A Survey of Machine Learning Approaches to Cyber Data Breach Detection: Current Challenges and Future Research Directions

Paul Ntim Yeboah, A. S. M. Kayes, Wenny Rahayu, Eric Pardede, and Syed Mahbub
La Trobe University, Melbourne, VIC 3086, Australia.

***Abstract*—Although several machine learning driven solutions are deemed to be effective at detecting data breaches, the recent proliferation in data breach incidents resulting from cyber attacks demands an updated, thorough analysis of machine learning (ML) based data breach countermeasures to identify research gaps and guide future studies. In view of this, this study employs a systematic approach and draws insight from 81 research articles to classify machine learning based data breach countermeasures using *six* criteria namely learning tasks, learning classifiers, proactive learning strategies, feature engineering methods and multimodal approaches. In classifying the studies, we: (a) propose a taxonomy of feature extraction and representation to classify studies using ten sub-criteria, (b) identify proactive learning techniques to categorise studies using four sub-criteria including self labelling, data augmentation, automated feature extraction and re-training, (c) classify multimodal machine learning approaches used in the studies into three fusion sub-criteria: namely early fusion, intermediate fusion and late fusion. To aid the literature identification, we analyse forty recent incidents and obtain prevalent cyber attack vectors of data breaches, which we present as the general workflow for data breaches due to cyber attacks. Finally, we highlight the research issues associated with existing ML-based data breach countermeasures and recommend future research directions.**

***Index Terms*—Cyber threats and incidents, data breach, machine learning, proactive learning, multimodal learning, learning classifier, feature engineering, and cyber data breach detection.**

I. Introduction

THE ubiquitous and indispensable nature of technology which impacts every sphere of human endeavour has resulted in a data explosion. Although individuals, businesses and governments harness the inherent potential of these volumes of generated data, incidents of data and privacy breach regularly dominate headlines [1]–[4]. Statistics over the past decade show an unprecedented increase in the number of recorded data breach cases. For example, a yearly research report conducted by Verizon [5] shows that, over the period from 2008 to 2022, the number of investigated data breach incidents skyrocketed from over 90 to 5,200, as shown in figure 1. This overwhelming number of incidents reflect changes in the sources, motivations and targets of data breaches. Verizon's 2022 report reveals espionage becoming the second motivation after financial, bypassing hacktivism and other motivators. In terms of the most targeted data types, personally identifiable information (PII) and credential information are the top two, as opposed to a 2008 report which found that payment card information was the most sought after data type [5].

A report by Statista [6] estimated that about 4,772 data compromises were recorded in the USA between 2020 to 2022, impacting millions of individuals. For the past 15 years, Ponemon, an IBM sponsored institute has periodically published yearly report on the cost of data breaches. A breakdown of their 2023 report [7] derived from about 553 breaches reveals that the average cost of data breaches reached a record level of $4.45 million, representing a 2.3% increase from the previous year. They further indicated that, despite tougher regulations, the health sector continues to record the costliest data breaches of all sectors reaching over $10.93 million and $10.10 million in 2023 and 2022 respectively.

Studies show that, these breaches are the result of either accidental or malicious actions [3], [8], [9]. Among the two sources, malicious cyber attacks are identified as the leading driver of recent data breach incidents. According to a report by the Office of the Australian Information Commissioner (OAIC) [10], cyber incidents accounted for 63% and 65% of the first and second halves of 2022 respectively, together yielding a total of 284 data breaches. Another 2022 report by the Identity Theft Resource Centre (ITRC) [11] also pointed to cyber attacks as the leading cause of data compromises in the USA, accounting for 90% of the 1802 breaches recorded. In their annual data breach investigations from 2008 - 2022, Verizon reports that hacking continues to be the most common form of attack [5], [12]. In 2023, the MOVEit cyber incident [13] attracted global attention as one of the largest and most devastating data breaches. The attack which was perpetrated by a ransomware group known as Clop, impacted over 2000 organisations across the globe and affected over 62 million individuals. In the attacks, the threat actor exploited zero-day SQL injection vulnerabilities in the MOVEit file transfer software and exfiltrated volumes of personal data.

Signature-based detection mechanisms (also referred as misuse detection) are used on networks for detection of data breaches due to cyber attacks. In the misuse detection approach, specific rules (i.e., signatures of attacks) derived from previous malicious network activities are pre-defined to detect cyber related breaches. The limitation is that, such a detection solution can easily be circumvented if the threat actor

This paper was produced by Paul Ntim Yeboah (p.yeboah@latrobe.edu.au), A. S. M. Kayes (a.kayes@latrobe.edu.au), Wenny Rahayu (w.rahayu@latrobe.edu.au), Eric Pardede (e.pardede@latrobe.edu.au), and Syed Mahbub (s.mahbub@latrobe.edu.au).

Manuscript received February 20, 2024; revised February 20, 2024.

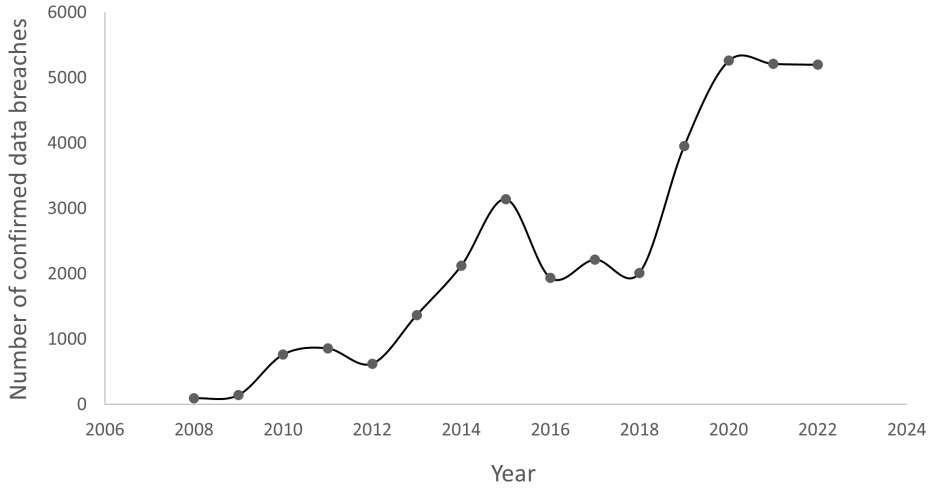

Fig. 1. The number of breaches analyzed in data breach investigation report (DBIR) over the period 2008 to 2022.

uses novel tactics in the hack. For this reason, a more resilient solution based on machine learning (ML) is proposed as countermeasure for data breaches resulting from cyber attacks. Unlike misuse detection, ML-based methods exploit malicious and benign data samples to train models which are used for future inference on unseen samples for attack detection. Other ML-based models are also trained single-handedly on normal traffic for anomaly detection.

Despite the contribution by the existing research in relation to ML-driven data breach countermeasures, the prevalence in recent data breach incidents resulting from cyber attacks calls for an updated review of the existing works to direct future studies. This study achieves this by conducting a systematic literature review of publications between 2014 and 2023, to synthesise and comparatively analyse the countermeasures which are being taken to identify the research gaps.

A. Contributions of this study

The main contributions of this study are as follows:

1) We investigate cyber incident data breaches and present a general attack structure employed to compromise data.
2) We review the studies on ML-based detection systems proposed as countermeasures for data breaches due to prevalent types of cyber attacks (such as phishing and ransomware).
3) We propose six criteria and twenty-four sub-criteria from the reviewed literature to classify the studies.
4) We discuss the limitations of the reviewed studies and recommend future research directions.

B. The Organisation of the Paper

The rest of this survey article is organised as follows. The background and motivation of the study are discussed in section II. The methodology followed in conducting the study is described in section III. In section IV, the criteria used in classifying the study is presented. From sections V to IX, we compare and classify the studies using the criteria previously presented, such as learning tasks and classifiers, feature extraction and representation, proactive learning strategies and multimodal learning approaches. The issues identified and future research directions are outlined in section X. Related reviews are discussed in section XI.

II. Background and Research Motivation

This section introduces the background information which motivated this study. First, we present data breaches with a definition and sources. We introduce cyber data breaches as one of the prevalent sources of data breaches and present a general cyber attack structure employed in malicious breaches. Then, we show the general workflow of ML-based detection systems.

A. Data Breaches

The proliferation in data breaches has impacted governments and various industries including healthcare, education, financial services, retail and many more [2], [14], [15]. The aftermath impact of data breaches has resulted in varying levels of damaging consequences such as financial loss, reputational damage to both the breached organisation and individual victims, identity fraud and operational downtime [3], [4], [16], [17]. In 2019, for example, LandMark, a property evaluation firm lost about $7 million in revenue due to a data breach which resulted in the compromise of 137,500 unique records comprising of contact information and other sensitive property related data [18], [19]. Another data breach incident which led to the exfiltration of about 3.9 million personal data records of Medibank (an Australian insurer) customers in 2022 [20], [21], was estimated to have cost the company between $25 to $35 million.

In an attempt to define *data breach*, we present the existing definitions from various data and privacy protection regulators and government bodies:

– “A data breach occurs when the data for which your company/organisation is responsible suffers a security

TABLE I
DATA BREACH DEFINITION AND EXAMPLES

Definition	Data Breach Examples	Breach Type
A data breach is an incident involving the unauthorised disclosure, unauthorised modification and unauthorised deletion of sensitive or personal information	Gloucester City Council [27]: Resulted in unauthorised disclosure and deletion of personally identifiable and contact information.	Availability data breach Confidentiality data breach
	Courts [28]: Led to unauthorised disclosure and modification of customer personal information.	Integrity data breach Confidentiality data breach

incident resulting in a breach of confidentiality, availability or integrity." - European Commission.[1]
– "A data breach happens when personal information is accessed or disclosed without authorisation or is lost." - OAIC.[2]
– "A data breach occurs when sensitive or personal information is accessed, disclosed or exposed to unauthorised people." - ACSC.[3]
– "A data breach refers to an incident exposing personal data in an organisation's possession or under its control to unauthorised access, collection, use, disclosure, copying, modification, disposal or similar risks." - PDPC.[4]

Amalgamating the above four definitions, we present our definition of a *data breach* (also referred to as a data leak, data exfiltration or data theft) as:

> ***An incident involving the unauthorised disclosure, unauthorised modification or unauthorised deletion of sensitive or personal information.***

An example of a data breach, occurred in 2021 when a cyber ransomware group compromised the systems of Gloucester City Council and maliciously copied and encrypted the personal information of members, denying them access to the data [27]. Similarly, in 2020, another data breach incident involving a website misconfiguration resulted in the unauthorised disclosure and modification of customer information of Courts, a retail company [28]. The cited examples like all other data breach incidents either resulted in unauthorised disclosure, unauthorised modification or unauthorised removal of personal information. Table I provides example of data breaches and breach type.

1) Data Breach Sources: The sources of data breaches can be broadly categorized into *accidental* and *malicious* causes [3], [8], [9], as summarized in Table II. *Accidental* data breaches are inadvertent incidents without any malicious intent resulting from human mistakes, slips and lapses such as sending an email with a wrong recipient address (misdelivery) [10], [29], inadvertent publication of information on a website (publication error) [26], an employee forgetting to shred a document containing sensitive information before throwing it into the trash (disposal error) [26], [30] and an employee mistakenly not applying the right configurations to an installed application (misconfiguration) [29], [30]. An example of an accidental data breach for instance occurred in 2018, where an employee of Strathmore secondary school inadvertently published the health records of over 300 students school's intranet [31], [32].

Conversely, *malicious* breaches result from the exploitation of weaknesses in systems and humans, with the intention of causing havoc [3], [10], [26]. Malicious breaches include physical breaches such as the theft of a disk drive containing sensitive information, or it may involve the installation of skimming devices on terminals to steal customer data [9], [33], [34]. For example, in 2021, the retail company Costco Wholesale Corporation notified customers of a possible data breach in relation to customers credit cards after discovering a skimming device on their payment terminals during a routine security check [35]. Malicious data breaches could also result from a disgruntled insider in the organisation [5], [9], [33]. Such insiders misuse their legitimate access to systems to expose data held by an organisation mostly for financial benefits or other motives like settling scores with employers.

In 2021, a medical centre in South Georgia reported a data breach involving the unauthorised copying of protected health data of over 34,344 patients, which investigations revealed that an employee had downloaded from the hospital's systems onto a USB drive [36], [37]. Another source of a malicious data breach is an impersonation attack where a malicious actor pretends to be someone else to steal sensitive data using social engineering tactics to gain a foothold to systems for data exfiltration. Such impersonation schemes mostly consist of pretexting attacks where a threat actor is disguised as a legitimate party and initiate a dialogue usually via phone call to demand for information such as login credentials.

Studies including the 2019 to 2022 reports by OAIC [10], [26] as shown in Fig. 2 and works such as [3], [5], [25], find that the majority of data breaches are linked to cyber attacks. Such cyber incidents leading to data breaches predominantly consist of phishing, malware, ransomware, credential stuffing and application exploitation attacks such as structured query language (SQL) injection [5], [9], [10], [24]. Perpetrators of malicious cyber-related data breaches belongs to various threat groups, from highly skilled organised criminals who employ sophisticated attack tactics and tools such as ransomware, to low-skilled individuals who may leverage common cyber attack vectors like phishing to exfiltrate data from network

[1]https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/what-data-breach-and-what-do-we-have-do-case-data-breach_en
[2]https://www.oaic.gov.au/privacy/notifiable-data-breaches
[3]https://www.cyber.gov.au/threats/types-threats/data-breaches
[4]https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Data-Breach-Management/Introduction-to-Managing-Data-Breaches-2-0.pdf

TABLE II
A SUMMARY OF DATA BREACH SOURCES

Sources	Definitions	Categories
Accidental data breaches	Inadvertent incidents of no malicious intents resulting from human mistakes, slips and lapses.	Misdelivery [5] Disposal errors [40] Misconfiguration [41] Publishing errors [40] Loss [3]
Malicious data breaches	A breach incident resulting from the exploitation of weaknesses in systems and humans, with the intentions of causing havoc.	Cyber attacks [25] Physical threats [40] Insider threats [25] Social engineering (impersonation) [42]

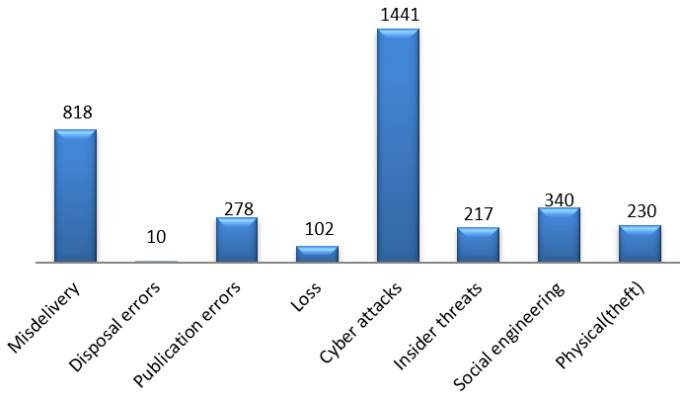

Fig. 2. Top data breach sources between 2019 to 2022.

systems. In this study, we classify data breaches resulting from a cyber-related incident involving the exploitation of vulnerabilities within a network connected system or application as a cyber data breach. Hence, our definition of a cyber data breach is:

> ***Cyber data breaches are a type of data breach (leak or exfiltration) resulting from cyber attacks such as phishing and ransomware.***

An example of a cyber data breach occurred in 2021 when a cybercriminal group known to be Uawrongteam exploited a vulnerability in FlexBooker's (an online booking system) Amazon web service (AWS) servers and installed malicious code leading to the compromise of PIIs belonging to over 3 million customers which later was sold on underground hacker forums [39]. Unfortunately, the trend is common and compromised data is regularly sold on illicit marketplaces after breached companies fail to agree on terms with cyber perpetrators.

2) Attack Structure for Cyber Data Breaches: To understand the general attack structure employed by cyber perpetrators in the compromise and exfiltration of data, we investigated data breach incidents from 2018 till now. The incidents were extracted from three data breach repositories namely, Webber insurance services [47], DataBreachDb [48] and DataBreaches [49]. DataBreaches publishes articles on data breaches, whereas Webber Insurance Services and DataBreachDb provide a list of data breaches with basic information and respective article links for further details. The criteria used for the inclusion of incident cases were based on the impact of the breach and availability of information sources about the breach form which to infer insights. Our information sources were based on reports from victim organisations, blog posts and articles produced by media outlets such as ZDNET [50], HealthItSecurity [51] and ARNET [52]. Table X in the appendix lists the forty cyber data breach incident cases which were selected from the three repositories and used to draw a general picture of the series of attack steps leveraged in data exfiltrations.

After investigating all forty cyber data breach incident cases, we identified three attack stages which is consistent with the Unified Kill Chain (UKC) [53]. In the first stage, as shown in Fig. 3, threat actors employed various attack vectors to compromise and gain an *initial foothold* into the target system. All the studied breach cases used at least one of the following initial attack vectors: *phishing* [50]–[52], *brute-force* [54], [55], *exploitation of vulnerable applications* [45], [56] or *stolen credentials* [21]. With regard to phishing-based breaches, threat actors employed social engineering schemes to lure victims mostly via email with a malicious link or attachment leading to the compromise of login credentials or the installation of malicious code. For instance in 2023, a ransomware group, Black Basta, breached Capita, a UK outsourcing firm, using phishing email leading to the installation of ransomware and the theft of personal information of pension members [57]. 55% of all investigated 40 data breach cases employed phishing for entry into target systems. The use of compromised credentials accounted for several incidents including Medibank data breach [21]. In cases like North Face data breach [54], hackers used compromised credentials obtained from previous data breaches, an attack known as credential stuffing, a variant of brute-force attack. Other forms of hacking incidents involving the exploitation of vulnerable servers and applications including web injection attacks, such

TABLE III
CYBER DATA BREACH DEFINITION AND EXAMPLES

Definition	Cyber Data Breach Example	Attack Type
Cyber data breaches are the type of data breaches (leaks or exfiltrations) resulting from cyber attacks such as phishing and ransomware.	Aveanna healthcare [43]	Phishing - Credential theft
	Norman Public schools [44]	Phishing + Ransomware
	Accellion data breach [45]	SQL injection + malware(web shell)
	Canva data breach [46]	Brute-force - Credential compromise

as structured query language injection (SQLi) and cross-site scripting (XSS) were used as vectors to gain foothold into systems [45], [56].

After initial entry, threat actors either explored and propagated through the network, or executed the final objective of compromising data in the last stage. In the *exploration and propagation* stage, threat actors either installs additional malware, execute arbitrary codes, escalate privileges or pivot to other files. In 2020 for example, cyber-criminals exploited SQLi vulnerabilities in the Accellion file transfer appliance (FTA), leading to a global data breach on over hundred companies using Accellion's legacy FTA [45]. After the initial foothold with SQLi, the threat actors further installed a malicious web shell known as DEWMODE, which extracted the files available on FTA's backend database. These files were subsequently used to exfiltrate data from the compromised system. MoveIT, a managed file transfer service [13], [58] fell victim to another global data breach when perpetrators gained entry using an SQLi vulnerability and further executed arbitrary codes resulting in privilege escalation and data exfiltration. In the majority of phishing-based credential theft incidents [43], [50]–[52] however, hackers performed the final action of *exfiltrating data* without exploration or propagation.

B. ML-Based Detection Systems

Intrusion detection systems play a pivotal role in network defense, especially in the era of IoT where connectivity has become ubiquitous. As technological advancement accelerates, cyber-attacks become more complex, hence the need for robust systems for the accurate detection of sophisticated and novel cyber intrusions resulting in data breaches. Usually, conventional detection methods which leverage on signatures of known attacks fall short to generalize. Such signature-based approaches rely on a predefined set of patterns such as the hashes of previously seen attacks to categorize network traffic. Conversely, anomaly-based systems (also known as behaviour-based system) detect intrusions using a modelled behaviour baseline, where any deviation from the baseline is considered an anomaly and regarded as an intrusion. This normal behaviour threshold could be modelled using statistical methods, machine learning or knowledge-based techniques [59]–[61].

Recently, many studies proposed as countermeasure for detection of cyber data breaches are powered by ML, given that ML models offer high accuracy and require less human input. Developing ML-based IDS requires the *collection and preparation of dataset* which may consists of real or artificially engineered data obtained from internal or external sources such as a company's internal network activities including packet flows and system logs or open source threat intelligence feeds [62]. The next phase involves *feature engineering*, which consists of the extraction, selection and representation of discriminant features to categorize classes such as legitimate versus intrusion. To select the best discriminating features, techniques such as correlation analysis, sequential search or embedded techniques like XG-Boost are applied to an initial set of features [63]–[65]. Here the notion is to select features with high entropy and variance. This feature engineering stage can however be automated using feature representation learning methods to avoid hand-crafting features for the detection system [66], [67].

Subsequently, the obtained features are leveraged to develop a *detection model*, which is either anomaly-based or classification-based [59], [60], [68]. The anomaly detection method employs a learning technique capable of identifying rare instances (intrusion) that deviate from a data's standard behaviour (such as normal network data behaviour). Well regarded anomaly models used for intrusion detection comprise Bayesian networks, K-nearest neighbour and K-means [63], [65]. On the other hand, classification-based detection uses techniques that learn simultaneously the data of two or more classes (e.g., both legitimate and malicious data) to model a borderline to separate the classes. Classification models include conventional classifiers such as Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) [63], [65]. Deep learning models such as Recurrent Neural Networks (RNN), Autoencoder NN and Convolutional NN (CNN) [59], [60] are another category of classification models which are mostly employed for representation learning.

III. SURVEY METHODOLOGY

To study the ML countermeasures proposed for cyber data breach detection, we follow a reproducible and systematic literature review approach [69], [70]. In conducting a systematic literature review, Kitchenham [69] proposed that research questions are defined first, followed by literature identification. The identified studies are checked against a defined criteria before inclusion for selection. Then, data is extracted from the collected literature, synthesised and the findings are reported using the appropriate channel. The next subsections, examine the relevant literature to address the following survey questions:

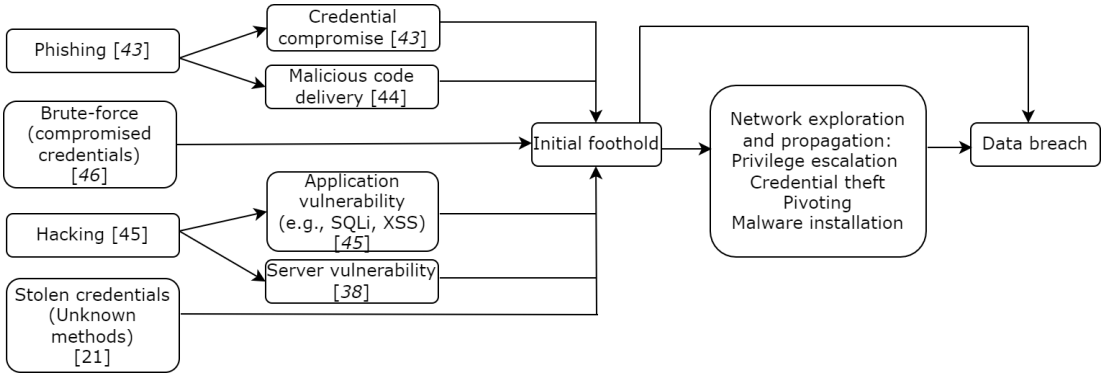

Fig. 3. A generic cyber attack structure for the forty cyber data breach incident cases.

- What modeling approaches in relation to learning tasks, classifiers, feature engineering and multimodal learning techniques are employed by ML-based cyber data breach countermeasures?
- What proactive learning strategies are used by such countermeasures?

A. Literature Identification Sources

The literature search incorporated multiple databases, with the notion that no database contains a complete set of all peer-reviewed literature. Three renowned digital databases were selected as the sources for literature collection, namely ACM library[5], IEEE Xplore[6] and Science Direct[7]. We found that both ACM and IEEE Xplore libraries can be queried using the same search syntax, while the same query string was slightly adjusted for Science Direct's search engine. To identify studies published in the last ten years, we limited the date filters associated with the library search engines to publications between 2014 to 2023.

B. Search Keywords

We constructed search terms from the research questions to identify relevant studies. The search keywords comprised four domains, namely, "detection", "cyber attacks", "data breach" and "machine learning". Cyber attacks portion in the search term was represented with prevalent attack vectors described in section II-A2 of the cyber data breach attack structure. Respective synonyms were used in conjunction with the other three domains. For example, "data breach" was expressed with other related terms like "data disclosure", "data leak" and "data exfiltration". Boolean operators "AND" and "OR" were used to join the main search terms and synonyms respectively. The following additional data sources relevant to our survey topic were searched to double check that the leading first quartile (Q1) journals and A* ranked conferences from the ACM and IEEE digital libraries had been included:

- IEEE Transactions on Dependable and Secure Computing[8].
- ACM Transactions on Privacy and Security (TOPS), formerly Transactions on Information and System Security (TISSEC)[9].
- ACM Conference on Computer and Communications Cecurity[10].
- IEEE Symposium on Security and Privacy[11].

The final search string constructed for our search was: *("intrusion detection" OR detection OR prevention) AND (ransomware OR phishing OR malware OR "sql injection" OR hacking OR "php injection" OR "brute force" OR "stolen credential" OR XSS) AND (data AND (exfiltration OR breach OR disclosure OR (leak AND (breach OR disclosure)))) AND ("machine learning" OR "deep learning")*

C. Literature Eligibility Criteria

Studies were considered for selection if they met the following criteria:

- The study proposed a ML-based countermeasure for detecting data breaches resulting from cyber attacks.
- The study was published between 2014 – 2023.
- The study was published in the English language.
- The study is available in full text version.
- The study was peer-reviewed.

D. Literature Identification Results

The identified studies were screened for relevance and quality before being included for data extraction. The titles and abstracts were reviewed to exclude studies that were not aligned to our overall survey objective. Further, the full texts of eligible studies were assessed for their true relevance. As shown in Fig. 4, our initial search resulted in 2,546 studies, comprising 1,275 papers from the ACM library, 82 from IEEE and 1189 from Science Direct. After screening for eligibility, 81 studies were finally selected for the review.

IV. Criteria for the Classification of Studies on ML-based Cyber Data Breach Detection

To classify ML-based studies on the detection of cyber-attacks, the criteria used as basis for comparison include

[5]https://dl.acm.org/
[6]https://ieeexplore.ieee.org/Xplore/home.jsp
[7]https://www.sciencedirect.com/
[8]https://www.computer.org/csdl/journal/tq
[9]https://dl.acm.org/journal/tops
[10]https://dl.acm.org/conference/ccs
[11]https://ieeexplore.ieee.org/xpl/conhome/1000646/all-proceedings

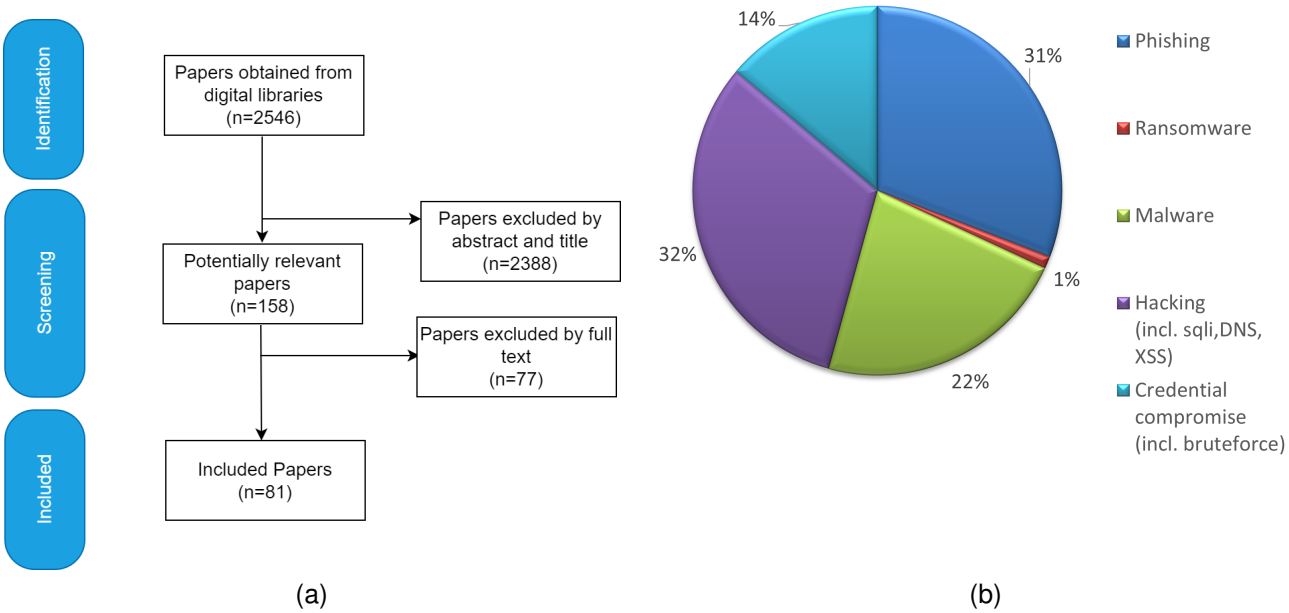

Fig. 4. Literature identification results. (a) Study selection flow diagram. (b) Distribution of selected studies based on attack vectors.

factors such as: the size and the quality of dataset, the features used as indicators of maliciousness, feature engineering approaches comprising the mode of feature extraction, feature selection and representation, learning algorithms and their structure, and model performance. These factors have become *de facto* criteria for the classification of ML-based studies in multidisciplinary research areas [71]–[73]. In addition to these factors, the proactiveness of a study is another important criterion to be considered in the evaluation of ML-based research. For example, ML-based studies can be classified by their ability to self-initiate (i.e., automate operations without human inputs) and their adaptability to environmental changes. Additionally, the data modalities used for training should not be overlooked in the classification of ML-based research. Multimodal learning has become an active area in ML-based studies, since the fusion of different types of information has the potential to enhance performance. In consideration of the above factors, we comprehensively classify ML-based cyber data breach detection studies based on the following six criteria, as shown in Fig. 5:

- **Learning task:** we classify studies by the inference type used, which usually depends on the problem being solved and the available data. Studies may approach the detection of cyber data breaches using one or more of the following five learning tasks:
  - *Supervised anomaly*: this inference technique aims to detect anomalies from an imbalanced labelled dataset comprising normal and abnormal samples [74].
  - *Unsupervised anomaly*: this task involves the identification of abnormal behaviours or outliers in an unlabelled dataset. Several studies [75], [76] employ this inference type in instances where outliers are rarely seen.
  - *Semi-supervised anomaly*: in this inference approach, the model is trained with a labelled but imbalanced dataset, to generate pseudo labels with the trained model to detect unusual behaviour from a given unlabelled data.
  - *Semi-supervised learning*: this method is used to also automate data labelling using a model trained on a small but balanced labelled dataset [77].
  - *Classification learning*: this inference technique uses a balanced labelled dataset to supervise learning to detect malicious activities [77], [78].
- **Feature extraction:** the classification of studies is based on approaches used in obtaining and converting discriminative features about data into formats suitable for ML algorithms. Studies may leverage one or more of these six feature extraction approaches:
  - *Context-based*: an approach of using the semantic relationships in data to extract features. Examples include, n-grams of word tokens or sub-word embedding features [82], [86].
  - *Statistical-based*: a method of extracting features based on the statistical properties of data, for

example, the frequency and mean of network traffic flows [87].
    - *Time-based*: this approach relies on time attribute to extract features, such as the round trip time (RTT) of a network packet [79], [88].
    - *Content-based*: a straightforward approach, where feature extraction is based on the occurrence of some information in data, e.g., the appearance of numbers or IP in a URL of an http request [86], [89].
    - *Environmental-based*: this category of feature extraction relates to where the attack transpired and the level of harm caused. For example, if an attack is launched via admin or staff web form, and the level of damage is either small, medium or low [90].
    - *Behavioural-based*: this method relies on program execution attributes, direction of data connection or flow, or the general behaviour of an application or user to extract features. This includes system call executions and inbound or outbound connection of network packet [74], [91].
- **Feature representation:** studies are classified based on how features are represented (modelled) as input for training with an ML model. The following are feature representation techniques:
    - *Image-based*: this approach models the extracted features as images for training or for further extraction of features from the input image. Image-based representations are mostly composed from raw content of the data (such as web page snapshot) or feature vectors stacked as 2 or 3-dimension [92], [93].
    - *Vector-based*: features are represented as vectors usually consisting of binary, statistical, categorically encoded and word embedded values [88], [94].
    - *Topology-based*: this method represents features as graphs composing of vertices(nodes) and edges to capture the complex and non-linear relationships in data [77].
    - *Sequence-based*: this representation technique models features as vector sequences prior to training with sequence based ML classifiers like recurrent neural networks [95].
- **Learning classifier:** the studies are classified by classifier type under the umbrella of conventional ML and deep learning:
    - *Conventional ML*: these are traditional ML algorithms which mostly rely on relatively few and human-engineered features to train with. Such algorithms include Support Vector Machine (SVM), Decision Trees (DT) and Naïve Bayes (NB) [79].
    - *Deep learning*: these are ML models with non-linear functions mostly employed to extract features on a large volume of data. Feed forward and convolutional neural networks are examples of deep learning [80], [81].
- **Proactive strategies:** we classify studies based on self-initiation strategies and adaptiveness to future changes (such as drifts in data). The following four proactive ML strategies are identified:
    - *Re-training strategies*: these strategies update ML-models to be adaptable to environmental changes such drift in feature distribution of data [82].
    - *Data augmentation*: a proactive method of addressing class imbalance through the creation of synthetic samples in dataset using techniques like Synthetic Minority Oversampling Technique (SMOTE) [83], [84].
    - *Feature extraction automation*: this approach uses methods like TF-IDF and word2vec to automate feature extraction [85].
    - *Self or automated labelling*: this strategy employs techniques to automate the annotation of unlabelled dataset, without requiring for human knowledge [77].
- **Multimodal approaches:** we classify studies by the approach employed in combining multiple modalities of data. Studies are classified as one of the following three (3) multimodal strategies:
    - *Early-fusion*: in this approach , multiple unimodal features are concatenated at the input level to create a single representation [96].
    - *Intermediate-fusion*: the output (representation) of models trained with separate unimodal features are combined, forming a shared intermediate representation, which is further used as input to another ML model [89].
    - *Late-fusion*: also referred to as decision-based fusion, an ensemble technique is employed to combine prediction probabilities of models trained with separate unimodal features [96].

V. Survey of studies by learning task

This section discusses the type of inferences used in the studies to mitigate cyber data breaches. The choice of inference type depends on the nature of task being solved, or the availability of data for the task. The learning tasks employed can be broadly classified into *anomaly* and *classification* detection. Figure 6 shows the statistics on classification and anomaly learning tasks, and their mapping to cyber data breach vectors.

Classification-based tasks. When balanced data (i.e., equal portions of malicious and benign samples) is available, the detection of cyber data breaches is likely to be approached as a classification problem. Our review shows 64 studies applied classification detection to data exfiltration related attacks. Phishing, malware and hacking (web and DNS) represent the highest percentage of attacks detected using classification approaches as shown on Fig. 6b. About 99% of phishing-related studies applied classification detection. This could be due to the pervasiveness of phishing attacks yielding to the availability of data to train ML models. This was however

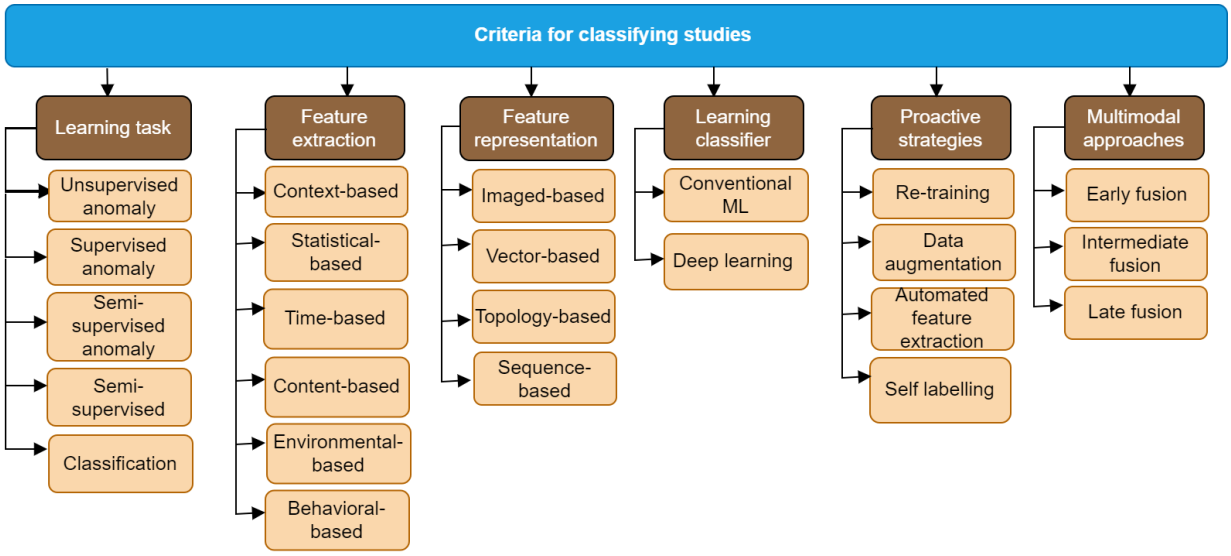

Fig. 5. Criteria for classifying studies on ML-based cyber data breach detection.

not the case for Saka et al. [86], who employed unsupervised clustering algorithms such as K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and agglomerative on unlabelled email data to identify phishing scams. Three studies [90], [97], [103] combined anomaly detection and classification for the detection of credential, web and malware related attacks using ML models trained with HTTP and network flow traffic. Han et al. [77] combined semi-supervised learning and classification, with the former using a graph based model to automate the annotation of unlabelled email into known and unknown phishing campaigns and the latter for classifying email into different campaign categories.

Anomaly-based task. Overall, 16 studies from a total of 81 applied anomaly detection for the purpose of identifying outliers or abnormal behaviours on networks, applications and hosts systems. *Unsupervised anomaly* is the most commonly used learning task (i.e., 14/16 studies) of the other anomaly detection techniques. This is because, malicious traffic is rarely seen on networks, hence studies employ unsupervised anomaly techniques to train ML models on unannotated data. Two studies [74], [97] leveraged *supervised anomaly* on an imbalanced labelled dataset containing a few abnormal network traffic samples to aid in the detection of data breaches resulting from bruteforce and other network attacks. Anomaly-based detection methods are predominantly used for tasks pertaining to the identification of credential theft (such as bruteforce) [76], [98] and web attacks [99], [100], but are least used for phishing related attack detection as shown on Fig 6b. Anomaly detection techniques are generally suitable in environments where the model relies on network traffic (be it flow or packet) and system logs for training and identification of attacks resulting to data exfiltration [75], [98], [101], [102], since in such settings anomalies are rarely encountered. To mimic a real-world scenario where anomalous traffic are rarely seen, Bohara et al. [101] adopted unsupervised anomaly and trained ML clustering models on an unlabelled network and system logs, comprising a majority of normal logs, to detect possible exfiltration attacks resulting from bruteforce on enterprise network.

VI. Survey of Studies by Feature Extraction and Representation Techniques

In this section, we cover feature engineering techniques, with a focus on the extraction and representation techniques employed by studies for modelling ML-based countermeasures for cyber data breach detection. We present a taxonomy of the various methods used for feature extraction, and also a taxonomy of the techniques used by studies for modelling the extracted features into suitable representations used as input to a ML classifier, as shown in Fig. 7.

A. Feature Extraction Techniques

Feature extraction is an important phase in the detection of cyber data breaches using ML, since the resulting discriminative features have the potential to lift the model's performance. From the reviewed studies, it is clear that feature extraction approaches can be classified as one of the following six methods: statistical-based, context-based, content-based, time-based, behavioural-based and environmental-based. As shown in Table V, studies may either use one extraction method or fuse several to extract the same kind features or hybrid features respectively.

Statistical-based. This feature extraction approach harnesses the statistical information in data to serve as discriminative features for detecting cyber data breaches. This mode of feature extraction was the most utilised extraction method from the surveyed studies (i.e., 51/81 studies). Statistical methods adopted by the studies for feature extraction include, averages, sizes or length, frequencies, standard deviation, summations, min-max, ratios, entropy and median of data [78], [85], [94], [101]. A total of 20 studies leveraged these techniques to solely extract statistical features, whereas 31 studies combined statistical techniques with other extraction methods. Verma et

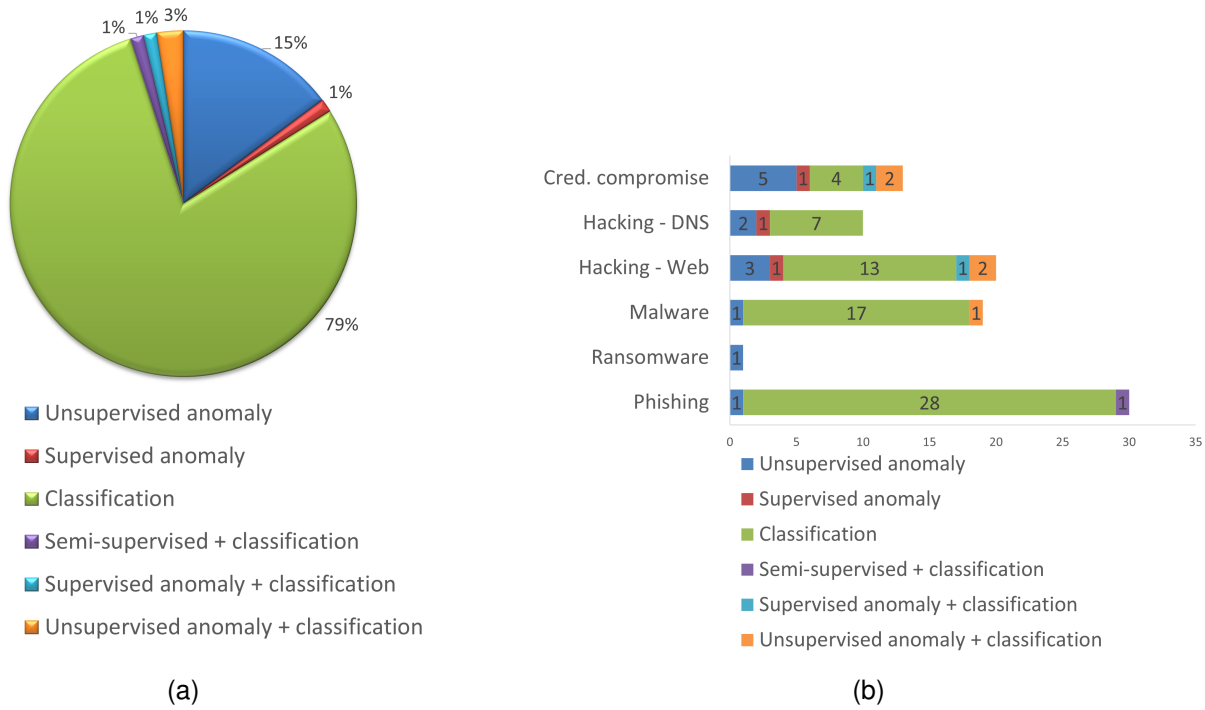

Fig. 6. Statistics on learning task of ML-based countermeasures for cyber data breach detection. (a) Distribution of learning tasks. (b) Mapping learning tasks to cyber data breach vectors.

TABLE IV
COMPARISON OF DIFFERENT STUDIES ON ML-BASED CYBER DATA BREACH DETECTION BY LEARNING TASK (√: YES, ×: NO)

Attack Vector	Research	Year	Unsupervised anomaly learning	Supervised anomaly learning	Semi-supervised learning	Classification learning
Hacking (incl. web and DNS)	[104]	2014	×	×	×	√
	[105]	2014	×	×	×	√
	[99]	2019	√	×	×	×
	[100]	2019	√	×	×	×
	[81]	2021	√	×	×	×
Ransomware	[75]	2021	√	×	×	√
Credential compromise (incl. bruteforce)	[101]	2016	√	×	×	×
	[98]	2019	√	×	×	×
	[106]	2021	√	×	×	×
	[107]	2022	×	×	×	√
	[103]	2022	√	×	×	√
	[97]	2023	×	√	×	√
Phishing	[77]	2016	×	×	√	√
	[108]	2019	×	×	×	√
	[109]	2020	×	×	×	√
	[110]	2021	×	×	×	√
	[86]	2022	√	×	×	×
	[111]	2022	×	×	×	√
	[112]	2023	×	×	×	√
	[113]	2023	×	×	×	√
Malware	[114]	2017	√	×	×	×
	[115]	2022	×	×	×	√
	[103]	2022	√	×	×	√
	[116]	2023	√	×	×	×

al. [78] proposed character frequency feature vectors obtained using statistical approaches such as Kolmogorov-Smirnov, Kullback-Leibler (KL) divergence, Euclidean distance, edit distance and normalised frequencies to train ML models to detect phishing related data breaches from a given URL. In the same work, they combined these statistically extracted features with content based features for the same detection task. The majority of the statistical-based studies extracted features manually, whereas four of them [85], [100], [117], [128] automated the process of extraction. All four studies leveraged Term Frequency - Inverse Document Frequency (TFIDF) in automating the mining of frequency information from data for cyber data breach detection using ML.

Context-based. A total of 11 studies used contextual rep-

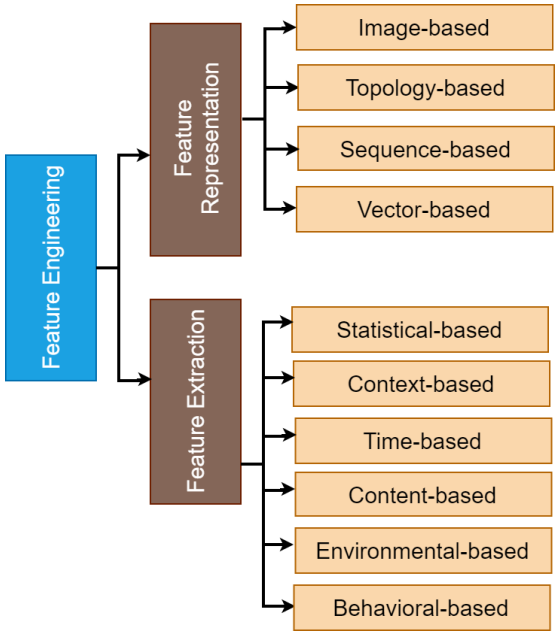

Fig. 7. Taxonomy of feature extraction and representation techniques for ML-based cyber data breach detection.

resentations in data as the basis for feature extraction. Such approaches rely on the sequence, semantics and structural properties in data to mine features. Contextually extracted features were used by 9 studies [82], [85], [92], [98], [99], [110], [113], [118], [119] as standalone features, while the remaining 2 studies [100], [120] combined them with statistical-based features, as detailed in Table V. In the work of Li et al. [99] where they proposed anomaly-based web attacks detection, context-based extraction was applied on HTTP URL logs to generate semantic feature vector using the word2vec embedding model. In contrast, Zhang et al. [120] fused semantically embedded HTTP request feature vectors (also obtained using word2vec) with statistical features obtained from network flow data to detect SQLI attacks leading to possible data breaches.

Content-based. A total of 34 out of the 81 studies utilised the content-based approach for extraction of features for cyber data breach detection. This was the second most used feature extraction approach utilised in the studies. The reason for this could be due to the ease of extraction. As shown in Table V, 11 studies used content-based features for data exfiltration detection, while the remainder concatenated these with features obtained from other extraction methods. Several studies including [2], [78], [104], [108] solely extracted content-based features from URLs, domain names and network flows to detect data exfiltration resulting from attacks such as SQLI, phishing, credential bruteforce and malware. On the other hand, a significant number of studies [77], [79], [121], [122] combined content-based extraction with time and statistical-based methods to complete the same task.

Time-based. This method extracts features based on some time component associated with the data. One study [105] used only time-based features to detect malicious domains from DNS traffic. Of the 81 studies, 20 leveraged time-based features merged them with features obtained from other extraction methods. Kondracki et al. [123] for example extracted time-based features comprising of the round trip time (RTT) of network packets, and fused them with content-based features from HTTP packets (such as the version of TLS), which together were used to train a random forest (RF) classifier to detect phishing related data breaches.

Other extraction methods. The remaining two feature extraction methods, environmental and behavioral based, were utilised to generate auxiliary features, since such features were not used alone. Behavioral-based features were used in 5 studies and are often combined with statistical, content and time-based features. Behavioral features such as the direction of network traffic flow and packet connection direction which were extracted by studies such as [74], [91], [124], were merged with features from other extraction techniques to detect data exfiltration resulting from malware, web application hacking and credential bruteforce. Environmental-based features were used once by Ivanova et al. [90] to extract features such as the environment where the attack took place (e.g., whether attack was performed via student or admin form), and the severity of damage on the environment. These features were fused with three other categories of features to aid in the detection of web attacks leading to possible data breaches.

B. Feature Representation Techniques

The next phase of feature engineering involves the modelling of the extracted features into representations suitable for input to ML classifiers. Classifiers may further extract features from these representations or be base on them for inference on attacks leading to possible data breaches. Studies may leverage any of the following four techniques to represent the extracted features: vector-based, sequence-based, topology-based and image-based. This section discusses the four feature

TABLE V
SUMMARY OF FEATURE EXTRACTION, TYPES AND REPRESENTATIONS FOR CYBER DATA BREACH DETECTION

Feature Extraction Technique	Feature Type	Topology-based	Image-based	Sequence-based	Vector-based	Papers	Description
Statistical-based	HTTP traffic (8)	1	-	-	7	[78], [81], [85], [125]–[129]	The statistical approach of feature extraction harnesses statistical information in data to serve as discriminative features, for example, the frequency and mean of network traffic flows.
	DNS traffic (1)	-	-	-	1	[105]	
	Host logs (1)	-	-	-	1	[101]	
	Network traffic (7)	1	-	1	5	[83], [84], [106], [107], [117], [130]	
	HPC events (1)	-	-	1	1	[95]	
	Network logs (1)	-	-	-	1	[101]	
	API calls (1)	-	-	-	1	[94]	
	App behaviour (1)	-	-		1	[114]	
Context-based	HTTP traffic (6)	-	2	1	3	[82], [85], [99], [113], [118], [119]	The context-based extraction approach involves the use of semantic relationships in data to extract features.
	Host log (1)	1	-	-	-	[98]	
	Opcode (1)	-	-	-	1	[92]	
	CT logs (1)	-	-	1	-	[110]	
Content-based	HTTP traffic (9)	-	4	1	4	[104], [108], [131], [132]	In the content-based method, feature extraction is based on the occurrence of some information in data, e.g., the appearance of numbers or IP in a URL.
	Network traffic (1)	-	1	-	-	[2]	
	binaries (1)	-	1	-	-	[92]	
Time-based	DNS traffic (1)	-	-	-	1	[105]	The time-based method extracts features based on some time component associated with the data.
Statistical + Context	Network traffic (1)	-	-	-	1	[120]	Consists of features extracted using statistical and context methods.
	HTTP log (1)	-	-	-	1	[100]	
Statistical + Time	HTTP traffic (1)	-	-	-	1	[109]	Made up of statistical and time extracted features.
	Network traffic (4)	-	-	-	4	[2], [87], [133]	
	DNS traffic (2)	-	1	-	1	[105]	
	Host log and process parameters (1)	-	-	-	1	[134]	
Statistical + Content	HTTP traffic (6)	-	-	-	6	[78], [132], [135]	Consists of dual features extracted using statistical and content-based methods.
	Network traffic (1)	-	-	-	1	[121]	
	DNS traffic (1)	-	-	-	1	[136]	
	CT logs (1)	-	-	-	1	[130]	
Content + Time	HTTP + Network traffic (1)	-	-	-	1	[123]	Made up of content and time based features.
Statistical + Content + Time	HTTP traffic (3)	1	-	-	2	[77], [79], [137]	Consist of the fusion of statistical, content and time-based features
	DNS traffic (2)	-	-	-	2	[88], [102]	
	HTTP log + network info (1)	-	-	-	1	[122]	
Stats + Cont + Time + Beh	Network traffic (3)	-	-	-	3	[74], [91], [124]	Combines statistical, content, time and behavioral based features.
	DNS traffic+ network information (1)	-	-	-	1	[138]	
Stats + Cont + Beh + Env	HTTP traffic (1)	-	-	-	1	[90]	Quadruple features consisting of statistical, content, behavioral and environmental based features.

representation techniques in relation to extraction approaches, the classifiers employed and the attack vectors.

Image-based. As shown in Table V, a total of 9 studies modelled extracted features as image representations. Such representations are mostly composed from the raw content of the data (such as a web page snapshot) or feature vectors stacked as 2 or 3-dimension. Hussain et al. [113] generated URL character embeddings and formed 2D image representation for URLs by stacking the semantic vector embeddings, which was used as input for a one-dimensional convolutional neural network (1D-CNN) for further feature extraction and the prediction of phishing URLs. Bac et al. [93] approached a similar phishing detection task using image representations of website snapshot (content-based extraction) as input to deep learning models. Similarly, the work by D'Angelo et al. [139] also constructed a 2D image representation of statistical features from DNS traffic and fed this as input to 2D-CNN for the detection of malicious DNS tunnels resulting in data breaches. As shown in Fig. 8, image-based representations are most employed for phishing detection (i.e., 5 studies), and are mostly used as input to CNN models as shown in Fig. 9 to extract features from the input image.

Vector-based. The vector-based technique was the most used form of feature representation (i.e., 57 studies) for the detection of cyber related data breaches using ML. Vectors usually consist of binary, statistical, categorically encoded and word or character embedded values, extracted using the 6 feature extraction approaches. Vector-based representations are predominantly used because the majority of ML classifiers as shown on Fig. 9 require feature vectors as their input.

Topology-based. As shown In Table V, four studies leveraged topological (graph-based) techniques to represent extracted features as input to varying graph-based ML models. For instance, Han et al. [77] tackle email phishing campaign attribution using a topology-based model. In their graph formulation, each email feature vector was used as a node feature and edges consisted of the similarity between the email node features. The resulting graph representations were passed as input to the K-nearest neighbouring (KNN) graph model for the detection and attribution of phishing campaigns. Liu et al. [98] also modelled user behaviour from network logs using heterogeneous graphs and employed random walk and word2vec models to generate vector embeddings from the graph representations. These node vector embeddings were used as input to the pair-wise similarity clustering algorithm to detect credential and other network attacks leading to data breaches.

Sequence-based. Five studies approached the detection of cyber data breach attacks as a sequence classification problem. These studies consider the sequential relationship within data (including network flows, HTTP traffic, certificate transparent logs and hardware performance counters, as shown in Table V) to model sequential events reflecting user or application behaviours. Sequence-based models [95], [110], [118] either represent extracted feature vectors as sequences or represent values in one feature vector as sequences and employ recurrent neural network models including Long-Short Term Memory networks (LSTM) and Echo States Networks (ESN), as shown in Fig. 9 to train such sequence representations to make inference on cyber data breaches.

VII. Survey of studies by learning classifiers

This section discusses the learning classifiers used by the reviewed ML-based studies for data breach detection resulting from cybersecurity attacks. As shown in Table VI, we categorise the classifiers into conventional and deep learning, and give insights into their mapping to the prevalent cyber data breach vectors. Fig. 9 shows the classifiers and their usage of the various feature representation methods.

Conventional Classifiers. Our review shows that in the selected studies, the most frequently used conventional classifiers include Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbour (K-NN) and Decision Tree (DT). RF was the classifier most commonly selected in the studies (i.e., used in 32 instances) for data breach detection, resulting from phishing, malware, web, DNS and bruteforce-based credential thefts. RF classifiers employ a mechanism to improve prediction by fitting and computing the averages of multiple DT classifiers. While RF was predominantly used as a base classifier by studies such as [77], [83], [94], [102], Verma et al. [78] instead built stacked learners using RF together with conventional counterparts like LR, Sequential Minimal Optimization (SMO) and J48 in their proposed phishing detection framework, and employed the ZeroR classifier as the meta-learner. Other works [85], [99], [140] trained ensemble learners using conventional classifiers like K-means, consensus clustering, spectral clustering, AdaBoost, DT and K-NN to detect data exfiltration attacks resulting from phishing and web-related attacks. Several studies including [84], [109], [133] also adopted the use of multiple conventional classifiers to perform the same data exfiltration detection task. For example, Sethi et al. [94] compared the performance of three classifiers (DT, RF and SMO) on their proposed malware (backdoor and spyware) detection framework and reported the best performing classifier (DT) attained 100% detection accuracy.

Deep learning. Common DL techniques employed for cyber data breach detection as shown in Table VI include convolutional neural networks (CNNs), recurrent neural networks (such as LSTM and ESN), transformer networks and the vanilla feed forward networks. CNN was used to detect 13 instances of cyber data breach vectors as shown in Table VI, mostly for the extraction of features from image represented cyber data. Such extracted features are further fed into conventional classifiers or deep learning models (mostly fully connected FFN) for inference on data exfiltration. Bac et al. [93] proposed a transfer learning technique for phishing detection using varying CNN models like VGG16/19, InceptionV3 and Xception, which were all explored for feature extraction for downstream classifiers including K-NN, NB, LR RF and FFN. Similarly, D'Angelo et al. [139] attempted DNS tunnel detection using a CNN-based feature extractor on DNS flow represented images with a fully connected FFN. Like CNNs, recurrent NN models such as Long Short-Term Memory (LSTM) were also employed to detect 8 instances

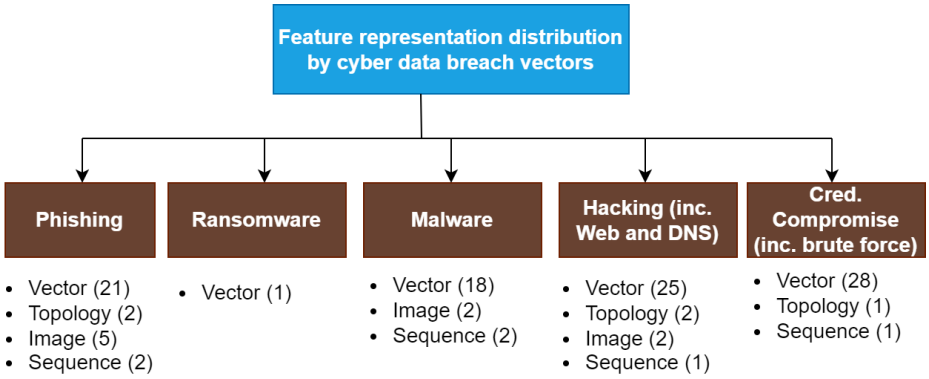

Fig. 8. Illustration of feature representation methods based on cyber data breach vectors.

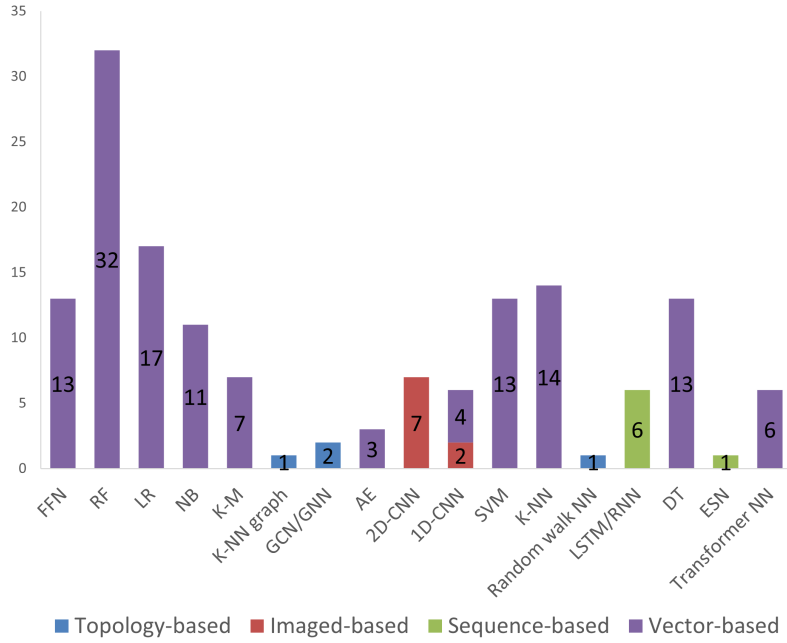

Fig. 9. Distribution of feature representation methods by learning classifiers.

of data exfiltration attacks, serving as a feature extractor for sequence represented cyber data. Elsayed et al. [130] studies the detection of web and credential bruteforce attacks and proposed an unsupervised anomaly detection model consisting of an LSTM-based autoencoder and OC-SVM. Likewise, Jishnu et al. [118] combined LSTM and transformer neural networks for URL phishing detection.

VIII. Survey of Studies by Proactive Learning Strategies

This section discusses the proactive methods employed by ML-based countermeasures for the detection of data exfiltration. Two categories of proactive strategies were identified from the reviewed studies. The first category describes self-initiation strategies which include how studies automate the extraction and annotation of features and labels respectively without any human input, and how studies resolve the issue of imbalanced dataset using data augmentation techniques. The second category describes the methods used by studies in adaptating to future environmental changes (such as a drift in data). Table VII summarises the proactive strategies used by the reviewed studies.

Feature extraction automation. A total of 13 studies explored varying models to automate the extraction of features for training ML-based data breach detection systems. The most commonly used FE techniques were TF-IDF (5 studies), word2vec embedding models (4 studies) and transformer models (3 studies). These automated FE models were applied predominantly for the detection of hacking (i.e., web) and phishing related data exfiltration attacks. In the work of Zhang et al. [120] where a deep belief network was proposed for SQL injection attacks detection, they leveraged on the word2vec to train an embedding layer to extract context-based features of HTTP POST and GET requests. For phishing detection,

TABLE VI
MAPPING OF ML CLASSIFIERS WITH CYBER DATA BREACH VECTORS

ML Types	Classifiers	Phishing	Ransomware	Malware	Hacking - Web/DNS	Credential. Comp	Papers
Conventional ML	Random Forest (32)	13	-	7	9	3	[77]–[79], [102], [114], [137]
	Logistic regression (17)	9	-	5	3	-	[93], [114], [122], [125], [134]
	Naive Bayes (13)	7	-	5	1	-	[112], [117], [125], [141]
	K-means (7)	1	1	-	4	1	[90], [99]–[101]
	DBSCAN (3)	1	-	-	1	1	[86], [101], [138]
	KNN-graph (1)	1	-	-	-	-	[77]
	Decision Tree (13)	6	-	4	2	1	[94], [112], [133], [134]
	MinibatchKmeans (1)	-	-	-	1	-	[99]
	SVM (17)	7	-	4	5	1	[84], [88], [135], [142]
	K-NN (16)	5	-	5	5	1	[83], [125], [131], [143]
	AdaBoost (8)	3	-	1	3	1	[74], [112], [117]
	OC-SVM (1)	-	-	-	-	1	[130]
	Gaussian NB (1)	-	-	-	1	-	[88]
	iForest (1)	-	-	-	1	-	[127]
Deep Learning	Feed forward network (13)	4	-	4	3	2	[81], [91], [92], [97]
	Echo state network (1)	-	-	1	-	-	[95]
	TCN (1)	-	-	-	1	-	[145]
	Transformer (6)	3	-	1	2	-	[2], [82], [85], [118]
	GCN/GNN (2)	1	-	1	-	-	[107], [128]
	LSTM/RNN(8)	3	-	1	2	2	[76], [103], [118], [131]
	Random walk NN (1)	-	-	-	-	1	[98]
	CNN (13)	5	-	2	4	2	[93], [108], [114], [119]
	Autoencoder(4)	-	-	1	2	1	[95], [100], [103], [130]
	DBN (1)	-	-	-	1	-	[120]

Bountakas et al. [85] exploited three different feature extraction models comprising TF-IDF, BERT and word2vec to automatically learn the representation of email content data.

Self labelling. Machine learning models generally require an annotated dataset for supervised learning. The manual annotation of volumes of data can be time consuming, costly and error prone. This shows the importance of automated (self) annotation. In automating the annotation of unlabelled email, Han et al. [77] trained a semi-supervised ML-model on a small portion of a manually labelled email campaign, which was further used to annotate a greater percentage of unlabelled email campaigns for spear phishing detection. Other studies including [82], [85], [118] have also exploited deep learning models like word2vec and BERT, which are largely regarded as self-labelling models. These models leverage on the input data as labels for supervision. A proposed framework by Gniewkowski et al. [82] exploited the BERT model on unlabelled HTTP server logs for the detection of data exfiltration resulting from anomalous HTTP and malicious URLs.

Data augmentation. A common issue faced by classification learning tasks is the imbalance in dataset labels. Often this issue results in a biasd model due to the dominance of one class. One approach to remediate the class imbalance problem is by augmenting the dataset, which involves either the modification of the existing data or the exploitation of inherent properties within the classes to increase the size of the minority class. Research [146] shows in addition to increasing the size of dataset, data augmentation techniques can enhance the quality of the data to boost a model's performance. Seven studies were identified having applied data augmentation in their proposed ML-based data breach detection system. All seven studies either have used SMOTE or border-SMOTE (a variant of SMOTE) techniques to generate synthetic samples for the minority class. Several studies including [74], [84] exploited the SMOTE oversampling strategy to augment network and HTTP request traffic for the detection of data exfiltration resulting from web and credential compromise attacks. Similarly, [119], [131] have leveraged varying SMOTE methods to augment the

TABLE VII
PROACTIVE LEARNING STRATEGIES EMPLOYED BY STUDIES.

Proactive learning strategy	Description	Approaches	Papers
Feature extraction automation	This approach leverages on algorithms (mostly deep learning) to automate the extraction of features without the need for human input.	Automated feature extraction with models including word2vec, TF-IDF, Glove, BERT and countvectoriser to detect phishing, credential bruteforce and hacking (web and DNS) related data breaches.	[76], [85], [98], [99], [120]
Self or automated labelling	This strategy employs techniques to automate the annotation of unlabelled datasets, without the need for human knowledge.	• Annotation automation of unlabelled email campaigns using semi-supervised learning model. • Self-annotation using word2vec and transformer models	[77], [100], [117], [144]
Data Augmentation	A method of enhancing dataset through the modification of existing data or the exploitation of inherent properties within data, to increase the size (i.e., addressing class imbalance) and quality of the dataset to boost the model's performance.	Data augmentation using SMOTE and borderline-SMOTE algorithms to generate synthetic HTTP and network traffic for data exfiltration detection.	[74], [84], [119], [131]
Re-training	A proactive approach of continuously training to update ML models to be robust against novel attacks.	Retrained ML-based phishing detection models using Learning Without Forgetting (LWF) and Elastic Weight Consolidation (EWC) techniques.	[147]

detection of malicious emails and URL related data breaches.

IX. SURVEY OF STUDIES BY MULTIMODAL APPROACHES

This section focuses on ML-based studies which employ multiple modalities of cyber data to detect data exfiltration. Although a predominate number of studies depend on one type of information for attack detection, a handful of studies exploit a mix of multiple feature types, as shown in Table VIII. Such studies may use one or more of the following three fusion techniques to combine various modalities of data to enhance performance: *early fusion, intermediate fusion and late fusion*.

Early fusion. This is the most used fusion technique (11 studies) for combining multiple information types of cyber data. The early fusion strategy simply concatenates features of two or more modalities of data or information types into a single feature vector. Such a technique of fusing multiple unimodal feature vectors is computationally efficient, hence the reason for its widespread adoption. Bohara et al. [101] combined frequency-based feature vectors extracted from network and system logs data, and used this joint representation to detect credential bruteforce attacks. Phishing detection studies such as [77], [79], [109], [137] have all exploited the early fusion technique to create a single feature representation from feature types including HTTP URLs, certificate transparency (CT) logs, webpages, email texts and external service information as shown on Table VIII. Oprea et al. [122] combined statistical, content and time based features extracted from HTTP log and network information data into a single feature vector to detect data breaches resulting from malware.

Intermediate fusion. This approach of combining multiple unimodal features was leveraged by three studies [110], [132], [144]. With the intermediate technique, each unimodal feature is trained separately using a single or multiple ML models. The resulting intermediate outputs from the models are then combined into a single representation and used as input for another model. Drichel et al. [110] for example exploited uni and bi-directional LSTM models to obtain intermediate features from domain and certificate based CT logs, which were further fed into a fully connected feed forward layer to detect data breaches resulting from phishing attacks. A similar phishing detection model proposed by Miao et al. [132] also employed LR as a meta-leaner for the merged features obtained from ML models trained with URLs and webpage feature vector.

Late fusion. The late fusion (also referred as decision-based fusion) approach trains multiple ML models using different unimodal features similar to the intermediate fusion approach but output predictive values instead of a feature representations. These predicted values are further combined using ensemble techniques such as voting or weighted average. Bountakas et al. [144] proposed a phishing detection approach leveraging two feature modalities of email (content and text-based features) to train two ML models, DT and KNN. As their late fusion strategy, they use a soft voting ensemble to average the prediction probabilities of the ML models to predict whether an email is associated with a phishing attack. Pierazzi et al. [141] also explored late fusion for malware detection using static and dynamic features including permissions, network activities, and authors to train separate models with conventional and deep learning classifiers, and employed a late fusion ensemble to combine prediction probabilities to detect data breaches resulting from spyware.

Performance. Different evaluation metrics, as shown in Table VIII are employed to analyse the performance of studies which used multiple feature types for data exfiltration detection. This include metrics such as accuracy (ACC), precision (PREC), recall (REC), Area Under Curve (AUC), F-score (F1) and False Positive Rate(FPR). Our analysis show that, the use of multiple modalities of cyber data enhances the detection of attacks. In [79], Hannousse et al. trained separate ML models with hybrid and unimodal features to detect data breaches resulting from phishing attacks. Using a random forest classifier trained with single features, they obtained the best accuracy score of 94.09%. However, by combining three modalities, namely URL, email and external service features,

TABLE VIII
COMPARISON OF DIFFERENT STUDIES BY MULTIMODAL APPROACHES SUCH AS FEATURE TYPES AND FUSION TECHNIQUES.

Attack Vector	Fusion Strategy	Feature Types	Description	Performance (%)	Research
Phishing	Early fusion	Email body text, attachments, recipient, origin	Combined all four feature categories into a single feature vector.	94.8(REC)	[77]
Malware	Early fusion	Network flow and packet traffic	A concatenation of statistical, behavioral, content and time based features extracted from network flow and packet traffic.	<0.6(FPR)	[124]
Hacking (web)	Early fusion	Network traffic (packet) and HTTP traffic (POST/GET data)	Extracted statistical and context-based features from network and HTTP traffic into a single feature vector.	96(ACC)	[120]
Phishing	Early fusion	ULR lexical and ULR's host-based	Extracted lexical and host based features into a single feature vector.	94.9(ACC)	[109]
Malware	Late fusion	static (permission,author, etc) and dynamic (network activities, etc)	Ensemble late fusion: combines predictions from conventional and deep leaning classifiers using the weighted sum technique.	98.2(AUC)	[141]
Phishing	Intermediate fusion	Domain-based and certificate based features of CT logs	Learned representations from CT logs using uni and bi-directional LSTMs, amalgamated their output and fed into fully connected layer.	NA	[110]
Malware	Early fusion	HTTP log and network info (WHOIS, geolocation)	Combined statistical, content and time-based features from HTTP log and network information data into a single feature vector.	97(PREC)	[122]
Phishing	Intermediate and late fusion	HTTP traffic (email content-based and email text-based)	*Intermediate, stack ensemble*: Exploits DT and KNN as base classifiers and employs FFN as meta-learner on their output. *Late Fusion*: Prediction probabilities of the two base learners are averaged using soft voting ensemble.	99.4(F1)	[144]
Phishing	Early fusion	HTTP traffic (URL and email content) and external service features	A unified feature vector obtained from URL and external services data(including WHOIS and openpagerank for page index).	96.6(ACC)	[79]
Phishing	Intermediate fusion	URL and Webpages	Stacking ensemble: Trained an LR on the output of two separate classifiers also trained using URL and webpage features.	98.6(ACC)	[132]
Hacking (DNS)	Early fusion	DNS traffic and network information (WHOIS and BGP)	Fused feature vectors extracted from DNS traffic and external network sources.	<4(FPR)	[138]
Credential compromise	Early fusion	Network and host logs	Concatenated network and host based features into a single feature vector.	NA	[101]
Malware	Early fusion	Network traffic (flow) and App behavior patterns	Combined network flow-based feature vectors with behavioral pattern weight vectors using Dynamic Time Wrapping (DTW).	99.1(PREC)	[114]
Phishing	Early fusion	HTTP traffic (URL) and External service features (WHOIS)	Concatenated URL features with external service features.	97.5(ACC)	[137]

the detection accuracy increased to 96.6%. Zhang et al. [120] also combined statistical features of network packet data with context-based features obtained from HTTP GET/POST data to detect data breaches due to SQL injection attacks. These combined features yielded an accuracy of 96% using a Deep Belief Network (DBN) classifier. However there was a 10% drop in performance when training without the statistical feature. A late fusion model proposed by Pierazzi et al. [141] for spyware detection also achieved the best AUC score of 98.2% when compared with other models trained with single features.

X. RESEARCH ISSUES AND FUTURE DIRECTIONS

This section delineates the present research issues regarding ML-based data breach countermeasures and outlines prospective directions for future investigations aimed at resolving these challenges.

Issue I: *Self-supervised data breach countermeasures required.* Various studies have approached data exfiltration detection using different learning tasks depending on available

data or the nature of task. Studies generally employ classification learning task in instances where balanced data are available, otherwise anomaly detection is used. While a high percentage of studies (79%) used the former approach, the requirement of annotated data for supervision can be time consuming, error prone, costly and may not be proactive at detecting novel cyber attacks since data need to be labelled before training. Automated techniques for the annotation of cyber data are therefore imperative for developing cost-effective and proactive ML-based data exfiltration countermeasures. Although reviewed works including [85], [120], [144] used self-supervised models such as word2vec and transformers, where labels were automatically generated from the input data, other self-supervision techniques should be explored, especially for the detection of phishing and web attacks which are the most prevalent vectors of cyber data breaches.

Issue II: *Graph-based models required*. Four studies attempted data exfiltration as a graph-based problem. This was the least represented technique in the reviewed studies, despite the recent breakthroughs of graph models in the detection of general cyber attacks [148]. Graph models are generally employed to automate the embedding of node features into low-dimension vectors which are further used by downstream models for training. One distinct property of graph-based models is their ability to capture non-linear relationships in data which might be undetected by other models like sequence models. For example, in the detection of network anomalies (such as password bruteforce attacks) using packet or flow data, several studies [76], [103], [130] consider the sequence of events to model a user's sequential behaviour, ignoring other non-linear relationships.

Issue III: *Automated feature extraction methods needed*. As detailed in section VIII of this review, 13 studies exploited automated feature extraction methods to detect cyber data breaches, hence the majority of the reviewed works used manual feature extraction approaches, which rely on human expertise to extract features which can be laborious and error prone. Also, in an era where novel cyber attacks are launched almost each day, the manual feature extraction approach may be outpaced. This could result in features quickly becoming outdated, leading trained ML models being easily circumvented. Conversely, automatically extracted features are more resilient for developing ML-based data breach countermeasures against these dynamic and fast-paced cyber attacks.

Issue IV: *Continual learning methods for model upgrades*. To survive the test of time, ML-based data breach countermeasures need to be proactive in relation to environmental changes, particularly to changes in feature distributions (concept drift) due to novel attacks. Hence, models need to be re-trained with samples that reflect current cyber attack trends to be effective at data exfiltration detection. When re-training with novel attack samples, models are likely to degrade in performance with respect to detecting old attacks. To prevent this, various studies employed incremental continual training strategies [147], [149] to upgrade ML-models with new sample data while maintaining the model's former performance. Given the benefit of model re-training, one study [147] implemented continual training algorithms to incrementally update the data breach detection model for phishing attacks.

Issue V: *The use of one data modality*. Section IX discussed that 14 studies considered multiple modalities of cyber data to propose ML-based data breach countermeasures. This reveals that most studies used one feature type for data exfiltration detection. However 7 out of these 14 multimodal studies addressed exfiltration due to phishing attacks, whereas 2 studies addressed the detection of web/DNS related breaches, as shown in Table VIII. Zhang et al. [120] proposed a multimodal strategy for web attacks detection, leveraging the early fusion approach to combine two modalities of data (network packet and HTTP traffic) which yielded an accuracy of 96% with deep belief network. However, their findings reveal that, the model's accuracy dropped by 10% when trained with a smaller number of features. This shows that the use of multiple modalities of data provide a better representation of cyber attacks, boost performance.

Future Directions: It is imperative to early detect and prevent data breaches as the aftermath is severe and often result to serious damages. Hence, we recommend that future research investigates ways to innovate ML-based data breach mitigation approaches by introducing automated, proactive and efficient mechanisms to address the issues identified in this review relating to feature engineering, learning supervision, model updates and data modalities.

Firstly, given the dynamic and fast-paced nature of cyber attacks, future research should adopt more adaptable methods such as the use of automated techniques (as shown in Table VII) for the extraction of features for data exfiltration detection. Other deep learning models such as CNNs and RNNs which have been successfully applied to NLP and computer vision tasks should be explored in the domain of cyber data breach detection to automate the extraction of attention features from initially extracted features, as seen in [139], [150], [151].

Classification and supervised anomaly learning tasks for cyber data breach detection require labelled data for supervision. Given that data annotation can be error prone, non-proactive and labour intensive, we recommend future studies explore self-supervised methods to automate the generation of labels from unlabelled input data, alleviating the need for manual annotation. Studies which attempt data exfiltration as image classification problem can leverage self-supervised learning techniques [152], [153] used in computer vision to pretrain models to solve downstream data exfiltration detection tasks.

Further work needs to also factor in efficient approaches of modelling benign and malicious behaviour to yield satisfactory performance of ML-based data breach countermeasures. Often, user behaviour is modelled as a sequence classification problem due to the sequential occurrence of cyber events. The drawback is that, sequential models fail to detect other substantial representations in data except for linear relationships. To harness the complex structures of cyber data and improve data exfiltration detection, we recommend future studies explore GNN models which are deemed to be capable of capturing non-linear representations in data.

Our review reveals the need for constant model updates due to the prevalence and dynamic nature of cyber attacks.

TABLE IX
COMPARISON OF EXISTING RELATED SURVEYS (√: YES, ×: NO, △: PARTIALLY)

Survey Scope	**[154]**	**[155]**	**[24]**	**[156]**	**[157]**	**[158]**	**This Survey**
Year	2023	2021	2019	2017	2016	2021	2023
Categorization of data breach countermeasures	√	×	√	√	√	√	×
Dataset for evaluating cyber data breach detection systems	×	√	×	×	×	√	×
ML/DL for cyber data breach detection	×	△	×	×	×	√	√
Feature extraction for cyber data breach detection	×	×	×	×	×	√	√
Feature representation for cyber data breach detection	×	×	×	×	×	×	√
Proactive techniques for cyber data breach detection	×	×	×	×	×	△	√
Multimodal approaches for cyber data breach countermeasures	×	×	×	×	×	×	√
Promising future research directions	√	×	×	×	√	√	√

We found that one study in the review implemented continual learning algorithms to incrementally update models, hence we recommend future studies incorporate model update mechanisms, as detailed in Table VII and [147], [149] to ensure ML-based data breach countermeasures are adaptable to novel attacks.

Our review also shows that most works employed one data modality to detect cyber data breaches. Combining multiple types of features however is proven to improve performance as demonstrated in works such as [120]. Given that web attacks are one of the leading vectors of cyber data breaches, we recommend future studies consider different web data modalities for future countermeasures. We also recommend future research explores late multimodal fusion strategies for the detection of phishing related breaches as only one study used this approach.

XI. PREVIOUS RELATED SURVEYS AND COMPARATIVE ASSESSMENTS

A large number of studies have proposed varying countermeasures for data breach detection. Several reviews have been conducted to synthesize and compare these research studies. This section presents the existing related surveys on data breach countermeasures, as shown in Table IX.

In [154], Chung et al. performed a review on data exfiltration threats and countermeasures. They used cyber threat frameworks including the Microsoft STRIDE model, cyber kill chain, and MITRE ATT&CK framework to highlight the stages of exfiltration campaigns. They also identified three categories of countermeasures for data exfiltration, namely perimeter defense, data protection and alert and monitoring. The findings of Chung et al. reveal the need to incorporate human expertise into the development of machine-based systems used in defending against data exfiltration threats. Avila et al. [155] presented a systematic review on data leak detection using security logs. They proposed four categories of personal data and their corresponding GDPR guidelines. They also categorised the security logs and datasets used in the academic literature. Additionally, they identified six potential attack vectors of information leak and machine learning algorithms which were used for the processing of logs for data leak detection. Khan et al. [24] primarily focused on the risks and resolutions associated with data breaches. With regard to risks, they classified the causes, locus and impact of data breaches. They further identified three data breach resolution categories to manage data breach risk, namely prevention, containment and recovery. Kaur et al. [156] produced a comparative analysis of data breach preventive approaches. They reported insiders, software failure, viruses and natural disasters as the main drivers of data leaks. Three core functionalities of a data leak countermeasure system specified in their work include to protect, monitor and discover. Furthermore, they identified two approaches (i.e., signature and learning-based) employed by countermeasures and described three data states requiring protection. Alneyadi et al. [157] reviewed software solutions for data exfiltration prevention and classified them by their analysis type (i.e., content or context), mitigation type (i.e., preventive or detective), by deployment (according to states of data) and by remedial action type (such as block or alert). They also discussed the pros and cons of methods such as policy and access rights, cryptographic approaches, data mining and text clustering used by data exfiltration prevention systems. Sabir et al. [158] conducted a systematic literature review on machine learning based data exfiltration detection and presented a taxonomy of ML-based data exfiltration countermeasures, feature engineering approaches, evaluation metrics and datasets used by the selected studies. Sabir et al. made several recommendations including the need for robust ML-models to combat adversarial evasions.

XII. CONCLUSION

This paper presents a systematic literature review to synthesise the existing research on ML-based models proposed as countermeasures for the detection of data breaches resulting from cyber attacks. Following our survey methodology, we retrieved and compared a total of 81 studies using six criteria: learning tasks, learning classifiers, proactive learning strategies, feature engineering methods, and multimodal learning approaches. We examined over 40 recent incidents to identify the most prevalent cyber attack vectors leading to data breaches. This information was presented as the general workflow of data breaches due to cyber attacks and served as the foundation for identifying the relevant literature for our survey.

We classified the studies based on learning tasks, finding that approaches to detecting cyber data breaches primarily utilised anomaly or classification detection techniques. Approximately 79% of studies employed classification detection, with a focus on breaches resulting from phishing attacks. We then presented a taxonomy of feature extraction and

representation, considering six approaches: statistical, context, time, content, environmental, and behavioral-based. Our comparative assessment revealed that the statistical-based method was the most commonly used form of feature extraction, followed by the content-based method. Additionally, our analysis unveiled that various features were incorporated into existing ML algorithms using representation techniques such as vector-based, topology-based, sequence-based, and image-based, with vector-based and topology-based being the most and least utilised forms of representations, respectively.

After analysing the studies based on learning classifiers, we found that RF was the most commonly used conventional classifier, while CNN and FFN were the most applied deep learning models for detecting data exfiltration. Additionally, studies were comparatively analysed based on four proactive learning strategies: self-labelling, re-training, feature extraction automation, and data augmentation. The final criteria for characterising studies focused on the number of feature types (or data modalities) used by learning models. We classified the studies based on the strategy used to combine multiple unimodal features, identifying three multimodal fusion methods: early fusion, late fusion, and intermediate fusion.

The survey concluded with a discussion of the research issues and offered recommendations to guide future research. In particular, we identified that most studies employed one data modality to detect cyber data breaches. Given the severe consequences of data breaches, we specifically recommend further studies employ automated, proactive, and robust ML-based approaches for feature engineering, learning supervision, model updates, and data modalities in the context of data breach countermeasures.

REFERENCES

[1] S. Khan, I. Kabanov, Y. Hua and S. Madnick, "A systematic analysis of the capital one data breach: Critical lessons learned," *ACM Transactions on Privacy and Security*, vol. 26, no. 1, pp. 1–29, 2022.

[2] I. Agrafiotis, J. R. Nurse, M. Goldsmith, S. Creese and D. Upton, "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate," *Journal of Cybersecurity*, vol. 4, no. 1, p. tyy006, 2018.

[3] F. Khan, J. H. Kim, L. Mathiasssen and R. Moore, "Data breach management:An integrated risk model," *Information & Management*, vol. 58, no. 1, p. 103392, 2021.

[4] S. Goode, H. Hoehle,V. Venkatesh and S. A. Brown, "User compensation as a data breach recovery action," *MIS Quarterly*, vol. 41, no. 3, pp. 703–A16, 2017.

[5] "Data Breach Investigations Report." Verizon. 2022. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/2022-data-breach-investigations-report-dbir.pdf

[6] "Annual number of data compromises and individuals impacted in the United States from 2005 to 2022." Statista. 2023. [Online]. Available: https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/

[7] "Cost of a data breach report." IBM Security & Ponemon Institute. 2023. [Online]. Available: https://www.ibm.com/downloads/cas/E3G5JMBP

[8] M. Guri, R. Puzis, K. K. R. Choo, S. Rubinshtein, G. Kedma and Y. Elovici, "Using malware for the greater good: Mitigating data leakage," *Journal of Network and Computer Applications*, vol. 145, p.102405, 2019.

[9] Y. Roumani, "Detection time of data breaches," *Computers & Security*, vol. 112, p.102508, 2012.

[10] "Notifiable data breaches report." OAIC. 2022. [Online]. Available: https://www.oaic.gov.au/__data/assets/pdf_file/0026/39068/OAIC-Notifiable-data-breaches-report-July-December-2022.pdf

[11] "Data breache report." ITRC. 2022. [Online]. Available: https://www.idtheftcenter.org/publication/2022-data-breach-report/

[12] "Data Breach Investigations Report." Verizon. 2020. [Online]. Available: https://www.verizon.com/business/resources/reports/2020-data-breach-investigations-report.pdf

[13] "MOVEit cyberattacks: keeping tabs on the biggest data theft of 2023." TheVerge. 2023. [Online]. Available: https://www.theverge.com/23892245/moveit-cyberattacks-clop-ransomware-government-business

[14] K. Thomas, F. Li, A. Zand, J. Barrette, L. Invernizzi and E. Bursztein, "Data breaches, phishing, or malware? Understanding the risks of stolen credentials," *In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, vol. 32, no. 2, pp. 1421–1434, 2017.

[15] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar. A. Agrawal, R. Kumar and R. Ahmad Khan, "Healthcare data breaches: insights and implications," *Healthcare*, vol. 8, no. 2, p. 133, 2020.

[16] C. A. Makridis, " Do data breaches damage reputation? Evidence from 45 companies between 2002 and 2018," *Journal of Cybersecurity*, vol. 7, no. 1, p. tyab021, 2021.

[17] R. Sen and S. Borle, "Estimating the contextual risk of data Breach: An empirical approach," *Journal of Management Information Systems*, vol. 32, no. 2, pp. 314–341, 2015.

[18] "Massive data breach costs valuer LandMark White $7m." Itnews. 2019. [Online]. Available: https://www.itnews.com.au/news/massive-data-breach-costs-valuer-landmark-white-7m-524716

[19] "Financial impact of LandMark White cyber attack revealed." Business news Australia. 2019. [Online]. Available: https://www.businessnewsaustralia.com/articles/financial-impact-of-landmark-white-cyber-attack-revealed.html

[20] R. Pilla, T. Oseni and A. Stranieri, "A Study Into the impact of data breaches of electronic health records," *In Proceedings of the 2023 Australasian Computer Science Week*, pp. 252–254, 2023.

[21] "Medibank confirms hacker had access to data of all 3.9 million customers." Theguardian. 2022. [Online]. Available: https://www.theguardian.com/technology/2022/oct/26/medibank-confirms-all-39-million-customers-had-data-accessed-in-hack

[22] A. Emmanuel, "Standing in the aftermath of a data breach," *Journal of Law & Cyber Warfare*, vol. 4, no. 4, pp. 150–209, 2015.

[23] "The Equifax data breach." Cpanjournal. 2017. [Online]. Available: https://www.cpajournal.com/2017/12/15/equifax-data-breach/

[24] F. Khan, J. H. Kim, L. Mathiasssen and R. Moore, "Data breach risks and resolutions: A literature synthesis," *Americas Conference on Information Systems, Cancun*, 2019.

[25] S. Hamza and N. Muhammad, "SoK: Anatomy of data breaches," *Proc. Priv. Enhancing Technol*, vol. 2020, no. 4, pp. 153–174, 2020.

[26] "Notifiable data breaches report." OAIC. 2022. [Online]. Available: https://www.oaic.gov.au/__data/assets/pdf_file/0020/23663/OAIC-Notifiable-Data-Breaches-Report-Jan-Jun-2022.pdf

[27] "Just a slap on the wrist for Gloucester council data breach which saw people's data fall into hands of criminals." Gloucestershirelive. 2023. [Online]. Available: https://www.gloucestershirelive.co.uk/news/gloucester-news/slap-wrist-gloucester-data-breach-8599619

[28] "Courts fined $9,000 for second data breach in two years." straitstimes. 2020. [Online]. Available: https://www.straitstimes.com/tech/courts-fined-9000-for-second-data-breach-in-two-years

[29] X. Zhang, M. M. Yadollahi, S. Dadkhah, H. Isah, D. P. Le and A. A. Ghorbani, " Data breach: analysis, countermeasures and challenges," *International Journal of Information and Computer Security*, vol. 19, no. 3-4, pp. 402–442, 2022.

[30] A. O. Adebayo and Y. O. O. A. J. Omotosho, " System and data capture framework insights into breach Data toward improved feedback," *System*, vol. 4, no. 3, 2013.

[31] "Melbourne student health records posted online in 'appalling' privacy breach." Theguardian. 2018. [Online]. Available: https://www.theguardian.com/australia-news/2018/aug/22/melbourne-student-health-records-posted-online-in-appalling-privacy-breach

[32] "Distressed Strathmore dad takes aim at school after private information leaked online." 3AW. 2018. [Online]. Available: https://www.3aw.com.au/distressed-strathmore-dad-takes-aim-at-school-after-private-information-leaked-online/

[33] M. Mahapatra, N. Gupta, R. Kushwaha and G. Singal, " Data breach in social networks using machine learning," *Communications in Computer and Information Science*, vol. 1528, 2022.

[34] "Common Types of Data Breaches." Shrednations. 2015. [Online]. Available: https://www.shrednations.com/blog/different-types-of-data-breaches/?cn-reloaded=1

[35] "Costco Hit by Card Skimming Attack Heading Into Holiday Season." Securityweek. 2021. [Online]. Available: https://www.securityweek.com/costco-hit-card-skimming-attack-head-holiday-season/
[36] "Former South Georgia Medical Center Employee Arrested Over 41K-Record Data Breach." HIPAA Journal. 2022. [Online]. Available: https://www.hipaajournal.com/former-south-georgia-medical-center-employee-arrested-over-41k-record-data-breach/
[37] "7 Examples of Real-Life Data Breaches Caused by Insider Threats." Ekrasystem. 2022. [Online]. Available: https://www.ekransystem.com/en/blog/real-life-examples-insider-threat-caused-breaches
[38] "NSW Transport agency extorted by ransomware gang after Accellion attack." Bleeping Computer. 2021. [Online]. Available: https://www.bleepingcomputer.com/news/security/nsw-transport-agency-extorted-by-ransomware-gang-after-accellion-attack/
[39] "Top data breaches and cyber attacks of 2022." TechRadar. 2022. [Online]. Available: https://www.techradar.com/features/top-data-breaches-and-cyber-attacks-of-2022
[40] "Data Breach Investigations Report." Verizon. 2019. [Online]. Available: https://www.verizon.com/business/resources/reports/2019-data-breach-investigations-report.pdf
[41] J. Guffey, and Y. Li, " Cloud service misconfigurations: Emerging threats, enterprise data breaches and solutions," *IEEE13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0806–0812, 2013.
[42] D. Chen, M. M. Chowdhury, and S. Latif, " Data breaches in corporate setting.," *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 01–06, 2021.
[43] "Aveanna Healthcare Reaches $425K Settlement After Healthcare Data Breach." HealthItSecurity. 2022. [Online]. Available: https://healthitsecurity.com/news/home-health-provider-reaches-425k-settlement-after-healthcare-data-breach
[44] "Norman Public Schools' employee and student leaked on dark web by ransomware gang ." DataBreaches.Net. 2022. [Online]. Available: https://www.databreaches.net/update-norman-public-schools-employee-and-student-leaked-on-dark-web-by-ransomware-gang/
[45] "Global Accellion data breaches linked to Clop ransomware gang." BleepingComputer. 2021. [Online]. Available: https://www.bleepingcomputer.com/news/security/global-accellion-data-breaches-linked-to-clop-ransomware-gang/
[46] M. H. N. Ba, J. Benneth, M. Gallagher and S. Bhunia, " A case study of credential stuffing attack: Canva data breach," *IEEE Access*, pp. 735–740, 2021.
[47] "The complete list of data breaches in Australia for 2018 - 2023." WebberInsurance. 2023. [Online]. Available: https://www.webberinsurance.com.au/data-breaches-list
[48] "Cyber Security Data Breaches." DataBreachesdb. 2020. [Online]. Available: https://databreachdb.com/
[49] "Breach Types." DataBreaches. 2023. [Online]. Available: https://www.databreaches.net/category/breach-types/
[50] "Company blames February 2019 security breach on phishing email received in July 2018." Zdnet. 2019. [Online]. Available: https://www.zdnet.com/article/bodybuilding-com-discloses-security-breach/
[51] "MA pharmacy falls victim to email phishing attack, results in PHI exposure." HealthItSecurity. 2023. [Online]. Available: https://healthitsecurity.com/news/ma-pharmacy-falls-victim-to-email-phishing-attack-results-in-phi-exposure
[52] "ACU systems compromised and details stolen after cyber breach." ARN. 2019. [Online]. Available: https://www.arnnet.com.au/article/663022/acu-systems-compromised-and-details-stolen-after-cyber-breach/
[53] "The unified kill chain." UnifiedKillChain. 2023. [Online]. Available: https://www.unifiedkillchain.com/
[54] "200,000 North Face accounts hacked in credential stuffing attack." Bleeping Computer. 2022. [Online]. Available: https://www.bleepingcomputer.com/news/security/200-000-north-face-accounts-hacked-in-credential-stuffing-attack/
[55] "Indian discount airline SpiceJet suffers data breach affecting 1.2M customers." Silicon Angle. 2020. [Online]. Available: https://siliconangle.com/2020/01/30/indian-low-cost-airline-spicejet-suffers-data-breach-affecting-1-2m-customers/
[56] "Group dating app found leaking basically everything about its users worldwide." The Verge. 2019. [Online]. Available: https://www.theverge.com/2019/8/9/20798290/3fun-data-breach-security-cybersecurity-group-dating-app
[57] "Capita Cyber Incident: FAQs." Dalriada. 2023. [Online]. Available: https://www.dalriadatrustees.co.uk/capita-cyber-incident/
[58] "MOVEit mass exploit timeline: How the file-transfer service attacks entangled victims." Cybersecurity Dive. 2023. [Online]. Available: https://www.cybersecuritydive.com/news/moveit-breach-timeline/687417/
[59] A. Khraisat and A. Alazab, " A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges," *Cybersecurity*, vol.4, pp. 1–27, 2021.
[60] P. Spadaccino and F. Cuomo, " Intrusion detection systems for IoT: opportunities and challenges offered by edge computing and machine learning," *ArXiv Preprint*, vol.4, pp. 1–25, 2020.
[61] M. Husák, J. Komárková, E. Bou-Harb and P. Čeleda, " Survey of attack projection, prediction, and forecasting in cyber security," *IEEE Communications Surveys & Tutorials*, vol.4, pp. 640–660, 2018.
[62] Y. D. Lin, Z. Y. Wang, P. C. Lin,V. L. Nguyen, R. H. Hwang and Y. C. Lai, "Multi-datasource machine learning in intrusion detection: Packet flows, system logs and host statistics," *Journal of Information Security and Applications*, p. 103248, 2022.
[63] M. A. Rahman, A. T. Asyhari, O. W. Wen, H. Ajra, Y. Ahmed and F. Anwar , " Effective combining of feature selection techniques for machine learning-enabled IoT intrusion detection," *Multimedia Tools and Applications*, pp. 1–19, 2021.
[64] A. Nazir and R. A. Khan, " A novel combinatorial optimization based feature selection method for network intrusion detection," *Computers & Security*, vol.102, p. 102164, 2021.
[65] K. Albulayhi, Q. Abu Al-Haija, S. A. Alsuhibany, A. A. Jillepalli, M. Ashrafuzzaman and F. T. Sheldon, " IoT intrusion detection using machine learning with a novel high performing feature selection method," *Applied Sciences*, vol. 12, no. 10, p. 5015, 2022.
[66] Y. Bengio, A. Courville and P. Vincent, " Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[67] W. Wang, M. Zhu, X. Zengm, X. Ye and Y. Sheng, " Malware traffic classification using convolutional neural network for representation learning," *International conference on information networking (ICOIN)*, pp. 712–717, 2017.
[68] V. P. Gandi, N. S. L. Jatla, G. Sadhineni, S. Geddamuri, G. K. Chaitanya and A. K. Velmurugan" "A comparative study of AI algorithms for anomaly-based intrusion detection," *7th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 530–534, 2023.
[69] B. Kitchenham, " Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.
[70] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, " Lessons from applying the systematic literature re- view process within the software engineering domain," *Journal of systems and software*, vol. 80, no. 4, pp. 571–583, 2007.
[71] M. Abdullahi, Y. Baashar, H. Alhussian, A. Alwadain, N, Aziz, L. F. Capretz and S. J. Abdulkadir, " Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review," *Electronics*, vol. 11, no. 2, pp. 198, 2022.
[72] N. Kleanthous, A. J. Hussain, W. Khan, J. Sneddon, A. Al-Shamma'a, and P. Liatsis, " A survey of machine learning approaches in animal behaviour," *Neurocomputing*, vol. 491, pp. 442–463, 2022.
[73] A. Garg, and V. Mago, " Role of machine learning in medical research: A survey," *Computer science review*, vol. 40, p. 100370, 2021.
[74] M. Komisarek, M. Pawlicki, M. Kowalski, A. Marzecki, R. Kozik, and M. Choraś, " Network intrusion detection in the wild-the Orange use case in the SIMARGL project," *In Proceedings of the 16th International Conference on Availability, Reliability and Security*, pp. 1–7, 2021.
[75] K. Cortial, and A. PachOt, " Sodinokibi intrusion detection based on logs clustering and random forest," *International Conference on Artificial Intelligence and Information Systems*, pp. 1–4, 2021.
[76] I. V. Ring, H. John, M. Colin, V. Oort, S, Durst, V. White, J. P. Near and C. Skalka, " Methods for host-based intrusion detection with deep learning," *Digital Threats: Research and Practice (DTRAP)*, vol. 2, no. 4, pp. 1–29, 2021.
[77] Y. Han, and Y. Shen, " Accurate spear phishing campaign attribution and early detection," *in Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pp. 2079–2086, 2016
[78] R. Verma, and K. Dyer, " On the character of phishing URLs: Accurate and robust statistical learning classifiers," *In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, pp. 111–122, 2015
[79] A. Hannousse, and S. Yahiouche, " Towards benchmark datasets for machine learning based website phishing detection: An experimental

study," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104347, 2021.

[80] J. V. Jawade, and S. N. Ghosh, " Phishing website detection using Fast. ai Library," *In 2021 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1–5, 2021.

[81] I. Jawad, J. Ahmed, I. Razzak, and R. Doss, " Identifying DNS exfiltration based on lexical attributes of query name," *In 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2021.

[82] M. Gniewkowski, H. Maciejewski, T. Surmacz, and W. Walentynowicz, " Sec2vec: Anomaly detection in HTTP traffic and malicious URLs," *In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pp. 1154–1162, 2023.

[83] M. Conti, G. Rigoni, and F. Toffalini, " ASAINT: a spy app identification system based on network traffic," *In Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–8, 2020.

[84] H. Chindove and D. Brown, " Adaptive machine learning based network intrusion detection," *In Proceedings of the International Conference on Artificial Intelligence and its Applications*, pp. 1–6, 2021.

[85] P. Bountakas K. Koutroumpouchos and C. Xenakis, " A comparison of natural language processing and machine learning methods for phishing email detection," *In Proceedings of the 16th International Conference on Availability, Reliability and Security*, pp. 1–12, 2021.

[86] T. Saka K. Vaniea and N. Kökciyan, " Context-based clustering to mitigate phishing attacks," *In Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, pp. 115–126, 2022.

[87] R. Rader K. Jerabek and O. Rysavy, " Detecting DoH-Based data exfiltration: FluBot malware case study," *In IEEE 48th Conference on Local Computer Networks (LCN)*, pp. 1–4, 2023.

[88] S. Mahdavifar H. A. Salem, P. Victor, A. H, Razavi, M. Garzon, N. Hellberg and A. H. Lashkari, " Lightweight hybrid detection of data exfiltration using dns based on machine learning," *International Conference on Communication and Network Security*, pp. 80–86, 2021.

[89] K. Tian, S. T. Jan, H. Hu, D. Yao and G. Wang, " Tracking down elite phishing domains in the wild," *In Proceedings of the Internet Measurement Conference*, pp. 429–442, 2018.

[90] M. Ivanova, and A. Rozeva, " Detection of XSS attack and defense of REST web service–machine learning perspective," *International Conference on Machine Learning and Soft Computing*, pp. 22–28, 2021.

[91] M. Piskozub, F. De Gaspari, F. Barr-Smith, L. Mancini and I. Martinovic, " Malphase: fine-grained malware detection using network flow data," *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1314–1325, 2021.

[92] Y. Ma, S. Liu, J. Jiang, G. Chen and K. Li, " A comprehensive study on learning-based PE malware family classification methods," *ACM Asia conference on computer and communications security*, pp. 774–786, 2021.

[93] N. B. Trinh, T. D. Phan,and V. H. Pham, " Leveraging Deep Learning Image Classifiers for Visual Similarity-based Phishing Website Detection," *International Symposium on Information and Communication Technology*, pp. 134–141, 2022.

[94] K. Sethi, S. K. Chaudhary, B. K, Tripathy and P. Bera, " A novel malware analysis framework for malware detection and classification using machine learning approach," *International conference on distributed computing and networking*, pp. 1–4, 2022.

[95] C. Liu, Z. Yang, Z. Blasingame, G. Torres and J. Bruska, " Detecting data exploits using low-level hardware information: A short time series approach," *First Workshop on Radical and Experiential Security*, pp. 41–47, 2018.

[96] D. Gibert, C. Mateu and J. Planes, " The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *Journal of Network and Computer Applications*, vol. 153, p. 102526, 2020.

[97] D. Kwon, R. M. Neagu, P. Rasakonda, J .T. Ryu and J. Kim, " Evaluating unbalanced network data for attack detection," *In Proceedings on Systems and Network Telemetry and Analytics*, pp. 23–26, 2023.

[98] F. Liu, Y. Weng, D. Zhang, X. Jiang, X. Xing and D. Meng, " Log2vec: A heterogeneous graph embedding based approach for detecting cyber threats within enterprise," *In Proceedings of ACM SIGSAC conference on computer and communications security*, pp. 1777–1794, 2019.

[99] B. Li, G. Yuan, L. Shen, R. Zhang and Y. Yao, " Incorporating URL embedding into ensemble clustering to detect web anomalies," *Future Generation Computer Systems*, vol 96, pp. 176–184, 2019.

[100] J. Zhan, X. Liao, Y. Bao, L. Gan, Z. Tan, M. Zhang, R. He and J. Lu, " An effective feature representation of web log data by leveraging byte pair encoding and TF-IDF," *In Proceedings of the ACM Turing Celebration Conference-China*, pp. 1–6, 2019.

[101] J. Bohara, U. Thakore and W. H. Sanders, " Intrusion detection in enterprise systems by combining and clustering diverse monitor data," *In Proceedings of the Symposium and Bootcamp on the Science of Security* , pp. 7–16, 2016.

[102] A. L. Buczak, P. A. Hanke, G. J. Cancro, M. K. Toma, L. A. Watkins and J. S. Chavis, " Detection of tunnels in PCAP data by random forests," *In Proceedings of the 11th Annual Cyber and Information Security Research Conference*, pp. 1–4, 2016.

[103] M. Alrehaili, A. Alshamrani and A. Eshmawi, " A hybrid deep learning approach for advanced persistent threat attack detection," *In The 5th International Conference on Future Networks & Distributed Systems*, pp. 78–86, 2021.

[104] N. M. Sheykhkanloo, "Employing neural networks for the detection of SQL injection attack," *In Proceedings of the 7th International Conference on Security of Information and Networks*, pp. 318–323, 2014.

[105] L. Bilge, S. Sen, D. Balzarotti, E. Kirda and C. Kruegel, " MExposure: A passive dns analysis service to detect and report malicious domains," *ACM Transactions on Information and System Security (TISSEC)*, vol. 16, no. 4, pp. 1–28, 2014.

[106] T. Zoppi, A. Ceccarelli, T. Capecchi and A. Bondavalli, " MExposure: Unsupervised anomaly detectors to detect intrusions in the current threat landscape," *ACM/IMS Transactions on Data Science*, vol. 2, no. 2, pp. 1–26, 2021.

[107] D. Pujol-Perich, J. Suárez-Varela, A. Cabellos-Aparicio and P. Barlet-Ros, " Unveiling the potential of graph neural networks for robust intrusion detection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 49, no. 4, pp. 111–117, 2022.

[108] Y. Lu, M. K. Kumar, N. Mohammed and Y. Wang "Homoglyph attack detection with unpaired data," *In Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pp. 377–382, 2019.

[109] S. S. Sirigineedi, J. Soni and H. Upadhyay "Learning-based models to detect runtime phishing activities using URLs," *In Proceedings of the 4th International Conference on Compute and Data Analysis*, pp. 102–106, 2020.

[110] A. Drichel, V. Drury, J. V. Brandt and U. Meyer, "Finding phish in a haystack: A pipeline for phishing classification on certificate transparency logs," *In Proceedings of the 16th International Conference on Availability, Reliability and Security*, pp. 1–12, 2021.

[111] P. Saraswat and M. S. Solanki, "Phishing detection in E-mails using machine learning," *In 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 420–424, 2022.

[112] J. Doshi, K. Parmar, R. Sanghavi and N. Shekokar, " A comprehensive dual-layer architecture for phishing and spam email detection," *Computers & Security*, vol. 133, p. 103378, 2023.

[113] M. Hussain, C. Cheng, R. Xu and M. Afzal, "CNN-Fusion: An effective and lightweight phishing detection method based on multi-variant ConvNet," *Information Sciences*, vol. 631, pp.328–345, 2023.

[114] Z. Cheng and X. Chen, Y. Zhang, S. Li and Y. Sang "Detecting information theft based on mobile network flows for android users," *In 2017 International conference on networking, architecture, and storage (NAS)* , pp. 1–10, 2017.

[115] S. K. Katherasala, V. S. Manvith, A. Therala and M. Murala, "NetMD-Network traffic analysis and malware detection," *In International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* , pp. 011–016, 2022.

[116] A. Kumar and I. Sharma, "Enhancing data privacy of IoT healthcare with keylogger attack mitigation," *In 4th International Conference for Emerging Technology (INCET)* , pp. 1–6, 2023.

[117] A. M, Al Badri and S. Alouneh, "Detection of malicious requests to protect web applications and DNS servers against SQL injection using machine learning," *In International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)* , pp. 5–11, 2023.

[118] K. S. Jishnu and B. Arthi, "Phishing URL detection by leveraging RoBERTa for feature extraction and LSTM for classification," *In 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)* , pp. 972–977, 2023.

[119] R. Alotaibi, I. Al-Turaiki and F. Alakeel, "Mitigating email phishing attacks using convolutional neural networks," *In International Conference on Computer Applications & Information Security (ICCAIS)* , pp. 1–6, 2020.

[120] H. Zhang, B. Zhao, H. Yuan, J. Zhao, X. Yan and F. Li, "SQL injection detection based on deep belief network," *In Proceedings of the 3rd International Conference on Computer Science and Application Engineering* , pp. 1–6, 2019.

[121] S. Shamshirband and A. T. Chronopoulos "A new malware detection system using a high performance-ELM method," *In Proceedings of the*

23rd international database applications & engineering symposium, pp. 1–10, 2019.

[122] A. Oprea, Z. Li, R. Norris and K. Bowers "Made: Security analytics for enterprise threat detection," *In Proceedings of the 34th Annual Computer Security Applications Conference*, pp. 124–136, 2018.

[123] B. Kondracki, B. A. Azad, O. Starov and N. Nikiforakis "Catching transparent phish: Analyzing and detecting MITM phishing toolkits," *In Proceedings of ACM SIGSAC Conference on Computer and Communications Security* , pp. 36–50, 2021.

[124] S. Wu, S. Liu, W, Lin, X. Zhao and S. Chen, " Detecting remote access trojans through external control at area network borders," *In ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, vol.102, pp. 131–141, 2017.

[125] B. Janet and R. J. A. Kumar, "Malicious URL detection: a comparative study," *In International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 11–1151, 2021

[126] P. Saraswat and M. S. Solanki, "Phishing detection in E-mails using machine learning," *In 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 420–424, 2022

[127] J. Ahmed, H. H. Gharakheili, Q. Raza, C. Russel and V. Sivaraman, " Monitoring enterprise DNS queries for detecting data exfiltration from internal hosts," *IEEE Transactions on Network and Service Management*, vol. 104, no. 1, pp. 265–279, 2019.

[128] A. Alhogai and A. Alsabih, " Applying machine learning and natural language processing to detect phishing email," *Computers & Security*, vol. 110, no. 1, p. 102414, 2021.

[129] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione and X. Chang, " A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 110, no. 1, pp. 47–57, 2021.

[130] S. M. Elsayed, N. .A. Le-Khac, S. Dev and A. D. Jurcut, " Network anomaly detection using LSTM based autoencoder," *In Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, pp. 37–45, 2020.

[131] V. Devalla, S. S. Raghavan, S. Maste, J. D. Kotian, and D. Annapurna, " A tool for detection of malicious urls and injection attacks.," *Procedia Computer Science*, vol. 215, pp. .662-676, 2022.

[132] M. Miao and B. Wu, " A flexible phishing detection approach based on software-defined networking using ensemble learning method," *In Proceedings of the 4th International Conference on High Performance Compilation, Computing and Communications*, pp. 70–73, 2020.

[133] K. S. Yin and M. A. Khine, " Network behavioral features for detecting remote access Trojans in the early stage," *In Proceedings of the VI International Conference on Network, Communication and Computing*, pp. 92–96, 2017.

[134] A. Mudgerikar, P. Sharma and E. Berttino, " Edge-based intrusion detection for IoT devices," *ACM Transactions on Management Information Systems (TMIS)*, vol. 11, no. 4, pp. 1–21, 2020.

[135] L. Gallo, A. Botta and G. Ventre, " Identifying threats in a large company's inbox," *In Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pp. 1–7, 2019.

[136] G. Ruiling, D. Jiawen, C. Xiang and S. Shouyou, " A DNS-based data exfiltration traffic detection method for unknown samples," *In 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)* , pp. 191–198, 2021.

[137] M. M. Alani and H. Tawfik, " PhishNot: a cloud-based machine-learning approach to phishing URL detection," *In Proceedings of the First Workshop on Radical and Experiential Security* , pp. 33–39, 2018.

[138] M. Weber, J. Wang, and Y. Zhou, " Unsupervised clustering for identification of malicious domain campaigns," *In 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)* , pp. 191–198, 2021.

[139] G. D'Angelo, A. Castiglione and F. Palmieri, " DNS tunnels detection via DNS-images," *Information Processing & Management*, vol. 59, no. 3, p.102930, 2022.

[140] A. Rahmadeyan, I. Ahmad, A. D. Alexander and A. Rahman, " Phishing website detection with ensemble learning approach using artificial neural network and AdaBoost," *In 2023 International Conference on Information Technology Research and Innovation (ICITRI)* , pp. 162–166, 2023.

[141] F. Pierazzi, G. Mezzour, Q. Han, M. Colajanni and V. S Subrahmanian, "A data-driven characterization of modern Android spyware," *ACM Transactions on Management Information Systems (TMIS)*, vol. 11, no. 1, pp. 1–38, 2020.

[142] J. Irungu, S. Graham, A. Girma and T. Kacem, " Artificial intelligence techniques for SQL injection attack detection," *In Proceedings of the 8th International Conference on Intelligent Information Technology* , pp. 38–45, 2023.

[143] T. C. Chen, T. Stepan, S. Dick and J. Miller, "An anti-phishing system employing diffused information," *ACM Transactions on Information and System Security (TISSEC)*, vol. 16, no. 4, pp. 1–31, 2014.

[144] P. Bountakas and C. Xenakis, "HELPHED: Hybrid Ensemble Learning PHishing Email Detection," *Journal of Network and Computer Applications*, vol. 210,p. 103545, 2023.

[145] I. Chiscop, F. Soro, and P. Smith, " AI-based detection of DNS misuse for network security," *In Proceedings of the 1st International Workshop on Native Network Intelligence*, pp. 27–32, 2022.

[146] C. Shorten, T. M. Khoshgoftaar and B. Furht, "Text data augmentation for deep learning.," *Journal of big Data*, vol. 8, pp. 1–34, 2021.

[147] A. Ejaz, A. N. Mian and S. Manzoor, "Life-long phishing attack detection using continual learning.," *Scientific Reports*, vol. 13, no. 1, p. 11488, 2023.

[148] M. Lv, C. Dong, T.Chen, T. Zhu, Q. Song and Y. Fan, "A Heterogeneous graph learning model for cyber-Attack detection.," *arXiv preprint arXiv:2112.08986*, 2021.

[149] B. Pfülb, "Continual learning with deep learning methods in an application-Oriented context.," *arXiv preprint arXiv:2207.06233*, 2022.

[150] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer and M. Woźniak, "Accurate and fast URL phishing detector: a convolutional neural network approach.," *Computer Networks*, vol. 178, p. 107275, 2020.

[151] W. B. Shahid, B. Aslam, H. Abbas, S. B. Khalid and H. Afzal, "An enhanced deep learning based framework for web attacks detection, mitigation and attacker profiling.," *Journal of Network and Computer Applications*, vol. 198, p. 103270, 2022.

[152] S. Gidaris, P. Singh and N. Komodakis, " Unsupervised representation learning by predicting image rotations.," *arXiv preprint arXiv:1803.07728*, 2018.

[153] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang and J. Tang, " Self-supervised learning: Generative or contrastive.," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp.857–876, 2021.

[154] M. H. Chung, Y. Yang, L. Wang, G. Cento, K. Jerath, A. Raman, D. Lie and M. H. Chignell, "Implementing data exfiltration defense in Situ: A survey of countermeasures and human involvement," *acm computing surveys*, vol. 55, no. 14s, pp. 303–340, 2023.

[155] R. Avila, R. Khoury, R. Khoury and F. Petrillo , " Use of security logs for data leak detection: a systematic literature review," *Security and communication networks*, vol. 2021, pp. 1–29, 2021.

[156] K. Kaur, I. Gupta, A. k. Singh , " A comparative study of the approach provided for preventing the data leakage.," *International Journal of Network Security & Its Applications*, vol. 2021, pp. 1–29, 2021.

[157] S. Alneyadi, E. Sithirasenan and V. Muthukkumarasamy, " A survey on data leakage prevention systems.," *Journal of Network and Computer Applications*, vol. 62, pp. 137–152, 2016.

[158] B. Sabir, F. Ullah, M. A. Babar and R. Gaire, "Machine learning for detecting data exfiltration: A review," *ACM Computing Surveys (CSUR)*, vol. 9, no. 5, pp. 21–33, 2017.

APPENDIX A
DATA BREACH INCIDENTS

TABLE X
FORTY INVESTIGATED DATA BREACH INCIDENTS.

Breached Organisation	Industry
NSW Transport Agency	Transportation
Medibank	Healthcare
Canva	Technology
North Face	Wholesale
Spirit Super	Finance and Insurance
Flexbooker	Transportation
South Australia	Government
Wawa	Retail
Avalon	Healthcare
Oregon Department of Human Services	Government
Magellan	Healthcare
Mednax	Administration
Unitypoint	Healthcare
Air Newzealand	Transportation
Australian Catholic University	Education
United Health Services of Delaware	Healthcare
bodybuilding.com	Wholesale
Gloucester City Council	Government
Capita	Technology
Allcare Plus Pharmacy	Healthcare
Norman Public schools	Education
Aveanna	Healthcare
Chegg	Technology
EyeMed	Healthcare
Cytometry Specialists	Professional
The Methodist Hospital	Healthcare
SpiceJet	Transportation
MoveIT	Technology
LA County Department of Mental Health	Government
Valley View Hospital	Healthcare
Red Robin	Accommodation
Spokane Regional Health District	Government
Sacremento County	Government
North American Dental Management	Healthcare
Health First, Inc	Healthcare
Women's care Florida	Healthcare
Centre for Computing History (CCH)	Education
Apunipima	Healthcare
3fun	Others
Accellion Global	Others

B BIOGRAPHIES

Paul N. Yeboah is currently a PhD student at the Department of Computer Science and Computer Engineering, La Trobe University, Australia. He completed his master's and bachelor degrees in computer science and cybersecurity. He works on various advanced topics in cybersecurity, such as cyber threat and data breach detection, prevention, and mitigation.

A. S. M. Kayes is a Senior Lecturer in Cybersecurity in the Department of Computer Science and Information Technology at La Trobe University, Australia. His research interests encompass various areas within cybersecurity, such as data security, context-aware access control, IoT and fog security, cyber incidents, and data/privacy breaches. Over the past 10 years, he has published more than 75 research articles in international journals and conference proceedings.

Wenny Rahayu is a Professor and Dean of the School of Computing, Engineering, and Mathematical Sciences at La Trobe University, Australia. Prior to this appointment, she was the Head of Department of Computer Science and Information Technology from 2012 to 2014. Her research interests include big data management, data privacy, integration of data from diverse sources, and access control. In the last 20 years, she has authored more than 260 research articles, including books, journals, and conference papers.

Eric Pardede is an Associate Professor and Associate Dean of the School of Computing, Engineering, and Mathematical Sciences at La Trobe University, Australia. His research interests include data analytics, modelling, privacy and security, higher education pedagogy and practices, and online social networks development and management. Over the past decade, he has published more than 100 research articles in international journals and conference proceedings.

Syed Mahbub is currently a Lecturer in the Department of Computer Science and Information Technology at La Trobe University, Australia. He worked in the industry as a software engineer for about 3 years. He completed his master's degree in Information Technology and his PhD in Information Systems from La Trobe University in 2016 and 2022, respectively. His research interests include data analytics, machine learning, and natural language processing within the context of improper behaviour detection in online social networks.