecasting framework constitutes a multi-horizon approach, wherein predictions are made for multiple future time steps simultaneously.

Our experimental findings support the utilisation of a non-autoregressive approach in our forecasting methodology. Training the model to predict multiple time steps concurrently, without dependence on previously generated predictions, yielded higher accuracy and more comprehensive pattern capture. Unlike autoregressive models, our approach solely utilises past observed values for forecasting the subsequent months. By avoiding reliance on prior predictions, our model mitigates the error propagation problem, leading to enhanced forecasting accuracy and efficacy [142].

4.3.5 Hyper-parameter Optimisation

We performed a random search with 60 iterations, in order to find the set of hyper-parameters that produces the model with the lowest validation error. Random search is a simple method for hyper-parameter optimisation, with several advantages including efficiency, flexibility, and robustness. Extensive research in the literature has demonstrated that this method outperforms grid search in numerous cases [111]. For each set of hyper-parameters, we trained the model using the mean absolute error (MAE) as the loss function, and while using ADAM as the optimisation algorithm [112]. The model then was validated by forecasting the graph 3 years in advance, and the average performance was recorded. Once the set of hyper-parameters with the minimum error was found, we assessed the model's performance on the testing set and recorded the corresponding error. As a last step, we employed the optimal hyper-parameter settings to train the model using the entire dataset, followed by generating forecasts for the forthcoming three years, extending up to December 2025.

The first group of hyper-parameters includes the learning rate with values that range from 1×10^{-4} to 1×10^{-2}, the number of epochs with values up to 200, the number of layers in the range 1 to 2, and the dropout value between 0.2 and 0.7. Other hyper-parameters are specific to the GNN model including the graph convolution depth in the range 1 to 3, the convolution channels in the range 4 to 16, the activation function controller α (see equation (4.4)) within the range of 0.05 to 9, and the information propagation controller β (see equation (4.6)) ranging from 0.05 to 0.8. The range of these values was obtained from the literature and online code repositories [27, 143].

4.3.6 Model Evaluation

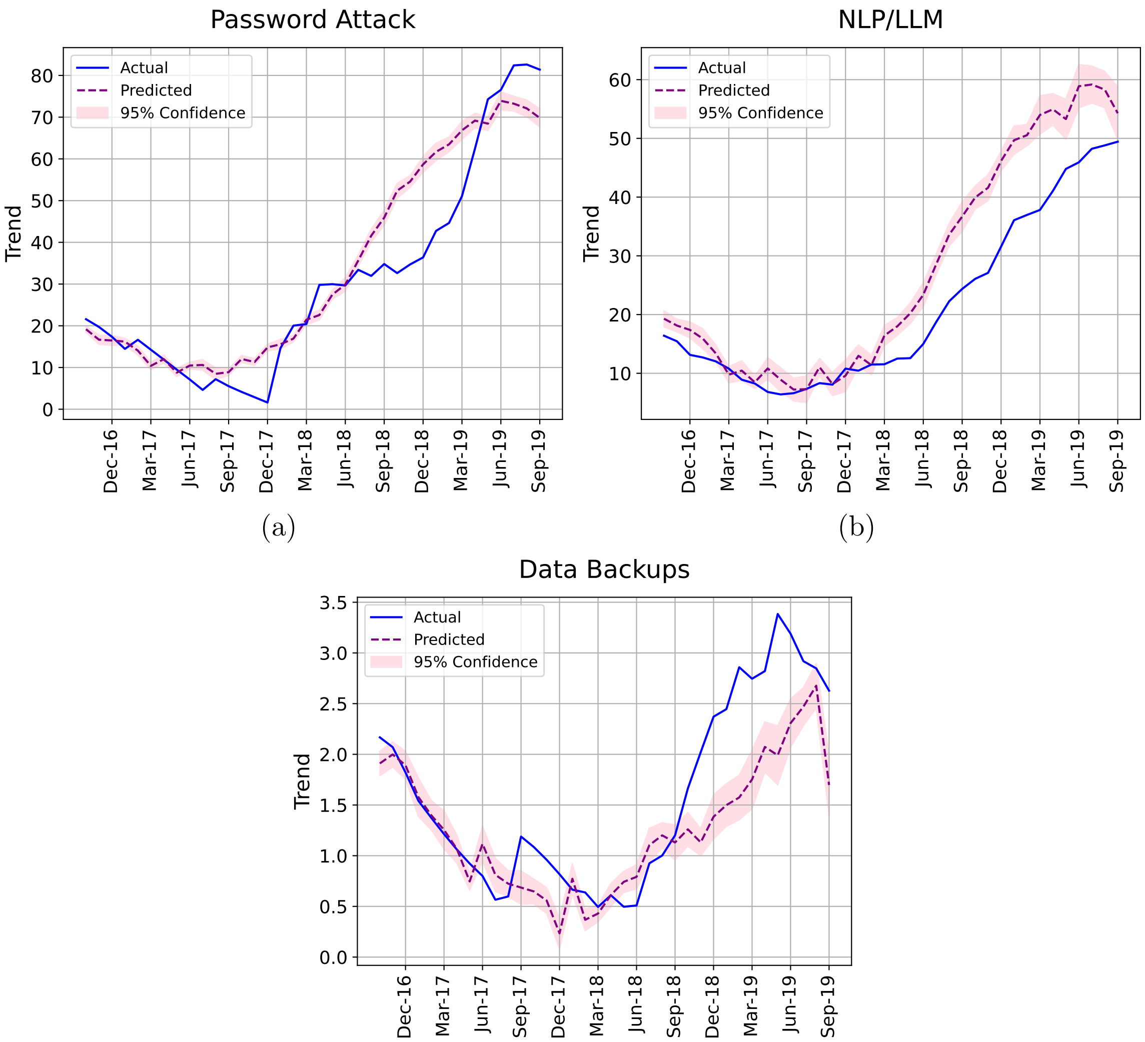
In the evaluation phase (validation and testing), we used two evaluation metrics namely the Root Relative Squared Error (RSE) and the Relative Absolute Error (RAE) [144]. These metrics compute the model's error relative to the error of a simple model that can predict the average trend of each node. Formally, let $Y_{j,m}$ denote the actual value in the test set of node j during month m, and $\hat{Y}_{j,m}$ denote the predicted value, where $\mathbf{Y}, \hat{\mathbf{Y}} \in \mathbb{R}^{N \times T_{\text{test}}}$. Then, RSE and RAE are given by the following formulas,

$$RSE = \frac{\sqrt{\sum_{(j,m)\in\Omega_{Test}} (Y_{j,m} - \hat{Y}_{j,m})^2}}{\sqrt{\sum_{(j,m)\in\Omega_{Test}} (Y_{j,m} - \text{mean}(\mathbf{Y}_j))^2}} \tag{4.10}$$

$$RAE = \frac{\sum_{(j,m)\in\Omega_{Test}} |Y_{j,m} - \hat{Y}_{j,m}|}{\sum_{(j,m)\in\Omega_{Test}} |Y_{j,m} - \text{mean}(\mathbf{Y}_j)|} \tag{4.11}$$

These metrics provide readable evaluation, regardless the scale of the data. For both metrics, the lower value is better.

The model validation results are provided in Figure 4.4. As shown in the figure, the predicted data points are aligned with the ground truth, and the model is able to capture the time series patterns effectively. For some nodes (*e.g.*, NLP/LLM), we notice a slight increase in the confidence interval as we move towards the later years, suggesting less certainty about the prediction in those years. This increase in the uncertainty can be reduced with more knowledge in terms of new features or more samples [6]. Overall, in terms of validation error, the average RSE computed over 142 nodes is 0.52, and the average RAE is 0.66, which provides a noticeable improvement over the benchmark model.



(a)

(b)

(c)

Figure 4.4: The B-MTGNN validation results of predicting threats and PATs from October, 2016 to September, 2019. (a) Password Attack with RAE = 0.37. (b) NLP/LLM with RAE = 0.53. (c) Data Backups with RAE = 0.51. The 95% confidence interval of the predicted distribution using the Bayesian approach is shown in pink colour.

4.4 Results

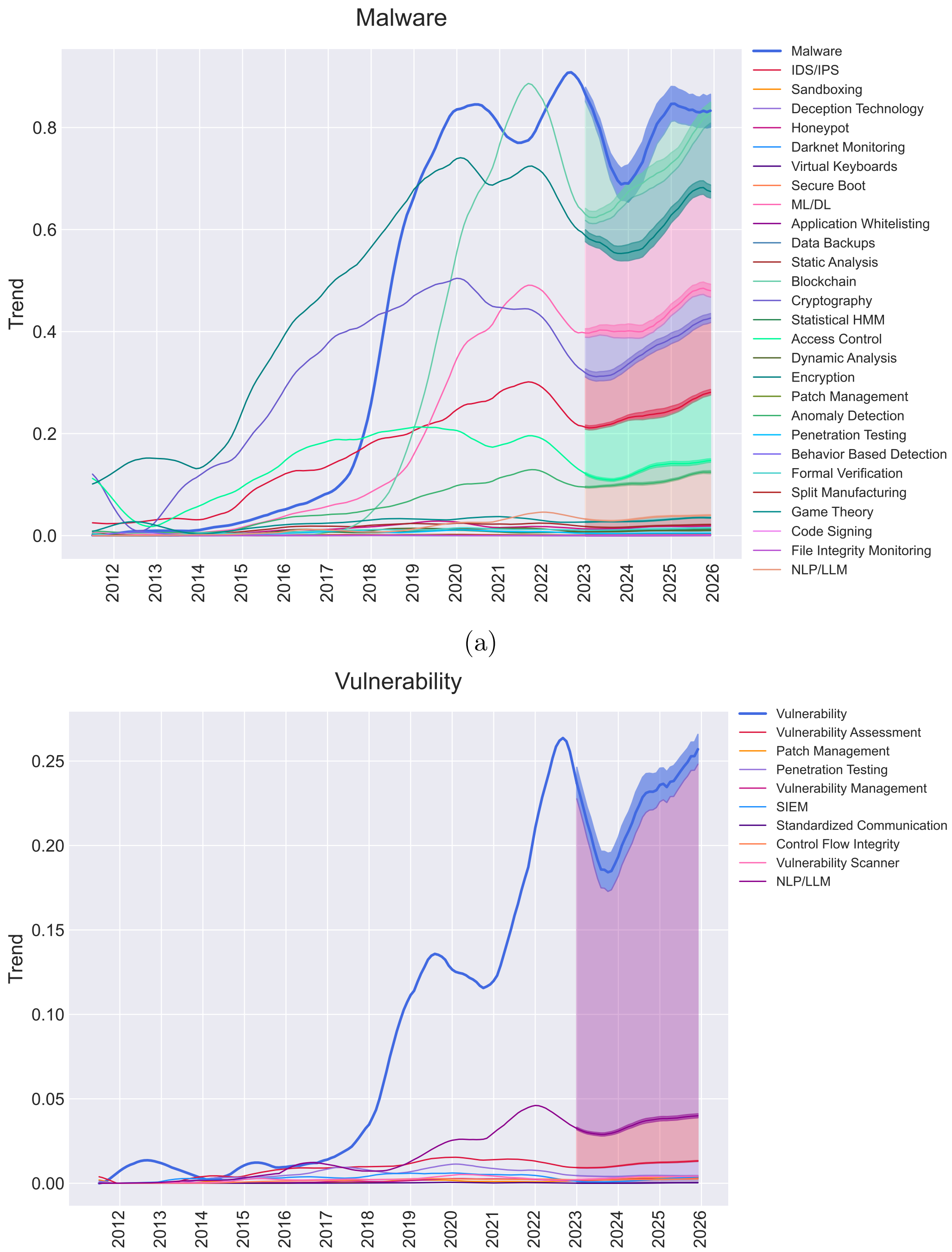
4.4.1 Trend Forecast

The forecast of the cyber threats and their PATs in the coming 3 years is provided in Figure 4.5. Here, we focus on the most important threats for which there will be a significant gap in the future with the respective PATs based on the forecast, while including threats from both categories (the rapidly increasing and emerging threats). We also focus on the PATs that will likely have a positive gap with the relevant threat. In other words, the PATs shown are those for which the trend was forecasted to be below the trend of the relevant threat. In Figure 4.5, the gap area is visually represented using the same colour as the corresponding PAT curve.

The malware attack stands out for having the most significant gaps with respect to its PATs compared to other types of attacks. The forecast illustrated in Figure 4.5a indicates a considerable disparity reaching a value of 0.8, and expected to persist over the next three years between malware and various PATs, including Application Whitelisting, File Integrity Monitoring, and Darknet Monitoring. Other PATs such as Blockchain, Anomaly Detection, and ML/DL are also expected to trail behind malware. However, the gaps of these PATs with respect to malware are comparatively smaller, narrowing clearly in the case of Blockchain, thanks to the recent growing body of research in these fields [145, 146, 147].

The next concern is the vulnerability related attacks shown in Figure 4.5b. Here, we observe a consistently widening gap with some PATs including Standardised Communication, Security Information and Event Management (SIEM), and Control Flow Integrity. Compared to these PATs, Vulnerability Assessment and NLP/LLM are expected to be more visible, even though the anticipated gaps are still large, exceeding a value of 0.2.

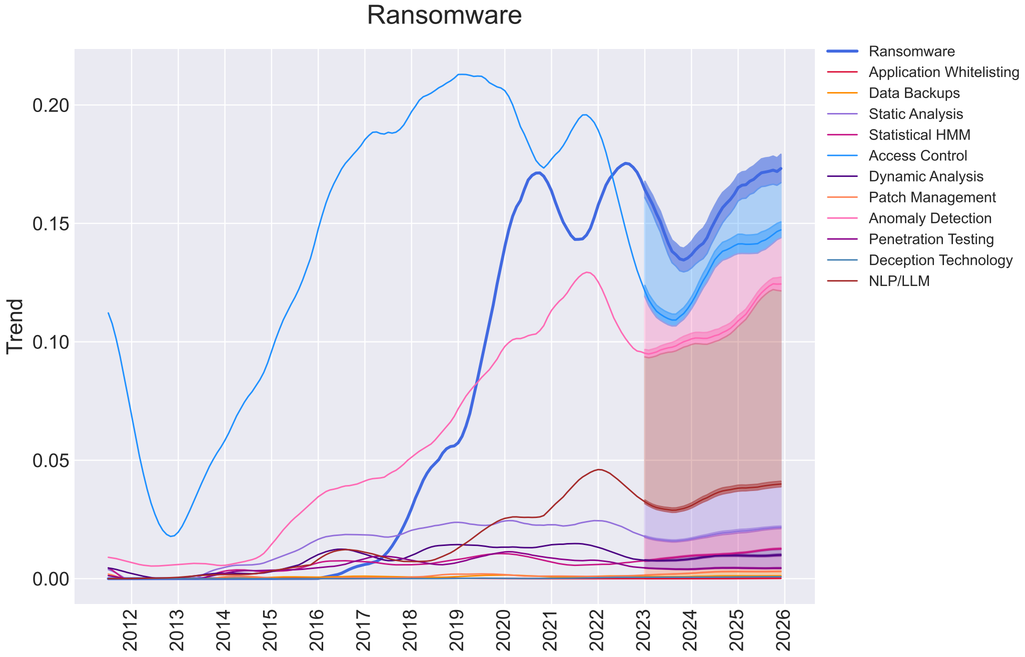
Concerning the more recently emerging threats (Figure 4.5c and Figure 4.5d), ransomware will likely exhibit gap values above 0.1 with respect to several PATs including Application Whitelisting, Deception Technology, and Data Backups, while having relatively smaller gaps with Access Control and Anomaly Detection (below 0.05). The adversarial attack is expected to have the largest gap of 0.09 with respect to Spatial Smoothing, Defensive Distillation, and Noise Injection, and the smallest gap with respect to NLP/LLM (around 0.05).



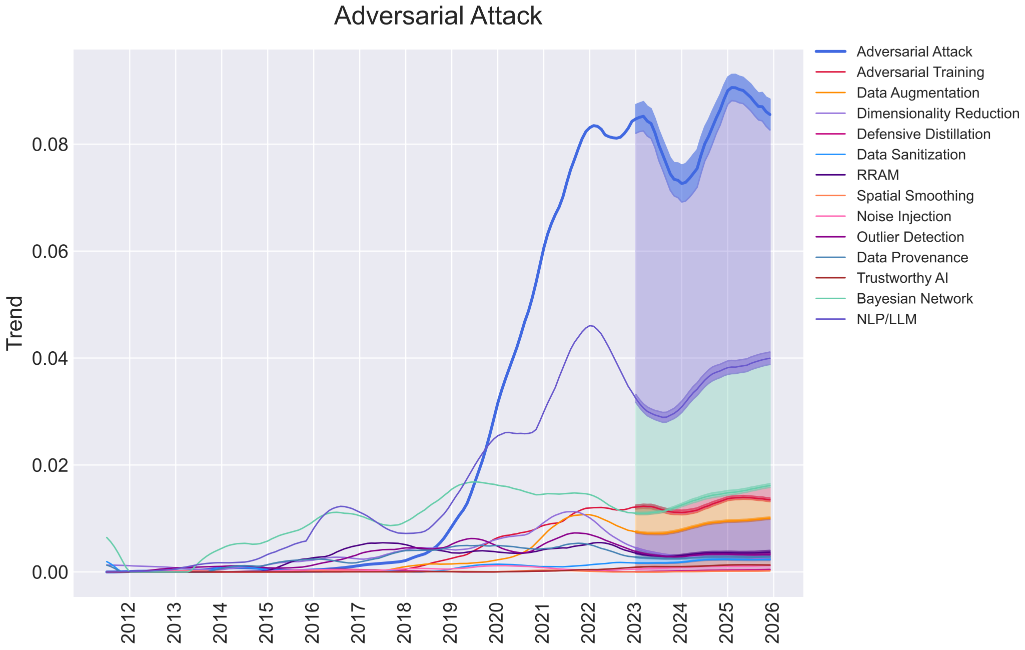
(a)

(b)

Figure 4.5: The forecast of the trend for two rapidly increasing threats and their PATs. The period of the trend plots is between July, 2011 and December, 2025, with the period between January, 2023 and December, 2025 forecasted using B-MTGNN. The shown PATs are those for which the trend is predicted to be lower than the trend of the corresponding threat. The gaps are highlighted in the same colour as the corresponding PAT curve. (a) Malware (b) Vulnerability. The curves are smoothed using exponential smoothing with $\alpha = 0.1$ to reduce the noise and capture the trend. The 95% confidence interval is shown for each trend prediction.



(c)



(d)

Figure 4.5: Continued...The forecast of the trend for two emerging threats and their PATs. The period of the trend plots is between July, 2011 and December, 2025, with the period between January, 2023 and December, 2025 forecasted using B-MTGNN. The shown PATs are those for which the trend is predicted to be lower than the trend of the corresponding threat. The gaps are highlighted in the same colour as the corresponding PAT curve. (c) Ransomware (d) Adversarial Attack. The curves are smoothed using exponential smoothing with $\alpha = 0.1$ to reduce the noise and capture the trend. The 95% confidence interval is shown for each trend prediction.

4.4.2 Trend Categories

Our analysis for the future gaps between the threats and PATs allowed us to categorise the gap trend into four main categories, as shown in Table 4.2 and Table 4.3. In these tables, PATs are listed in descending order of the gap, while considering different types of threats and threat categories. Here, we computed the average gap in each year and recorded the result for each of the three years (2023 to 2025).

The first category is the Strictly Widening Gaps (SWG) shown in the first half of Table 4.2. These are the gaps that are predicted to be consistently increasing between the years 2023 and 2025. Examples of such gaps include the gaps between the vulnerability related threats and each of Standardised Communication (SC), SIEM, and Control Flow Integrity (CFI). Similarly, the gaps for IoT Device Attack with respect to Merkle Signature (MS), Secure Boot (SB), and Multi-Factor Authentication (MFA) are consistently widening, even though they exhibit smaller values.

The second category is the Overall Widening Gaps (OWG) shown in the second half of Table 4.2. These gaps are anticipated to increase in the year 2025 compared to 2023, with expected fluctuations in between. Among the top in the list are the gaps between malware and Application Whitelisting (AW), File Integrity Monitoring (FIM), and Darknet Monitoring (DM). Other examples with smaller gap values include the gaps between deepfake and each of 3 Dimensional Face Reconstruction (3DFR) and Digital Watermark (DW).

Third is the Overall Narrowing Gaps (ONG) illustrated in the upper part of Table 4.3. These gaps are likely to decrease in the year 2025 compared to their values in 2023, despite the expected fluctuations in between. Among the top in the list are the gaps between malware and Cryptography (CR) and between ransomware and Access Control (AC). Examples with much smaller gap values include the gaps between APT and Deception Technology (DT), as well as between APT and Least Privilege (LP).

Table 4.2: Widening Gaps

Strictly Widening Gaps					
Threat	**PAT**	**Gap Forecast**			**GD**
		2023	**2024**	**2025**	
Vulnerability	SC	0.202	0.218	**0.244**	↑↑
Vulnerability	SIEM	0.201	0.217	**0.241**	↑↑
Vulnerability	CFI	0.200	0.216	**0.241**	↑↑
Account Hijacking	LP	0.186	0.199	**0.229**	↑↑
Account Hijacking	SM	0.186	0.199	**0.229**	↑↑
Account Hijacking	MFA	0.182	0.195	**0.226**	↑↑
Ransomware	AW	0.146	0.149	**0.170**	↑↑
Ransomware	DT	0.146	0.149	**0.169**	↑↑
Ransomware	DB	0.146	0.148	**0.169**	↑↑
IoT Device Attack	MS	0.043	0.050	**0.055**	↑↑
IoT Device Attack	SB	0.043	0.049	**0.054**	↑↑
IoT Device Attack	MFA	0.039	0.046	**0.052**	↑↑
Overall Widening Gaps					
Threat	**PAT**	**Gap Forecast**			**GD**
		2023	**2024**	**2025**	
Malware	AW	0.766	0.763	**0.837**	↓↑
Malware	FIM	0.766	0.763	**0.836**	↓↑
Malware	DM	0.766	0.763	**0.836**	↓↑
Ransomware	NLP/LLM	0.116	0.114	**0.131**	↓↑
Adversarial Attack	SS	0.080	0.079	**0.088**	↓↑
Adversarial Attack	DD	0.080	0.079	**0.088**	↓↑
Adversarial Attack	NI	0.079	0.078	**0.087**	↓↑
Account Hijacking	AC	0.074	0.068	**0.086**	↓↑
Phishing	AC	0.062	0.049	**0.068**	↓↑
Ransomware	AD	0.049	0.046	**0.051**	↓↑
Deepfake	3DFR	0.047	0.046	**0.051**	↓↑
Deepfake	DW	0.046	0.045	**0.049**	↓↑

Items are displayed in descending order of the gap. GD refers to the Gap Directions. Please refer to Figure 4.6 for the PAT abbreviations.

Table 4.3: Narrowing Gaps

Overall Narrowing Gaps					
Threat	**PAT**	**Gap Forecast**			**GD**
		2023	**2024**	**2025**	
Malware	CR	**0.449**	0.401	0.429	↓↑
Ransomware	AC	**0.033**	0.018	0.027	↓↑
Deepfake	NLP/LLM	**0.017**	0.011	0.012	↓↑
MITM	SSP	**0.014**	0.013	0.013	↓ →
MITM	PT	**0.010**	0.009	0.009	↓ →
MITM	VPN	**0.007**	0.006	0.006	↓ →
APT	UBA	$\mathbf{14.2\times10^{-4}}$	7.1×10^{-4}	9.5×10^{-4}	↓↑
APT	NS	$\mathbf{12.9\times10^{-4}}$	6.5×10^{-4}	7.8×10^{-4}	↓↑
APT	DLP	$\mathbf{11.2\times10^{-4}}$	5×10^{-4}	8.2×10^{-4}	↓↑
Disinformation	CA	$\mathbf{7.2\times10^{-4}}$	4.1×10^{-4}	5.5×10^{-4}	↓↑
APT	DT	$\mathbf{8.1\times10^{-4}}$	2×10^{-4}	4.1×10^{-4}	↓↑
APT	LP	$\mathbf{7.8\times10^{-4}}$	2.2×10^{-4}	3.9×10^{-4}	↓↑
Strictly Narrowing Gaps					
Threat	**PAT**	**Gap Forecast**			**GD**
		2023	**2024**	**2025**	
Malware	EN	**0.199**	0.184	0.174	↓↓
Malware	BC	**0.129**	0.064	0.057	↓↓
MITM	PKI	**0.006**	0.004	0.003	↓↓

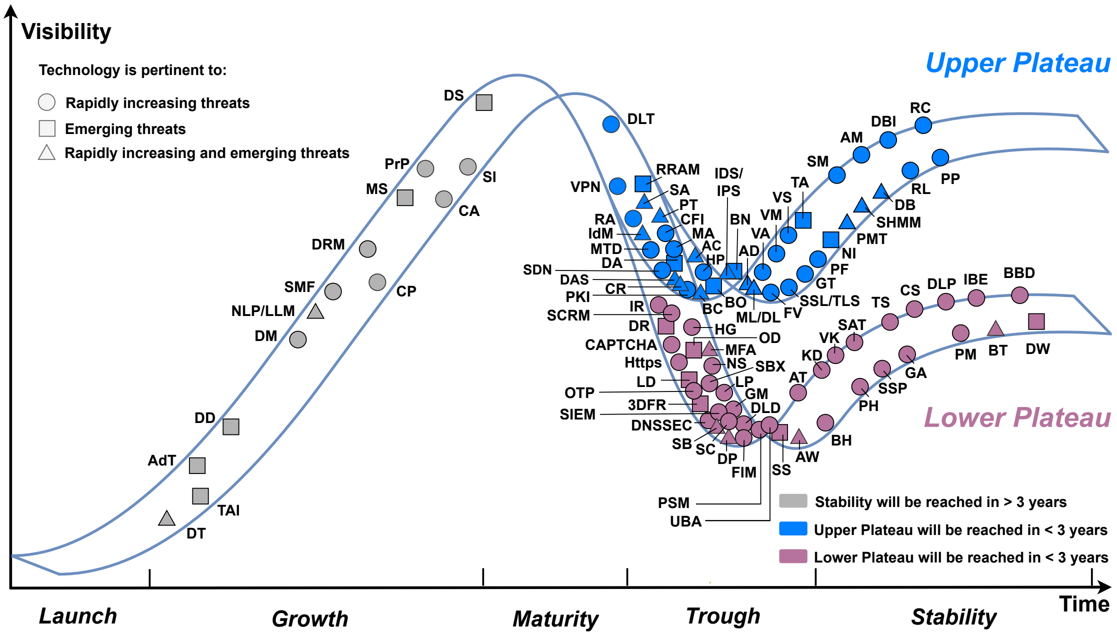
Items are displayed in descending order of the gap. GD refers to the Gap Directions. Please refer to Figure 4.6 for the PAT abbreviations.

The fourth and last category is the Strictly Narrowing Gaps (SNG). As shown in the lower part of Table 4.3, these gaps are consistently decreasing between the years 2023 and 2025. Examples include the gaps between malware and each of Encryption (EN) and Blockchain (BC), and the gap between MITM and Public Key Infrastructure (PKI). It is worth noting that this category comprises the fewest items, indicating the rarity of these gaps.

4.4.3 Alleviation Technologies Cycle

Our large scale analysis for the PATs' historical data and future predictions spanning three years facilitated the development of a generalisable model that provides a comprehensive understanding of the progression of these PATs as they transition through 5 phases, namely the launch, growth, maturity, trough, and stability. This model is referred to as the Alleviation Technologies Cycle (or ATC), which is depicted in Figure 4.6. During the launch phase, a new technology emerges and is adopted by few agencies for a brief period. Subsequently, there is a rapid surge in both the frequency and prominence of the technology as more security agencies become acquainted with and adopt the new PAT. Typically, PATs exhibit numerous variations in terms of speed of progression. For most of the PATs, we observe a slow progression during the growth phase compared to other types of technologies. This is due to the presence of various challenges in the world of cyber security including the resistance of attackers to the new security solution [148]. As the visibility reaches its peak, the PAT enters the maturity phase, characterised by a sustained and stable pattern for a short period of time. This is followed by a temporary decline into the trough where enthusiasm diminishes as trials and executions fall short of expectations. Based on the forecast, we identified two possible troughs that the PAT can reach. One of these troughs is deeper than the other, depending on the usability of the PAT and the demand for it. Eventually, the PAT recovers and moves to either a higher or lower plateau, depending on which trough it originated from. This recovery takes place as additional examples showcasing how the technology can advantage the organisation begin to solidify and gain broader comprehension. Within the plateaus, mainstream adoption accelerates as the criteria for evaluating viability become more distinct, showcasing the technology's widespread market utility and effectiveness [115].

As depicted in Figure 4.6, the positioning of the PATs on the cycle is determined by analysing their current trend slope, their historical patterns, and their future projections. During the trough phase, PATs exhibit either a trajectory towards the upper plateau or the lower plateau. By leveraging the predicted trends, illustrated in Figure 4.5, we were able to indicate the future destination for some PATs near the trough using distinct colours (blue or purple). For instance, Distributed Ledgers Technology (DLT), Resistive Random-Access Memory (RRAM), and Virtual Private Network (VPN) are displayed in blue colour, indicating their likelihood of transitioning toward



A – C		D – I		K – R		S – V	
3DFR	3D Face Reconstruction	DA	Data Augmentation	KD	Keystroke Dynamics	SA	Static Analysis
AC	Access Control	DAS	Dynamic Analysis	LD	Liveness detection	SAT	Strong Authentication
AD	Anomaly Detection	DB	Data Backups	NLP/LLM	Natural Language Processing/ Large Language Models	SB	Secure Boot
AM	Activity Monitoring	DBI	Dynamic Binary Instrumentation	LP	Least Privilege	SBX	Sandboxing
AT	Attack Tree	DD	Defensive Distillation	MA	Mutual Authentication	SC	Standardised Communication
AW	Application Whitelisting	DLD	Data Leakage Detection/Prevention	MFA	Multi-factor Authentication	SCRM	Supply Chain Risk Management
AdT	Adversarial Training	DLP	Data Loss Prevention	ML/DL	Machine Learning/Deep Learning	SDN	Software Defined Network
BBD	Behaviour-based Detection	DLT	Distributed Ledgers Technology	MS	Merkle Signature	SHMM	Statistical Hidden Markov Model
BC	Blockchain	DM	Darknet Monitoring	MTD	Moving Target Defence	SI	Source Identification
BH	Blackholing	DNSSEC	Domain Name System Security Extensions	NI	Noise Injection	SIEM	Security Information and Event Management
BN	Bayesian Network	DP	Data Provenance	NS	Network Segmentation	SM	Session Management
BO	Biometrics	DR	Dimensionality Reduction	OD	Outlier detection	SMF	Split Manufacturing
BT	Blacklisting	DRM	Dynamic Resource Management	OTP	One Time Password	SS	Spatial Smoothing
CA	Continuous Authentication	DS	Data Sanitisation	PF	Packet Filtering	SSL/TLS	Secure Sockets Layer/Transport Layer Security
CAPTCHA	Public Turing Comp Human	DT	Deception Technology	PH	Password Hashing	SSP	Secure Simple Pairing
CFI	Control Flow Integrity	DW	Digital Watermark	PKI	Public Key Infrastructure	TA	Taint Analysis
CP	Certificate Pinning	FIM	File Integrity Monitoring	PM	Password Management	TAI	Trustworthy AI
CR	Cryptography/Encryption	FV	Formal Verification	PMT	Patch Management	TS	Traffic Shaping
CS	Code Signing	GA	Graphical Authentication	PP	Password Policy	UBA	User Behaviour Analytics
		GM	Graphical Model	PSM	Password Strength Meters	VA	Vulnerability Assessment
		GT	Game Theory	PT	Penetration Testing	VK	Virtual Keyboards
		HG	Hypergame	PrP	Privacy Preserving	VM	Vulnerability Management
		HP	Honeypot	RA	Risk Assessment	VPN	Virtual Private Network
		Https	Hypertext Transfer Protocol Secure	RC	Rank Correlation	VS	Vulnerability Scanner
		IBE	Identity-based Encryption	RL	Rate Limiting		
		IDS/IPS	Intrusion Detection/Prevention System	RRAM	Resistive Random Access Memory		
		IR	Image Recognition				
		IdM	Identity Management				

Figure 4.6: The Alleviation Technologies Cycle (ATC). The PATs go through 5 stages, namely, launch, growth, maturity, trough, and stability. ATC captures the state of each PAT in 2023, where the colour of the PAT indicates which slope it would follow based on the model prediction until 2025 (*e.g.*, blue: upper plateau or purple: lower plateau). The PATs with unknown final destination are coloured in grey.

the upper plateau. It is important to note that during the initial three phases, the ultimate destination of a particular PAT, whether it will reach the upper or lower plateau, often remains uncertain and challenging to predict, thus denoted in grey. In addition, we distinguish each PAT by employing distinct shapes, indicating their relevance to either rapidly increasing or emerging threats (or possibly to both categories).

The ATC is similar to the well-known GHC [115], with some important differences. The ATC has a slower rate of growth compared to GHC given that it is specific to the challenging field of cyber security, as previously mentioned. Another notable distinction is the presence of two distinct troughs (and two plateaus) in the ATC instead of a single trough observed in GHC. This difference arises because the ATC is a specialised variant of GHC designed specifically for the cyber security domain.

In the early stage of the growth phase, PATs are mostly related to the emerging threats, as can be observed in Figure 4.6. These PATs include Defensive Distillation (DD), Deception Technology (DT), Trustworthy AI (TAI), and Adversarial Training (AdT). In the later stages of the growth phase, different types of PATs can be observed including those relevant to the rapidly increasing threats. Examples include NLP/LLM, Split Manufacturing (SMF), Certificate Pinning (CP), and Continuous Authentication (CA). After the peak, and into the upper trough, we find a combination of PATs (relevant to threats from different categories) sliding down, including Distributed Ledgers Technology (DLT), Control Flow Integrity (CFI), Static Analysis (SA), Dynamic Analysis (DAS), and Data Augmentation (DA). On the upper plateau, most of the PATs are relevant to the rapidly increasing threats including Session Management (SM), Rate Limiting (RL), Activity Monitoring (AM), Rank Correlation (RC), and Password Policy (PP). Many PATs are falling into the lower trough, and those are mostly relevant to the rapidly increasing threats. They include Supply Chain Risk Management (SCRM), One Time Password (OTP), Domain Name System Security Extensions (DNSSEC), and File Integrity Monitoring (FIM). On the lower plateau, most of the PATs are relevant to the rapidly increasing threats. These include Password Management (PM), Code Signing (CS), Data Loss Prevention (DLP), Identity-based Encryption (IBE), and Behaviour-based Detection (BBD).

4.5 Comparative Analysis

4.5.1 Ablation Study

In this section, we show experimentally the effect of our proposed external features (NoM of attacks, ACA, and PH) on the performance of the MTGNN model. In addition, we demonstrate the effectiveness of the graph convolution layers and the graph learning layer. To this end, we conducted multiple experiments to evaluate the performance of eight different variations of the MTGNN model in predicting the trends up to 3 years in advance, while using unseen data. For each model variant, we split the dataset into 70% training/validation and 30% testing. Each model undergoes random search with 60 iterations to optimise the set of hyper-parameters, and the final testing errors RSE and RAE are averaged over 10 experiments.

The first four models (Table 4.4) do not utilise our external features during the prediction and rather rely on the ground truth. The first model does not include any graph convolution layer and only performs temporal convolution. In the next three variations, we experimented with models that utilise graph convolution layers including two models that use a predefined adjacency matrix (uni-directional and bi-directional variants), and one model that uses the adaptively learned adjacency matrix through the graph learning layer. Intuitively, within the uni-directional predefined adjacency matrix, the threat node points to the relevant PAT node, since the threat often precedes the security measure. In the case of bi-directional adjacency matrix, both types of nodes point to each other. In the case of adaptive learning, we allow the model to learn these relationships. We note that in the case of predefined adjacency matrix, the edge weight is set to 1 or 0 (depending on whether two nodes are connected), since it is challenging to identify the level of relationship, which can be rather learned adaptively. This relationship weight is only used during model training, and not to be confused with the edge weight in the original graph, which represents the gap (equation 4.1).

The rest of four models utilise the external features with the following variations. The first model does not include any graph convolution layer and only performs temporal convolution. The second and third models utilise graph convolution layers with a predefined adjacency matrix (uni-directional and bi-directional variants). In the uni-directional variant, the feature node, such as ACA points to the threat nodes (*e.g.*, wars and conflicts precede the attack), and the threat node points to the relevant PAT node. The fourth model employs the graph learning layer along with the graph convolution layers to adaptively learn the relationships in the graph.

The evaluation results are presented in Table 4.4. The use of the external features made a significant difference, reducing the relative error to a value below 1, which provides an improvement over the simple model. The results also show that using graph convolution leads to a lower error compared to relying solely on the temporal convolution. In addition, we observe that the use of uni-directional predefined adjacency matrix consistently resulted in a better performance compared to the use of bi-directional variant. This is consistent with the findings in [27]. However, the use of graph learning layer to learn the adjacency matrix resulted in a better performance than using any predefined adjacency matrix. This is explained by the fact that the graph structure is not optimal and should be updated during training [27]. Overall, the best performance was obtained when combining the graph convolution layers (in addition to the temporal convolution layers), the graph learning layer, and the external features. This justifies the use of these layers along with our proposed features in our future forecast.

4.5.2 Comparative Evaluation

4.5.2.1 MTGNN

We conducted a comprehensive comparative evaluation to assess the performance of MTGNN against four established baseline models. These are ARIMA, Vector AutoRegression (VAR), LSTM, and Transformer Encoder-Decoder. Both ARIMA and VAR are statistical models commonly used for time series analysis and forecasting [149]. However, ARIMA is a univariate model, while VAR operates in a multivariate context. LSTM and Transformer are ML models commonly used for sequence-to-sequence prediction [150, 151], and were evaluated both as univariate and multivariate models. In contrast, MTGNN inherently operates as a multivariate model, leveraging its capacity to capture spatial relationships among all features and adaptively learns their hidden relationships.

Table 4.4: Comparative Evaluation for 8 variations of MTGNN

Model	**RSE**	**RAE**
TCN	3.75	3.31
TCN, GCN (PDAM - Bi-directional)	3.25	2.98
TCN, GCN (PDAM - Uni-directional)	3.25	2.97
TCN, GCN (ALAM)	3.20	2.89
TCN, external features	0.83	0.93
TCN, GCN (PDAM - Bi-directional), external features	0.76	0.88
TCN, GCN (PDAM - Uni-directional), external features	0.75	0.88
TCN, GCN (ALAM), external features	**0.73**	**0.85**

TCN: Temporal Convolution. GCN: Graph Convolution. PDAM: Pre-Defined Adjacency Matrix. ALAM: Adaptively Learned Adjacency Matrix.

For each of the four baseline models, we trained separate models to predict each feature in the dataset. This method aimed to facilitate easier learning and convergence by reducing dimensionality. Univariate models relied solely on the ground truth data for prediction (the single feature at hand), while multivariate models integrated additional features. Here, we employed a domain-driven feature selection approach, leveraging prior knowledge and assumptions to determine which features to include in addition to the ground truth. For instance, in models predicting NoI, additional features included external factors (NoM, ACA, PH) alongside pertinent technologies (PATs). Conversely, models predicting PATs incorporated relevant attack types as additional features. In our experiment, each model underwent standardised data partitioning, with approximately 70% allocated for training/validation and 30% for testing. Model performance was assessed on the testing set (unseen data). Random search with 30 iterations was employed to optimise hyper-parameters for each model, and the final performance was averaged over 5 experiments.

Analysis of the results, as depicted in Table 4.5, reveals MTGNN as the top performer in terms of both RSE and RAE. With an RSE of 0.77 and RAE of 0.83, MTGNN demonstrates superior forecasting accuracy compared to ARIMA, VAR, LSTM, and Transformer models, across both univariate and multivariate settings. The notable performance enhancement of MTGNN can be primarily attributed to its ability to adaptively learn and capture intricate spatial relationships among features, while effectively leveraging external information. While ARIMA and VAR models display reasonable performance, LSTM and Transformer models exhibit comparatively higher errors, indicating challenges in capturing the underlying temporal dependencies. This underscores the advantage of incorporating graph-based adaptive learning mechanisms, particularly in MTGNN, for time series forecasting tasks. Moreover, the integration of external features further bolsters MTGNN's predictive capabilities, underscoring its versatility and effectiveness in real-world forecasting scenarios.

Table 4.5: Comparative Evaluation for MTGNN and 4 baseline models

Model	**RSE**	**RAE**
LSTM (M)	1.42	1.38
Transformer Encoder-Decoder (M)	1.40	1.39
Transformer Encoder-Decoder (U)	1.40	1.36
LSTM (U)	1.40	1.34
VAR (M)	1.20	1.32
ARIMA (U)	1.00	0.87
MTGNN (M)	**0.77**	**0.83**

U stands for univariate model and M stands for multivariate model.

4.5.2.2 B-MTGNN

We additionally conducted a quantitative evaluation to justify the inclusion of the Bayesian module. Here, we evaluated the performance of the MTGNN model compared to five variations of the B-MTGNN model, where each variation uses a different number of iterations in the range 10-50 to approximate a Bayesian model. The number of iterations is denoted as it, where $it > 1$. Similar to our previous experiment, we divided the dataset into 70% for training/validation and 30% for testing. Additionally, we employed random search with 30 iterations to optimise the hyper-parameters of each model, and the final performance was averaged over 5 experimental runs.

The evaluation results are illustrated in Table 4.6. The results indicate that the inclusion of the Bayesian module significantly impacts the model's performance, particularly as the number of iterations increases. Specifically, the B-MTGNN model with 30 iterations ($it = 30$) outperforms all other models including the MTGNN model, achieving the lowest RSE of 0.67 and the lowest RAE of 0.78. This suggests that a higher number of iterations in the Bayesian approximation improves the model's accuracy and generalisation capability. However, it is also noteworthy that increasing the iterations beyond 30 does not yield further improvements, as observed with the B-MTGNN models having 40 and 50 iterations, where the performance slightly declines. This phenomenon highlights the presence of an optimal range for the number of iterations, beyond which the model's accuracy may not continue to increase and may even decrease due to factors such as computational inefficiencies or diminishing returns in model complexity. Therefore, the B-MTGNN model with 30 iterations stands out as the most effective configuration for balancing performance and computational cost, underscoring the value of Bayesian methods in enhancing predictive accuracy in complex models like MTGNN.

Table 4.6: Comparative Evaluation for MTGNN and 5 variations of B-MTGNN

Model	**RSE**	**RAE**
MTGNN	0.77	0.83
B-MTGNN ($it = 10$)	0.75	0.85
B-MTGNN ($it = 20$)	0.73	0.81
B-MTGNN ($it = 30$)	**0.67**	**0.78**
B-MTGNN ($it = 40$)	0.72	0.82
B-MTGNN ($it = 50$)	0.71	0.82

it stands for the number of iterations in the Bayesian model.

Overall, our experimental results in Table 4.6 indicate that the Bayesian model (B-MTGNN) with 30 iterations provides 20-30% improvement over the benchmark model. More specifically, it outperforms the benchmark model that can predict the average trend with 100% accuracy, by 33% in terms of RSE and 22% in terms of RAE.

This shows that a simple modification of enabling dropout during both training and inference, while performing multiple forward passes to average the resulting distribution for prediction, significantly improves performance compared to the benchmark model. Additionally, based on the results in Tables 4.5 and 4.6, we showed that the B-MTGNN model outperformed several other models including MTGNN, ARIMA, VAR, LSTM, and Transformer Encoder-Decoder.

4.6 Discussion

4.6.1 Findings

4.6.1.1 Gap Analysis

The trend forecast in Figure 4.5 reveals substantial gaps between several cyber threats and PATs over the next three years. Malware attacks are projected to maintain significant disparities with various PATs, while vulnerability-related attacks are anticipated to widen gaps with specific technologies. Emerging threats like ransomware and adversarial attacks also exhibit notable gaps with PATs, underscoring the importance of proactive measures to address evolving cyber threats effectively.

In light of these identified gaps between cyber threats and PATs, we believe that policymakers should prioritise investments in specific technologies to mitigate the vulnerabilities effectively. For malware attacks, which demonstrate the most significant gaps compared to other threats, policymakers should focus on enhancing technologies such as Application Whitelisting, File Integrity Monitoring, and Darknet Monitoring. Vulnerability-related attacks highlight the need for investments in technologies such as Standardised Communication, SIEM, and Control Flow Integrity, to address widening gaps. While Vulnerability Assessment and NLP/LLM are anticipated to exhibit some visibility, additional resources should be allocated to these areas to improve their effectiveness in combating vulnerability-related attacks.

Emerging threats like ransomware and adversarial attacks require targeted investments in technologies tailored to their unique characteristics. For ransomware, policymakers should prioritise technologies such as Application Whitelisting, Deception Technology, and Data Backups to narrow the gap. Similarly, adversarial attacks necessitate investment in technologies like Spatial Smoothing, Defensive Distillation, and Noise Injection to mitigate vulnerabilities effectively.

The above analysis enabled us to identify technologies worthy of investment by visualising past and projected gaps between each threat and its PATs. However, we recognise that incorporating gap categorisation and tabulation enhances this process by introducing a systematic approach to prioritise investments more effectively. This approach considers not only the magnitude of the gap but also its category, leading to more informed decision-making. It follows that categorising gaps into four distinct categories (SWG, OWG, ONG, and SNG) enables policymakers to prioritise investments in mitigation technologies more efficiently. Therefore, based on the results in Table 4.2 and Table 4.3, we recommend the investment in the research and development of the PATs with widening gaps with respect to the relevant threats, which are listed in Table 4.2. These PATs can be prioritised in the order of the table so that the PATs with wider gaps are given higher priority. Similarly, the PATs in the SWG group should receive higher attention compared to the PATs in the OWG group, since they are more likely to persist this widening trend. It follows that the investment in Standardised Communication, SIEM, Control Flow Integrity, Least Privilege, and Session Management is highly recommended (top five PATs in the SWG group). At the same time, it is also important to consider the significant gap values observed in the OWG group, hence to invest in Application Whitelisting, File Integrity Monitoring, Darknet Monitoring, NLP/LLM, and Spatial Smoothing. We note that the decision to invest in the top five technologies in each category is only an example. Policymakers may adjust this number according to their capacity and resources.

On the other hand, it is recommended that the PATs in the ONG and SNG groups (Table 4.3) be given less priority when making an investment decision, especially if they did not appear in the SWG or OWG groups. Here, less priority can be given to the PATs with smaller gap values and PATs with gaps that are consistently narrowing (SNG). Examples include Encryption, Blockchain, and Public Key Infrastructure.

While these PATs play an important role in cyber security, the forecast suggests that they are catching up with the trend of relevant threats and it is time to consider additional technologies to effectively combat evolving cyber threats.

These findings highlight the critical need for proactive cyber security measures to bridge the identified gaps between the threats and their PATs. Addressing the disparities requires strategic investments in research and development to enhance the efficacy of existing technologies and develop novel solutions capable of mitigating evolving cyber threats effectively. Additionally, collaborative efforts between industry stakeholders, policymakers, and cyber security experts are essential to facilitate knowledge sharing and promote the adoption of best practices in cyber security defence strategies. Failure to address these gaps adequately may leave organisations vulnerable to cyber-attacks, potentially resulting in significant financial losses, reputational damage, and disruptions to critical infrastructure and services.

4.6.1.2 Alleviation Technologies Cycle - Analysis

The analysis of PATs over a three-year span, culminating in the development of the ATC, has significant implications for cyber security preparedness. By understanding the life cycle stages of PATs (launch, growth, maturity, trough, and stability), security agencies can strategically align their investment and adoption efforts.

The ATC provides insights into the progression of PATs and their relevance to emerging and rapidly increasing threats. This understanding allows agencies to anticipate trends and prioritise resources accordingly. For example, during the growth phase, where PATs are often related to emerging threats, agencies can focus on early adoption and experimentation. As PATs mature and reach stability, agencies can assess their effectiveness and make informed decisions about long-term integration. Furthermore, the identification of trough phases in the ATC highlights potential challenges and areas for improvement in PAT deployment. Agencies can use this information to proactively address issues such as declining enthusiasm or performance gaps. By recognising these patterns, agencies can better navigate the complexities of cyber security technology adoption and ensure continuous improvement in their defence strategies.

Furthermore, the ATC presents policymakers with a comprehensive framework to strategically allocate resources and align defence mechanisms with the evolving landscape of cyber threats. For instance, Figure 3.6 indicates that malware is currently peaking, while in Figure 4.6, the PAT *File Integrity Monitoring* is situated in the lower trough. In response, policymakers should prioritise advancing File Integrity Monitoring to the plateau swiftly. This action would help bridge the gap between this technology and the evolving trend of malware, potentially facilitating a decline in malware incidents. Similarly, ransomware exhibits rapid growth, while Application Whitelisting is in the process of recovering from a trough phase. To address this gap, policymakers should focus on elevating the trend of Application Whitelisting to the plateau, thereby aligning its efficacy with the escalating trend of ransomware.

We advocate for policymakers to prioritise advancing the PATs towards the upper plateau rather than the lower plateau. PATs positioned on the upper plateau offer greater visibility and are better aligned with relevant threats, reducing the likelihood of significant gaps. Achieving this entails closely monitoring the trend of PATs and enhancing their usability as they enter the trough phase. By encouraging investment in these technologies during this phase, increased effort and experimentation can raise awareness and illustrate how the technology benefits organisations, facilitating a quicker recovery from the trough. This concerted effort ultimately propels the PATs towards the upper plateau, where they are better positioned to effectively address emerging cyber threats.

Overall, the ATC framework offers a systematic approach to understanding the evolution of cyber security technologies and their alignment with threat landscapes. This enables security agencies to make data-driven decisions, optimise resource allocation, and enhance their overall cyber security posture in the face of evolving threats.

4.6.1.3 Model Performance

The model's validation results showed that the average RSE computed over 142 nodes is 0.52, and the average RAE is 0.66. These metrics represent noticeable improvements over the benchmark model, indicating that the B-MTGNN model performs significantly better in terms of predictive accuracy. The validated performance allows organisations to make informed decisions regarding risk management strategies. By understanding the potential future trajectories of cyber-attacks and technologies, organisations can allocate resources more effectively and prioritise investments in cyber security measures.

The results from the ablation study highlighted the significant impact of integrating our external features on enhancing the model's predictive performance, particularly when dealing with previously unseen data. By incorporating these external factors, we were able to achieve a relative error below 1, indicating a notable improvement compared to the benchmark model. The inclusion of external factors such as attacks' mentions in the literature (NoM), tweets about wars and political conflicts (ACA), and public holidays (PH) proved to be instrumental in reducing relative errors and improving forecasting accuracy. These external features offer valuable contextual information that can influence the occurrence of cyber incidents. For instance, fluctuations in the frequency of attacks' mentions in research papers (NoM) may reflect shifts in the attention and scrutiny given to specific cyber threats within academic discourse. Similarly, tweets about wars and conflicts (ACA) can serve as proxies for geopolitical tensions, which may in turn impact the likelihood of cyber-attacks originating from or targeting regions affected by such conflicts. Furthermore, the occurrence of cyber incidents may exhibit temporal patterns correlated with public holidays (PH), suggesting potential opportunities for threat actors to exploit vulnerabilities during periods of reduced security vigilance or target entities during anniversaries or commemorative events.

Moreover, the integration of graph convolution layers enabled our model to capture spatial dependencies between nodes in the graph, including attacks, PATs, and external features. By leveraging graph structures to represent relationships between entities, the model can effectively capture complex interactions and dependencies that may exist within the cyber threat landscape. Additionally, the better performance when using a uni-directional predefined adjacency matrix over a bi-directional one suggests that certain relationships within the graph may be inherently one-directional, aligning with previous research findings [27]. The utilisation of graph learning layer further enhanced the model performance by enabling the adaptive learning of relationships within the graph during training. This is particularly important as the relationships between entities in the cyber threat landscape may not be easily quantifiable or predefined by human experts, highlighting the importance of leveraging ML techniques to uncover and adaptively learn optimal graph structures. Overall, the optimal model configuration, which combines graph convolution layers, temporal convolution layers, graph learning layer, and external features, underscores the synergistic effect of integrating diverse components to improve forecasting accuracy. By effectively capturing both spatial and temporal dependencies within the cyber threat landscape while leveraging contextual information from external features, our model demonstrates promising capabilities in forecasting cyber incidents and PATs and assisting policymakers in proactive threat mitigation strategies.

A notable trend observed from the comparative evaluation is the consistent outperformance of univariate approaches over their multivariate counterparts across the four baseline models, as evident in Table 4.5. With the exception of the MTGNN model, multivariate models, including multivariate LSTM and Transformer, consistently exhibited higher RSE and RAE compared to their univariate counterparts, and the univariate model ARIMA outperformed the multivariate model VAR. This discrepancy is attributed to pre-assumed feature relationships that may not necessarily be optimal, underscoring the importance of learning these interdependencies among multiple variables for accurate time series forecasting. MTGNN explicitly learns and quantifies these relationships. Additionally, representing these features as nodes in a graph provides the opportunity to capture hierarchical relationships. Moreover, in cases where no such relationships exist between nodes, the graph convolution layer can adapt and preserve the original node's self-information [27].

The superior performance of the Bayesian model (B-MTGNN) compared to its deterministic counterpart (MTGNN) can be attributed to its ability to aggregate predictions from multiple iterations. By taking the mean of the distribution as the prediction, the Bayesian model leverages the collective knowledge encoded in these iterations, resulting in a more comprehensive and stable forecast. This approach helps mitigate the effects of overfitting and variability, leading to improved generalisation ability and enhanced predictive accuracy. We note that the benefits of the Bayesian model are not limited to improved accuracy; it also provides a measure of uncertainty, offering confidence in its predictions. Overall, the Bayesian model's capacity to capture uncertainty information and its robust averaging mechanism enable it to outperform its deterministic counterpart in terms of both performance and reliability.

4.6.1.4 Confidence Intervals - Analysis

In Bayesian modelling with Monte Carlo dropout, achieving 95% coverage accuracy with the 95% confidence interval is an ideal objective, indicating that in repeated experiments, the interval should encompass the true value 95% of the time. However, deviations from this ideal can occur due to several reasons. Model imperfections in learning due to insufficient samples or noisy data can lead to over-confidence or uncertainties that are not accurately captured by the model [152]. This could result in confidence intervals that do not always achieve the desired coverage. Such outcomes are possible in Bayesian modelling especially when "approximating" such models, reflecting the inherent uncertainty and complexity of real-world data. While potential solutions include integrating additional data samples, the challenge lies in the limited availability of cyber security data due to its confidentiality and sensitivity.

While the confidence intervals shown in Figure 4.4 may have limited overlap with the true values, the overall trend captured by the model aligns well with the actual data. This alignment indicates that the model effectively captures the underlying patterns and dynamics of the data, which is a crucial aspect of forecasting, particularly in the context of cyber threats. The primary goal of our model is to identify and understand the general trends in cyber threat activity over time, which is more valuable for strategic planning and proactive measures than precise numerical predictions. In the cyber security domain, where threat patterns are often complex and rapidly evolving, capturing these trends provides significant foresight and helps inform effective response strategies. For example, a model predicting an upward trend in certain types of cyber-attacks can be crucial for guiding investments in defensive technologies and shaping policy decisions [6]. While precise predictions may not always be possible, capturing the broader trends allows for better preparedness and response to emerging threats.

Despite the limited overlap between confidence intervals and actual values, the provided confidence intervals still serve an essential purpose in quantifying the uncertainty of the model's predictions [47]. The fundamental principle of Bayesian inference is to provide a probabilistic range of potential outcomes based on the data and prior information, which does not guarantee complete overlap with actual values at every point in time. The primary goal of including confidence intervals is to provide a sense of how confident the model is in its predictions and to account for the inevitable uncertainty in the data.

Overall, the ability to capture various cyber security trends suggests that our model, even with a small dataset, is robust and capable of providing meaningful insights. As noted earlier, our model's performance shows a noticeable improvement over the benchmark, reinforcing its effectiveness in predicting cyber threat trends. Our work represents the initial effort to forecast cyber security trends years in advance, acknowledging the inherent challenges and limitations in achieving perfect predictions in this evolving field. The model can be further refined by incorporating additional data samples (when they become available) and employing more advanced techniques to better estimate uncertainty and improve predictive accuracy [152]. These steps will enhance the model's robustness and reliability, ensuring it continues to provide valuable insights into the dynamic landscape of cyber threats.

4.6.2 Practical Concerns

4.6.2.1 Adversarial Adaptation

One concern with our forecasting approach is that once the key building blocks of our cyber threat prediction model are made explicit (*e.g.,* Figure 4.2), adversaries could potentially reconstruct or reverse-engineer the model to invalidate the effectiveness of the alleviation technologies, by altering their strategies accordingly. This possibility highlights the need for a robust framework that not only forecasts threats but is also resilient against being compromised by the very actors it seeks to defend against.

As a solution, it is possible to consider strategies that enhance the model's resilience against potential reconstruction and exploitation. The implementation of our framework could involve mechanisms for continuous learning and dynamic updates, ensuring that the model adapts to the evolving threat landscape. Integration of real-time threat intelligence and adjustment of the model's parameters based on the latest trends could make it more challenging for adversaries to develop effective countermeasures [153]. However, it is crucial to ensure that the model's learning and adaptation capabilities outpace the rate at which attackers can develop countermeasures, thereby maintaining a strategic advantage. By regularly updating the model with the latest information, we can ensure that it remains responsive to emerging threats and retains a first-mover advantage in the cyber defence landscape [154]. This approach allows the model to adapt to new challenges more swiftly than adversaries can develop counter-strategies, ensuring its ongoing effectiveness.

Incorporating game-theoretic principles into the model's decision-making processes could further enhance its resilience. By introducing elements of unpredictability and strategic ambiguity, we can make the model's behaviour more difficult for attackers to predict and counteract [155]. For instance, by varying the types and sources of data and employing randomisation techniques, we can create an environment where the model's behaviour is less predictable and more robust against exploitation. This strategic variability makes it challenging for attackers to anticipate and counteract the model's forecasts, thus maintaining the integrity and effectiveness of the alleviation technologies.

We would like to highlight that our approximated Bayesian model which generates a distribution of possible outcomes contributes to this strategic ambiguity. This is achieved by allowing for a range of potential future scenarios to be considered. The probabilistic output means that the system's responses can vary based on the predicted likelihood of different threats, creating a defence strategy that is dynamic and harder for attackers to anticipate. This variability and adaptive response framework, driven by the inherent uncertainty in predictions, makes it difficult for adversaries to predict and exploit specific defensive actions, thereby maintaining a strategic advantage.

Additional techniques include exploring the use of differential privacy methods which could be another effective way to safeguard the model. By implementing differential privacy, the model can ensure that its internal parameters and data inputs are obscured, making it challenging for attackers to reverse-engineer or gain meaningful insights into the model's structure and functionality [156]. This would help to maintain the confidentiality of the model's operations, even if its high-level architecture is known. Moreover, employing ensemble learning techniques, which involve the combination of multiple models to produce a collective output, could add complexity to the system, making it harder for adversaries to anticipate and exploit the model's predictions [157]. While these techniques are not currently implemented in our model, they represent a potential avenue for future enhancements aimed at improving security and resilience.

4.6.2.2 Data Centralisation

Centralising sensitive data to enhance the model's performance poses significant risks, particularly regarding the potential for data breaches. If an industry consortium were to contribute confidential data, the centralisation could create a new vulnerability. It is essential to find a balance between aggregating data for model improvement and maintaining data confidentiality. Here, it is possible to adopt federated learning, which allows for collaborative model training without centralising the data. Each organisation could train a local model on its confidential data and share only aggregated updates with a central server. This approach maintains the confidentiality of sensitive data while still allowing for collaborative improvements to the model [158]. Moreover, the use of secure enclave technologies for data processing can also mitigate these risks. Secure enclaves provide a protected environment that ensures data confidentiality during computation, even on shared infrastructure. This technology could be explored to ensure that data used for model training remains secure, even in a collaborative setting [159]. The implementation of the model can also consider incorporating differential privacy techniques to anonymise and obscure data contributions, ensuring that sensitive information remains protected. This approach would enable secure data collaboration without compromising the confidentiality of the data provided by contributing organisations [156]. These strategies collectively offer a pathway to secure data collaboration that minimises vulnerabilities while enhancing the model's robustness and accuracy.

4.6.2.3 Overall Defence Strategy

Bringing together the strategies discussed, the overall defence strategy for our cyber threat forecasting model is designed to be both adaptive and resilient against potential adversarial exploitation. By anticipating and addressing practical concerns such as adversarial adaptation and data centralisation, we aim to maintain a robust defence posture while continuously enhancing the model's effectiveness. Our overall defence strategy combines continuous learning, strategic modelling, and secure data collaboration to create an adaptive and resilient cyber defence framework. By staying ahead of adversaries with dynamic and unpredictable responses and ensuring data confidentiality through advanced privacy-preserving technologies, we aim to build a cyber threat prediction model that not only anticipates emerging threats but also withstands the challenges posed by adversarial exploitation.

4.6.3 Highlights and Contributions

This work pioneers a proactive approach in cyber security using ML for long-term prediction of cyber threats and the PATs. It represents a step forward in the field of cyber security, aligning with the growing body of literature advocating for proactive defence strategies [5, 17, 160]. By proposing the long-term prediction of cyber threats and PATs, this research addresses a critical gap identified in prior studies [6]. The integration of advanced ML techniques, particularly Bayesian graph learning, builds upon existing literature on predictive modelling approaches from different domains such as traffic forecasting [27, 43]. Furthermore, the improved model's performance when using the proposed features echoes findings from prior research [15, 17, 55], highlighting the crucial role of feature engineering in enhancing predictive models' performance.

The implications of this work on research include advancing the research on proactive cyber security. It sets a precedent for future research to explore and refine predictive models, incorporating evolving ML techniques to foresee cyber threats effectively as well as the relevant technologies. The demonstrated improved performance when using the Bayesian model indicates a potential shift towards employing advanced techniques in graph analytics. Future research may delve into optimising and customising graph-based algorithms for cyber threat prediction, thereby enhancing the accuracy and efficiency of predictive models. The proposed effective data features can be also utilised and extended to further improve the performance. Furthermore, by highlighting the use of extensive global data and coverage of 36 countries, this work underlines the importance of comprehensive data analysis for a more holistic understanding of the cyber threat landscape. Future research could explore further enhancements in data collection, analysis, and representation for an even broader international scope.

In practice, this work enhances cyber security preparedness and planning. The proactive approach advocated in this work emphasises the need for organisations to establish early-stage communication with potential cyber threats and the PATs. This suggests that real-world applications should invest in proactive planning, enabling them to develop optimal defensive measures well in advance. This optimality results from the reduced uncertainty which leads to the prioritisation of the security measures by considering future threat gaps. Furthermore, this shift towards automated, data-driven methodologies aims to minimise subjective biases. In practice, this implies a transition towards quantitatively-driven decisions, reducing reliance on human judgment. Organisations should consider integrating automated, data-centric approaches to ensure consistency and impartiality in threat analysis and decision-making processes. Finally, the noted improvement in performance using Bayesian GNN and the proposed features suggests that incorporating advanced ML techniques, especially those suited for graph-based data, can significantly enhance predictive capabilities. Organisations should explore and implement such techniques to improve the accuracy and efficacy of their cyber threat prediction systems.

4.6.4 Limitations

The current dataset provides valuable insights into high-level attack types and PATs, offering a foundational understanding of cyber security threats and mitigation strategies. The applicable scope of the dataset primarily encompasses strategic and tactical analysis for cyber security professionals, serving as a basis for developing broad-based defence mechanisms against a spectrum of cyber threats. It is particularly valuable for organisations seeking to establish a foundational cyber security posture by understanding prevalent threats and corresponding preventive technologies. However, as the cyber security landscape continues to evolve rapidly, there is a growing need to explore the possibility of extending the dataset to encompass more fine-grained attack types. This expansion presents an opportunity to delve deeper into specific attack vectors and enhance the effectiveness of cyber security defences.

For example, consider the category of "Malware" within the current dataset. While it provides a broad overview of malicious software threats, including viruses, worms, and ransomware, a more granular approach could distinguish between different variants and functionalities of malware. By categorising malware based on behaviour, propagation methods, and targeted platforms, organisations could tailor their defence mechanisms more precisely to combat specific threats. For instance, distinguishing between fileless malware [161], which operates solely in memory, and traditional file-based malware could inform strategies for endpoint detection and response.

Similarly, the category of "Adversarial Attack" highlights the diverse range of techniques employed by threat actors to subvert ML models and AI systems. However, a finer-grained classification could differentiate between adversarial attacks targeting image recognition systems, natural language processing models, and reinforcement learning algorithms. Indeed, the profoundly different nature of the training algorithms applicable to these categories of AI systems suggests differentiation among their adversarial attacks. Finer-grained attack classification would enable researchers and practitioners to develop specialised countermeasures, such as robustness enhancements [162], data augmentation techniques [133], and adversarial training strategies [163], tailored to each specific threat context.

Incorporating more fine-grained attack types into the dataset also opens avenues for exploring emerging threats and vulnerabilities. For instance, the rise of deepfake technology poses novel challenges in detecting and mitigating manipulated media content. By analysing different types of deepfake attacks, such as facial manipulation [138], voice synthesis [164], and video impersonation [165], cyber security professionals can develop innovative detection algorithms and authentication mechanisms to combat the spread of disinformation and fraudulent content.

Moreover, extending the dataset to include fine-grained attack types facilitates cross-domain analysis and correlation studies. For example, correlating specific malware families with targeted industries or geographical regions could reveal patterns of cybercriminal activity and inform proactive defence strategies [166]. Similarly, identifying commonalities between adversarial attacks in different domains, such as image recognition and natural language processing, could lead to the development of holistic defence frameworks that address underlying vulnerabilities across diverse application areas.

Other limitations of this work include its reliance on a limited dataset that encompasses data since 2011 only. This is due to the challenges encountered in accessing confidential and sensitive information. Extending the prediction period necessitates the model to forecast further ahead into the future, requiring increased data samples and informative features. Also, a notable limitation stems from the lack of a systematic approach for the evaluation of the E-GPT algorithm, which is instrumental in extracting the PATs and constructing the graph. Moreover, such evaluation often depends on subjective and potentially biased human judgment. As a result, ensuring an optimal graph structure becomes challenging, particularly in the absence of a mechanism to quantify the assumed relationships between nodes in the graph. The subjectivity issue is also observed in the placement of PATs on the cycle, where a fully automated approach would lead to a more efficient process and more reliable results.

4.7 Summary and the Road Ahead

In this chapter, we introduced a proactive approach based on ML for long-term prediction of cyber threats and PATs. The goal is to establish an effective communication with the future disparity between the potential attacks and relevant security measures at an early stage, enabling proactive planning for the future. By adopting this approach, there is an increased chance to prevent incidents by allowing more time for the development of optimal defensive actions and tools, thereby bridging the gap between cyber threats and PATs. Moreover, our automated approach shows promise in addressing the widely recognised challenges associated with human-based analysis. By eliminating the reliance on human judgment and adopting a purely quantitative methodology driven by data, our approach aims to minimise subjective biases and promote consistency within the subject matter.

With access to extensive data sources encompassing a vast volume of information and global geographic coverage, our study contributes to the construction of a comprehensive dataset encompassing different cyber security trends, which can be utilised for various purposes. We used this dataset to construct a novel Bayesian GNN model which was utilised to provide 3 years forecast for the future gaps between several cyber threats and PATs. Based on the future forecast, we categorised the gap trends, and recommended future investment decisions accordingly. Following a large-scale analysis for past and future trends, we proposed the ATC model identifying the life cycle phases in the trend of 98 alleviation technologies. This cycle serves as a robust foundation for raising awareness when investing in security measures aimed at preventing cyber-attacks. It presents policymakers with a comprehensive framework to strategically allocate resources and align defence mechanisms with the evolving landscape of cyber threats.

We have demonstrated the efficacy of our Bayesian model, outperforming several baseline models while also providing a measure of confidence through the articulation of epistemic uncertainty. Additionally, our incorporation of external features has demonstrated tangible improvements in model performance, further bolstering the reliability and utility of our predictive framework. Overall, our work not only advances the theoretical understanding of cyber threat prediction but also furnishes practical insights for managerial decision-making. By offering a proactive, data-driven approach to cyber security planning, we aspire to equip policymakers with the foresight and tools necessary to navigate an increasingly complex threat landscape with confidence and efficacy.

In the next chapter, we will explore the intrinsic aspects of explainability within our GNN-based prediction model. As we navigate the complexities of the cyber threat landscape, the focus will be on elucidating how the interpretability provided by saliency maps and attention scores becomes instrumental in empowering decision makers. This exploration aims to demystify the predictive black box, offering stakeholders a transparent and actionable understanding of the model's decision-making processes. By unravelling the intricacies embedded within these components, we seek to equip organisations with the knowledge necessary to navigate and fortify their defences with confidence and foresight.

Code and Data Availability

The code and dataset used in this chapter are available at the following link:
https://github.com/zaidalmahmoud/Cyber-trend-forecasting.

Table 4.1: Cyber Threats and Pertinent Alleviation Technologies in our study

Threat	Type	Pertinent Alleviation Technologies (PATs)
Account Hijacking	RI	AC, AD, CAPTCHA, CR, IDS/IPS, IdM, LP, MFA, ML/DL, NLP/LLM, PT, SM
Adversarial Attack	E	AD, AdT, BN, DA, DD, DP, DR, DS, ML/DL, NI, NLP/LLM, OD, RRAM, SS, TAI
APT	RI	AC, DLP, DRM, DT, GT, IDS/IPS, LP, MFA, ML/DL, NLP/LLM, NS, PT, RA, UBA
Backdoor	RI	AD, DAS, IDS/IPS, ML/DL, PT, SA
Botnet	RI	AD, BC, BH, BT, CAPTCHA, GM, GT, HP, IDS/IPS, ML/DL, NLP/LLM, PF, PT, RC, RL, SDN, TS
Brute Force Attack	RI	CAPTCHA, CR, DBI, IDS/IPS, MFA, ML/DL, OTP, PH, PT
Cryptojacking	E	BT, ML/DL, PT, TA
DDoS	RI	BC, BH, BT, IDS/IPS, ML/DL, NLP/LLM, PF, PT, RC, RL, TS
Data Poisoning	E	AD, AdT, BN, DP, DS, ML/DL, NLP/LLM, OD, TAI
Deepfake	E	3DFR, AD, BO, DW, LD, ML/DL, NLP/LLM
Disinformation	RI	BC, CA, DLT, DP, DT, GT, HG, IR, ML/DL, NLP/LLM, SI
DNS Spoofing	RI	BC, CR, DNSSEC, ML/DL, PT, RA
Dropper	RI	AW, CS, FIM, IDS/IPS, ML/DL, NLP/LLM, PT, SBX
Insider Threat	RI	AC, AD, AM, AT, CR, DLD, IDS/IPS, KD, LP, ML/DL, MTD, NLP/LLM, PT, UBA
IoT Device Attack	E	AD, BC, CR, IDS/IPS, IdM, MFA, ML/DL, MS, PT, SB
Malware	RI	AC, AD, AW, BBD, BC, CR, CS, DAS, DB, DM, DT, FIM, FV, GT, HP, IDS/IPS, ML/DL, NLP/LLM, PMT, PT, SA, SB, SBX, SHMM, SMF, VK
MITM	RI	BC, CAPTCHA, CP, CR, ML/DL, PKI, PT, SSL/TLS, SSP, VPN
Password Attack	RI	CAPTCHA, CR, GA, IDS/IPS, MA, MFA, ML/DL, NLP/LLM, OTP, PH, PM, PP, PSM, PT
Phishing	RI	AC, BT, CR, DT, MA, MFA, ML/DL, NLP/LLM, PKI
Ransomware	E	AC, AD, AW, BC, CR, DAS, DB, DT, IDS/IPS, ML/DL, NLP/LLM, PMT, PT, SA, SHMM
Session Hijacking	RI	AD, CA, CR, Https, IBE, ML/DL, PT, SAT, SM, SSL/TLS
Supply Chain Attack	RI	AC, AD, BC, CR, IdM, ML/DL, NLP/LLM, PT, SCRM
Targeted Attack	RI	AC, DRM, DT, GT, IDS/IPS, LP, MFA, ML/DL, NLP/LLM, NS, PT, RA, UBA
Trojan	RI	AD, BBD, CR, FV, GT, IDS/IPS, ML/DL, NLP/LLM, PT, SMF
Vulnerability	RI	CFI, IDS/IPS, ML/DL, NLP/LLM, PMT, PT, SC, SIEM, VA, VM, VS
Zero-day	RI	AD, DT, FIM, GT, IDS/IPS, ML/DL, NLP/LLM, PrP, VM, VPN

The list of attack types in the **Threat** column are the emerging and rapidly increasing threats identified in [6] based on past and future analysis. These threats require the highest attention when investing in related technologies, compared to the other declining threats. The list of PATs for each attack type was extracted using Algorithm 1. In the column **Type**, RI refers to the rapidly increasing threats and E refers to the emerging threats. The PATs abbreviations table can be found in Figure 4.6.

5 | On the Explainability of Cyber Trend Forecasting

5.1 Background

In an era where the digital landscape is increasingly susceptible to cyber threats, the ability to forecast and comprehend the dynamics of cyber security trends has become imperative for organisations striving to fortify their defences. This chapter serves as a critical exploration into the multifaceted landscape of predicting cyber security trends, recognising the indispensable nature of deciphering model predictions. Our focus is centred on unraveling the intricacies embedded within our GNN-based prediction model, where precision is not only a measure of accuracy but also a conduit for informed decision-making and strategic planning.

The two correlational explainability methods that we study in this chapter, namely *saliency maps* and *attention scores*, contribute to advancing the interpretability and real-world applicability of our predictive model. Through the lens of saliency maps, we embark on a visual journey that sheds light on the temporal dependencies crucial for anticipating cyber threats with an unprecedented lead time of 36 months. These maps not only unravel the significance of past time steps in predicting the future but also bridge the gap between historical data points and the model's foresight, providing stakeholders with a tangible means to comprehend the rationale behind each prediction.

Simultaneously, the exploration of attention scores takes us deeper into the inner workings of our B-MTGNN model. Beyond mere predictive capabilities, this aspect unravels the model's adeptness at adaptively learning hidden adjacency matrices, unveiling nuanced relationships between various nodes within the cyber security landscape. These relationships, depicted through attention scores, not only guide the model in making accurate predictions but also offer a unique perspective for decision makers seeking to comprehend the interconnected dynamics of cyber threats.

As we embark on this exploration, our objective is to demystify the predictive black box, empowering stakeholders with actionable insights gleaned from the intricate interplay of data and algorithms. In doing so, we aim to equip organisations with the knowledge and understanding necessary to navigate the evolving cyber security landscape with confidence and foresight.

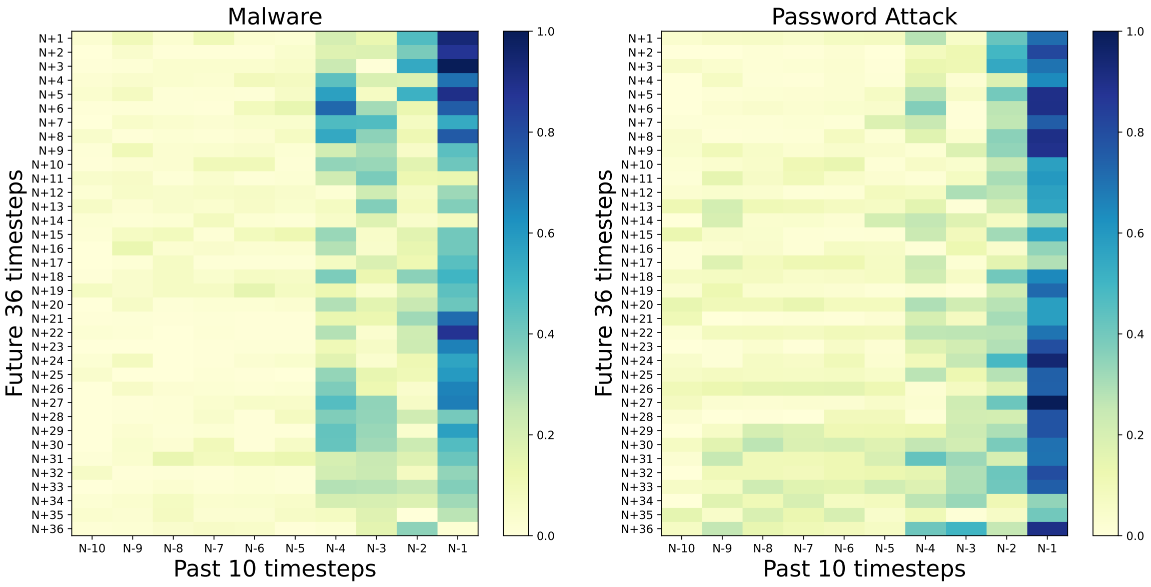
5.2 Saliency Maps

A leading correlational explainability technique in the domain of time series forecasting is the *contrastive gradient-based saliency maps* [77]. This method offers a straightforward yet powerful approach to understanding how input variables influence the model's predictions. By computing gradients of the model output with respect to the input data, saliency maps effectively generate heat maps that visually depict the importance of each input variable.

Through this technique, the norm of the gradient signifies the relative importance of input variables in shaping the model's forecasts. This allows practitioners to identify which time steps have the most significant impact on the model's decisions. Notably, the direction of the maximum positive rate of change in the model output aligns with the resultant gradient in the input space. Consequently, components associated with negative gradient values are disregarded, selectively preserving input variables that positively contribute to the solution.

Figure 5.1 shows examples for two saliency maps produced by our GNN model which predicts cyber-attack trends 36 months in advance (3 years). The saliency maps illustrate the importance of each of the previous 10 time steps when predicting the future 36 time steps for two attack types namely malware and password attack.

Figure 5.2 shows a more comprehensive example. Here, The saliency map can be used along with the prediction heat map to justify the prediction. For example, in Figure 5.2, the predicted value in month X is close to the past data points that have high importance when predicting X according to the saliency map. In the case of malware, the predicted value during the first month in the future window (Jan-23) is close to



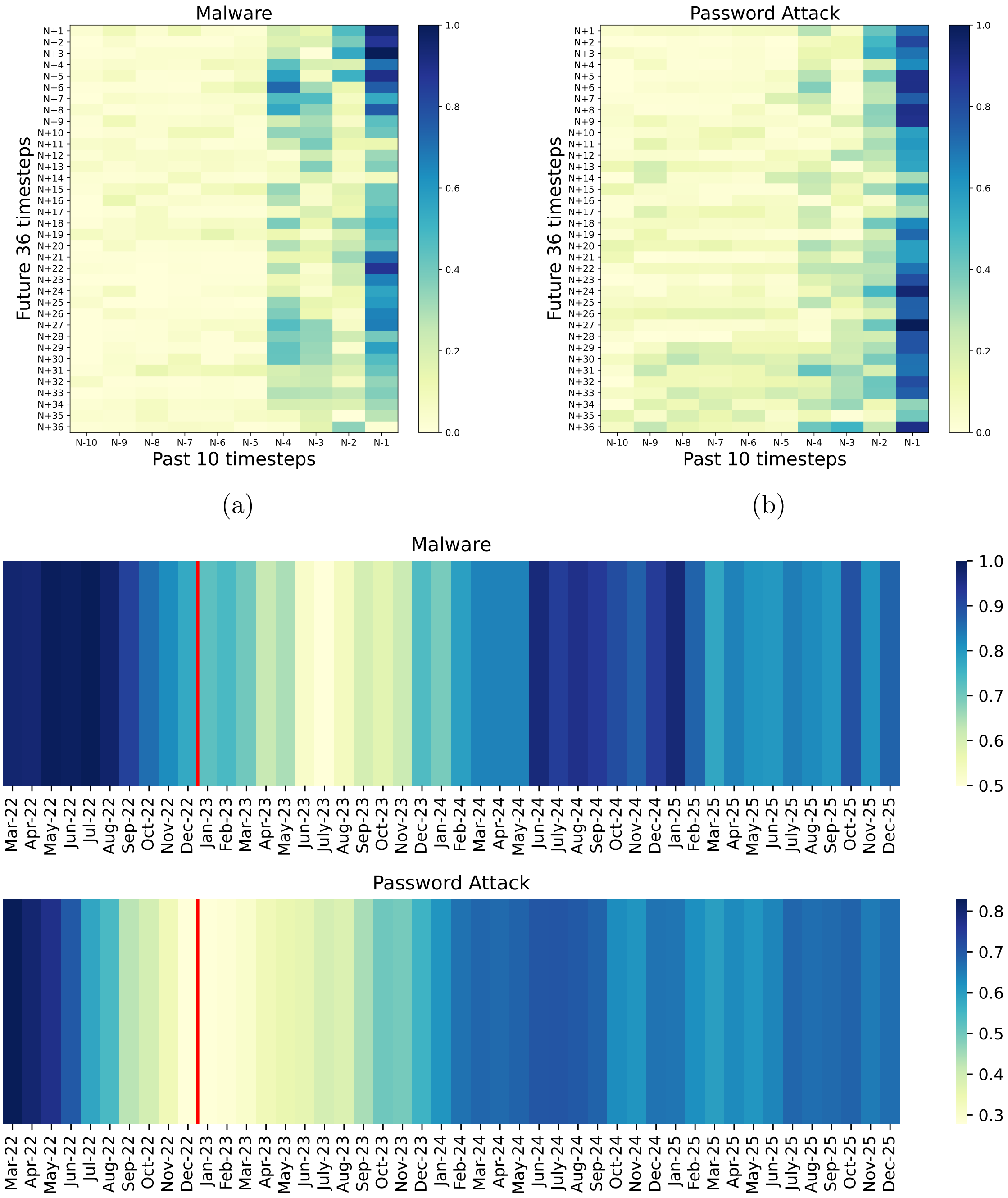
(a) (b)

Figure 5.1: Saliency maps illustrating the importance of each of the past 10 time steps when predicting the future 36 time steps while using a GNN model. (a) Malware (b) Password attack.

the value of the last month in the lookback window (Dec-22), since according to the saliency map, the last month's value is highly important for predicting the the next month's value.

5.3 Revealing the Hidden Relationships

Our B-MTGNN model adaptively learns the hidden adjacency matrix such that the resulting edge weights minimise the prediction error of node values. Consequently, we can visualise these learned relationships between nodes, which we will refer to as *attention scores*. The visualisation of these scores allows the model to provide explanations for its prediction decisions. This is because the graph convolution layer which aids in the prediction utilises the learned adjacency weights when fusing the node's value with its neighbourhood. For a comprehensive example, Figure 5.3 illustrates the attention scores between attacks and PATs along with different prediction heat maps. In the figure, it is evident that adversarial attack and Adversarial Training (AdT) exhibit a high attention score, and therefore, they were predicted proportionally. A



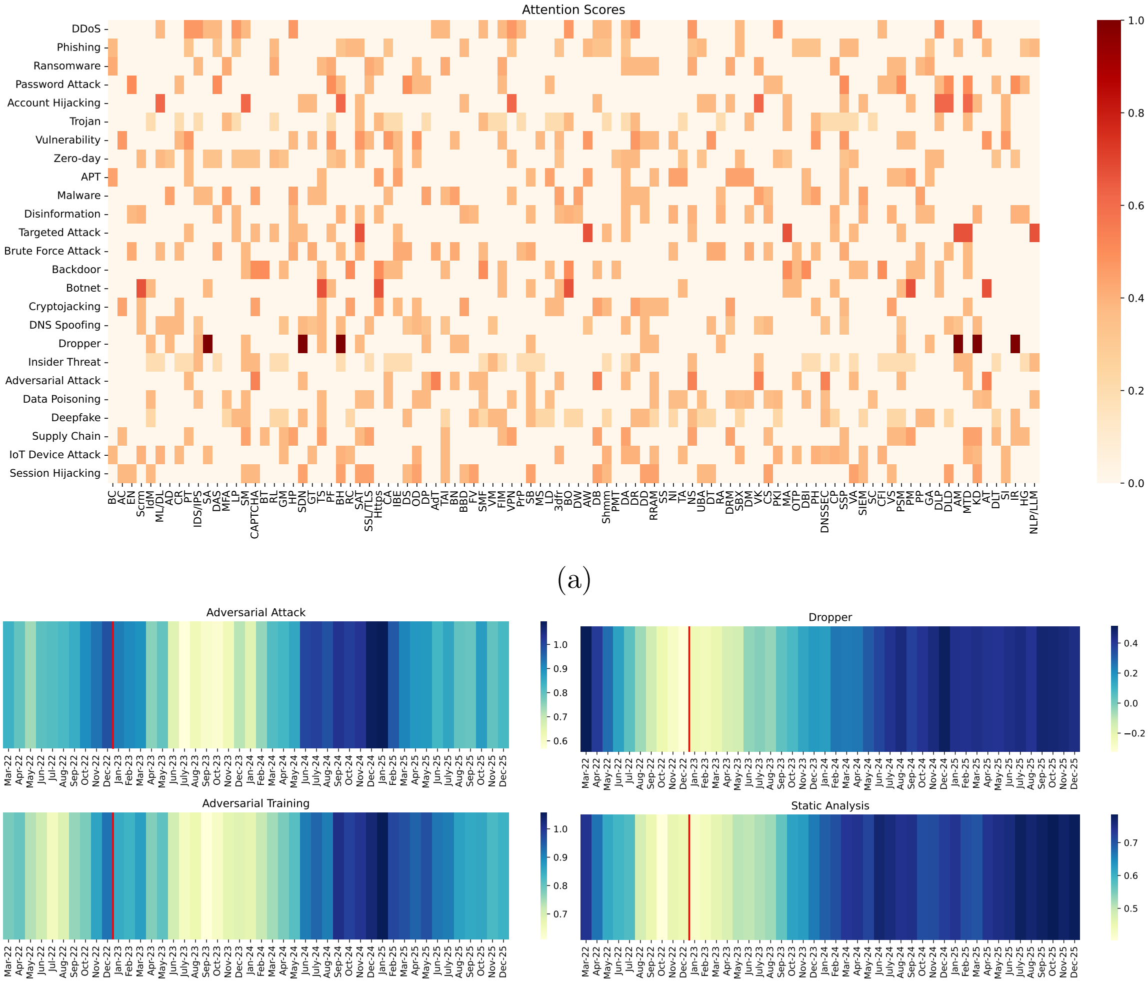
(a)

(b)

(c)

Figure 5.2: Saliency maps and heat maps for past and predicted data. The saliency maps illustrate the importance of each of the past 10 time steps when predicting the future 36 time steps while using a GNN model. (a) Malware saliency map (b) Password attack saliency map. (c) Past data and forecast of malware and password attack. The red lines in (c) mark the start of the forecast period until the end of 2025. In (c), the predicted value in month X is close to the past data points that have high importance when predicting X according to (a) and (b).

similar observation can be made between dropper and Static Analysis (SA).



(a)

(b)

(c)

Figure 5.3: Attention scores and forecast heat maps. (a) Attention scores of attack nodes vs PAT nodes. (b) and (c) Past data and forecast of (b) Adversarial attack and Adversarial Training (AdT) (c) Dropper and Static Analysis (SA). In (b) and (c), the compared nodes in each figure carry high attention scores according to (a) and thus were predicted proportionally. The red lines in (b) and (c) mark the start of the forecast period until the end of 2025. Please refer to Figure 4.6 for the PATs' abbreviations.

The learned adjacency matrix does not only facilitate explainability, but importantly, it also reveals hidden relationships between the different nodes. This revelation empowers decision makers to leverage such insights for investment decisions or the formulation of defence strategies, allowing them to discern and respond to underlying relationships within the system. For example, according to the attention scores in Figure 5.4, there is a relationship between wars and conflicts and the occurrence of

APT as well as brute force attack. Identifying such threats in times of war allows decision makers to focus on mitigating these threats and prioritise the defences accordingly. Similarly, it can be observed in Figure 5.4 that public holidays are relevant to targeted attacks. In Figures 5.5 and 5.6, more attention scores were produced by the model.

5.4 Discussion

5.4.1 Findings

5.4.1.1 Saliency Maps - Analysis

Based on the saliency maps shown in Figures 5.1 and 5.2, we find that the last four months in the past are especially important for predicting future months. We confirmed this across various saliency maps generated for different cyber trends. Additionally, it is noticeable that the saliency value gradually decreases as we move back in time but does not reach zero. Overall, from Figure 5.1, it is evident that the last 10 months provide a reasonable lag period containing the most important information for forecasting.

It follows that the utility of contrastive gradient-based saliency maps extends beyond mere understanding; it offers tangible benefits for improving forecasting accuracy and model interpretability. Understanding the importance of different features or time steps facilitates informed feature selection and engineering efforts, enabling practitioners to focus on the most influential aspects of the data. Moreover, the interpretability provided by saliency maps enhances stakeholders' trust in the forecasting model, as it offers transparency into the rationale behind the model's predictions. Armed with this knowledge, stakeholders can make more informed decisions based on the forecasts, ultimately leading to better outcomes in various domains including cyber security and beyond.

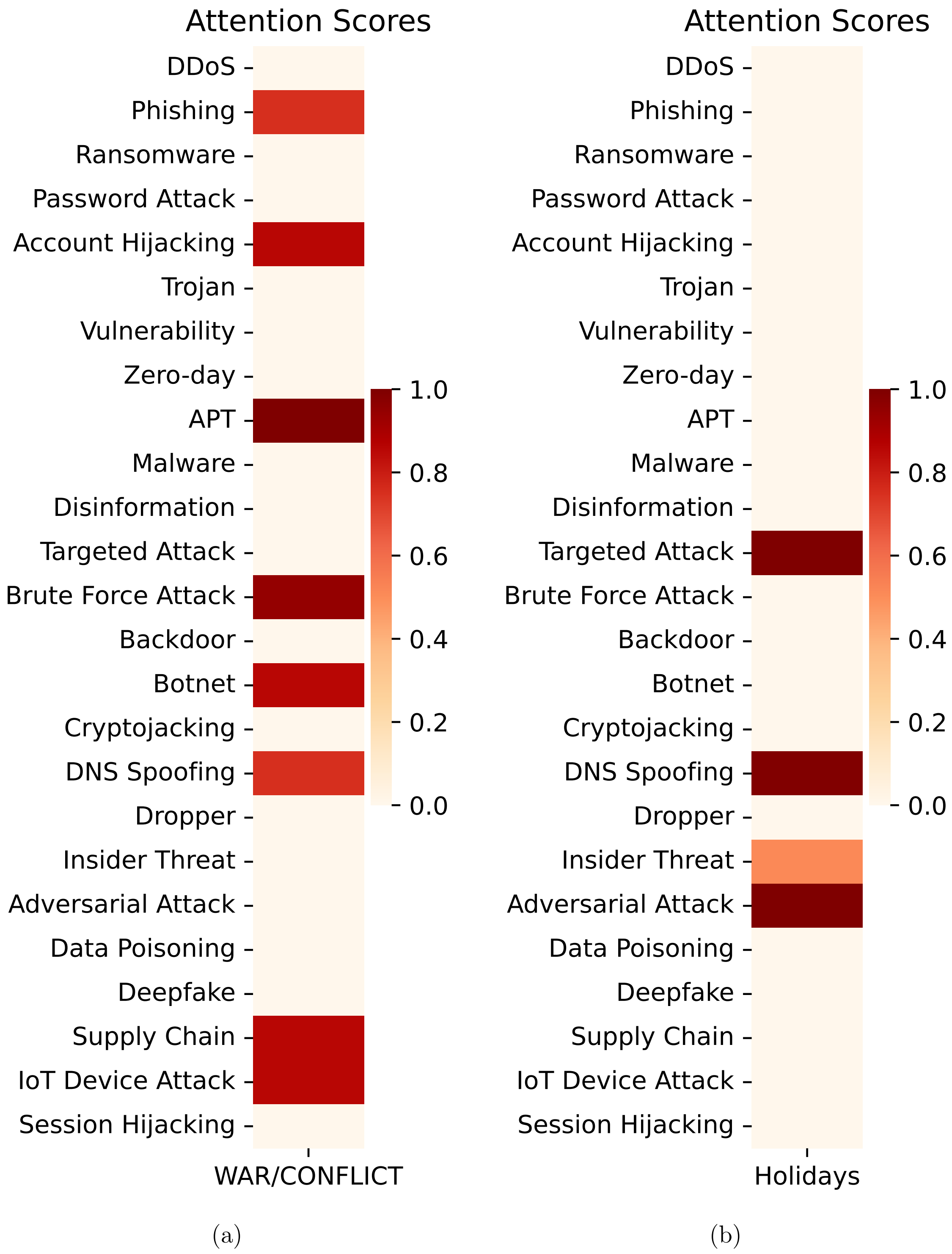


Figure 5.4: Attention scores heat maps. (a) Attacks vs Wars. (b) Attacks vs Holidays.

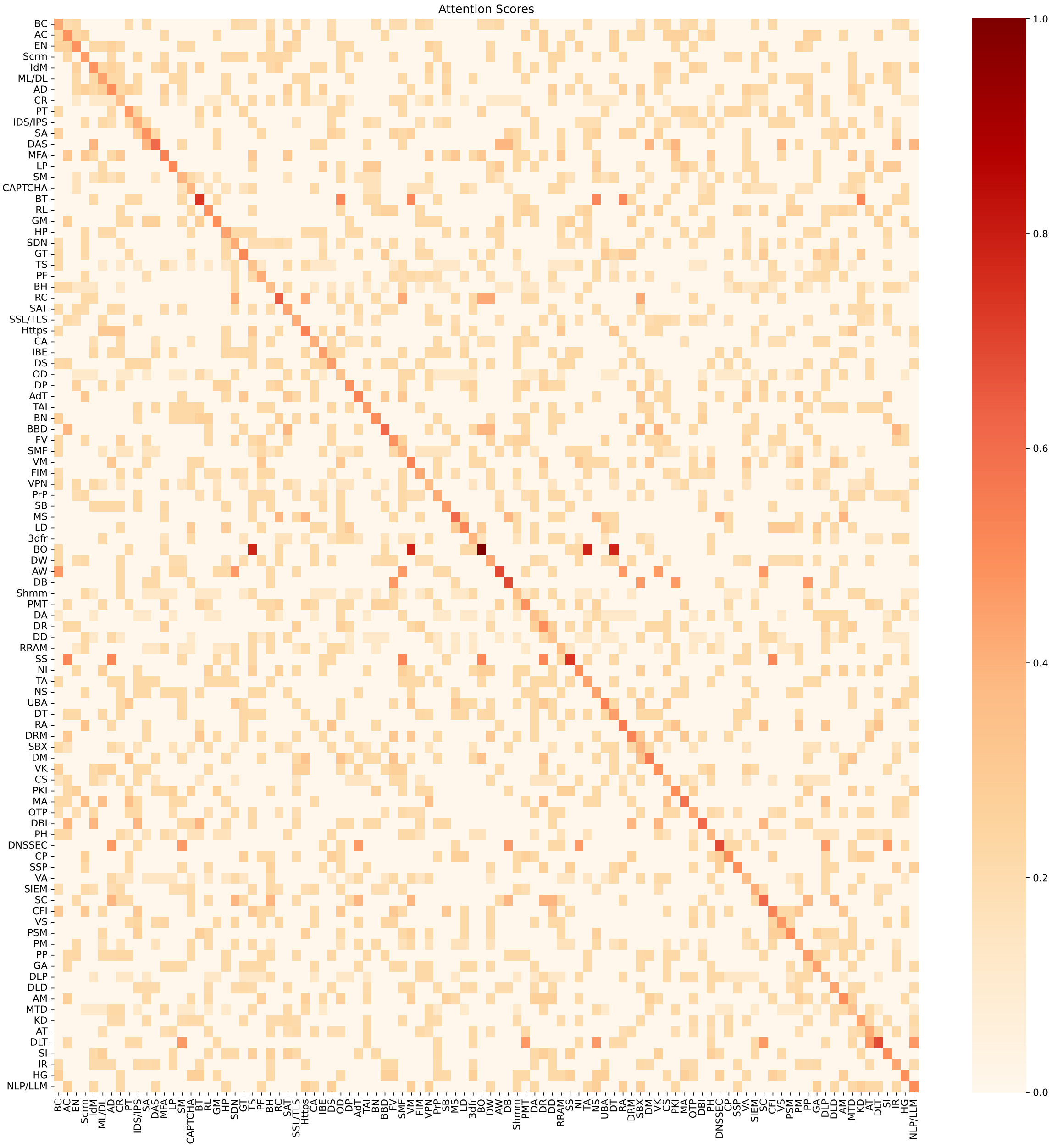
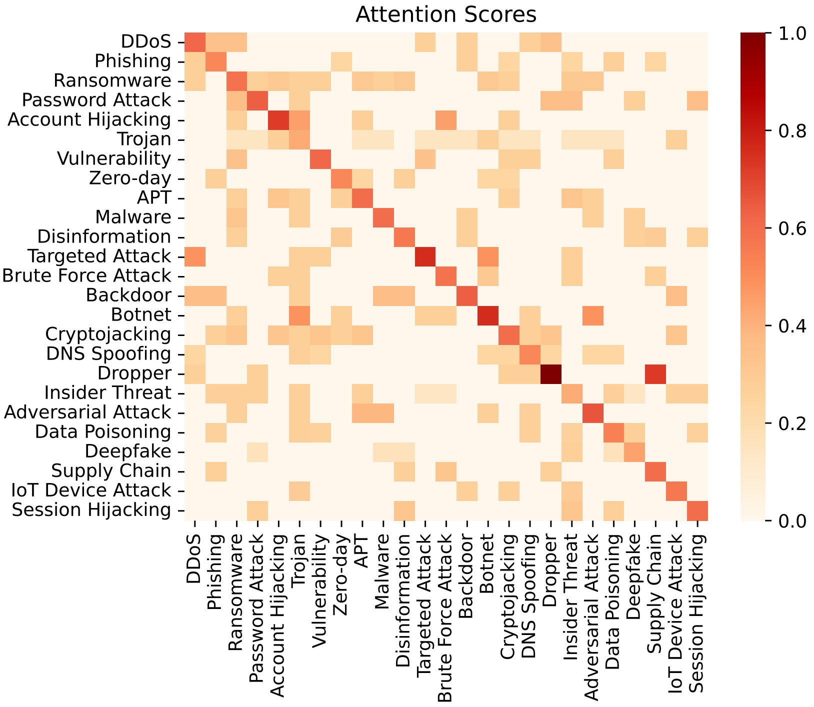


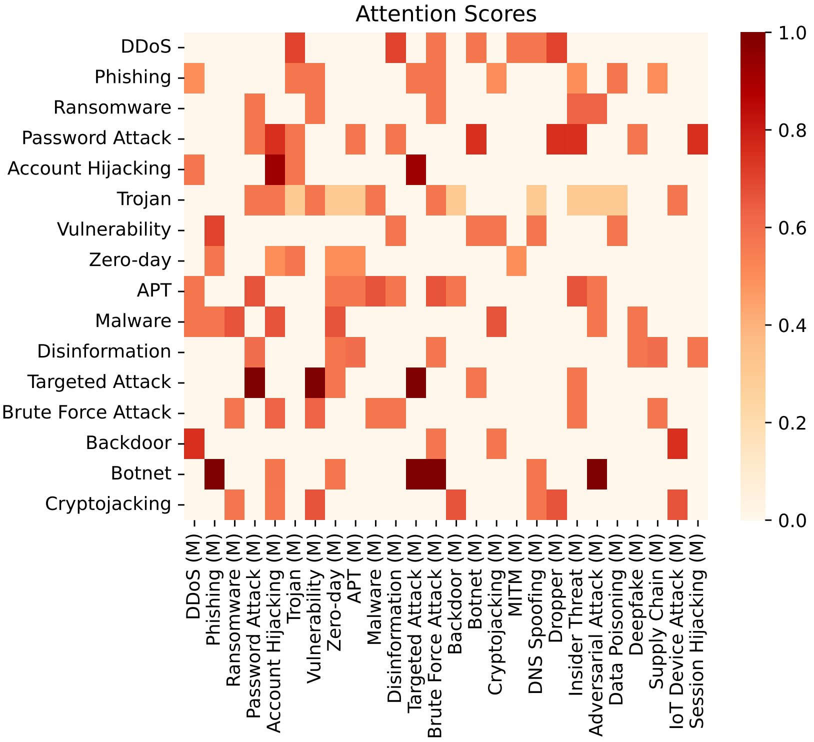
Figure 5.5: Attention scores heat maps - PATs vs PATs.

5.4.1.2 Attention Scores - Analysis

Based on the results of Figure 5.3, some relationships learned by the model are consistent with our early assumptions, as exemplified by the high attention scores between adversarial attack and Adversarial Training (AdT), and between insider threat and Activity Monitoring (AM). However, the model also learned other relationships that we failed to capture in the TPT graph and that are not easily perceived by humans. This includes the high attention scores between targeted attack and Moving Target



(a)



(b)

Figure 5.6: Attention scores heat maps. (a) Attacks vs Attacks (b) Attacks vs Attack Mentions.

Defence (MTD), and between botnet and Supply Chain Risk Management (SCRM). This implies that predefined relationships may not be optimal for accurate prediction, and it is more effective to allow the model to capture these relationships in a way that leads to more accurate predictions [27]. Moreover, machines provide the additional benefit of quantifying these relationships, which can be challenging for human experts to achieve. Overall, we observe that combining the attention scores with the prediction heat maps provides effective explainability for the machine's prediction decisions.

The heat map analysis depicted in Figure 5.4 reveals noteworthy patterns in the attention scores assigned by the GNN model to different cyber threats in relation to specific contextual nodes, such as wars and conflicts, and public holidays. Notably, during periods of heightened geopolitical tension or conflicts (represented by the WAR/-CONFLICT node), the GNN model indicates high attention scores for attack types like phishing, account hijacking, and APT. These attacks are often associated with exploiting vulnerabilities during chaotic periods, making them critical focus areas for policymakers seeking to safeguard sensitive data and infrastructure.

Conversely, during public holidays, the model assigns high attention scores to targeted attacks, insider threats, and DNS spoofing. These attacks may capitalise on reduced cyber security vigilance or altered behaviour patterns during holidays, warranting heightened security measures and awareness campaigns. The lower attention scores for attack types like disinformation or deepfake during these periods suggest a potential shift in threat actor tactics or priorities, which policymakers should monitor and adapt their strategies accordingly. Overall, these findings offer policymakers actionable insights into prioritising cyber security efforts based on contextual factors like geopolitical events or public holidays. By understanding the varying degrees of threat relevance during different periods, policymakers can tailor response strategies, allocate resources effectively, and enhance cyber security resilience in a dynamic threat landscape.

Further relationships can be inferred from Figures 5.5 and 5.6 between the alleviation technologies, between the attacks, and between the attacks and the attacks' mentions in the literature. Such relationships represented by attention scores can potentially aid in feature engineering to enhance the predictive accuracy of ML models. Researchers can explore these relationships in greater depth when developing forecasting models, aiming to enhance the model's performance and adopt a proactive stance in cyber trend forecasting.

5.4.1.3 Comparison with the Human Approach

While the advent of advanced ML models, such as our GNN-based prediction model, has revolutionised the realm of cyber threat forecasting, it is crucial to draw parallels with human prediction and explainability. Human intuition and experience have long been the bedrock of decision-making, particularly in the complex landscape of cyber security. Unlike ML models, human prediction often relies on a rich tapestry of contextual understanding, domain expertise, and a nuanced comprehension of the intricate relationships within the cyber threat ecosystem. Human analysts bring a wealth of tacit knowledge and the ability to factor in qualitative elements that might elude algorithmic approaches.

However, where human prediction excels in intuition and qualitative judgment, it may falter in processing vast datasets, identifying subtle patterns, and making predictions with extended lead times. This is where ML models, like our GNN-based model, step in to complement human capabilities. Saliency maps offer a visual aid, providing analysts with a transparent representation of the model's decision-making process, akin to how a human might mentally weigh the importance of various historical events when predicting future outcomes. Revealing hidden relationships, facilitated by adaptive learning of adjacency matrices, contributes an additional layer of interpretability, allowing stakeholders to glean insights into connections that might elude human intuition alone.

It is also important to re-emphasise the inherent subjectivity and potential for bias in human judgment compared to the objectivity of machine-based approaches. Human analysts may bring valuable intuition and qualitative judgment to the table, leveraging their domain expertise and contextual understanding. However, human predictions can be influenced by cognitive biases, personal experiences, and individual perspectives, leading to inconsistent or subjective assessments. In contrast, machine-based approaches offer objectivity and quantitative analysis, relying on data-driven techniques to make predictions.

Overall, the symbiosis of ML models and human expertise holds the potential to elevate cyber threat prediction to new heights. By combining the innate strengths of human intuition with the computational prowess of advanced models, organisations can foster a collaborative and dynamic approach to cyber security decision-making. This hybrid model aims to bridge the gap between the explainability of human judgment and the analytical capabilities of ML, providing a holistic and robust framework for navigating the ever-evolving landscape of cyber threats.

5.4.2 Highlights and Contributions

This work presents a significant contribution to the field of cyber threat forecasting by focusing on the critical aspect of model explainability. Through the exploration of saliency maps and attention scores, this work unveils the inner workings of the predictive model, offering insights that contribute to informed decision-making within organisations. By elucidating the importance of time steps through saliency maps, practitioners can make strategic choices regarding feature selection and engineering, ultimately enhancing the precision and accuracy of cyber threat predictions. This aspect is pivotal in empowering stakeholders with actionable insights derived from the model's decision-making process, thereby enabling them to make informed decisions based on data-driven analysis.

Furthermore, the work emphasises the role of explainability in enhancing the trustworthiness of predictive models. By providing transparent visual representations of how the model arrives at its predictions, such as through saliency maps, stakeholders gain a deeper understanding of the rationale behind the model's outputs. This transparency not only fosters trust in the model but also facilitates collaboration and buy-in from decision makers who rely on accurate and interpretable insights for strategic planning and risk mitigation.

A notable contribution of this work lies in the discovery and visualisation of hidden relationships within the cyber threat landscape. Through attention scores that reveal adaptive learning of adjacency matrices, the model uncovers nuanced connections between various nodes, offering a holistic view of the interconnected dynamics of cyber threats. This discovery empowers policymakers to formulate targeted defence strategies and investment decisions, prioritising resources based on a comprehensive understanding of threat interdependencies. Overall, this work contributions extend beyond predictive accuracy to encompass informed decision-making, enhanced trustworthiness in the model, and the discovery of hidden relationships critical for effective cyber threat management.

5.4.3 Limitations

One limitation of the explainability methods discussed in this chapter is the absence of evaluation criteria for assessing the model's explainability. For instance, the model may erroneously learn that a past time step is crucial for predicting a future time step. Similarly, the model may presume that two nodes should possess a high attention score or exhibit a strong relationship, despite potentially being irrelevant. These errors stem from the model's imperfections in predicting the future, thus the assumptions regarding feature importance or attention score values may not be optimal. While the model's prediction performance can be assessed using various evaluation metrics, such as RSE and RAE, evaluating explainability can pose challenges.

Another limitation of the described explainability techniques is the difficulty in ensuring that the model's predictions are solely based on the observed explainability. For instance, it could be coincidental that the model predicted two nodes similarly or proportionally, and their attention score is high. Perhaps, another factor, such as the importance of a time step, could influence this decision rather than attention scores alone. Therefore, it is a complex task to ensure that the assumed explainability holds true in all scenarios.

Additionally, one limitation pertains to the interpretability versus causality dilemma. While saliency maps and attention scores offer valuable insights into feature importance and relationship dynamics, they do not always elucidate causality. The model may identify correlations and patterns without necessarily understanding the underlying causal mechanisms. This distinction is crucial, especially in complex domains like cyber security, where causality can dictate effective mitigation strategies.

5.5 Summary and the Road Ahead

In this chapter, we studied correlational explainability techniques in cyber trend forecasting, focusing on two key components: saliency maps and attention scores. Saliency maps provide a visual representation of the importance of input variables over time, aiding in understanding the model's decision-making process. Through these maps, we were able to understand the significance of past time steps in predicting future trends, with a gradual decrease in importance as we move back in time. This understanding aids in improving forecasting accuracy and enhances model interpretability, fostering trust among stakeholders. Attention scores, on the other hand, reveal hidden relationships between nodes in the cyber security landscape, justifying the prediction, and offering insights into strategic decision-making. Importantly, these scores unveiled nuanced relationships between cyber threats and contextual factors like geopolitical events or public holidays, empowering decision makers to prioritise defences effectively. Overall, this work contributes to advancing the field of cyber threat forecasting by bridging the gap between model predictability and human interpretability.

In the next chapter, we will provide a high-level discussion about the results of this thesis, highlighting our findings and contributions, and addressing the limitations of this work. This discussion will cover the key research topics of this thesis namely cyber threat forecasting, alleviation technologies forecasting, and the explainability of cyber trend forecasting.

6 | Discussion

6.1 Overview of Key Findings

Our research spanned various dimensions of the cyber threat landscape, providing insights and actionable recommendations for proactive and effective defence strategies. Through meticulous analysis and innovative methodologies, we have classified threat trends, highlighted the gaps with mitigative technologies, introduced cycle models for understanding trend evolution, explored model explainability, and conducted extensive experimental evaluations. This chapter serves as a culmination of our efforts, offering a holistic perspective on cyber security readiness and the challenges and opportunities in forecasting and mitigating cyber threats effectively. We will dive into the key findings and contributions of this research journey and highlight the research limitations.

Our research illuminates various aspects of the evolving threat landscape in cyber security, offering actionable insights for proactive defence strategies. The analysis of threat trends and their gaps with the trend of the alleviation technologies categorises future trends into important categories. These categories are vital for prioritising future investments and identifying strategic defence decisions. Furthermore, the introduced cycle models namely TTC and ATC offer a structured framework for understanding threat and technology evolution, guiding strategic resource allocation based on the phases of the trends and the anticipated gaps. The explainability methods further compliment the proposed approach, allowing for model's transparency and revelation of hidden dependencies. Moreover, the extensive experimental evaluation validates the predictive capabilities of the proposed models and the effectiveness of the proposed data features.

Overall, this research stands as a seminal contribution to the field of cyber security, pioneering a multifaceted approach that addresses the challenges of modern cyber threats. At its core is the development of an ML proactive approach for long-term prediction of cyber threats and PATs. This proactive approach facilitates effective vision into future threats and allows for the development of optimal defensive measures based on the future gaps with cyber threats. By prioritising resources to prevent rapidly growing and emerging attacks and attacks with large gaps with its mitigation technologies, this research advocates for a proactive shift in cyber security strategies, moving away from reactive measures.

A key contribution of this research is the proposal of a fully automated approach, aimed at reducing the influence of human subjective bias and expertise scarcity, which can often lead to inconsistent and unreliable predictions. This automation ensures reproducibility of results and facilitates the application of recent advancements in XAI, making the decision-making process more transparent and comprehensible. This shift towards automation represents a paradigm shift in cyber security practices, enabling organisations to adopt a proactive stance and develop pre-emptive strategies based on data-driven insights, thereby staying ahead of evolving cyber threats and fortifying their defences effectively.

Furthermore, the research encompasses a wide geographic scope, constructing and analysing data from 36 countries to gain a holistic understanding of the global cyber threat landscape. This comprehensive analysis includes diverse data sources such as social media discourse, temporal factors like public holidays, and scholarly attention, highlighting the need for a multifaceted approach to threat prediction. Insights from this analysis inform targeted mitigation strategies and support the development of more effective defensive measures.

Moreover, this research integrates advanced ML techniques, notably Bayesian LSTM and Bayesian GNN, into its predictive models. It also explores and proposes novel external features to boost the models' accuracy. The incorporation of these techniques significantly enhances the models' performance and underscores the critical role of feature engineering in improving predictive accuracy. By automating the analytic process and leveraging quantitative procedures, the research ensures a lower degree of subjectivity in predictions, leading to more consistent and reliable outcomes.

Additionally, the research provides a model explainability approach through saliency maps and attention scores, providing stakeholders with transparent insights into the decision-making process. Furthermore, the research uncovers hidden relationships within the cyber threat landscape, empowering policymakers to formulate targeted defence strategies and investment decisions. By prioritising resources based on a comprehensive understanding of trend interdependencies, organisations can enhance their cyber security resilience and effectiveness. Overall, this research's cohesive framework encompasses proactive prediction models, advanced ML techniques, model explainability, global data analysis, and strategic planning, setting a precedent for proactive cyber security practices and effective cyber threat management.

6.2 Interpretation of Results

The categorisation of threat trends into distinct patterns as proposed in chapter 3, allows for a nuanced understanding of the cyber threat landscape. This classification facilitates targeted mitigation efforts and underscores the need for dynamic defence mechanisms that can adapt to evolving threat landscapes. Rapidly increasing threats like zero-day exploits require immediate attention and investment in technologies aimed at mitigating their impact. Meanwhile, overall increasing threats like APTs may necessitate continuous monitoring and adaptation of defensive measures, and emerging threats such as IoT attacks demand innovative approaches and proactive measures to stay ahead of evolving attack vectors. By understanding the trajectory of threats across these categories, decision makers can prioritise resources effectively and enhance their cyber security resilience. Additionally, the TTC model provides a structured framework for contextualising threat evolution, encompassing stages from launch to stability. This model not only captures the life cycle of cyber threats but also offers insights into their current states and future destinations. By delineating the phases of threat evolution, organisations can better anticipate shifts in threat landscapes and allocate resources accordingly. The model also guides investment priorities, emphasising the need to transition emerging threats to stability phases and prevent the plateauing of mature threats. This adaptive approach ensures that cyber security strategies remain dynamic and responsive to the changing threat landscape, ultimately enhancing overall resilience against cyber-attacks.

Moreover, feature importance analysis highlights the predictive power of multivariate approaches, showcasing the significance of features such as tweets related to conflicts, public holidays, and attack mentions in scientific literature. These features emerge as crucial informative factors capable of enhancing the B-LSTM model's performance and forecasting potential cyber threats accurately. The correlation between these features and the incidence of cyber-attacks underscores their importance in understanding the evolving threat landscape. Monitoring these features can provide valuable insights into cyber threats' prominence and guide proactive cyber security measures effectively.

Our results in chapter 4 include a comprehensive future gap analysis between cyber threats and the alleviation technologies, the development of the ATC model, and an assessment for the B-MTGNN model's performance, shedding light on critical aspects of cyber security preparedness. The gap analysis reveals significant disparities between cyber threats and several alleviation technologies highlighting the need for proactive measures to bridge these gaps effectively. Recommendations include prioritising investments in technologies such as Application Whitelisting, SIEM, Control Flow Integrity, and Deception Technology to mitigate vulnerabilities and enhance cyber security resilience. Categorising gaps into distinct categories enables policymakers to prioritise investments more efficiently, focusing on technologies with widening gaps and strategic importance such as Darknet Monitoring and NLP/LLM. The ATC model offers insights into the progression of PATs and their alignment with emerging and rapidly increasing threats. Policymakers can strategically allocate resources based on the ATC phases, advancing the progress of technologies to be positioned on the upper plateau to address relevant threats effectively. This approach facilitates data-driven decision-making, optimises resource allocation, and enhances overall cyber security posture amidst evolving threats.

The evaluation of the B-MTGNN model used for forecasting threats and technology gaps underscores the efficacy of integrating graph learning and graph convolutional layers in improving forecasting accuracy. The utilisation of these layers enhances the model's ability to capture complex relationships within the cyber threat landscape, contributing to more accurate forecasts and proactive threat mitigation strategies. Also, notable improvements in relative errors re-emphasises the impact of contextual information such as attacks' mentions in literature, tweets about conflicts, and public holidays on predicting cyber incidents. The comparative evaluation showcases the superiority of the Bayesian MTGNN model in cyber trend forecasting. The Bayesian model's ability to aggregate predictions, capture uncertainty information, and provide reliable forecasts outperforms deterministic counterparts, offering a robust framework for decision-making and risk management strategies.

Chapter 5 focuses on the analysis of saliency maps and attention scores, highlighting their role in enhancing forecasting accuracy and model interpretability. Saliency maps reveal that the trend in the last four months significantly impact future predictions, with a gradual decrease in importance as we move back in time. This understanding aids in informed feature selection and engineering, boosting forecasting accuracy and stakeholders' trust in the model. Attention scores further elucidate relationships learned by the model, showcasing consistent patterns and uncovering hidden associations that human intuition may overlook. The combination of attention scores and prediction heat maps offers effective explainability for the model's prediction decisions. Also, the attention scores provide actionable insights for policymakers to prioritise cyber security efforts based on contextual factors like geopolitical events or public holidays. Additionally, attention scores between alleviation technologies, attacks, and attacks' mentions in literature can aid in feature engineering, enhancing ML model performance in cyber trend forecasting.

While ML models like the GNN-based prediction model revolutionise cyber threat forecasting, our findings also underscores the complementary nature of human prediction and explainability. Human intuition and expertise excel in qualitative judgment and contextual understanding, whereas ML models excel in processing vast datasets and identifying subtle patterns. The symbiosis of human judgment and ML capabilities bridges the gap between explainability and analytical prowess, offering a holistic framework for effective cyber security decision-making. This collaborative approach aims to elevate cyber threat prediction by leveraging the strengths of both human analysts and advanced ML models, ensuring a dynamic and resilient defence against evolving cyber threats. Overall, the integrated findings provide a holistic perspective on cyber security readiness, highlighting the importance of proactive measures, strategic investments, and data-driven decision-making in mitigating cyber threats effectively.

6.3 Contextualisation within Existing Literature

Our study introduces a novel proactive approach in the field of cyber security, leveraging big data and state of the art ML models to forecast cyber threats and PATs over extended periods. This approach represents a substantial progression within the field of cyber security, aligning closely with the burgeoning body of research advocating for proactive measures in combating cyber threats [5, 17, 160]. By proposing the long-term forecast of cyber threats and PATs, our research directly addresses a significant research gap identified in previous studies [5, 6]. While existing studies are limited to short-term and midterm prediction spanning hours, days, or months at best [14, 15, 16, 17], our research extends the forecast horizon to span 3 years, allowing security agencies sufficient time to develop optimal defensive measures.

Additionally, the research addresses a crucial research gap by introducing a paradigm shift in the long-term forecasting of cyber security trends, eliminating the traditional reliance on security experts [18, 19]. By adopting a fully automated and quantitative approach grounded in ML techniques, we mitigate the inherent subjectivity, bias, and scarcity associated with human expertise in the cyber security domain [20]. This departure from conventional methods enables us to develop robust predictive models capable of forecasting cyber security trends over extended periods, offering reproducible and transparent results. Our approach not only enhances the reliability and consistency of cyber security predictions but also broadens accessibility to such forecasting capabilities, democratising the process and empowering organisations to proactively manage cyber threats without being hindered by expertise limitations.

Also, our exploration into predicting the future disparity between cyber threats and the corresponding alleviation technologies represents a clear departure from existing approaches in the domain of cyber security. While traditional efforts have primarily focused on forecasting cyber threats alone [14, 15, 16], our research extends this scope by forecasting the trends of both cyber threats and their PATs. By incorporating both aspects into our predictive framework, we explore uncharted territory, aiming to anticipate not only the trajectory of cyber threats but also the evolving dynamics between threats and defensive measures. This pioneering approach goes beyond the confines of reactive security practices by embracing a proactive stance towards cyber defence. By forecasting the trends of both threats and PATs, we aim to identify future gaps and vulnerabilities in the cyber landscape, thereby enabling policymakers to pre-emptively address emerging risks. This forward-looking perspective not only enhances our ability to anticipate and mitigate cyber threats but also empowers us to strategically allocate resources and prioritise defensive measures.

Furthermore, our research represents a significant departure from traditional cyber security analyses, which typically focus on current threat assessments and immediate defensive responses. By forecasting future disparities between threats and PATs, we shift the focus towards strategic planning and long-term resilience building. This shift enables organisations to stay ahead of evolving cyber risks and adapt their defensive strategies accordingly. In essence, our exploration into predicting the future disparity between cyber threats and alleviation technologies opens up new avenues for proactive cyber defence. By forecasting the trends of both threats and PATs, we pioneer a holistic approach to cyber security that enables organisations to anticipate, adapt, and mitigate emerging risks effectively. Through this endeavour, we pave the way for a more resilient cyber landscape better equipped to withstand the challenges of an increasingly complex threat environment.

Moreover, the integration of advanced ML techniques, notably Bayesian models, builds upon established literature pertaining to predictive modelling methodologies across diverse domains, including but not limited to traffic forecasting [27, 43]. Notably, approximating the Bayesian variation of these models enhances model's interpretability by quantifying uncertainty of the model which leads to enhanced confidence in the model's prediction. Furthermore, quantifying the epistemic uncertainty effectively mitigates the ongoing challenge of long-term forecasting with limited data, a current hindrance in the cyber security domain [6]. In fact, Bayesian modelling allows the incorporation of prior domain knowledge and assumptions about the model parameters to enhance the model performance. Therefore, it represents a promising approach to deal with the issue of data scarcity in the domain of cyber security [45, 46].

Finally, the observed enhancement in model performance when incorporating our proposed external features supports findings from prior scholarly work [15, 17, 55], underscoring the pivotal role of feature engineering in augmenting the effectiveness of predictive models. An illustration of such a feature is the inclusion of data related to wars and conflicts in cyber threat prediction. Such feature represents an essential factor capable of boosting the performance of ML models and guiding subsequent analyses for cyber threat prediction.

6.4 Implications for Future Research

This research holds significant implications for advancing the field of proactive cyber security. By pioneering a multifaceted approach to predicting cyber threats, it sets a precedent for future research initiatives. One avenue for further exploration lies in refining and extending predictive models to encompass even longer forecast horizons. While our research extends the forecast horizon to span three years, future studies could push this boundary further, exploring the feasibility of forecasting cyber threats and PATs over even longer timeframes. This would enable security agencies to develop strategic defensive measures with greater foresight, thereby enhancing preparedness and resilience against cyber threats.

Another notable implication lies in the exploration and refinement of predictive models, emphasising the integration of evolving ML techniques to effectively anticipate cyber threats and relevant technologies. The demonstrated performance enhancement achieved through the Bayesian model suggests a potential paradigm shift towards leveraging advanced techniques in graph analytics. Future research may explore optimising and tailoring graph-based algorithms specifically for cyber trend prediction, aiming to enhance the accuracy and efficiency of predictive models. Additionally, future research could delve deeper into the development of hybrid models that combine the strengths of different ML techniques. For example, integrating graph-based models with deep learning approaches could yield more robust predictive models capable of capturing complex relationships within the cyber threat landscape. By leveraging the complementary strengths of different ML techniques, such hybrid models could further enhance predictive accuracy and reliability, offering valuable insights for proactive cyber defence strategies.

The effective data features proposed in this research offer avenues for further exploration and extension to enhance model performance. By identifying key data features that significantly impact predictive accuracy, future research can focus on refining these features and exploring additional variables to improve forecasting capabilities. Additionally, the emphasis on extensive global data coverage, spanning 36 countries, underscores the importance of comprehensive data analysis for a holistic understanding of the cyber threat landscape. Future research initiatives could explore further enhancements in data collection methodologies, analysis techniques, and representation approaches to achieve an even broader international scope.

Moreover, there is potential for future research to explore the integration of real-time data streams and dynamic updating mechanisms into predictive models. Incorporating real-time data feeds from diverse sources such as social media, news outlets, and cyber threat intelligence platforms could enable more timely and responsive threat forecasting. Dynamic updating mechanisms could also allow predictive models to adapt to evolving threat landscapes in real-time, ensuring that defensive measures remain effective and up-to-date in the face of rapidly changing cyber threats.

Another avenue for future research lies in the development of tailored predictive models for specific industries or sectors. While our research offers a general framework for forecasting cyber threats and PATs, future studies could focus on customising predictive models to address the unique challenges and threat landscapes faced by different industries. By tailoring predictive models to specific sectors such as finance, healthcare, or critical infrastructure, researchers could provide valuable insights and guidance for sector-specific cyber defence strategies.

Concerning explainability, integrating attention mechanisms directly into the model architecture represents a promising avenue for future research aimed at enhancing model interpretability in cyber security. By incorporating attention mechanisms into the model design, researchers can develop models that not only make accurate predictions but also provide insights into the factors driving those predictions. One approach could involve modifying existing ML architectures to include attention layers that dynamically weigh the importance of different input features or time steps during the prediction process [43]. By visualising these attention weights, researchers can gain valuable insights into which features or temporal patterns are most influential in predicting cyber security trends.

Moreover, future research could explore novel visualisation methods or diagrams to effectively communicate the insights gleaned from attention mechanisms to stakeholders. For example, researchers could develop interactive visualisation tools that allow users to explore the attention weights assigned to different features or time steps in the data. These tools could provide visual representations of how attention shifts over time or across different features, enabling stakeholders to gain a deeper understanding of the underlying patterns driving model predictions. Additionally, researchers could explore the use of heat maps, line graphs, or other graphical representations to visualise the attention weights in an intuitive and informative manner.

Furthermore, integrating attention mechanisms into the model architecture opens up opportunities for developing novel explainability techniques tailored specifically to cyber security applications. For instance, researchers could explore methods for quantifying the contribution of individual features or feature combinations to model predictions based on attention weights. By analysing these contributions, researchers can identify which features are most relevant for predicting specific types of cyber threats and tailor defensive strategies accordingly. Additionally, researchers could explore the use of attention-based anomaly detection techniques to identify unusual patterns or deviations in cyber security data that may indicate potential security breaches or emerging threats.

6.5 Practical Applications

The findings of our research offer valuable insights and practical applications for various stakeholders in industry and government settings. In industry, organisations can leverage the proactive approach proposed in this research to enhance their cyber security posture and effectively mitigate cyber threats. By incorporating the long-term forecast of cyber threats and PATs into their strategic planning, companies can anticipate future challenges and allocate resources accordingly. For example, organisations can prioritise investments in technologies that address future gaps with cyber threats as identified in this research. This enables companies to stay ahead of evolving cyber risks and adapt their defensive strategies to mitigate potential vulnerabilities effectively.

Moreover, the fully automated and quantitative approach proposed in this research reduces reliance on human subjective bias and expertise scarcity, ensuring reproducible and transparent results. Organisations can implement automated predictive models grounded in ML techniques to enhance the accuracy and reliability of cyber threat forecasts. By leveraging recent advancements in XAI, companies can improve decision-making processes and enhance stakeholders' trust in predictive models. This shift towards automation represents a paradigm shift in cyber security practices, empowering organisations to adopt proactive strategies based on data-driven insights and stay ahead of evolving cyber threats.

Furthermore, the comprehensive global data analysis conducted in this research underscores the importance of holistic data analysis for understanding the cyber threat landscape. Companies can utilise diverse data sources such as social media discourse, temporal factors like public holidays, and scholarly attention to gain a comprehensive understanding of cyber threats' prominence and evolution. Insights from this analysis can inform targeted mitigation strategies and support the development of more effective defensive measures tailored to specific industry sectors and geographical regions.

In government settings, policymakers can leverage the insights from this research to formulate strategic cyber security policies and initiatives. By understanding the trajectory of cyber threats and PATs across different categories and phases, governments can prioritise investments in cyber security technologies and infrastructure. For example, governments can allocate resources to enhance the development and adoption of technologies that address emerging threats or bridge anticipated gaps with cyber threats. This strategic approach enables governments to mitigate cyber risks effectively and safeguard critical infrastructure and national security interests.

Moreover, the integration of advanced ML techniques, such as Bayesian models, into predictive models offers governments a robust framework for forecasting cyber threats and PATs accurately. By incorporating these techniques into their cyber security strategies, governments can improve the effectiveness of threat detection and response efforts. Additionally, the model explainability approach through saliency maps and attention scores provides policymakers with transparent insights into the decision-making process behind cyber threat forecasts. This transparency enables governments to make informed decisions and prioritise cyber security measures effectively.

Overall, the findings of this research have significant implications for practical applications in industry and government settings. By adopting proactive approaches, leveraging automated predictive models, conducting comprehensive data analysis, and prioritising strategic investments, organisations and governments can enhance their cyber security resilience and effectively mitigate cyber threats in an increasingly complex threat landscape.

6.6 Limitations

From a methodological standpoint, it is essential to acknowledge the limitations inherent in the dataset used for this research. The restricted dataset spanning only 12 years poses challenges in capturing long-term trends and forecasting further into the future accurately. Accessing confidential and sensitive information presents hurdles in acquiring a more extensive dataset, which could enhance the model's forecasting capabilities and generalisation. Additionally, while Bayesian models are employed to address epistemic uncertainty, a more comprehensive dataset would reduce this uncertainty which further bolsters the model's predictive accuracy and reliability.

Additionally, methodological limitations manifest in the absence of systematic evaluation approaches for algorithms such as the E-GPT algorithm, which poses challenges in optimising the graph structure. Also, the reliance on human analysis in the development of cycle models may lead to inefficiencies and subjectivity in placing attacks and technologies on these cycles. Moreover, limitations in the explainability methods discussed raise concerns about evaluating model explainability and ensuring predictions are solely based on observed explainability. The complexity of ensuring causality in interpreted relationships and the potential for coincidental predictions based on attention scores alone also underscore challenges in fully understanding and leveraging model explainability.

Regarding the scope of work, it is crucial to recognise the research's limitations in exploring fine-grained attack types due to dataset granularity constraints. While the current dataset provides insights into high and medium-level attack types, distinguishing between specific variants of cyber-attacks remains challenging. This limitation restricts the development of tailored defence mechanisms against nuanced threats, highlighting the need for datasets with finer granularity to address this aspect comprehensively.

Moreover, the current scope of work lacks prediction for the onset of cyber-attacks or anticipation of novel attack types. The onset of a cyber-attack marks the initiation of malicious actions targeting computer systems or networks, signifying the commencement of unauthorised activities by threat actors [48]. Furthermore, anticipating novel attack types is imperative for staying ahead of evolving cyber threats. These emerging tactics often exploit unaddressed vulnerabilities or employ innovative methods undetectable by traditional defence mechanisms. Early identification and understanding of these novel attack types empower organisations to fortify their defences proactively and mitigate potential risks. Perhaps, integrating additional features capturing attackers' intentions, motivations, and short-term indicators like network traffic patterns could significantly enhance the model's ability to forecast imminent attacks and promptly identify emerging threat vectors.

In summary, while the research makes significant contributions to the field of cyber security, addressing methodological limitations concerning model's performance and evaluation frameworks, alongside considerations of the scope of work including attack type granularity and prediction of imminent or novel attacks is essential for advancing the efficacy and reliability of cyber trend forecasting approaches.

7 | Conclusion

7.1 Summary

In conclusion, this work represents a significant contribution to the field of cyber security, proposing a comprehensive approach that addresses key challenges in cyber threat prediction and mitigation. Our research has introduced an ML-based proactive approach for long-term prediction of cyber-attacks, facilitating early communication with potential threats years in advance to enable strategic planning for optimal defensive measures. By leveraging automated processes and big data sources, we have demonstrated the potential to overcome limitations associated with human-based analysis, such as expertise scarcity and subjective bias, resulting in more consistent and reproducible predictions. The study has shown the ability of our proposed Bayesian LSTM model to produce the future forecast for the trend of 42 attack types 3 years ahead while capturing epistemic uncertainty. Our extensive analysis facilitated the categorisation of cyber threat trends and the development of the TTC model, capturing key phases in the life cycle of 42 attack types. These results provide a roadmap for future research endeavours and practical cyber security strategies. Tailoring defence mechanisms to anticipated threats and leveraging collective defence efforts globally are essential steps towards proactive mitigation strategies and enhanced cyber security preparedness.

Furthermore, our work expanded on this proactive approach by incorporating novel ML techniques for long-term prediction of cyber threats and PATs. The introduction of the Bayesian GNN model (B-MTGNN) along with our external features has enabled us to forecast future gaps between cyber threats and PATs 3 years in advance, categorise these gaps, and recommend investment and strategic defence decisions accordingly. Our large scale analysis culminated in the development of the ATC model, identifying the life cycle phases in the trend of 98 mitigation technologies. This comprehensive framework equips policymakers with valuable insights and tools to strategically allocate resources and align defence mechanisms with evolving cyber threats effectively. Our extensive experiments have shown that our proposed features boost the performance of the GNN model, and that the proposed Bayesian GNN model outperforms several baseline models including its deterministic counterpart. Lastly, we equipped our GNN model with explainability features, offering insights through saliency maps and attention scores. These methods have improved model interpretability and fostered trust among stakeholders, bridging the gap between model predictability and human interpretability. Moreover, the attention scores revealed hidden associations between cyber trends allowing for improved decision-making and effective prioritisation of defences. Overall, our research not only advances theoretical understanding but also provides practical insights for managerial decision-making, aiming to navigate the complex and evolving landscape of cyber threats with confidence and efficacy.

7.2 Review of Chapters

Chapter 1 starts by describing the background of our research establishing the context of the study and setting the stage for readers to understand the motivation behind our work. This includes a description of contemporary methods for addressing cyber threats, which are predominantly reactive in nature, despite the escalating frequency of cyber incidents and their damaging impacts. This is followed by the proposal for our proactive approach to forecasting cyber security trends, describing the research scope and highlighting the research objectives including the long-term forecast of cyber threats and PATs along with the exploration of explainability techniques. We then described the rationale behind our work including the promising shift from reactive to proactive approach and extending the forecast horizon for effective future planning and resource allocation. We also described the benefits of our automated approach in addressing the issue of subjectivity and bias in human judgement and the scarcity of human security experts. We additionally highlighted the importance of forecasting the disparity between cyber threats and PATs for prioritising defences and allocating resources more effectively. Exploring explainability was also justified by emphasising its key benefits including model's transparency and the facilitation of informed decision-making process. We ended this chapter by outlining the research questions and contributions including the proposal of novel Bayesian models for a holistic and long-term forecast of cyber threats and PATs, the provision of 3 years forecast for cyber security trends, the categorisation of future trends along with future recommendations, the proposal of TTC and ATC models offering structured frameworks for contextualising threat and technology evolution, and the incorporation of explainability features into our final model.

Chapter 2 provides a comprehensive literature review of the existing work related to this thesis. It starts by describing the emergence and complexity of modern cyber threats and emphasising the need for theoretical and analytical frameworks to forecast these threats in the long-term. This is followed by describing our data-driven adaptation of PMT as the foundation of our research framework offering a thorough comprehension of individuals' reactions to cyber threats. As a next step, we reviewed key concepts in the field of time series analysis. This includes the definition of univariate and multivariate time series, describing time series forecasting, conceptualising time series as a feature, and modelling multivariate time series in a graph structure. The review also includes an introduction to Bayesian modelling and Bayesian approximation approaches. Next, we categorised cyber threat forecasting into long-term, midterm, and short-term categories, outlining the key benefits and challenges of each approach. This is followed by identifying several research gaps including the lack of research on the long-term forecast of cyber threats and the focus of existing methods on reactive approaches and restricted settings. Moreover, we highlighted the absence of a completely automated methodology for predicting cyber threat-related technologies, the underutilisation of big data and GNNs in forecasting cyber trends, and the oversight in addressing epistemic uncertainty and explainability within such models.

Chapter 3 outlines our approach to cyber threat forecasting. After introducing key concepts in cyber threat prediction, we described our methodologies, which include our framework for forecasting cyber threats using unstructured big data from diverse sources such as news and government advisories, Elsevier, Twitter, and Python APIs. We also discussed our methods for extracting key features from these sources, including NoI, NoM, ACA, and PH, which together comprise our dataset. Next, we detailed our development for the encoder-decoder architecture of the B-LSTM model, which we employed for forecasting cyber threats three years in advance while addressing epistemic uncertainty. Additionally, we proposed the M-SMAPE evaluation metric to account for trend direction in multi-step forecasting, while providing a comprehensible metric. We then described our optimisation for the B-LSTM models, which involves utilising random search to optimise the set of hyper-parameters and smoothing constants, as well as performing stochastic selection of features to obtain the model with the least error for each cyber-attack type.

Our results in chapter 3 include validation analysis of our model, categorisation of threat trends into four main categories, and the proposal of the TTC model, which identifies a unified life cycle for the trend of 42 cyber-attack types. The discussion section at the end of this chapter highlights our key findings, including a set of emerging and rapidly increasing threats worthy of consideration, the leverage of the proposed TTC model for effective resource allocation, and the significance of our proposed features in cyber threat forecasting. Furthermore, we highlighted our key contributions, which include the successful proposal of an ML-based proactive approach to cyber threat prediction and the addressing of human limitations through a quantitative and fully automated approach. We also recommended the adaptation of the proposed TTC model to prevent the escalation of emerging and rapidly increasing threats. At the end of the chapter, we addressed the limitations of our approach, including its reliance on a limited dataset and its inability to forecast the onset of cyber-attacks.

Chapter 4 extends the research scope by studying the forecast of the gap between the trend of cyber threats and the PATs. It starts by describing the extended version of the problem representing the problem as a graph to capture the relationships between threats and PATs. The methods section describes the extended dataset and framework to incorporate the trend of alleviation technologies and develop a GNN-based model. This is followed by detailing the graph construction process leveraging GPT-3 and Elsevier to extract the PATs of each cyber threat. Then, we detailed our development for the B-MTGNN model combining temporal and graph convolutional layers to capture temporal and spatial dependencies in the graph, while adaptively learning the graph structure and quantifying epistemic uncertainty. We then described our experimental settings and hyper-parameter optimisation utilising the random search method. We also described the model evaluation process using RSE and RAE evaluation metrics for validating and testing our model.

The results of chapter 4 include the provision of three years forecast for the disparity between cyber threats and PATs, categorising the gap trends into four main categories, and introducing the ATC model describing the evolution stages of 98 alleviation technologies. Our comparative analysis includes an ablation study to justify the inclusion of the graph learning and graph convolutional layers in our models along with our external features to boost the model's performance. We also provided a comparative evaluation to show that the proposed Bayesian model outperforms several baseline models including ARIMA, VAR, Transformer, as well as its deterministic (non-Bayesian) counterpart. The discussion section highlights our key findings including significant anticipated gaps in the future between several threats and PATs and the recommendation of investment decisions accordingly. The findings also explain how the ATC model can be leveraged to make informed decisions for allocating cyber security resources effectively and to align the defences with the anticipated threats. The superior model performance was also highlighted as a promising finding that can enhance the predictive modelling in the context of cyber trend forecasting. The discussion additionally provided analysis of the Bayesian model's confidence intervals, emphasising their role in quantifying uncertainty and guiding strategic planning despite occasional deviations from ideal coverage. Furthermore, practical concerns such as adversarial adaptation and data centralisation were addressed, proposing resilient strategies to safeguard the model against potential exploits and maintain data confidentiality.

Our key contributions in chapter 4 was highlighted including pioneering a proactive approach to cyber trend forecasting that addresses a critical gap in the literature and advances the research in proactive cyber security. We also highlighted the practical implications of this work which enhances cyber security agencies' preparedness and future planning. We then described several limitations of this work including the dataset's lack of fine granularity in terms of attack types, the lack of a systematic evaluation method to assess the performance of the graph construction approach, and the subjectivity in the manual placement of the PATs on the ATC.

In chapter 5, we explored correlational explainability techniques for cyber trend forecasting. First, we introduced two explainability methods namely saliency maps and attention scores. This is followed by visualising saliency maps for multiple attack types to highlight the importance of previous time steps in the prediction and interpret the prediction decisions accordingly. Then, we visualised attention score heat maps to interpret the GNN model's predictions and reveal the hidden dependencies in the graph. In the discussion, we highlighted our key findings including the importance of the past four months in predicting the future trends and the importance of attention scores for learning relationships that elude human intuition alone. This is exemplified by noteworthy patterns learned by the model between cyber threats and contextual nodes such as wars and political conflicts. Overall, in contrast to human-based methods, we found that while human prediction demonstrates strengths in intuition and qualitative judgment, machine-based approaches complement human capabilities through their capacity to analyse large datasets, detect subtle patterns, and provide objective predictions with extended lead times.

We highlighted our contributions in chapter 5 including our exploration for the inner workings of our B-MTGNN model to offer insights that lead to informed decision-making process. We emphasised that elucidating the importance of time steps through saliency maps allows practitioners to make strategic choices regarding feature selection and engineering to enhance the accuracy of cyber threat prediction. We also highlighted that such explainability techniques enhance trust and confidence in ML models by obtaining transparent visual representations of how models arrive at their predictions. An important contribution was also highlighted which is the ability of these methods to reveal hidden associations between cyber security trends allowing policymakers to formulate targeted defence strategies and investment decisions. Chapter 5 ends by acknowledging the limitations of our work, including the absence of evaluation criteria for assessing the model's explainability, the challenge of ensuring that predictions align with observed explainability, and the oversight in addressing causality through the presented methods.

Chapter 6 presents a comprehensive discussion of our thesis. First, it provides an overview of our key findings including the categorisation of future trends, the identification of critical technology gaps with several cyber threats, and the understanding of threats and PATs evolution through the proposed cycle models. The overview also highlights the key contributions of this work as detailed previously. Chapter 6's interpretation of results highlights the importance of threat evolution understanding aided by threat categorisation. Moreover, the interpretation of results explains how the TTC and ATC models facilitate resource allocation, and how the proposed features and our model analysis lead to accurate threat forecasting. The interpretation of results also highlights how the analysis of saliency maps and attention scores enhances model interpretability.

In chapter 6, we additionally summarised our efforts to address the identified research gaps by proposing the proactive approach to cyber trend prediction and the shifting from human expertise to automated ML techniques. This is followed by a discussion for the implications of our research for advancing proactive cyber security, highlighting avenues for future exploration and refinement. This includes extending the forecast horizon and enhancing model's explainability through attention mechanisms and novel visualisation methods. The discussion also offers practical applications for industry and government stakeholders, enabling them to enhance their cyber security posture and effectively mitigate cyber threats by leveraging the proposed proactive approach and advanced models. Finally, chapter 6 re-emphasises the limitations of this work including those related to model's performance, evaluation of the proposed methods, and data granularity and scope.

7.3 Future Work

7.3.1 Data Expansion

Expanding the dataset is essential for enhancing the robustness and efficacy of cyber trend forecasting models. One avenue for future research involves enriching the dataset with a broader range of fine-grained attack types. For instance, DDoS attacks can be divided into volumetric attacks, application layer attacks, and protocol-based attacks [167]. Similarly, insider threats can be classified based on the motives of the insiders such as disgruntled employees, malicious insiders, or unintentional insiders [168]. By including a more diverse array of attack categories, researchers can capture a more comprehensive spectrum of potential threats, enabling the model to make more nuanced and specific predictions. This approach facilitates the development of tailored security measures that are specifically designed to mitigate the unique characteristics of each type of cyber threat.

Furthermore, the dataset can be augmented by incorporating additional data samples, thereby enriching the diversity and depth of the available information. Increasing the size of the dataset allows for a more comprehensive exploration of the underlying patterns and trends in cyber security data. By including a larger number of samples, researchers can capture a more representative sample of the cyber security landscape, reducing the risk of sampling bias and improving the generalisability of the model. Overall, this should lead to an improvement in the model's forecasting accuracy.

In addition, augmenting the dataset with informative features can further enhance the performance of ML models in cyber trend forecasting. In our work, we have demonstrated that features, such as the count of tweets about wars and conflicts, the frequency of attack mentions in literature, and the count of public holidays, provide valuable context to cyber threat instances. Future research can focus on studying additional features like network traffic patterns, economic indicators, and social media sentiment analysis [169]. This could offer valuable insights into cyber threats and relevant technologies. Incorporating these diverse features can enrich the predictive capabilities of the model and enable more accurate forecasting of security trends.

7.3.2 Methods Evaluation and Automation

Evaluating the proposed methodologies is also crucial for a successful forecasting process. Therefore, we suggest the establishment of a systematic approach for the evaluation of the E-GPT algorithm, used for the extraction of the PATs (*i.e*, the graph construction). This can be done by interviewing and consulting security experts, in order to validate the algorithm's outputs, and possibly to contribute to the adjustment of the graph connectivity (semi-automated approach). This may require careful design of expert panels to tune inter-rater variance of evaluations. Alternatively, and perhaps more promisingly, a fully automated approach is feasible through the utilisation of the graph learning layer [27]. This layer adaptively learns and quantifies the relationships between the threats and PATs. The obtained relationship values can then be leveraged, in conjunction with the predicted gaps, to prioritise the PATs effectively. Additionally, the positioning of the PATs on the ATC can also be automated by computationally measuring the slope of the PAT curves and placing each PAT on the cycle accordingly. Overall, the automation of these phases maximises the machine's involvement in the process, thereby reducing the reliance on human bias and subjectivity.

Moreover, addressing the explainability limitations described in section 5.4.3 involves a multi-faceted approach. It includes developing standardised evaluation metrics tailored for assessing model explainability, conducting thorough validation studies to ensure that explainability aligns with actual model behaviour across diverse scenarios, and advancing techniques that bridge the gap between interpretability and causality. These efforts will contribute to enhancing the reliability and trustworthiness of explainable models in cyber trend forecasting, ultimately empowering organisations to make more informed decisions and bolster their defences effectively.

7.3.3 Correlational Explainability via Attention Mechanism

Concerning correlational explainability, a promising direction is to augment the proposed B-MTGNN model with an advanced attention mechanism, focusing specifically on improving the model's explainability. MTGNN has demonstrated proficiency in capturing spatial and temporal dependencies in multivariate time series data. By infusing it with an attention mechanism, the goal is to not only maintain its predictive prowess but also to significantly enhance its transparency and interpretability, which are crucial in unbiased forecasting of cyber threats and PATs. The proposed integration targets the development of a model that is both accurate and interpretable as follows:

- Development of the enhanced model: The aim is to create a hybrid B-MTGNN model that incorporates an attention mechanism. This mechanism is tailored to highlight the model's focus on specific temporal sequences and spatial nodes, making the internal decision-making process more transparent.
- Focus on explainability: The attention mechanism will serve as a tool for visualising and understanding the model's predictions. It will allow users to see which parts of the data were most influential in the model's decision-making process, thereby demystifying the predictions.
- Optimisation and evaluation: The model should undergo extensive training and evaluation. Emphasis should be placed on how well the attention mechanism elucidates the model's decision-making, in addition to its predictive accuracy.
- Case studies for demonstrating explainability: The objective is to apply the model to the time series features extracted from diverse big data sources and present case studies demonstrating how the attention mechanism provides insights into the model's predictions, thereby enhancing trust and reliability in the model's outcomes.

To further elucidate the model's decision-making process, it is possible to integrate two additional visualisation tools: Attention Flow Diagrams [170] and Attention Distribution Plots [171]. Attention Flow Diagrams graphically depict the model's shifting focus over time across different data nodes, providing an intuitive understanding of temporal dynamics in the data. Simultaneously, Attention Distribution Plots offer a statistical perspective of the attention allocation across various features, clarifying which aspects of the data are most influential in the model's predictions. These tools are designed to demystify the model's internal workings, making it more accessible to users with varying levels of technical expertise and enhancing its diagnostic capabilities.

The integration of these visualisation tools into the B-MTGNN model with attention mechanism will make significant strides in the field of explainable AI. By offering clear, visual representations of how the model processes and prioritises information, the aim is to bridge the gap between complex ML techniques and their practical, interpretable applications. This advancement promises to not only heighten trust and transparency in AI-driven decision systems but also to facilitate a broader understanding and adoption of these technologies across various critical sectors, paving the way for more informed and transparent decision-making processes.

7.3.4 Causal Explainability

Future work could also consider integrating causal explainability techniques into cyber trend forecasting methodologies. Causal explainability goes beyond merely correlating input variables with model predictions and aims to uncover the causal relationships between variables, providing deeper insights into the underlying mechanisms driving cyber threat trends.

One potential approach for achieving causal explainability is through the use of causal inference techniques, such as causal Bayesian networks [172] or structural equation modelling [173]. These methods allow researchers to model and infer causal relationships between variables by explicitly representing the causal dependencies among them. By incorporating causal inference techniques into cyber trend forecasting models, researchers can identify not only which variables are correlated with cyber security trends but also understand the causal pathways through which they influence each other.

Another avenue for future work is the development of counterfactual analysis techniques [65]. Counterfactual analysis involves simulating hypothetical scenarios in which certain variables are modified or controlled to assess their impact on the predicted outcomes. By conducting counterfactual analyses, researchers can identify the causal factors driving cyber threat trends and evaluate the effectiveness of potential interventions or mitigation strategies. For example, researchers could simulate the impact of implementing specific security measures or policy changes on the likelihood of cyber-attacks occurring.

Overall, the integration of causal inference techniques with ML models, such as causal graphical models or causal inference frameworks, holds promise for enhancing the interpretability and reliability of cyber trend forecast. These models allow for the explicit representation of causal relationships within the data, enabling researchers to disentangle spurious correlations and identify true causal drivers of cyber security trends, leading to effective and informed decision-making process.

References

[1] S. Ghafur, S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin, “A retrospective impact analysis of the wannacry cyberattack on the nhs,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–7, 2019.

[2] M. Al-Asli and T. A. Ghaleb, “Review of signature-based techniques in antivirus products,” in *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–6, IEEE, 2019.

[3] S. K. Hassan and A. Ibrahim, “The role of artificial intelligence in cyber security and incident response,” *International Journal for Electronic Crime Investigation*, vol. 7, no. 2, 2023.

[4] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, “A comparative study of anomaly detection schemes in network intrusion detection,” in *Proceedings of the 2003 SIAM international conference on data mining*, pp. 25–36, SIAM, 2003.

[5] “Anticipating cyber attacks: There’s no abbottabad in cyber space.” *Infosecurity Magazine* https://www.infosecurity-magazine.com/white-papers/anticipating-cyber-attacks, 2015.

[6] Z. Almahmoud, P. D. Yoo, O. Alhussein, I. Farhat, and E. Damiani, “A holistic and proactive approach to forecasting cyber threats,” *Scientific Reports*, vol. 13, no. 1, p. 8049, 2023.

[7] O. Kebir, I. Nouaouri, L. Rejeb, and L. B. Said, “Atipreta: An analytical model for time–dependent prediction of terrorist attacks,” *International Journal of Applied Mathematics and Computer Science*, vol. 32, no. 3, pp. 495–510, 2022.

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, “Highly accurate protein structure prediction with alphafold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[9] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021.

[10] P. Passeri, "Hackmageddon data set." *Hackmageddon* https://www.hackmageddon.com, 2022.

[11] J. Bharadiya, "Machine learning in cybersecurity: Techniques and challenges," *European Journal of Technology*, vol. 7, no. 2, pp. 1–14, 2023.

[12] T. Kuwahara, "Technology forecasting activities in japan," *Technological Forecasting and Social Change*, vol. 60, no. 1, pp. 5–14, 1999.

[13] J. R. S. Alrzini and D. Pennington, "A review of polymorphic malware detection techniques," *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 12, pp. 1238–1247, 2020.

[14] G. Werner, S. Yang, and K. McConky, "Time series forecasting of cyber attack intensity," in *Proceedings of the 12th Annual Conference on cyber and information security research*, pp. 1–3, 2017.

[15] P. Goyal, K. Hossain, A. Deb, N. Tavabi, N. Bartley, A. Abeliuk, E. Ferrara, and K. Lerman, "Discovering signals from web sources to predict cyber attacks," *arXiv preprint arXiv:1806.03342*, 2018.

[16] G. Werner, S. Yang, and K. McConky, "Leveraging intra-day temporal variations to predict daily cyberattack activity," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 58–63, IEEE, 2018.

[17] A. Okutan, S. J. Yang, K. McConky, and G. Werner, "Capture: cyberattack forecasting using non-stationary features with time lags," in *2019 IEEE Conference on Communications and Network Security (CNS)*, pp. 205–213, IEEE, 2019.

[18] G. Stephens, "Cybercrime in the year 2025," *Futurist*, vol. 42, no. 4, p. 32, 2008.

[19] A. Adamov and A. Carlsson, “The state of ransomware. trends and mitigation techniques.,” in *EWDTS*, pp. 1–8, 2017.

[20] A. Shoufan and E. Damiani, “On inter-rater reliability of information security experts,” *Journal of information security and applications*, vol. 37, pp. 101–111, 2017.

[21] Y.-O. Cha and Y. Hao, “The dawn of metamaterial engineering predicted via hyperdimensional keyword pool and memory learning,” *Advanced Optical Materials*, vol. 10, no. 8, p. 2102444, 2022.

[22] “Elsevier research products apis.” *Elsevier Developer Portal* https://dev.elsevier.com, 2022.

[23] “Twitter api v2.” *Developer Platform* https://developer.twitter.com/en/docs/twitter-api, 2022.

[24] “holidays 0.15.” *PyPI Â· The Python Package Index* https://pypi.org/project/holidays/, 2022.

[25] M. Visser, N. J. van Eck, and L. Waltman, “Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic,” *Quantitative Science Studies*, vol. 2, no. 1, pp. 20–41, 2021.

[26] “Gpt-3 model.” *OpenAI Platform* https://platform.openai.com/docs/models/gpt-3, 2023.

[27] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, “Connecting the dots: Multivariate time series forecasting with graph neural networks,” in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 753–763, 2020.

[28] J. A. Lewis, “Economic impact of cybercrime: No slowing down,” *Center for Strategic and International Studies*, 2018.

[29] J. Kose, “Cyber warfare: An era of nation-state actors and global corporate espionage,” *ISSA Journal*, vol. 19, no. 4, 2021.

[30] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2015.

[31] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, pp. 305–316, IEEE, 2010.

[32] C. L. A. Navarro, J. A. Damen, T. Takada, S. W. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. Moons, *et al.*, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review," *bmj*, vol. 375, 2021.

[33] H.-y. S. Tsai, M. Jiang, S. Alhabash, R. LaRose, N. J. Rifon, and S. R. Cotten, "Understanding online safety behaviors: A protection motivation theory perspective," *Computers & Security*, vol. 59, pp. 138–150, 2016.

[34] P. Norman, H. Boer, E. R. Seydel, and B. Mullan, "Protection motivation theory," *Predicting and changing health behaviour: Research and practice with social cognition models*, vol. 3, pp. 70–106, 2015.

[35] R. W. Rogers, "A protection motivation theory of fear appeals and attitude change1," *The journal of psychology*, vol. 91, no. 1, pp. 93–114, 1975.

[36] A. Loukaka and S. Rahman, "Discovering new cyber protection approaches from a security professional prospective," *International Journal of Computer Networks & Communications (IJCNC) Vol*, vol. 9, 2017.

[37] F. Oggier and M. J. Mihaljević, "An information-theoretic security evaluation of a class of randomized encryption schemes," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 158–168, 2013.

[38] R. Vinayakumar, K. Soman, and P. Poornachandran, "Evaluation of recurrent neural network and its variants for intrusion detection system (ids)," *International Journal of Information System Modeling and Design (IJISMD)*, vol. 8, no. 3, pp. 43–63, 2017.

[39] R. Sharma and S. Thapa, "Cybersecurity awareness, education, and behavioral change: strategies for promoting secure online practices among end users," *Eigenpub Review of Science and Technology*, vol. 7, no. 1, pp. 224–238, 2023.

[40] J. D. Cryer, *Time series analysis*, vol. 286. Duxbury Press Boston, 1986.

[41] C. OâReilly, K. Moessner, and M. Nati, "Univariate and multivariate time series manifold learning," *Knowledge-Based Systems*, vol. 133, pp. 1–16, 2017.

[42] M. E. Athanasopoulou, J. Deveikyte, A. Mosca, I. Peri, and A. Provetti, "A hybrid model for forecasting short-term electricity demand," in *Proceedings of the Second ACM International Conference on AI in Finance*, pp. 1–6, 2021.

[43] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 922–929, 2019.

[44] E. Gibney *et al.*, "Where is russia's cyberwar? researchers decipher its strategy," *Nature*, vol. 603, no. 7903, pp. 775–776, 2022.

[45] B. L. Kozyrskiy, D. Milios, and M. Filippone, "Imposing functional priors on bayesian neural networks.," in *ICPRAM*, pp. 450–457, 2023.

[46] B.-H. Tran, S. Rossi, D. Milios, and M. Filippone, "All you need is a good functional prior for bayesian deep learning," *Journal of Machine Learning Research*, vol. 23, no. 74, pp. 1–56, 2022.

[47] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *arXiv preprint arXiv:1506.02142v6*, 2016.

[48] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu, "Cloudy with a chance of breach: Forecasting cyber security incidents," in *24th USENIX Security Symposium (USENIX Security 15)*, pp. 1009–1024, 2015.

[49] M. Husák and J. Kašpar, "Aida framework: real-time correlation and prediction of intrusion detection alerts," in *Proceedings of the 14th international conference*

on availability, reliability and security, pp. 1–8, 2019.

[50] M. Husák, V. Bartoš, P. Sokol, and A. Gajdoš, "Predictive methods in cyber defense: Current experience and research challenges," *Future Generation Computer Systems*, vol. 115, pp. 517–530, 2021.

[51] L. Bilge, Y. Han, and M. Dell'Amico, "Riskteller: Predicting the risk of cyber incidents," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 1299–1311, 2017.

[52] A. Salih, S. T. Zeebaree, S. Ameen, A. Alkhyyat, and H. M. Shukur, "A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection," in *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)*, pp. 61–66, IEEE, 2021.

[53] K. Ren, Y. Zeng, Z. Cao, and Y. Zhang, "Id-rdrl: a deep reinforcement learning-based feature selection intrusion detection model," *Scientific Reports*, vol. 12, no. 1, pp. 1–18, 2022.

[54] X. Liu and J. Liu, "Malicious traffic detection combined deep neural network with hierarchical attention mechanism," *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[55] B. Munkhdorj and S. Yuji, "Cyber attack prediction using social data analysis," *Journal of High Speed Networks*, vol. 23, no. 2, pp. 109–135, 2017.

[56] X. Qin and W. Lee, "Attack plan recognition and prediction using causal networks," in *20th Annual Computer Security Applications Conference*, pp. 370–379, IEEE, 2004.

[57] J. Malik, A. Akhunzada, I. Bibi, M. Imran, A. Musaddiq, and S. W. Kim, "Hybrid deep learning: An efficient reconnaissance and surveillance detection mechanism in sdn," *IEEE Access*, vol. 8, pp. 134695–134706, 2020.

[58] J. GRAY, "Futuristic forecast of tools and technologies," *Communications of the ACM*, vol. 44, no. 3, p. 29, 2001.

[59] G. Adomavicius, J. Bockstedt, A. Gupta, and R. J. Kauffman, “Understanding evolution in technology ecosystems,” *Communications of the ACM*, vol. 51, no. 10, pp. 117–122, 2008.

[60] X. Li, Q. Xie, T. Daim, and L. Huang, “Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology,” *Technological Forecasting and Social Change*, vol. 146, pp. 432–449, 2019.

[61] R. Chandra and S. Collis, “Digital agriculture for small-scale producers: challenges and opportunities,” *Communications of the ACM*, vol. 64, no. 12, pp. 75–84, 2021.

[62] “Gartner Website.” https://www.gartner.co.uk/en. Accessed: October 17, 2023.

[63] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” *arXiv preprint arXiv:1709.04875*, 2017.

[64] D. Cao, Y. Wang, J. Duan, C. Zhang, X. Zhu, C. Huang, Y. Tong, B. Xu, J. Bai, J. Tong, *et al.*, “Spectral temporal graph neural network for multivariate time-series forecasting,” *Advances in neural information processing systems*, vol. 33, pp. 17766–17778, 2020.

[65] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications,” *Information Fusion*, vol. 81, pp. 59–83, 2022.

[66] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, “On image classification: Correlation vs causality,” *arXiv preprint arXiv:1708.06656*, 2017.

[67] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, “Statistical interpretation of machine learning-based feature importance scores for biomarker discovery,” *Bioinformatics*, vol. 28, no. 13, pp. 1766–1774, 2012.

[68] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106584, 2022.

[69] G. Szepannek and K. Lübke, "How much do we see? on the explainability of partial dependence plots for credit risk scoring," *Argumenta Oeconomica*, no. 1 (50), 2023.

[70] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, "Causality learning: A new perspective for interpretable machine learning," *arXiv preprint arXiv:2006.16789*, 2020.

[71] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry," *Accounts of Materials Research*, vol. 3, no. 6, pp. 597–607, 2022.

[72] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.

[73] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang, "Xinsight: explainable data analysis through the lens of causality," *Proceedings of the ACM on Management of Data*, vol. 1, no. 2, pp. 1–27, 2023.

[74] J. Rao, S. Zheng, Y. Lu, and Y. Yang, "Quantitative evaluation of explainable graph neural networks for molecular property prediction," *Patterns*, vol. 3, no. 12, 2022.

[75] F. Wu, Y. Long, C. Zhang, and B. Li, "Linkteller: Recovering private edges from graph neural networks via influence analysis," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2005–2024, IEEE, 2022.

[76] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE transactions on pattern analysis and machine intelli-*

gence, vol. 45, no. 5, pp. 5782–5799, 2022.

[77] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10772–10781, 2019.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[79] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[80] J. Henry and S. Habib, "Bridging the gap: Making ai understandable with explainable artificial intelligence," tech. rep., EasyChair, 2024.

[81] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, *et al.*, "A survey of data-driven and knowledge-aware explainable ai," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 29–49, 2020.

[82] O. Kammouh and G. P. Cimellaro, "Cyber threat on critical infrastructure: A growing concern for decision makers," in *Routledge handbook of sustainable and resilient infrastructure*, pp. 359–374, Routledge, 2018.

[83] R. McCusker, "Transnational organised cyber crime: distinguishing threat from reality," in *Transnational Financial Crime*, pp. 415–432, Routledge, 2017.

[84] V. S. Sathyanarayan, P. Kohli, and B. Bruhadeshwar, "Signature generation and detection of malware families," in *Information Security and Privacy: 13th Australasian Conference, ACISP 2008, Wollongong, Australia, July 7-9, 2008. Proceedings 13*, pp. 336–349, Springer, 2008.

[85] X. Fang, M. Xu, S. Xu, and P. Zhao, "A deep learning framework for predicting cyber attacks rates," *EURASIP Journal on Information security*, vol. 2019, pp. 1–11, 2019.

[86] C. Whyte and B. Mazanec, *Understanding cyber-warfare: Politics, policy and strategy.* Routledge, 2023.

[87] I. Yaqoob, E. Ahmed, M. H. ur Rehman, A. I. A. Ahmed, M. A. Al-garadi, M. Imran, and M. Guizani, "The rise of ransomware and emerging security challenges in the internet of things," *Computer Networks*, vol. 129, pp. 444–458, 2017.

[88] M. A. Mos and M. M. Chowdhury, "The growing influence of ransomware," in *2020 IEEE International Conference on Electro Information Technology (EIT)*, pp. 643–647, IEEE, 2020.

[89] S. Goel, "Cyberwarfare: connecting the dots in cyber intelligence," *Communications of the ACM*, vol. 54, no. 8, pp. 132–140, 2011.

[90] S. Svetoka *et al.*, *Social media as a tool of hybrid warfare.* NATO Strategic Communications Centre of Excellence, 2016.

[91] C. Point, "The season of cyber vulnerability: How the holidays become a hacker's playground–global security mag online," 2024.

[92] S. Wass, S. Pournouri, and G. Ibbotson, "Prediction of cyber attacks during coronavirus pandemic by classification techniques and open source intelligence," in *Cybersecurity, Privacy and Freedom Protection in the Connected World: Proceedings of the 13th International Conference on Global Security, Safety and Sustainability, London, January 2021*, pp. 67–100, Springer, 2021.

[93] A. Dalton, B. Dorr, L. Liang, and K. Hollingshead, "Improving cyber-attack predictions through information foraging," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4642–4647, IEEE, 2017.

[94] "2021 trends show increased globalized threat of ransomware." *Cybersecurity & Infrastructure Security Agency* https://www.cisa.gov/uscert/ncas/alerts/aa22-040a, 2022.

[95] N. Öztürk and S. Ayvaz, "Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136–

147, 2018.

[96] K. K. Lai, L. Yu, S. Wang, and W. Huang, "Hybridizing exponential smoothing and neural network for financial time series predication," in *International Conference on Computational Science*, pp. 493–500, Springer, 2006.

[97] B. Huang, Q. Ding, G. Sun, and H. Li, "Stock prediction based on bayesian-lstm," in *Proceedings of the 2018 10th international conference on machine learning and computing*, pp. 128–133, 2018.

[98] Y. Mae, W. Kumagai, and T. Kanamori, "Uncertainty propagation for dropout-based bayesian neural networks," *Neural Networks*, vol. 144, pp. 394–406, 2021.

[99] "Scopus preview." *Scopus* https://www.scopus.com/home.uri, 2022.

[100] P. Jia, H. Chen, L. Zhang, and D. Han, "Attention-lstm based prediction model for aircraft 4-d trajectory," *Scientific reports*, vol. 12, no. 15533, 2022.

[101] R. Chandra, S. Goyal, and R. Gupta, "Evaluation of deep learning models for multi-step ahead time series prediction," *IEEE Access*, vol. 9, pp. 83105–83123, 2021.

[102] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[103] A. Sagheer and M. Kotb, "Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.

[104] L. P. Swiler, T. L. Paez, and R. L. Mayes, "Epistemic uncertainty quantification tutorial," in *Proceedings of the 27th International Modal Analysis Conference*, 2009.

[105] F. Chollet, *Deep Learning with Python.* Manning Publications, 2 ed., 2017.

[106] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," *Advances in neural information processing systems*, vol. 30, 2017.

[107] K. Sedor, "The law of large numbers and its applications," *Lakehead University: Thunder Bay, ON, Canada*, 2015.

[108] T. Papamarkou, M. Skoularidou, K. Palla, L. Aitchison, J. Arbel, D. Dunson, M. Filippone, V. Fortuin, P. Hennig, A. Hubin, *et al.*, "Position paper: Bayesian deep learning in the age of large-scale ai," *arXiv preprint arXiv:2402.00809*, 2024.

[109] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012.

[110] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, V. A. Kamaev, *et al.*, "A survey of forecast error measures," *World applied sciences journal*, vol. 24, no. 24, pp. 171–176, 2013.

[111] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[112] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[113] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[114] Y. Shifferaw and S. Lemma, "Limitations of proof of stake algorithm in blockchain: A review," *Zede Journal*, vol. 39, no. 1, pp. 81–95, 2021.

[115] O. Dedehayir and M. Steinert, "The hype cycle model: A review and future directions," *Technological Forecasting and Social Change*, vol. 108, pp. 28–41, 2016.

[116] F. Abri, S. Siami-Namini, M. A. Khanghah, F. M. Soltani, and A. S. Namin, "Can machine/deep learning classifiers detect zero-day malware with high accuracy?," in *2019 IEEE international conference on big data (Big Data)*, pp. 3252–3259, IEEE, 2019.

[117] E. James and F. Rabbi, "Fortifying the iot landscape: Strategies to counter security risks in connected systems," *Tensorgate Journal of Sustainable Technology and Infrastructure for Developing Countries*, vol. 6, no. 1, pp. 32–46, 2023.

[118] S. Tedeschi, C. Emmanouilidis, J. Mehnen, and R. Roy, "A design approach to iot endpoint security for production machinery monitoring," *Sensors*, vol. 19, no. 10, p. 2355, 2019.

[119] S. Zobaed, F. Rabby, I. Hossain, E. Hossain, S. Hasan, A. Karim, and K. Md Hasib, "Deepfakes: Detecting forged and synthetic media content using machine learning," *Artificial Intelligence in Cyber Security: Impact and Implications: Security Challenges, Technical and Ethical Issues, Forensic Investigative Challenges*, pp. 177–201, 2021.

[120] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *International journal of computer vision*, vol. 130, no. 7, pp. 1678–1734, 2022.

[121] A. Paudice, L. Muñoz-González, A. Gyorgy, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *arXiv preprint arXiv:1802.03041*, 2018.

[122] S. Venkatesan, H. Sikka, R. Izmailov, R. Chadha, A. Oprea, and M. J. De Lucia, "Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems," in *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)*, pp. 874–879, IEEE, 2021.

[123] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," *Computers & security*, vol. 72, pp. 212–233, 2018.

[124] M. Imthiyas, S. Wani, R. A. A. Abdulghafor, A. A. Ibrahim, and A. H. Mohammad, "Ddos mitigation: A review of content delivery network and its ddos defence techniques," *International Journal on Perceptive and Cognitive Computing*, vol. 6, no. 2, pp. 67–76, 2020.

[125] F. Wang, X. Hu, and J. Su, "Unfair rate limiting for ddos mitigation based on traffic increasing patterns," in *2012 IEEE 14th International Conference on Communication Technology*, pp. 733–738, IEEE, 2012.

[126] Z. Chen, B. Li, S. Wu, K. Jiang, S. Ding, and W. Zhang, "Content-based unrestricted adversarial attack," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[127] "Global ransomware threat expected to rise with ai." *National Cyber Security Centre* https://www.ncsc.gov.uk/news/global-ransomware-threat-expected-to-rise-with-ai, 2024.

[128] "Fcdo annual report and accounts 2022 to 2023." *UK Foreign, Commonwealth Development Office*, https://www.gov.uk/government/publications/fcdo-annual-report-and-accounts-2022-to-2023, 2023.

[129] N. K. Samia, *Global Cyber Attack Forecast using AI Techniques*. PhD thesis, The University of Western Ontario (Canada), 2023.

[130] M. Kumar, S. S. Darshan, V. Yarlagadda, *et al.*, "Introduction to the cybersecurity landscape," in *Malware Analysis and Intrusion Detection in Cyber-Physical Systems*, pp. 1–21, IGI Global, 2023.

[131] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631–636, 2003.

[132] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *Journal of Big Data*, vol. 11, no. 1, p. 18, 2024.

[133] Y. Zeng, H. Qiu, G. Memmi, and M. Qiu, "A data augmentation-based defense method against adversarial attacks in neural networks," in *Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II 20*, pp. 274–289, Springer, 2020.

[134] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: an overview," in *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, pp. 557–566, Springer, 2021.

[135] C. Wueest, "The continued rise of ddos attacks," *White Paper: Security Response, Symantec Corporation*, 2014.

[136] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Computers & Security*, vol. 104, p. 102221, 2021.

[137] M. National Academies of Sciences, Engineering *et al.*, *Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies: Proceedings of a Workshop.* National Academies Press, 2019.

[138] R. Shao, T. Wu, and Z. Liu, "Detecting and recovering sequential deepfake manipulation," in *European Conference on Computer Vision*, pp. 712–728, Springer, 2022.

[139] M. Singh, B. Mehtre, and S. Sangeetha, "Insider threat detection based on user behaviour analysis," in *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30-31, 2020, Proceedings, Part II 2*, pp. 559–574, Springer, 2020.

[140] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[141] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[142] S. B. Taieb, A. Sorjamaa, and G. Bontempi, "Multiple-output modeling for multi-step-ahead time series forecasting," *Neurocomputing*, vol. 73, no. 10-12, pp. 1950–1957, 2010.

[143] J. Kim, G. Lee, S. Lee, and C. Lee, "Towards expert–machine collaborations for technology valuation: An interpretable machine learning approach," *Technological Forecasting and Social Change*, vol. 183, p. 121940, 2022.

[144] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

[145] A. Kosmarski, "Blockchain adoption in academia: Promises and challenges," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 4, p. 117, 2020.

[146] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE access*, vol. 8, pp. 222310–222354, 2020.

[147] Y. K. Dwivedi, A. Sharma, N. P. Rana, M. Giannakis, P. Goel, and V. Dutot, "Evolution of artificial intelligence research in technological forecasting and social change: Research topics, trends, and future directions," *Technological Forecasting and Social Change*, vol. 192, p. 122579, 2023.

[148] G. N. Reddy and G. Reddy, "A study of cyber security challenges and its emerging trends on latest technologies," *arXiv preprint arXiv:1402.1842*, 2014.

[149] D. D. Thomakos and J. B. Guerard Jr, "Naive, arima, nonparametric, transfer function and var models: A comparison of forecasting performance," *International Journal of Forecasting*, vol. 20, no. 1, pp. 53–67, 2004.

[150] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[151] X. Song, Y. Wu, and C. Zhang, "Tstnet: a sequence to sequence transformer network for spatial-temporal traffic prediction," in *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artifi-*

cial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30, pp. 343–354, Springer, 2021.

[152] S. Hernández, D. Vergara, M. Valdenegro-Toro, and F. Jorquera, "Improving predictive uncertainty estimation using dropout–hamiltonian monte carlo," *Soft Computing*, vol. 24, no. 6, pp. 4307–4322, 2020.

[153] E. Kim, K. Kim, D. Shin, B. Jin, and H. Kim, "Cytime: Cyber threat intelligence management framework for automatically generating security rules," in *Proceedings of the 13th International Conference on Future Internet Technologies*, pp. 1–5, 2018.

[154] B. S. Rawal, S. Liang, A. Loukili, and Q. Duan, "Anticipatory cyber security research: An ultimate technique for the first-move advantage.," *TEM Journal*, vol. 5, no. 1, 2016.

[155] Z. Han, D. Niyato, W. Saad, T. BaÅar, and A. HjÃ¸rungnes, "Game theory in wireless and communication networks: theory, models, and applications," 2012.

[156] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and TrendsÂ® in Theoretical Computer Science*, vol. 9, no. 3â4, pp. 211–407, 2014.

[157] T. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[158] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[159] V. Costan and S. Devadas, "Intel sgx explained," tech. rep., IACR Cryptology ePrint Archive, 2016.

[160] M. Husák, J. Komárková, E. Bou-Harb, and P. Čeleda, "Survey of attack projection, prediction, and forecasting in cyber security," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 640–660, 2018.

[161] Sudhakar and S. Kumar, "An emerging threat fileless malware: a survey and research challenges," *Cybersecurity*, vol. 3, no. 1, p. 1, 2020.

[162] L. Tong, Z. Chen, J. Ni, W. Cheng, D. Song, H. Chen, and Y. Vorobeychik, "Facesec: A fine-grained robustness evaluation framework for face recognition systems," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13254–13263, 2021.

[163] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[164] D. Bilika, N. Michopoulou, E. Alepis, and C. Patsakis, "Hello me, meet the real me: Audio deepfake attacks on voice assistants," *arXiv preprint arXiv:2302.10328*, 2023.

[165] C. R. Gerstner and H. Farid, "Detecting real-time deep-fake videos using active illumination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 53–60, 2022.

[166] G. Mezzour, K. M. Carley, and L. R. Carley, "An empirical study of global malware encounters," in *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, pp. 1–11, 2015.

[167] B. B. Gupta and A. Dahiya, *Distributed Denial of Service (DDoS) Attacks: Classification, Attacks, Challenges and Countermeasures.* CRC press, 2021.

[168] I. Homoliak, F. Toffalini, J. Guarnizo, Y. Elovici, and M. Ochoa, "Insight into insiders and it: A survey of insider threat taxonomies, analysis, modeling, and countermeasures," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–40, 2019.

[169] L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using social media to predict the future: a systematic literature review," *arXiv preprint arXiv:1706.06134*, 2017.

[170] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[171] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[172] A. De Waal and J. W. Joubert, "Explainable bayesian networks applied to transport vulnerability," *Expert Systems with Applications*, vol. 209, p. 118348, 2022.

[173] J. Li, T. Sawaragi, and Y. Horiguchi, "Introduce structural equation modelling to machine learning problems for building an explainable and persuasive model," *SICE Journal of Control, Measurement, and System Integration*, vol. 14, no. 2, pp. 67–79, 2021.