Journal of Computer Science

Review

Cybercrime in the AI Era: Definitions, Classification, Severity Assessment and the Role of AI in Combating Threats

Norah Sultan Alshahrani, Fahd Saleh Alotaibi, Khaled Hamed Alyoubi and Muhammad Sher Ramzan

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Article history
Received: 07-11-2024
Revised: 11-12-2024
Accepted: 18-12-2024

Corresponding Author:
Norah Sultan Alshahrani,
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
Email: nalshahrani0099@stu.kau.edu.sa

Abstract: The growing prevalence, complexity and financial impact of cybercrimes pose significant challenges for law enforcement agencies worldwide. Many organizations are utilizing different technologies such as cloud technology to improve speed, accuracy, and reliability. However, without proper security measures, the risk is still vital against cyberattacks. Cybercrime can lead to various negative outcomes, including theft, fraud, financial losses, a decline in customer trust, and emotional consequences like fear, anger, and insecurity. This research aims to explore multiple definitions of cybercrime and identify the most effective categories for assessing its severity, focusing on key characteristics found in cybercrime descriptions. It also examines regulations designed to combat and prevent cybercrime, which Saudi Arabia and the UK are considering as Case studies. Additionally, the study explores the role of artificial intelligence, machine learning, deep learning, transformer models, and generative models in fighting cybercrime, especially their application in classification tasks. The research evaluates datasets used in previous studies, highlighting their features and providing insights into the nature and trends of cybercrime. The findings demonstrate how transformer models and generative AI approaches, such as BERT and GPT, have driven significant advancements in natural language processing tasks, improving cybercrime classification and severity assessment. Furthermore, the review underscores the importance of detailed datasets with case descriptions, demographic information, and clear labels, offering valuable insights into prevalent cybercrime methods and trends. It offers actionable recommendations for future research, emphasizing the need for interdisciplinary collaboration, robust datasets, and innovative AI approaches to address the evolving landscape of cybercrime.

Keywords: Cybercrime, Severity, Classification, Artificial Intelligence, Transformer Models

Introduction

Cybercrime has become a significant challenge for law enforcement agencies worldwide, as cyberattacks continue to grow in number, complexity, and cost. In the past two years, 415 million adults across 10 countries (Australia, Brazil, France, Germany, India, Italy, Japan, New Zealand, the UK and the US) fell victim to cybercrime, according to a 2022 Norton survey (Norton *et al*., 2015). Many organizations are migrating their workloads to the cloud to enhance speed, accuracy, and reliability; however, neglecting adequate security measures exposes them to cyberattacks (Andleeb *et al*., 2019). Cybercrime can be categorized into four aspects: Communication media, target devices, attack methodology, and countermeasures (Basit *et al*., 2021). It can result in theft, fraud, financial losses, and a decline in consumer confidence. Additionally, cyberattacks can provoke emotional responses such as fear, anger, and insecurity. Understanding the complexities of cybercrime is essential for developing effective prevention and mitigation strategies. Various Artificial Intelligence (AI) techniques, including Machine Learning (ML), Deep Learning (DL), Transformer models, and generative models, are employed to identify and analyze different forms of cybercrime. ML techniques such as Decision Trees (DTs), Random Forests (RFs), and Support Vector Machines (SVMs) enable computers to learn from data without explicit programming (Tsakalidis and Vergidis, 2019). DL, a subset of ML, is based on artificial neural networks

and belief networks, which analyze and classify inputs through layers, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks (Vaswani *et al*., 2017). The Transformer method, which uses two stages (encoder and decoder), processes large datasets and analyzes them iteratively (Vaswani *et al*., 2017), with Bidirectional Encoder Representations from Transformers (BERT) being a notable example. Generative AI, unlike the discriminative modeling used for data-driven decision-making (Feuerriegel *et al*., 2024), employs generative models such as GPT-4, Llama 3, and Copilot.

Advancements in AI, ML, DL, transformer models, and generative AI models have improved cybercrime classification, but there is a gap in the diversity and representativeness of the datasets used. Most existing datasets fail to account for region-specific cybercrime characteristics, such as cultural, legal, and technological variations. Additionally, they lack detailed descriptions of the emotional, social, and economic impacts of cybercrimes, which are crucial for developing robust classification models.

Advanced AI models like Transformers and GPT have shown promise in classification tasks, but their performance on diverse datasets is limited due to imbalanced or biased data. Comparative analysis is needed to assess their performance on diverse datasets and address these limitations in real-world scenarios. By bridging these gaps, future work can improve the reliability and applicability of AI techniques in combating cybercrime globally.

This study focuses on identifying the most effective categories for assessing the severity of cybercrime. It aims to pinpoint key features in cybercrime case descriptions that help determine the type of crime and assess its severity. The study provides a comprehensive overview of the cybercrime landscape, beginning with an examination of various definitions and classifications. It explores crucial aspects of cybercrime, such as offense descriptions, types of victims, and observed severity levels across different regions, with a particular focus on Saudi Arabia and the UK. A significant portion of the review is dedicated to the role of AI techniques such as ML, DL, transformer models, and generative models in addressing cybercrime. Additionally, a comparative analysis of AI techniques was conducted, assessing methods, datasets, and accuracies reported in prior studies, alongside their limitations. It discusses how these advanced methods can be used to identify cybercrime features and develop robust classification models. Special attention is given to Transformer models, which have shown promising results in classification tasks. The review synthesizes findings from numerous studies that have applied feature extraction techniques, ML, DL, transformer models, and generative models to classify cybercrime. This provides readers with a comprehensive understanding of the state-of-the-art in this rapidly evolving field. Lastly, the paper examines the datasets used in the reviewed studies, highlighting their parameters and the insights they offer into the nature and patterns of cybercrime. This analysis will help researchers and practitioners identify gaps in the current knowledge base and guide future research efforts.

Methodology

This review employs a structured approach to gain a comprehensive understanding of cybercrime, its classification, features, regulatory frameworks, and technological advancements in AI-based classification techniques. The investigation concentrates on recent research that was published between 2017 and 2024, with a particular emphasis on works that prioritize AI techniques, such as ML, DL, Transformer models, and generative approaches, for the classification of cybercrime.

A detailed examination of cybercrime across a variety of legal, technological, and socio-cultural contexts was facilitated by the analysis of studies from a variety of countries and cultures, thereby achieving a global perspective. It begins by examining the evolution of the concept across different eras and perspectives, focusing on early definitions and contemporary interpretations. The paper subsequently examines a variety of cybercrimes in a variety of disciplines, including academia, law enforcement, and cybersecurity. It also explores the features and characteristics of cybercrimes, analyzing studies from computer science, psychology, and law enforcement to comprehend their intricacies. The article subsequently analyzes the regulatory frameworks of various countries, with a particular emphasis on their methods for identifying and punishing cybercrime, as well as evaluating its severity.

Subsequently, it contrasts the advancement of AI technologies utilized in classification, the transition to deep learning techniques, transformer models, and generative models, and the evolution of AI-based classification techniques for cybercrime from 2017-2024. It also addresses the constraints that have been identified in these methodologies, including computational overhead, dataset biases, and scalability issues. The report ultimately concentrates on the influence of datasets on AI-based classification techniques and provides suggestions for enhancement. It assesses the extent, diversity, and representativeness of datasets utilized in reviewed studies.

Furthermore, it suggests strategies for developing datasets that are more comprehensive and balanced in order to enhance the generalizability and accuracy of classification. Finally, the paper summarizes findings from all stages to provide an integrated understanding of cybercrime, its classification and AI-based techniques.

Cybercrime Definitions

Cybercrime has become a significant challenge for law enforcement agencies worldwide, as these crimes continue to increase in number, sophistication, and cost. In the past 12 months, 415 million adults across 10 countries (Australia, Brazil, France, Germany, India, Italy, Japan, New Zealand, the UK, and the US) fell victim to cybercrime, according to a 2022 Norton survey (1428 AH, 2007).

Al-Khater *et al*. (2020) defines cybercrimes as criminal acts committed in cyberspace using electronic access and communication devices, intended to instill fear and anxiety in individuals or to cause damage, harm, or destruction to property (Al-Khater *et al*., 2020). In 2021 similarly Goni defines cybercrime as any illegal behavior where a computer or the internet serves as either the means, the target, or both (Goni, 2021).

However, obtaining a universally accepted and comprehensive definition of cybercrime is difficult, as highlighted in 2016 by Ibrahim (2016). This challenge arises from two key factors. First, there are numerous synonyms for cybercrime that convey similar meanings. Second, cybercrime spans a broad spectrum of offenses across multiple domains (Ibrahim, 2016).

Sarkar and Shukla (2023) offer a more detailed definition, describing cybercrime as illicit activities occurring in cyberspace that are considered unlawful within the relevant jurisdiction. These acts cause socioeconomic and psychological harm to affected individuals. Their definition hinges on five criteria related to the criminal consequences of cybercrime, including jurisdiction, the intended victim, the online nature of the activity, and distinguishing cybercrime from cyberattacks (Sarkar and Shukla, 2023).

In 2024 Wall's updated book (Wall, 2024) argues that the term “cybercrime” is inherently ambiguous, often referring to harmful virtual activities that are driven by emotions and sensationalized by media narratives (Wall, 2024). Wall emphasizes the cultural, scientific, legal, and political dimensions of cybercrime, pointing out that its complexity makes it difficult to gather and harmonize data on cybercrime across different countries (Wall, 2024).

Finally, in 2024 AllahRakha presents a comprehensive definition that considers motives, skills, intended harm, and types of criminal activity. This definition describes cybercrime as a broad spectrum of illegal activities involving computers, networks, and digital devices. While financial motives often drive cybercrime, it can also include politically, or personally motivated attacks aimed at damaging systems or compromising data. Perpetrators range from highly skilled organized groups to less experienced individuals, reflecting the diverse landscape of cybercriminal activity. This study adopts this definition to enhance understanding and prevention efforts (AllahRakha, 2024).

This research explores these different perspectives on cybercrime to provide clearer insights into its complex nature.

Cybercrime Categories

After conducting an extensive review of numerous articles on cybercrime, it is evident that cybercrimes can be categorized into various distinct types. Each publication provides a unique perspective on how these crimes are classified.

The most common classification is based on three categories proposed (Wall, 2024) which were later adopted by the European Commission in 2013 with a similar approach (Tsakalidis and Vergidis, 2019). The first category, computer integrity crimes, includes offenses specific to computers and information systems, such as Denial of Service (DoS) attacks and malware, as recognized by the European Commission. The second category, computer-assisted crimes (referred to as traditional offenses by the European Commission), involves crimes like identity theft and scams. The third category, computer content crimes, refers to content-related offenses, such as those involving pornography, as classified by the European Commission.

According to Ibrahim (2016), cybercrimes can be classified into three categories: Geopolitical cybercrime, which involves activities such as cyber espionage by government agents or state actors; psychosocial cybercrime, which is driven by psychological and emotional factors, such as cyberbullying; and socioeconomic cybercrime, which includes internet fraud, forgery and impersonation leading to financial losses (Ibrahim, 2016).

In 2020 Al-Khater *et al*. classifies cybercrimes into two main categories: Those that use computers and those that target them (Al-Khater *et al*., 2020). Cybercrimes facilitated by computers include activities such as child pornography, fraud, money laundering and cyberstalking. In contrast, cybercrimes that target computers involve hacking, phishing and website defacement (Al-Khater *et al*., 2020).

In their study Al-Khater *et al*. (2020), examined various forms of cybercrime targeting specific demographics, considering factors such as age and gender. This includes criminal activities like child pornography and cyberbullying, as well as database-related offenses such as SQL injection attacks. The study also addresses computer system threats, including phishing and DoS attacks. Additionally, the authors highlight the harmful effects of state-sponsored activities, such as cyber terrorism, cyber warfare, and cyber espionage.

In addition, Goni (2021) categorizes cybercrime into three types: Data crime, which involves attempts to harm personal data; network crime, which pertains to

unauthorized access to software and other systems; and related crime, which includes aiding or facilitating criminal activities (Goni, 2021). Goni specifically addresses cyber assaults targeting devices, systems, and networks, such as malware, phishing emails, hacking, virus propagation, carding, and similar methods (Goni, 2021).

In 2023, the study by Khyioon Abdalrdha *et al*. (2023) discusses various cyber threats that target similar groups, including malware, identity theft and online fraud.

In 2021 Basit *et al*. provide an in-depth analysis of cybercrime, focusing particularly on its impact on social media platforms (Basit *et al*., 2021). They highlight various forms of cybercrime, such as abuse within social settings, intelligent voice responses, collaboration within social networks, and deceptive social engineering. The authors also cover cybercrime targeting emails, websites, and smart device browsers, including email spoofing, malicious attachments, URL spoofing, website spoofing, spoofed mobile internet browsers, and the installation of malicious web content (Basit *et al*., 2021).

In 2019, Tsakalidis and Vergidis (2019) examined the surveillance of computer-related crimes, focusing on organizations such as EUROPOL, the FBI, the European Union Agency for Network and Information Security, and the US Department of Justice. They identified five distinct categories of cybercrimes based on a range of features and characteristics, including the offender, target, means of attack, victim, and resulting harm. Type A includes crimes that involve violations of the integrity, availability, and confidentiality of computers, systems, and data. Examples are data interference, illegal access, data espionage, illegal interception, system interference, and misuse of devices. Type B covers computer-related crimes such as computer-related forgery, computer-related fraud, and identity theft. Type C focuses on content-related crimes, including the dissemination of pornographic material, cyberbullying, racism and hate speech online, spam, and other related threats. Type D pertains to crimes that infringe upon copyright and related rights, including copyright and trademark offenses. Finally, Type E represents a mixture of offenses from the previous categories, encompassing complex tactics such as phishing, cyber laundering, cyber warfare, and the use of the internet for terrorism.

The study conducted in 2021 by Matveev *et al*. (2021) identifies four distinct classes of cybercriminal actions. The classification is based on the social context and the level of public danger associated with each action. The first class includes crimes that specifically target the confidentiality, integrity, and availability of data and computer systems. The second class focuses on crimes directly involving computers. The third class pertains to illegal activities related to material goods. Finally, the fourth class addresses violations of copyright and associated rights.

In 2022, Phillips *et al*. developed a comprehensive categorization system for cybercrime types, based on multiple taxonomies (Phillips *et al*., 2022). This classification framework is layered, beginning with numeric types 1, 2, and 3 and then expanding into two main categories: Cyber-dependent and cyber-enabled crimes. Within these categories, there are four interrelated classes: Crimes against machines, crimes using machines, crimes involving machines, and cyber-assisted crimes. Each of these classes includes several subclasses. For example, crimes against machines are further divided into three subclasses: Data and system assaults, attacks on individuals and organizations (such as illegal access and ransomware), and attacks targeting states and countries. Crimes using machines involve property theft or damage, including copyright infringement and spam. Crimes involving machines have four subclasses of violence: Interpersonal (such as cyberbullying and harassment), sexual, group-related (such as terrorism and attacks based on protected characteristics), and general violence. The cyber-assisted class involves crimes that utilize additional technologies. Additionally, the framework includes cross-category crimes, such as those related to deep web markets or cybercrime-as-a-service (Phillips *et al*., 2022).

The following tables summarize the most common categorizations and types of cybercrimes discussed throughout this review. Table (1) outlines various cybercrime categories, while Table (2) lists specific cybercrimes frequently mentioned in the literature.

Table 1: Common cybercrime categorizations

1	Computer integrity and availability crimes	8	Crimes targeting computers
2	Computer-assisted crimes	9	Data crime
3	Computer content crimes	10	Network crime
4	Geopolitical cybercrime	11	Related crime
5	Psychosocial cybercrime	12	Impact on Social Websites
6	Socioeconomic cybercrime	13	Cybercrime targeting emails and websites
7	Crimes using computers	14	Copyright violation crimes

Table 2: Common cyber crimes considered in the literature

1	Denial of service	11	Identity theft
2	Cyberstalking	12	Cyber laundering
3	Data breach	13	Cyber espionage
4	Cyber terrorism	14	Illegal gambling and online games
5	Malware	15	Computer-related forgery
6	Cyberbullying	16	Copyright-related offenses
7	Computer-related fraud	17	Misuse of devices
8	Cyber warfare	18	Spam
9	Phishing	19	Illegal access
10	Pornography	20	Ransomware

Cybercrime Features

Quantifying cybercrime is challenging due to the rapidly evolving nature of digital victimization and the lack of uniform criteria or measurement methods (Breen *et al*., 2022). Tsakalidis and Vergidis identified eight key features that can explain cybercrime (Tsakalidis and Vergidis, 2019). These features include a comprehensive description of the crime, an assessment of whether the identified offenses qualify as criminal acts, and the identification of those responsible. Additionally, it is essential to pinpoint the specific computer or network access violations involved in the crime. This is followed by identifying potential victims, such as individuals (including children and women), organizations, and government entities. An evaluation of the damage caused by the crime, including financial losses and privacy breaches, is also necessary. Finally, relevant policies, actions, and measures should be determined to address and prevent such crimes.

A computational method proposed in 2020 by Ch *et al*. involves categorizing cybercrime charges by extracting content-based features and applying ML techniques (Ch *et al*., 2020). This method considers variables such as the criminal, the extent of damage inflicted, breach of access, the year of the incident, and the affected individual.

Matveev *et al*. used social motivation and the level of danger associated with the offense as essential attributes to distinguish different types of cybercrime (Matveev *et al*. 2021).

In 2022 Breen *et al*. suggested a technique for assessing the scope and financial impact of cybercrime incidents by tracking the prevalence of scams across various platforms, including banking systems, email providers, and social networking sites. A significant challenge arises when an attack spans multiple platforms, starting on a social media platform and ending with a bank system to deceive users and steal their money. In such cases, the social media platform may fail to recognize the event as an attack (Breen *et al.* 2022).

In 2023 Sarkar and Shukla proposed five features to provide a comprehensive understanding of cybercrime: Effect, jurisdiction, intended target, methods, and the distinction between cybercrime and cyberattacks. They examined over 20 definitions from various organizations and perspectives. The “impact” feature refers to the consequences of cybercrime, including socioeconomic, psychosocial, and geopolitical effects. “Jurisdiction” pertains to the lawful authority over a specific geographical area. The “target” feature addresses whether the victim is an individual or an organization. The “means” feature concerns the actions carried out through cyberspace. Lastly, they differentiate between cybercrime and cyberattacks (Sarkar and Shukla, 2023).

To evaluate the type and severity of a crime, consider the following features:

1. Description of the incident
2. Who is committing the crime?
3. Is it a criminal action?
4. Type of violation
5. Who are the victims?
6. Level of public danger
7. The harm sustained
8. Assessing the scope and financial impact
9. Social motivation
10. level of danger associated with the offense
11. Distinguish between cybercrime and cyberattacks
12. Regional rules, activities, and procedures apply to this form of crime

Cybercrime Severity Levels

The Saudi Arabian government has enacted the Anti-Cyber Crime Law (1428 AH, 2007) to address cybercrimes and protect the security of digital information and networks within the country (Table 3). This legislation covers a broad range of cyber offenses, including hacking, unauthorized access to computer networks, distribution of malicious software, identity theft, internet-based fraud, and the facilitation of terrorist activities online. The law specifies what constitutes an offense, prescribes appropriate penalties based on the severity of the crime, and outlines legal measures to combat cybercrime. Penalties and severity levels for cybercrime are regulated starting with Article 3. Nevertheless, the primary goals of Articles 1 and 2 are to provide a comprehensive definition of cybercrime, to outline its underlying ideas, and to outline the specific situations in which an activity is considered a cybercrime. In addition, the primary aims of the legislation on cybercrime regulation and enforcement are laid forth in these introductory articles.

Additionally, the National Cyber Security Centre (NCSC) and UK law enforcement agencies use a categorization strategy to classify cyber incidents (Table 4). The Incident Management team within the NCSC is responsible for evaluating and classifying incidents based on their severity and potential impact on the UK. This approach helps prioritize resource allocation to manage the most significant cyber events affecting the nation.

Table 3: Saudi Arabian anti-cybercrime law

Article No.	Penalty	Offenses
3	Up to one year in prison and a fine not exceeding 500,000 riyals, or either penalty	Spying on intercepting or receiving data transmissions Unauthorized access with intent to threaten or blackmail hacking a website Invasion of privacy using a camera Defamation and damage using technology
4	Up to three years in prison and a fine not exceeding 2,000,000 riyals, or either penalty	Acquisition of movable property Illegal access to bank or credit data, funds, or services
5	Up to four years in prison and a fine not exceeding 3,000,000 riyals, or either penalty	Unauthorized access to private data Disruption of computer networks, programs, or data Obstructing access to services by any means
6	Up to five years in prison and a fine not exceeding 3,000,000 riyals, or either penalty	Production, preparation, transmission, or storage of content that harms public order, religious values, public morals, or privacy through an information network or computer Creation or publication of websites that promote or facilitate human trafficking. Preparation, publication, or promotion of material related to pornography or gambling that violates public morals Creation or publication of websites for trading, distributing, or facilitating narcotics and psychotropic drugs
7	Up to ten years in prison and a fine not exceeding 5,000,000 riyals, or either penalty	Establishing or publicizing a website to support terrorist organizations by facilitating communication, financing, promoting ideologies, publicizing methods for making incendiary devices or explosives, or other terrorist activities Unauthorized access to a website or information system to obtain data threatening national security or the state's economy

Table 4: Categorization of UK Cyber Incidents

Category	Consequences	Response
1 National cyber emergency	A cyberattack causing prolonged disruption of essential UK services, impacting national security with significant economic, social, or life-threatening consequences	Involves a swift, coordinated government response led by ministers and the Cabinet Office Briefing Room (COBR). The NCSC and law enforcement provide technical cooperation, with the NCSC advising and managing government outreach to victims and offering incident response expertise
2 Highly significant incident	Cyber harm with severe consequences for the central government, key UK services, a large segment of the population, or the UK economy	The NCSC leads the response, which may escalate to COBR if needed, and works closely with law enforcement, often the National Crime Agency (NCA)
3 Significant incident	A cyberattack with a substantial impact on a large corporation, local or wider government, or posing a considerable threat to central government or critical services in the UK	The NCSC takes the lead in organizing response actions and collaborates with law enforcement, including the NCA, as required
4 Substantial incident	A cyberattack that significantly affects a medium-sized firm or poses a substantial threat to a major business or local/wider government	The response is led by either the NCSC or law enforcement (NCA or Regional Organized Crime Unit, ROCU), depending on the nature of the incident
5 Moderate incident	Involves a cyberattack targeting a small organization or posing a significant risk to a medium-sized organization. It may also indicate early signs of cyber activity against a large organization or government	Law enforcement, usually the Regional Organized Crime Unit (ROCU) or local police, leads the response with input from the NCA as necessary
6 Localized Incident	A cyberattack targeting an individual or showing early signs of cyber activity against a small or medium-sized organization	The local police force leads the response, with input from the NCA as needed.

To determine cybercrime severity levels, begin by reviewing key features that help classify cybercrimes. These features include the nature of the offense, its impact on victims, the execution method, the target, and the motive behind the crime. Assess the relevance of each feature and its potential impact on determining severity. Next, categorize cybercrimes into distinct types using the identified features and established frameworks to ensure consistency. Refer to regional regulations and guidelines to understand the criteria for severity determination. Evaluate each cybercrime type based on these features and regulations, considering factors such as the extent of

harm, the scale of the attack, the value of affected assets, and the intent of the perpetrator. Apply a severity scale that aligns with the regulations of each country. The proposed approach involves using features outlined in Table (2) and cybercrime classifications from Table (1) to evaluate severity levels. The methodology integrates regional regulations to determine severity. Tables (3-4) present severity levels for each crime, based on punishment and damage, as observed in two different countries.

To facilitate this approach, various AI techniques will be employed. Initially, AI will be used to extract features from cybercrime cases. Following this, AI methods for text classification will be applied to categorize cybercrime cases into appropriate types and determine their severity levels. The next section will detail the specific AI techniques used for feature extraction and text classification to ensure accurate classification of cybercrime cases.

Artificial Intelligent Feature Extraction Techniques

Feature extraction begins with transforming a raw dataset into a set of useful features for further analysis and learning (Dara and Tumma, 2018; Wang *et al*., 2020c). Feature selection, on the other hand, involves removing unnecessary and redundant properties from a dataset to enhance the performance efficiency of AI algorithms (Odhiambo Omuya *et al*., 2021) and address the issue of excessive dimensionality in the feature space (Wang *et al*., 2020c). Table (5) provides a comprehensive list of AI-based feature extraction techniques, which can be applied to various domains, including cybercrime classification, to extract relevant features from data. Principal Component Analysis (PCA) is a sophisticated feature extraction technique (Suhaidi *et al*., 2021). PCA reduces data dimensionality by creating a lower-dimensional feature set from the original data. It requires selecting the appropriate number of principal components to represent the data effectively.

The Bag of Words (BoW) technique is essential for feature extraction, as it categorizes features and stores results in a vector based on term frequency rather than term position (Suhaidi *et al*., 2021). However, the BoW approach may overlook terms with minimal informative value due to its focus on frequently repeated words, which can reduce accuracy, especially with longer documents. The Term Frequency-Inverse Document Frequency (TF-IDF) technique addresses these limitations (Tabassum and Patil, 2020). TF-IDF evaluates the importance of a term within a document by combining Term Frequency (TF) and Inverse Document Frequency (IDF). TF calculates how often a word appears in a document, while IDF assesses the significance of the word across the document collection (Tabassum and Patil, 2020).

Named Entity Recognition (NER) is vital in information retrieval as it identifies and retrieves proper nouns such as individuals, countries, and institutions, creating document entity highlights (Tabassum and Patil, 2020). Additionally, feature extraction can be achieved by selecting n consecutive tokens from text or audio using n-grams (Das *et al*., 2023). During n-gram extraction, segments of size n move across the corpus, extracting successive words or characters. N-grams identify continuous composite features while excluding punctuation and stop words to reduce ambiguity (Suhaidi *et al*., 2021).

In DL, feature extraction involves machine learning algorithms with non-linear layers that extract and convert features. Each layer utilizes the output from the previous layer as its input (Dara and Tumma, 2018). Understanding different levels of representation involves comparing various layers of abstraction (Dara and Tumma, 2018). Word embeddings are crucial in DL models Word embeddings are continuous real-number (Wang *et al*., 2020a). Word embeddings can be either context-independent vectors that map words to a latent vector space using syntactic and semantic information.

Word embeddings can be either context-independent (Wang *et al*., 2020a) or static (Selva Birunda and Kanniga Devi, 2021). Context-independent embeddings, such as global vectors for word representation (GloVe), FastText, and Word2Vec, use shallow neural networks trained on generic text collections to determine word meanings without considering the context. GloVe, developed in 2014 by Pennington, Socher, and Manning, uses global word-occurrence data to understand semantic relationships between words, benefiting tasks like word similarity and analogy completion (Selva Birunda and Kanniga Devi, 2021).

Table 5: List of AI feature extraction techniques

AI feature extraction techniques	References
Principal Component Analysis (PCA)	Suhaidi *et al*. (2021)
Bag of Words (BoW)	Suhaidi *et al*. (2021); Wang *et al*. (2020a)
Named Entity Recognition (NER)	Tabassum and Patil (2020)
N-grams	Suhaidi *et al*. (2021); Das *et al*. (2023)
Term Frequency-Inverse Document Frequency (TF-IDF)	Tabassum and Patil (2020)
Word2Vec	Wang *et al*. (2020a)
FastTex	Bojanowski *et al*. (2017)
GloVe	Selva Birunda and Kanniga Devi (2021)
ELMo	Selva Birunda and Kanniga Devi (2021)
BERT	Wang *et al*. (2020a); Selva Birunda and Kanniga Devi (2021); Ganesh *et al*. (2021); Devlin (2018)

FastText, introduced by Facebook's AI Research team in 2016, enhances Word2Vec by representing each word as a collection of character n-grams. This approach captures morphological structures in more detail, improving the representation of semantic and syntactic relationships between words (Bojanowski *et al*., 2017). Word2Vec, a widely used method, analyzes semantic similarities by reducing features through a graph search approach that groups similar properties. It employs two model architectures: The continuous BoW model, which predicts words based on their context, and the skip-gram model, which predicts context words based on the main word (Wang *et al*., 2020a).

Alternatively, word embeddings can be context-dependent (Wang *et al*., 2020a), capturing the semantic representation of a word based on its surrounding context within a text. Contextualized word embeddings, such as those generated by ELMo, use a bidirectional language model trained on extensive text data to anticipate word representations considering both left and right context (Selva Birunda and Kanniga Devi, 2021). BERT, a cutting-edge language representation model, employs the transformer architecture to pretrain word embeddings using diverse text data. Unlike previous approaches, BERT uses a Masked Language Model (MLM) strategy during pretraining, where words in a sentence are randomly masked and the model learns to predict these masked words based on contextual clues (Ganesh *et al*., 2021). BERT also includes a Next Sentence Prediction (NSP) task (Wang *et al*., 2020a), which predicts whether two given sentences follow sequentially in the original text (Devlin, 2018). By pretraining on a broad text corpus, BERT captures nuanced contextual embeddings of words and sentences, which can be fine-tuned for specific tasks such as text classification, named entity recognition, and question answering (Selva Birunda and Kanniga Devi, 2021).

Role of Artificial Intelligence in Cybercrime Classificatione

Machine Learning Techniques

To handle various cybercrime scenarios, researchers have explored different data mining and ML classification approaches, including SVMs, naive Bayes, K-Nearest Neighbors (KNNs), K-means, logistic regression, association rule learning, DTs, and RFs. Logistic regression identifies the most suitable model to elucidate the relationship between dependent and independent variables by producing coefficients that estimate changes in the likelihood of an event (Matias *et al*., 2022; Prabakaran and Mitra, 2018). SVMs divide the feature space into two subspaces to classify new objects into distinct categories (Matias *et al*., 2022; Onyekpeze *et al*., 2021). Naive Bayes is an efficient classification model that provides probability distributions to achieve the best results (Prabakaran and Mitra, 2018; Matias *et al*., 2022). K-means is an unsupervised learning technique used for clustering unlabeled data into groups. This method is known for its speed, reliability, simplicity, and effective outcomes (Prabakaran and Mitra, 2018). KNNs are a classification method that predicts the label of a data point based on the labels of its nearest neighbors in the training data (Matias *et al*., 2022; Onyekpeze *et al*., 2021). Association rule learning employs weighted support and confidence to make predictions. The model updates the rule base with new data, which helps in predicting the class label based on established rules (Prabakaran and Mitra, 2018). RF is flexible for classification and regression. It builds numerous DTs during training and aggregates their results via bootstrapping. Handling missing data, minimizing overfitting, and identifying categorical variables are benefits (Prabakaran and Mitra, 2018; Onyekpeze *et al*., 2021; Venkatesan, 2023).

DTs classify data by recursively splitting it into subgroups based on decision rules. The algorithm evaluates each attribute and divides the tree based on its explanatory power, growing the tree top-down for predictions (Onyekpeze *et al*., 2021; Venkatesan, 2023).

In a 2017 study, Altaher (2017) proposed a phishing detection approach that combines two algorithms: SVM and KNNs. The KNN algorithm is used in the initial phase for its efficiency in handling noisy data, while the SVM is applied as the classification tool, resulting in an accuracy of 90.04%. This approach leverages the strengths of both algorithms SVM's reliability and KNN's clarity despite their limitations when used independently. However, KNN has a notable limitation in that it treats all features equally, which can reduce the effectiveness of cybercrime detection by failing to prioritize more significant or relevant variables (Al-Khater *et al*., 2020).

In a separate study, Lekha and Prakasam (2017) proposed a methodology for categorizing cybercrime in the financial industry using K-means clustering and influenced association classification with prediction tree J48, an advanced variant of DT C4.5. The K-means clustering technique, being an unsupervised learning algorithm, does not guarantee correct results as the correct answers are unknown (Al-Khater *et al*., 2020). Additionally, this research is limited to the investigation of financial crimes .

In (Andleeb *et al*., 2019), the authors utilized sentiment analysis with the Natural Language Toolkit for feature extraction and employed two classifiers, SVM and Bernoulli naive Bayes, to develop a system for detecting cyberbullying. The system was designed to analyze three categories of characteristics: Textual features, behavioral features, and demographic features, all extracted from XML files of conversations. This research is specifically focused on cyberbullying, aiming to identify incidents that meet the criteria for classification as such. However, it is important to note that the study addresses only one specific form of cybercrime, despite the broad range of cyberbullying manifestations.

In 2020, the paper (Ch *et al*., 2020) introduced an approach that combines naive Bayes, K-means, and TF-IDF vectors for classification, clustering, and feature extraction, respectively. The validation results demonstrated that this method correctly categorizes cybercrime offenses with 99% accuracy evaluated using a confusion matrix to measure the model's classification performance. However, the research relies on binary classification to determine the presence or absence of cybercrime and utilizes a dataset limited to incidents in India between 2012 and 2017, which may limit its relevance to worldwide cybercrime trends. Given the fast growth of cybercrime tools and regional legal definitions, this time and geographical constraint may limit the results' generalizability.

According to Pandey *et al*. (2021), a hybrid framework utilizing model stacking as an ensemble learning technique is being developed. This framework integrates four algorithms SVM, logistic regression, DT, and RF to construct a dual-level architecture. For feature extraction, the framework employs the BoW model, a popular natural language processing method. The ensemble learning technique aims to address issues of poor accuracy and provide high reliability in cybercrime classification. The adoption of various machine learning and data mining methods necessitates the creation of clear data representations, involving the organization and preparation of data to enable accurate processing by these algorithms.

In the research presented in Venkatesan (2023), an effective Intrusion Detection System (IDS) for cybercrime was developed using machine learning techniques, including DT, RF, and SVM. The system employs feature selection methods such as the ANOVA F-test and recursive feature elimination to identify essential attributes, score them, and exclude irrelevant data. The results indicate that the IDS features and the RF method work effectively together .

Han *et al*. (2019) combined static and dynamic API call sequences to identify and classify multiclass malware using RF, DT, KNNs, and XGBoost. The research achieved a detection rate of 97.8% and a classification accuracy of 94.4% with RF. However, the accuracy of the results was considered poor in some contexts. The three studies previously mentioned employed DTs for detection but faced issues such as insufficient information and noise in the training data, which could negatively impact the classification results.

Since the KNNs algorithm treats all features equally, it cannot prioritize certain features as more significant or relevant, which limits its effectiveness in identifying cybercrimes. K-means clustering, being an unsupervised learning algorithm, does not guarantee correct results because it lacks predefined labels (Al-Khater *et al*., 2020). DTs can be used for cybercrime detection; however, they face limitations such as inadequate information and noise in the training data, which can negatively affect machine learning algorithms (Al-Khater *et al*., 2020). Recent evaluations of various classifiers across different datasets indicate that RF and SVMs are more likely to yield effective results compared to KNN and DT algorithms (Cakir and Dogdu, 2018; Basit *et al*., 2021). Unlike these earlier classification methods, naive Bayes assumes that features are independent and uncorrelated (Do *et al*., 2022). Table (6) presents a comprehensive review of recent studies that have utilized machine learning (ML) techniques for cybercrime classification, highlighting their performance and limitations in this domain.

Table 6: Summarizes recent studies on cybercrime classification using ML techniques

Ref.	Research title	Classification techniques	Cybercrime type	Limitations
Andleeb *et al*. (2019)	Identification and classification of cybercrimes using text mining technique	Naive Bayes, K-Means, TF-IDF Vectors	Categories cybercrime offenses	Focuses solely on classifying incidents as cyberbullying or not, despite the diverse forms and manifestations of cyberbullying
Ch *et al*. (2020)	Computational system to classify cybercrime offenses using machine learning	Support vector machine, Bernoulli naive Bayes	Cyberbullying detection	Uses binary classification to determine the presence or absence of cybercrime; the dataset is limited to incidents in India between 2012 and 2017, which may be outdated this time and geographical constraints may limit the results' generalizability
Venkatesan (2023)	Design an intrusion detection system based on feature selection using ML algorithms	Decision tree, random forest, support vector machine	Intrusion detection system	Focuses exclusively on intrusion detection and uses a restricted set of features
Altaher (2017)	Phishing website classification using hybrid support vector machine and k-nearest neighbors approach	Support vector machine, k-nearest neighbors	Website phishing detection	K-Nearest Neighbors cannot prioritize features, which reduces the effectiveness of cybercrime detection
Lekha and Prakasam (2017)	Data mining techniques in detecting and predicting cybercrimes in the banking sector	K-Means clustering, influenced association, decision tree j48	Cybercrime in the financial industry	K-means, being an unsupervised learning algorithm, lacks a guarantee for correct results; the study is limited to financial crimes
Pandey *et al*. (2021)	ENSEM-SLDR: Classification of cybercrime using an ensemble learning technique	Support vector machine, logistic regression, decision tree, random forest	Cybercrime classification	Requires clear data representation and organization for accurate understanding and processing by algorithms
Han *et al*. (2019)	DMaldae: detecting and explaining malware based on correlation and fusion of static and dynamic characteristics	Random forest, decision tree, k-nearest neighbors, xgboost	malware detection	Decision trees were impacted by insufficient information and noise in training data, resulting in poor accuracy

Deep Learning Techniques

DL techniques, which are based on artificial neural networks with multiple layers, analyze and classify input data by feeding the output of one layer as input to the next (Das and Nayak, 2013) Recent studies have focused on DL and hybrid DL approaches for cybercrime classification, particularly using CNNs, LSTM and bidirectional LSTM (BiLSTM). According to Do *et al*. (2022), LSTM and BiLSTM are the most prevalent DL approaches for phishing detection, with LSTM holding a 34% market share and BiLSTM 30%, while CNN is the second-most-used DL method.

CNNs are widely used in supervised learning for classification, prediction and recognition tasks, analyzing input patterns and using labeled data for forecasting outcomes (Do *et al*., 2022), demonstrating strong feature learning performance (Yang *et al*., 2021).

LSTM models effectively tackle the gradient issues of classic RNNs, are suitable for handling time-series data (Do *et al*., 2022), capture long-term dependencies, and exhibit strong performance in various time series and sequence forecasting tasks (Yazi *et al*., 2019)

BiLSTMs improve LSTM networks by encoding information in both forward and backward directions and when combined with attention mechanisms, they provide more direct dependencies between time points (Dadvar and Eckert, 2018).

Xiaofeng *et al*. (2018) employed a hybrid DL and ML model for malware behavior analysis, utilizing binary classification (benign vs. malignant). API call characteristics were extracted using RF. The combined RF and LSTM model achieved an accuracy of 96.7%. However, this study is limited to binary classification of malware behavior and the accuracy may be considered relatively low for certain applications.

In Wang *et al*. (2020b), a proposed model for phishing classification distinguishes between legal and phishing websites following feature extraction. It uses two hybrid classification techniques: RF and BiLSTM. The BiLSTM-based phishing detection model demonstrated a 95.47% identification rate, outperforming the traditional RF-based approach. However, BiLSTM models can suffer from overfitting, which may negatively impact performance on real-world data due to challenges in generalizing beyond the training set. Additionally, the high computational demands and complexity of RF and BiLSTM models can make them difficult to interpret, particularly in the context of cybercrime detection.

This study (Yang *et al*., 2021) introduces an integrated phishing website detection method combining CNNs and RFs. It converts URLs into fixed-size matrices using character embedding techniques, extracts features at various levels with CNN models, and classifies multilevel data using different RF classifiers. While CNNs can automatically learn features from raw data, RFs typically require manually created features. The extraction of relevant information for classification can be challenging and the reliance on website URLs alone may not provide sufficient data for accurate classification.

Using Windows API call sequences, (Yazi *et al*., 2019) created a dataset with eight different types of malware and applied Long Short-Term Memory (LSTM) classification models to detect these malware types. The model achieved an accuracy rate of 97.5%, the highest reported in their study. In (Zhang *et al*., 2019), a feature-hybrid malware variant detection technique was developed by integrating multiple criteria. This approach used a bi-gram model and encoded API calls into a frequency vector with CNNs, achieving a classification accuracy of 90%. Schofield *et al*. (2021) employed a one-dimensional CNN to categorize malware types based on the API call stream. While CNNs and RFs are effective, they require substantial amounts of annotated malware data for training. Acquiring a broad and well-annotated dataset for specific detection tasks can be challenging and costly.

A hybrid pipeline combining CNN and LSTM image processing techniques was proposed in Vinayakumar *et al*. (2019) as a robust DL model for malware classification. This approach, which avoids feature engineering and is relatively fast, focuses solely on malware, neglecting other threats such as phishing, distributed denial of service attacks and data breaches. Additionally, this model demands substantial computational resources for execution.

Dadvar and Eckert (2018) investigated cyberbullying across several social media platforms, including Formspring, Wikipedia, and Twitter. They used four deep neural network-based models: CNN, LSTM, BiLSTM, and BiLSTM with attention. While CNNs are effective for both text and image classification, LSTMs are specific to text classification. BiLSTM encodes data in both forward and backward directions. The study utilized models that allowed for transfer learning across various levels, such as entire, feature, and model-level transfer. The CNN model demonstrated superior performance compared to the others. However, the research is limited to cyberbullying and may overlook other cyber threats like hacking, online fraud, and data breaches. The diversity of social media platforms used for data collection makes consistent detection and identification challenging. Additionally, the focus on determining whether a behavior qualifies as cyberbullying may not fully capture the broader impacts of cyberbullying.

The widely used LSTM has shown potential for time-series prediction and sequence analysis, but big data analysis requires significant time and resources. Transformers were designed to address these challenges. Table (7) provides an overview of recent studies that have applied Deep Learning (DL) techniques to cybercrime classification, showcasing their effectiveness in accurately identifying and categorizing various types of cybercrimes.

Table 7: Summarizes recent studies on cybercrime classification using DL techniques

Ref.	Research title	Classification techniques	Cybercrime type	Limitations
Xiaofeng *et al*. (2018)	ASSCA: API-based sequence and statistics features combined with malware detection architecture	Random forest and long short-term memory	Malware detection	Only analyzes malware behavior with a binary classification approach; its accuracy is relatively low.
Ch *et al*. (2020)	Cyberbullying detection in social networks using deep learning-based models: A reproducibility study	Convolutional neural network, long short-term memory, bidirectional long short-term memory	Cyberbullying detection	Draws data from diverse social media platforms, making it challenging to establish consistent detection and identification; determining whether a behavior constitutes cyberbullying may not capture the full scope of its effects.
Dara and Tumma (2018)	Classification of metamorphic malware with deep learning (LSTM)	Long short-term memory	Malware classification	Focuses exclusively on malware
Vinayakumar *et al*. (2019)	Robust intelligent malware detection using deep learning	Convolutional neural network and long short-term memory	Malware classification	Limited to malware; the model requires substantial computational resources for execution.
Wang *et al*. (2020b)	Deep learning-based efficient model development for phishing detection using random forest and bidirectional long short-term memory classifiers	Random forest and bidirectional long short-term memory	Phishing website classification	High computational demands and complexity in Random Forest and Bidirectional Long Short-Term Memory models make them difficult to interpret.
Wang *et al*. (2020c)	Phishing website detection based on deep convolutional neural network and random forest ensemble learning	Convolutional neural networks and random forests	Phishing website detection	Convolutional Neural Networks learn automatically from raw data, while Random Forests require manually created features; extracting relevant information and using website URLs may not provide sufficient data for accurate classification.
Zhang *et al*. (2019)	Convolutional neural network for malware classification based on API call sequence	Convolutional neural network	Malware classification	Convolutional Neural Networks and Random Forests require large amounts of annotated malware data for effective training and obtaining a broad, well-annotated dataset can be challenging in terms of difficulty and cost.

Transformer Model (BERT)

In recent years, RNNs have gained popularity in supervised natural language processing for regression and classification tasks, but their recurrent nature limits their ability to handle long text. To address these limitations, the transformer model was introduced (Vaswani *et al*., 2017). Unlike RNNs, Transformers eliminate the recurrent architecture and rely solely on the attention mechanism (Gillioz *et al*., 2020). The attention function maps a query and a set of key-value pairs to an output, using vectors. Multi-head attention allows the model to simultaneously focus on different representation subspaces, whereas a single-attention head would only provide a limited perspective (Vaswani *et al*., 2017). This design overcomes issues like gradient vanishing and complex parallelization, making it easier and faster to train larger networks for processing long text data. The transformer model operates in two main phases: The encoder and the decoder (Vaswani *et al*., 2017). BERT, developed in 2017 by Google researchers based on the "transformers" algorithm (Vaswani *et al*., 2017), is a deep learning technique designed for sequence transduction tasks such as language interpretation and machine translation (Ganesh *et al*., 2021). Introduced in 2018 by Devlin BERT

has achieved exceptional performance in various NLP tasks (Devlin, 2018). As a bidirectional model, BERT captures context from both directions in a sentence, allowing for a more nuanced understanding of word relationships and improved language representation. The BERT architecture involves two main phases: Pretraining and fine-tuning:

- Pretraining: This is a process used to train a machine learning model, such as BERT, on unlabeled input. It involves two steps: MLM, which predicts missing words in sentences, and NSP, which understands sentence relationships (Ganesh *et al*., 2021; Devlin, 2018). MLM predicts the original words based on the surrounding context, while NSP predicts the relationship between sentences. Both methods work together to optimize BERT's parameters as shown in Fig. (1) (Han *et al*., 2021). BERT comes in two main pretrained forms: BERT-large and BERT-base, with BERT-large having 24 encoder layers and a hidden size of 1024 and BERT-base having 12 encoder layers and a hidden size of 768 (Ganesh *et al*., 2021)
- Fine-tuning: Involves initializing a model with pre-trained parameters and adjusting it to fit specific labeled data for downstream tasks, requiring each task to have its own customized model (Devlin, 2018)

Preprocessing, tokenization, fine-tuning, and inference are the key phases involved in using BERT for text categorization (Ganesh *et al*., 2021).

In the article by Demirkıran *et al*. (2022), the authors introduced a novel model called CyberBERT that addresses two distinct problems: Malware classification based on API calls and session-based recommendation. CyberBERT employs a bidirectional transformer architecture. The results demonstrate that CyberBERT outperforms standard algorithms such as LSTM and transformer encoders in terms of F1 scores for both binary and multiclass classification tasks. However, CyberBERT exclusively focuses on malware, which limits its applicability to other categories of cybercrime. Additionally, working with multiple datasets from various sources requires extra data preparation, potentially affecting the effectiveness of the results.

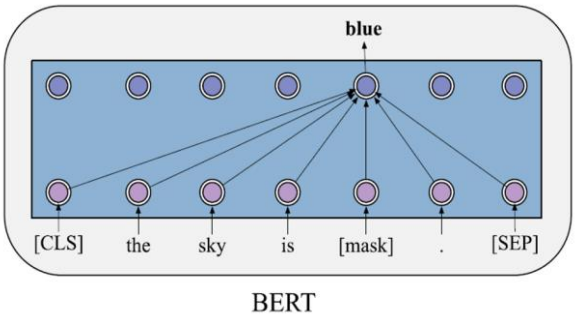

Fig. 1: BERT self-attention mechanisms and pretraining objectives (Ghourabi and Alohaly, 2023)

The study by Guo *et al*. (2022), presents a unique architecture called Augmented BERT designed to detect cyberbullying. To address the issue of limited annotated texts on cyberbullying, the authors use data augmentation techniques, including generative adversarial networks and autoencoders, to generate additional annotated data. The augmented data is used to improve BERT, particularly HateBERT, which is pretrained to recognize the abusive language. Experimental results show that this approach performs better than existing models for cyberbullying detection. However, the study focuses exclusively on cyberbullying, specifically hate speech, and uses a relatively small dataset with 3,000 entries from various sources.

In Sahmoud and Mikki (2022), researchers developed a highly effective spam detector capable of identifying spam emails and SMS messages. The model was trained on several corpora, including Enron, SpamAssassin, LingSpam, and SMS spam collections, achieving accuracy rates of 98.62, 97.83, 99.13 and 99.28%, respectively. The model's strong performance is attributed to the use of BERT, a pretrained model that enhances understanding of message context, leading to better spam detection accuracy. However, the study has limitations, including a small evaluation dataset and a focus solely on email content without considering email semantics and headers. A more comprehensive analysis that includes these factors could provide a deeper understanding of email-related phenomena.

The research by Giri *et al*. (2022) compares the effectiveness of two deep learning models for phishing email detection. The first model uses CNNs with GloVe embeddings, while the second uses BERT with fine-tuning. Several widely used datasets (lingSpam, EnronSpamSubset, complete SpamAssassin, Jose Nazario's phishing dataset, and the Enron email dataset) were used to evaluate the models. The findings indicate that GloVe embeddings achieve a higher accuracy rate of 98% compared to the BERT model. Despite these positive results, the study has limitations, including variability across different datasets, which affected the efficiency of the results. Additionally, the study did not analyze individual words or types of words, which could have provided more insights into the underlying linguistic patterns.

In Cao and Lai (2020) M-BERT was used to generate token-level vectors for training a TextCNN model for spam detection. This approach resulted in a BERT-CNN model with an accuracy of 96%. The combination leverages the feature extraction capabilities of CNNs with the language representation capabilities of BERT. However, the study faces several limitations, including a small dataset that may not adequately represent all features in each language, potential biases or inaccuracies from using an external feature extraction system, and time-consuming preprocessing required for multilingual data. These limitations could impact the quality and generalizability of the findings. Effective use of diverse deep learning

techniques requires careful construction of precise data representations to ensure the algorithms can accurately understand and analyze the information. BERT, trained on large-scale datasets, understands the contextual relationships between words and sentences, capturing subtle nuances and semantic meanings in cybercrime-related texts (Demirkıran *et al*., 2022). Table (8) presents a survey of recent studies that have employed transformer models, particularly BERT, for cybercrime classification, highlighting their performance and potential in this domain. This contextual understanding significantly enhances the effectiveness of cybercrime detection systems.

Generative AI (GPT)

Over the past year, Generative AI has gained significant popularity on the internet (Gupta *et al*., 2023). This technology employs computational algorithms and deep neural networks to generate meaningful content by learning patterns and structures from training data, which can include text, images, and audio. Examples of Generative AI models include GPT-3 and Copilot (Ahmed *et al*., 2023). These large models can generate output across various domains and data sources (Feuerriegel *et al*., 2024). Generative AI, a statistical approach, has potential applications in natural language processing, visual recognition, and data generation (Agrawal *et al*., 2024), potentially replacing knowledge workers, providing advice, and managing IT help desks (Feuerriegel *et al*., 2024). However, the integration of AI and generative models has raised concerns about security (Metta *et al*., 2024). In cybersecurity, generative models like ChatGPT and Google Bard are used for both defensive and offensive purposes (Falade, 2024).

Generative Pretrained Transformer (GPT) is a pretrained DL model that has been fine-tuned for various tasks, including language synthesis, sentiment analysis, machine translation, and text categorization. Utilizing a transformer architecture, GPT outperforms earlier NLP approaches such as RNNs and CNNs (Yenduri *et al*., 2024). The self-attention mechanism in GPT enhances language understanding and generation by considering the context of the entire phrase when predicting the next word (Yenduri *et al*., 2024). This mechanism enables the model to focus on relevant sections of the input text, generating coherent and contextually appropriate outputs (Saka *et al*., 2024).

GPT is a deep transformer architecture with generative techniques (Han *et al*., 2021). It is utilizing a multi-headed self-attention mechanism within a 12-layer decoder-only transformer. It generates output distributions across target tokens from input context tokens (Ghourabi and Alohaly, 2023). It optimizes conditional probabilities of words using preceding words as context (Han *et al*., 2021). The model includes generative pretraining and discriminative fine-tuning phases.

Generative pretraining: GPT pretrains using autoregressive language modeling, omitting cross-attention mechanism from Transformer decoder output layers. Masked multi-head self-attention operations compute conditional probability distributions on preceding words as shown in Fig. (2) (Han *et al*., 2021).

Discriminative fine-tuning: The GPT fine-tuning method adjusts pretrained parameters to downstream jobs. This involves feeding the input sequence through the GPT Transformer, obtaining outputs from the final layer, and optimizing conventional objectives using additional output layers (Han *et al*., 2021). BERT focuses on language interpretation, whereas GPT excels at natural language creation (Han *et al*., 2021)

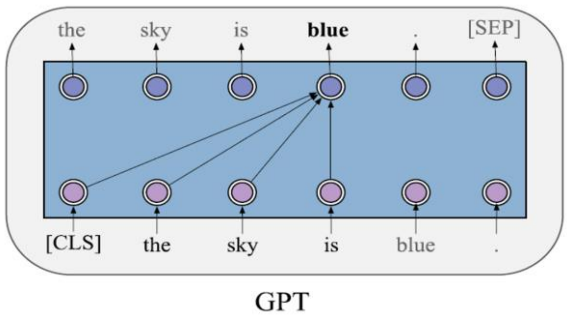

Fig. 2: GPT self-attention mechanisms and pre-training objectives (Ghourabi and Alohaly, 2023)

Table 8: Summarizes recent studies focused on cybercrime classification using transformer models (BERT)

Ref.	Research title	Classification techniques	Cybercrime type	Limitations
Cao and Lai (2020)	A bilingual multi-type spam detection model based on M-BERT	M-BERT-convolutional neural network	Spam detection	Small datasets may not adequately represent all features in each language; time-consuming preprocessing is required for multilingual data; need for explicit data representations
Demirkıran *et al*. (2022)	An ensemble of pretrained transformer models for imbalanced multiclass malware classification	Cyber-BERT	Malware classification	Exclusively focuses on malware; working with multiple datasets from different sources requires additional data preparation
Guo *et al*. (2022)	Cyberbullying detection using bert with augmented texts	Augmented BERT	Cyberbullying detection	Focuses only on cyberbullying and specifically hate speech; and uses a small dataset for training
Sahmoud and Mikki (2022)	Spam detection using BERT	BERT	Spam detection	Small evaluation dataset; focuses solely on email type and body without considering email semantics and headers
Giri *et al*. (2022)	Comparative study of content-based phishing email detection using global vectors and bidirectional encoder Representations from Transformers (BERT) word embedding models	Convolutional neural networks with global vectors and BERT	Phishing Email Detection	Efficiency is reduced due to variability across datasets; lacks categorization or emphasis on specific word types

Microsoft-backed firm OpenAI released ChatGPT, a generative AI tool, on November 30, 2022 (Gupta *et al*., 2023). Powered by the GPT-3 language model which has significantly advanced natural language processing and generation. These models excel at Natural Language Understanding (NLU), analyzing and comprehending text, and identifying entities and relationships within sentences (Yenduri *et al*., 2024). ChatGPT can simulate human-like conversations and has quickly gained popularity, amassing over 100 million users in just two months (Krishnamurthy, 2023). It is now widely used internationally. ChatGPT has been noted by (Ahmed *et al*., 2023) for its potential to rapidly and dramatically transform the field of AI. OpenAI's ethical and governance guidelines restrict ChatGPT's outputs (Gupta *et al*., 2023). Due to these ethical constraints, ChatGPT cannot process phishing email prompts, which criminals might exploit (Falade, 2024). ChatGPT uses autoregressive language generation with GPT-3 models to produce natural-sounding responses. Generative AI products like ChatGPT, which use LLMs trained on cyber threat intelligence data, can help cyber defenders better protect their systems from malicious hackers (Gupta *et al*., 2023). However, users often attempt to circumvent ChatGPT's constraints to prevent them from engaging in unlawful, unethical, or potentially dangerous behavior (Gupta *et al*., 2023).

Based on the GPT-3 Transformer, the research in (Ghourabi and Alohaly, 2023) provides a text embedding model for SMS categorization and spam detection. The model produces high-quality vector representations of text, improving classification results with a state-of-the-art accuracy of 99.91%.

A novel method called ChatPhishDetector was proposed by Koide *et al*. (2023), employing GPT-3.5 and GPT-4 to detect phishing sites with 98.3% accuracy and 98.4% recall. This method uses a web crawler to collect website data, create LLM prompts, and extract detection results from LLM responses. GPT-4V has shown the greatest accuracy (98.7%) and recall (99.6%) compared to other LLMs and systems, making it highly effective in improving software development safety by addressing various attacks in different languages (Gupta *et al*., 2023). While GPT-4 could be misused by cybercriminals for various attacks, its careful implementation may reduce the risk of individuals and organizations falling victim to such threats (Ferrag *et al*., 2024). It is essential to focus on developing robust, trustworthy, secure, multilingual, and multimodal solutions while deploying GPT models, ensuring they are resource-efficient and tailored to specific domains or user needs (Yenduri *et al*., 2024).

The study by Si *et al*. (2024) analyzes the spam email detection performance of ChatGPT compared to baseline models such as SVM, logistic regression, naive Bayes, and BERT. Evaluations were conducted using the Email Spam Detection (ESD) dataset and a Chinese spam dataset. The findings reveal that while ChatGPT excels in spam categorization, its overall performance on the English ESD dataset is weaker compared to supervised models. However, ChatGPT maintains strong spam detection capabilities across languages, achieving accuracy between 0.84 and 0.89 and an F1 score ranging from 0.76-0.80.

The goal of earlier studies on the identification and classification of cybercrime has been to enhance accuracy using various techniques. DL models have been found to outperform traditional machine learning classifiers in terms of accuracy. However, there is still potential for improvement using feature reduction techniques and ensemble models (Basit *et al*., 2021; Guo *et al*., 2022). One of the key shortcomings of traditional ML-based cybercrime detection systems is their reliance on feature engineering, feature learning, and feature representation methodologies (Cakir and Dogdu, 2018). These techniques require comprehensive domain knowledge (Cakir and Dogdu, 2018), which can be both challenging and time-consuming to acquire. To address this challenge, this research aims to leverage transformer models like BERT or generative AI models like GPT to enhance the understanding and extraction of meaningful features from text. By incorporating these models into the feature extraction process, it is possible to achieve more accurate results in cybercrime detection tasks. These models are adept at capturing complex patterns and relationships within the text, thereby enabling more comprehensive and accurate detection (Devlin, 2018). Table (9) provides a comprehensive summary of recent studies on cybercrime classification using Generative AI models, highlighting the effectiveness of GPT-based approaches in this domain.

Table 9: Summarizes recent studies on cybercrime classification using Generative AI models (GPT)

Ref.	Research Title	Classification Techniques	Cybercrime Type	Limitations
Ghourabi and Alohaly (2023)	Enhancing spam message classification and detection using transformer-based embedding and ensemble learning	GPT-3-based Embedding	SMS categorization and spam detection	The ensemble learning approach, which includes four classification algorithms, may increase computation and memory usage for large text models
Koide *et al*. (2023)	Detecting phishing sites using ChatGPT	GPT-3.5, GPT-4, and GPT-4V	Phishing site detection	Results may be impacted by ChatGPT models incorrectly classifying phishing and non-phishing sites, particularly for services established after September 2021
Si *et al*. (2024)	Evaluating the Performance of ChatGPT for Spam Email Detection	ChatGPT	Spam email detection	The study uses an inadequate Chinese dataset of short texts under 100 words

Dataset Parameters

The process of selecting datasets emerged as a pivotal phase in acquiring all necessary parameters aligned with the study objectives. A thorough review of various studies on cybercrime datasets revealed that the available resources were often insufficient. Some datasets were outdated, while others lacked the essential content needed for effective classification and understanding of the crime situation. Many datasets, such as those used in (Andleeb *et al*., 2019; Pandey *et al*., 2021) or found on platforms like Kaggle (2024); and GitHub (2023), primarily provide statistics about cybercrime types or occurrences in specific locations or timeframes, which do not meet the research goals.

The existing studies and datasets present several limitations based on the requirements of this research. First, they predominantly focus on phishing attack datasets (Wang *et al*., 2020b-c), which cover only one specific type of cybercrime targeted by this research.

Additionally, (Gillioz *et al*., 2020; Ch *et al*., 2020) describe cyberbullying texts collected from social media platforms, but they only indicate the words associated with cyberbullying. For this research, it is crucial to obtain detailed case information, such as the type of defendant and the victim's age, to accurately assess the severity of the cyberbullying incident.

Furthermore, (Do *et al*., 2022) discuss financial crimes related to fraud and identity theft, limiting the scope to financial cybercrimes. The available data is insufficient to accurately evaluate the severity of these crimes. Therefore, a more comprehensive dataset encompassing various types of cybercrimes, detailed case information, and severity indicators is essential for this research. In brief, the available dataset types used in the mentioned studies are:

1. Website phishing URLs
2. API call sequences for malware
3. SMS messages for spam detection
4. Email spam
5. Cybercrime statistics (based on location, time, or type)
6. Cyberbullying texts from social media

Therefore, it is essential to include specific parameters in the dataset of cybercrime cases. Based on insights from Section 5, these necessary features include the crime description, types of offenders and victims and the nature of the crime. These parameters assist researchers and models in accurately identifying the crime type and its severity level.

Moreover, it is recommended to have an independent platform to collect a comprehensive dataset from officially published cases within the countries and from websites related to crimes. One good example of that is the Sharing Electronic Resources and Laws on Crime (SHERLOC) knowledge management portal (United Nations Office on Drugs and Crime, 2023), developed and maintained by the United Nations Office on Drugs and Crime (UNODC) (UNODC, 2023). This platform can provide detailed and relevant case information crucial for our study. Cases example:

1. Unauthorized entry and hacking into an electronic newspaper website, altering its data, vandalizing it, destroying it, and changing the design to include explicit images
2. Promoting and selling CDs containing explicit sexual content (pornography)
3. Creating a page on the social media platform Facebook to publicly shame employees of a medical facility, along with photographing and publishing confidential official documents
4. Blackmailing a girl with the threat of publishing her photos, engaging in illicit relationships with multiple girls, and possessing pornographic videos and images
5. Promoting and using drugs and psychotropic substances through social media and information networks to spread and endorse prohibited items

A criminal group committed fraud by posing as bank employees. The victim provided her cell phone number and area code for advice and received calls from unidentified individuals. The victim's information was used to obtain a loan, which was then transferred to the defendant's account. Subsequently, the money was transferred to other parties.

Conclusion

Cybercrime, as defined in the literature, covers a broad spectrum of illegal activities conducted through digital means, emphasizing the need for precise categorization to effectively address and mitigate its impact. The examination of cybercrime categories highlighted the diversity of offenses, ranging from cyberbullying and phishing to more complex threats like botnets and malware attacks. Each category presents unique challenges and necessitates tailored approaches for prevention and enforcement.

Key features of cybercrime, including detailed descriptions of offenses, types of perpetrators, and profiles of victims, are crucial for understanding the nature and scope of these crimes. Furthermore, assessing the severity levels of cybercrime across different regions, particularly comparing Saudi Arabia and the UK, reveals how cultural, economic, and legal contexts influence the manifestation and impact of cybercrime.

The integration of advanced technologies, specifically AI, ML, and DL, plays a pivotal role in enhancing our ability to identify and classify cybercrime features. This review has highlighted how these technologies can be

leveraged to extract relevant features from cybercrime case descriptions, enabling more accurate classification and severity assessment. The use of transformer models and generative AI models, such as BERT and GPT, has shown promising results in natural language processing tasks, further improving the effectiveness of cybercrime classification models.

Suitable datasets should include detailed case descriptions, demographic information, and clear classification labels. Analyzing the datasets used in the reviewed studies provided valuable insights into common cybercrime methods, victim profiles, and evolving trends.

This review paper has explored the multifaceted domain of cybercrime, offering a comprehensive analysis that encompasses its definitions, categories, and distinctive features. Through the synthesis of existing studies, this study has provided insights into the methodologies and challenges associated with using AI and ML in cybercrime classification. The review of feature extraction techniques, ML, DL, transformer models, and generative AI models revealed their potential in advancing the field but also highlighted the need for continued innovation and refinement of these approaches. The discussion on datasets emphasized the importance of comprehensive and well-structured data for cybercrime research.

To advance the field of cybercrime research, Future research should focus not only on enhancing the accuracy of classification models but also on evaluating their practical implications in real-world contexts. Understanding how these models perform in actual cybercrime investigations and prevention efforts will provide valuable insights for refining their utility. Cross-regional studies are also critical, as they can uncover significant variations in the nature and severity of cybercrime across different regions, cultures, and legal frameworks. Such research can guide the development of strategies that are more adaptable to diverse environments.

Emerging AI technologies, including generative models like GPT-4 and advanced Transformer architectures, should be leveraged to improve the precision and interpretability of cybercrime classification systems. These tools can analyze complex datasets and uncover hidden patterns in cybercrime trends, offering a more comprehensive understanding of the field. Additionally, future studies should adopt multiclass classification methods rather than relying on binary models, as these can provide more detailed categorizations of cybercrime types and severity, enabling more effective and tailored interventions.

Another priority for future research is the collection of datasets with a broader spectrum of cybercrime categories, detailed descriptions, and demographic variables. These datasets should be sourced from a variety of regions, cultures, and industries to better reflect the global scope of cybercrime and mitigate the biases inherent in localized or outdated datasets. By addressing these aspects, researchers can contribute to the development of more robust and effective approaches to combating cybercrime.

Acknowledgment

The authors would like to thank King Abdulaziz University, which has supported and sponsored their studies and works.

Funding Information

The authors should acknowledge the funders of this manuscript and provide all necessary funding information.

Author's Contributions

Norah Alshahrani: Gathered papers and data from many sources. Wrote a summary of each paper represented in a method used and, the result of the paper. Designed the outline of the paper tables and figures. Conducted literature search and performed review paper.

Fahad Alotaibi: Provided valuable notes and comments to enhance the paper's quality. Reviewed the paper in general, sections and subsections correlation, verified method and strategy used.

Khalid Alyoubi: Performed an extensive review of the paper, contributing to improvements in its organization and quality, and validated the methods and techniques employed.

Muhammad Ramzan: Offered constructive feedback and insights to improve the paper's quality. Conducted a general review, assessed the coherence between sections and subsections, and validated the methods and strategies applied.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all co-authors have reviewed and approved the manuscript and affirms that no ethical issues are associated with its content.

References

1428 AH. (2007). *Anti-Cyber Crime Law*. https://laws.boe.gov.sa/BoeLaws/Laws/LawDetails/25df73d6-0f49-4dc5-b010-a9a700f2ec1d/2

Agrawal, G., Kaur, A., & Myneni, S. (2024). A Review of Generative Models in Generating Synthetic Attack Data for Cybersecurity. *Electronics*, *13*(2), 322. https://doi.org/10.3390/electronics13020322

Ahmed, I., Roy, A., Kajol, M., Hasan, U., Datta, P. P., & Reza, Md. R. (2023). ChatGPT vs. Bard: a comparative study. *Authorea*. https://doi.org/10.22541/au.168923529.98827844/v1

Al-Khater, W. A., Al-Maadeed, S., Ahmed, A. A., Sadiq, A. S., & Khan, M. K. (2020). Comprehensive Review of Cybercrime Detection Techniques. *IEEE Access*, *8*, 137293–137311. https://doi.org/10.1109/access.2020.3011259

AllahRakha, N. (2024). Transformation of Crimes (Cybercrimes) in the Digital Age. *International Journal of Law and Policy*, *2*(2). https://doi.org/10.59022/ijlp.156

Altaher, A. (2017). Phishing Websites Classification using Hybrid SVM and KNN Approach. *International Journal of Advanced Computer Science and Applications*, *8*(6), 90–95. https://doi.org/10.14569/ijacsa.2017.080611

Andleeb, S., Ahmed, R., Ahmed, Z., & Kanwal, M. (2019). Identification and Classification of Cybercrimes using Text Mining Technique. *2019 International Conference on Frontiers of Information Technology (FIT)*, 227–275. https://doi.org/10.1109/fit47737.2019.00050

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Breen, C., Herley, C., & Redmiles, E. M. (2022). A Large-Scale Measurement of Cybercrime Against Individuals. *CHI Conference on Human Factors in Computing Systems*, 1–41. https://doi.org/10.1145/3491102.3517613

Cakir, B., & Dogdu, E. (2018). Malware classification using deep learning methods. *Proceedings of the 2018 ACM Southeast Conference*, 1–5. https://doi.org/10.1145/3190645.3190692

Cao, J., & Lai, C. (2020). A Bilingual Multi-type Spam Detection Model Based on M-BERT. *GLOBECOM 2020-2020 IEEE Global Communications Conference*, 1–6. https://doi.org/10.1109/globecom42002.2020.9347970

Ch, R., Gadekallu, T. R., Abidi, M. H., & Al-Ahmari, A. (2020). Computational System to Classify Cyber Crime Offenses using Machine Learning. *Sustainability*, *12*(10), 4087. https://doi.org/10.3390/su12104087

Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *ArXiv:1812.08046*. https://doi.org/10.48550/arXiv.1812.08046

Dara, S., & Tumma, P. (2018). Feature Extraction By Using Deep Learning: A Survey. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1795–1801. https://doi.org/10.1109/iceca.2018.8474912

Das, M., Kamalanathan, S., & Alphonse, P. J. A. (2023). A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. *ArXiv:2308.04037*. https://doi.org/10.48550/arXiv.2308.04037

Das, S., & Nayak, T. (2013). Impact of cybercrime: Issues and challenges. *International Journal of Engineering Sciences & Emerging Technologies*, *6*(2), 142–153.

Demirkıran, F., Çayır, A., Ünal, U., & Dağ, H. (2022). An ensemble of pre-trained transformer models for imbalanced multiclass malware classification. *Computers & Security*, *121*, 102846. https://doi.org/10.1016/j.cose.2022.102846

Devlin, J. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv:1810.04805*.

Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2022). Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions. *IEEE Access*, *10*, 36429–36463. https://doi.org/10.1109/access.2022.3151903

Falade, P. V. (2024). Deciphering ChatGPT's impact: exploring its role in cybercrime and cybersecurity. *International Journal of Scientific Research in Computer Science and Engineering*, *12*(2), 15–24.

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., & Tihanyi, N. (2024). *Generative AI and large language models for cyber security: all insights you need*. https://doi.org/10.2139/ssrn.4853709

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. In *Business & Information Systems Engineering* (Vol. 66, pp. 111–126). https://doi.org/10.1007/s12599-023-00834-7

Ganesh, P., Chen, Y., Lou, X., Khan, M. A., Yang, Y., Sajjad, H., Nakov, P., Chen, D., & Winslett, M. (2021). Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. *Transactions of the Association for Computational Linguistics*, *9*, 1061–1080. https://doi.org/10.1162/tacl_a_00413

Ghourabi, A., & Alohaly, M. (2023). Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning. *Sensors*, *23*(8), 3861. https://doi.org/10.3390/s23083861

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the Transformer-based Models for NLP Tasks. *Annals of Computer Science and Information Systems*, 179–183. https://doi.org/10.15439/2020f20

Giri, S., Banerjee, S., Bag, K., & Maiti, D. (2022). Comparative Study of Content-Based Phishing Email Detection Using Global Vector (GloVe) and Bidirectional Encoder Representation from Transformer (BERT) Word Embedding Models. *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–6. https://doi.org/10.1109/iceeict53079.2022.9768612

GitHub. (2023). Build and ship software on a single, collaborative platform. *GitHub.Com*. https://github.com

Goni, O. (2021). Cyber Crime and Its Classification. *International Journal of Electronics Engineering and Applications*, *10*(2), 01–17. https://doi.org/10.30696/ijeea.x.i.2022.01-17

Guo, X., Anjum, U., & Zhan, J. (2022). Cyberbully Detection Using BERT with Augmented Texts. *2022 IEEE International Conference on Big Data (Big Data)*, 1246–1253. https://doi.org/10.1109/bigdata55660.2022.10020581

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, *11*, 80218–80245. https://doi.org/10.1109/access.2023.3300381

Han, W., Xue, J., Wang, Y., Huang, L., Kong, Z., & Mao, L. (2019). MalDAE: Detecting and explaining malware based on correlation and fusion of static and dynamic characteristics. *Computers & Security*, *83*, 208–233. https://doi.org/10.1016/j.cose.2019.02.007

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., & Yao, Y. (2021). Pre-trained models: past, present and future. *AI Open*, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Ibrahim, S. (2016). Social and contextual taxonomy of cybercrime: Socioeconomic theory of Nigerian cybercriminals. *International Journal of Law, Crime and Justice*, *47*, 44–57. https://doi.org/10.1016/j.ijlcj.2016.07.002

Kaggle. (2024). Datasets. *Kaggle.Com*. https://www.kaggle.com/datasets

Khyioon Abdalrdha, Z., Mohsin Al-Bakry, A., & K. Farhan, A. (2023). A Survey on Cybercrime Using Social Media. *Iraqi Journal for Computers and Informatics*, *49*(1), 52–65. https://doi.org/10.25195/ijci.v49i1.404

Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2023). Detecting phishing sites using ChatGPT. *ArXiv:2306.05816*. https://doi.org/10.48550/arXiv.2306.05816

Krishnamurthy, O. (2023). Enhancing cyber security enhancement through generative AI. *International Journal of Universal Science and Engineering*, *9*(1), 35–50.

Lekha, K. C., & Prakasam, S. (2017). Data mining techniques in detecting and predicting cyber crimes in banking sector. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 1639–1643. https://doi.org/10.1109/icecds.2017.8389725

Matias, S. M. M., Costales, J. A., & De Los Santos, C. M. (2022). A Framework for Cybercrime Prediction on Twitter Tweets Using Text-Based Machine Learning Algorithm. *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 235–240. https://doi.org/10.1109/prai55851.2022.9904212

Matveev, V., Khrypko, S., Nykytchenko, O. E., Stefanova, N., Ishchuk, A., Ishchuk, O., & Bondar, T. (2021). Cybercrime in the Economic Space: Psychological Motivation and Semantic-Terminological Specifics. *IJCSNS International Journal of Computer Science and Network Security*, *21*(11), 135–142. https://doi.org/10.22937/ijcsns.2021.21.11.18

Metta, S., Chang, I., Parker, J., Roman, M. P., & Ehuan, A. F. (2024). Generative AI in cybersecurity. *ArXiv:2405.01674*. https://doi.org/10.48550/arXiv.2405.01674

Norton, S., Schumacher, M., & Brasse, M. (2015). A pragmatic approach for integrating safety design and operational security. *10th IET System Safety and Cyber-Security Conference 2015*, Bristol, UK. https://doi.org/10.1049/cp.2015.0285

Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*, *174*, 114765. https://doi.org/10.1016/j.eswa.2021.114765

Onyekpeze, O., Owolabi, Olumide, & Ibrahim, Bisalla Hashim. (2021). Classification of cybersecurity incidents in Nigeria using machine learning methods. *Covenant Journal of Informatics and Communication Technology*, *9*(2), 1–16.

Pandey, H., Goyal, R., Virmani, D., & Gupta, C. (2021). Ensem_SLDR: Classification of Cybercrime using Ensemble Learning Technique. *International Journal of Computer Network and Information Security*, *14*(1), 81–90. https://doi.org/10.5815/ijcnis.2022.01.07

Phillips, K., Davidson, J. C., Farr, R. R., Burkhardt, C., Caneppele, S., & Aiken, M. P. (2022). Conceptualizing Cybercrime: Definitions, Typologies and Taxonomies. *Forensic Sciences*, *2*(2), 379–398. https://doi.org/10.3390/forensicsci2020028

Prabakaran, S., & Mitra, S. (2018). Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning. *Journal of Physics: Conference Series*, *1000*, 012046. https://doi.org/10.1088/1742-6596/1000/1/012046

Sahmoud, T., & Mikki, Mohammad. (2022). Spam detection using BERT. *ArXiv:2206.02443*. https://doi.org/10.48550/arXiv.2206.02443

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, *17*, 100300. https://doi.org/10.1016/j.dibe.2023.100300

Sarkar, G., & Shukla, S. K. (2023). Behavioral analysis of cybercrime: Paving the way for effective policing strategies. *Journal of Economic Criminology*, *2*, 100034. https://doi.org/10.1016/j.jeconc.2023.100034

Schofield, M., Alicioglu, G., Binaco, R., Turner, P., Thatcher, C., Lam, A., & Sun, B. (2021). Convolutional Neural Network for Malware Classification Based on API Call Sequence. *Computer Science & Information Technology (CS & IT)*, 85–98. https://doi.org/10.5121/csit.2021.110106

Selva Birunda, S., & Kanniga Devi, R. (2021). A Review on Word Embedding Techniques for Text Classification. *Innovative Data Communication Technologies and Application*, 267–281. https://doi.org/10.1007/978-981-15-9651-3_23

Si, S., Wu, Y., Tang, L., Zhang, Y., & Wosik, J. (2024). Evaluating the performance of ChatGPT for spam email detection. *ArXiv:2402.15537*. https://doi.org/10.48550/arXiv.2402.15537

Suhaidi, M., Abdul Kadir, R., & Tiun, S. (2021). A review of feature extraction methods on machine learning. *Journal of Information System and Technology Management*, *6*(22), 51–59.

Tabassum, A., & Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, *7*(6), 4864–4867.

Tsakalidis, G., & Vergidis, K. (2019). A Systematic Approach Toward Description and Classification of Cybercrime Incidents. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(4), 710–729. https://doi.org/10.1109/tsmc.2017.2700495

United Nations Office on Drugs and Crime. (2023). Environmentally sustainable practices and approaches. *Practical Guide: Alternative Development and the Environment*, 22–38. https://doi.org/10.18356/9789213585672c008

UNODC. (2023). United Nations Office on Drugs and Crime. *UNODC*. https://www.unodc.org

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Venkatesan, S. (2023). Design an intrusion detection system based on feature selection using ML algorithms. *Mathematical Statistician and Engineering Applications*, *72*(1), 702–710. https://doi.org/10.17762/msea.v72i1.2000

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019). Robust Intelligent Malware Detection Using Deep Learning. *IEEE Access*, *7*, 46717–46738. https://doi.org/10.1109/access.2019.2906934

Wall, D. S. (2024). *Cybercrime: The Transformation of Crime in the Information Age*.

Wang, C., Nulty, P., & Lillis, D. (2020a). A Comparative Study on Word Embeddings in Deep Learning for Text Classification. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 37–46. https://doi.org/10.1145/3443279.3443304

Wang, S., Khan, S., Xu, C., Nazir, S., & Hafeez, A. (2020b). Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers. *Complexity*, *2020*, 1–7. https://doi.org/10.1155/2020/8694796

Wang, D., Su, J., & Yu, H. (2020c). Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language. *IEEE Access*, *8*, 46335–46345. https://doi.org/10.1109/access.2020.2974101

Xiaofeng, L., Xiao, Z., Fangshuo, J., Shengwei, Y., & Jing, S. (2018). ASSCA: API based Sequence and Statistics features Combined malware detection Architecture. *Procedia Computer Science*, *129*, 248–256. https://doi.org/10.1016/j.procs.2018.03.072

Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2021). Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning. *Sensors*, *21*(24), 8281. https://doi.org/10.3390/s21248281

Yazi, A. F., Çatak, F. Ö., & Gül, E. (2019). Classification of Methamorphic Malware with Deep Learning (LSTM). *2019 27th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/siu.2019.8806571

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer) A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/access.2024.3389497

Zhang, J., Qin, Z., Yin, H., Ou, L., & Zhang, K. (2019). A feature-hybrid malware variants detection using CNN based opcode embedding and BPNN based API embedding. *Computers & Security*, *84*, 376–392. https://doi.org/10.1016/j.cose.2019.04.005