International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Securing Continual-Learning Network Intrusion Detection Systems: A Review of Concept Drift, Data Poisoning, Backdoors, and Catastrophic Forgetting

Kuchala Navanandu

G. Saiganesh, B. Pravarshika, D. Nancy Sherena

Abstract:

Network intrusion detection systems based on machine learning (ML-NIDS) are trained using datasets. In real life network patterns, network configurations and even the types of attacks keep changing all the time. In this paper we look at concept drift, adversarial machine learning and continual learning. We explore how these three areas intersect. Affect ML-NIDS. We begin by studying the reasons why concept drift happens. We also look at the methods used to adapt models when drift occurs. We consider how an attacker might exploit the system during adaptation or after adaptation has taken place. In this paper we present a security-focused classification framework. This framework helps us deal with drift detection and model adaptation data poisoning and backdoor threats. We also examine the issue of forgetting, where models forget old knowledge when learning new things. We go on to review benchmark datasets and evaluation techniques. We highlight the need to use data that is arranged by time data to make evaluations more realistic. This is especially important when testing how well models adapt under attacks. There is also a need for evaluation methods that can handle network setups. We analyze existing benchmark datasets. Assess how well they meet these evaluation needs. The final contribution of this paper is an evaluation framework. This framework is designed with security, in mind. It looks at whether ML-NIDS can maintain its performance, robustness and operational reliability when the network changes and when adversaries are actively trying to break the system.

Keywords:

Network intrusion detection; machine learning; continual learning; concept drift; adversarial machine learning; data poisoning; backdoor attacks; catastrophic forgetting; network security.

Introduction:

Network Intrusion Detection Systems (NIDS) are used to detect malicious behaviour on computer networks. With the rise of cloud computing, the Internet of Things and other

geographically dispersed online services many organisations now have networks whose traffic can be influenced by user actions, infrastructure changes, new applications, attacks and new software. Because of this there is now a need for intrusion detection which can use Machine Learning (ML) to detect malicious behaviour based on learned patterns in network traffic and separate that from normal network traffic. A major problem with machine learning based intrusion detection systems (ML-NIDS) is that the patterns learned during training may not translate well to the real world. A system trained on past network traffic may perform well in an experimental setting, but may fail when presented with network traffic that is not an exact representation of the training data. This may result in normal traffic being flagged as malicious due to a shift in application usage or a new attack that was not present during the training phase going undetected. This leads to the problem of how intrusion detection can adapt to the ever changing nature of attacks and networks. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).

Another important factor related to changing network environments is concept drift, which relates to 'changes in the statistical properties of the target variable which the model is trying to predict, or the behaviour of the input variables . Gama et al. (2014) explored in depth the causes, detection mechanisms, and model adaptation methods for concept drift, forming an important basis for work on learning from drifting data. Another common reason for concept drift is when the legitimate traffic distribution changes, a network change occurs, or the attacker creates new attack signatures that do not match the previously observed signatures. Such changes can result in subsequent low detection accuracy if the model is not updated appropriately to capture the change. Concept-drift detection and adaptation techniques attempt to cope with this problem by recognising that the incoming data has changed, and subsequently updating the model as required. However, detecting that the data has changed does not necessarily lead to conclusions that this change is due to a phenomenon that would require an adaptive response. Unnaturally induced change can also require adaptation, but in some cases, it can be indistinguishable from natural change. This leads to the conclusion that in ML-NIDS, concept drift must be considered not only in terms of effect on the predictor's accuracy but also in terms of effect on the security and robustness of model adaptation. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

Continual learning offers a possible solution to these scenarios. Rather than a machine learning system that is trained once and tested against a relatively stationary distribution, systems that

use continual learning remain exposed to new data throughout a sequence of learning episodes, which can allow an intrusion detector to cope with new traffic behaviour and newly emerging attack behaviour without having to be retrained again on all available data after a new update. Another issue of continual learning is that the model should learn new knowledge without forgetting the pre-existing knowledge. This problem is known as catastrophic forgetting. Parisi et al. (2019) defined the challenge of retaining previously learned knowledge as a defining challenge of sequential learning approaches in their review of continual and lifelong learning. In the context of ML-NIDS, this has implications that extend beyond reducing the overall predictive accuracy of a model. If a model learns to detect one particular attack, but subsequently becomes less capable of identifying a previously known attack, then this update may improve performance on the most recently encountered traffic, but reduce it again against an attack of a different type that remains relevant. A continual learning based detector should therefore be able to adapt to new conditions without sacrificing any earlier performance. This consideration makes long-term performance evaluation especially critical, as a model may seem to have excellent accuracy following an update, but not to have memorized the ability to reliably detect an earlier attack. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

The security implications increase once continual learning is considered with adversarial machine learning. If a network environment is dynamic, an attacker could attempt to evade the detector's current predictions, or target the data used to update the detector. Data poisoning consists of poisoning the data used to train or update a model in order to impact behaviour of a learning algorithm. Backdoor attacks aim to insert a covert association in the model, such that specific inputs will lead to the attacker desired output while other normal inputs are given seemingly correct outputs. With regards to adversarial machine learning, Biggio & Roli (2018) considered the general security gaps of learning algorithms to adversarially created inputs and training procedures. As a survey into adversarial attacks and defence in network intrusion detection, He et al. (2023) noted the general relevance of adversarial machine learning in regard to NIDS security. When considering such threats in continual learning settings, it is no longer simply a concern of model robustness, because a detector that adapts with training may instead become the focus of an attacker attempting to poison the update data, influence drift decisions or disrupt the adaptation process. Such strategies could cause the detector to fail to identify attack types of interest or induce a person-specific backdoor effect, contingent upon attacker

power and the update method used. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

There are also connections between these concepts and some of the terms and ideas from the literature review. Concept drift refers to system or input environment changes, continual learning refers to model performance as information is added, data poisoning and backdoors refer to perpetrator attacks on the learning process, and catastrophic forgetting to the loss of previously learned knowledge. While exploring each of these separate models can provide insight into individual processes and effects, it cannot fully explain what happens when combined within a functioning intrusion detection system. For example, an attacker might introduce traffic variations which trigger a detector decision to adapt, an ML NIDS might subsequently update on these data, thus producing an attack that exploits this vulnerability without an apparent weakness in the system. Conversely, a poisoning defence may minimise the influence of anomalous samples without considering whether the model update causes it to forget previously learned attack characteristics, a drift adaptation method might operate effectively against natural traffic variation but can still be susceptible to artificially induced shifts. This is why the next logical question for the security of a continually learning ML NIDS to answer is how all of these functions operate in conjunction in the course of the model's learning lifecycle and whether they compromise the overall system security. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

The existing literature provides important foundations to studying these issues. Gama et al. (2014) surveyed concept drift adaptation techniques, Parisi et al. (2019) reviewed continual and lifelong learning methods including catastrophic forgetting mitigations, Biggio and Roli (2018) studied adversarial machine learning, and He et al. (2023) investigated adversarial attacks and defences in the context of network intrusion detection. Research on NIDS datasets also implies that how the evaluation data reflects the characteristic behaviours of network traffic and attacks in the field must be taken into account. For instance, Ring et al. (2019) surveyed network based intrusion detection datasets, discussing the variations in their properties and their implications for evaluation. More recent work has explicitly studied concept drift resilient intrusion detection, indicating that adapting to changes in the network environment is a common research area. Collectively, the cited studies provide a wealth of knowledge about the individual aspects of adaptive and secure intrusion detection. However, the fact that there is relevant work in these different realms does not necessarily mean that they

have been integrated to yield a comprehensive understanding of their combined security implications. A review must therefore consider not only what approaches have been suggested, but the conditions under which they have been tested, the threats they aim to address, and whether their results can be related to one another. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

Assessing methods is also problematic. ML NIDS work often relies on benchmark datasets split randomly into a test set and a train set, evaluating under a relatively static experimental distribution. While such experiments facilitate comparison, they may only limit knowledge about a detector that continually adapts over time to a more realistic experimental distribution. Random splits may even include similar traffic conditions in the train and test sets, making it unclear whether any method can truly capture a changing environment. Chronologically ordered evaluation captures updates of the detector based on earlier traffic and evaluation on later traffic, often with the same dataset. However, alone, chronological evaluation cannot verify security. Detection may adapt successfully to real changes but be vulnerable to poisoning during the update step. The detector may be resistant to an adversary performing small perturbations, but not retain knowledge of previously learned attacks, as indicated by an initial drop in accuracy following an update. Performance under one network dataset does not guarantee similar results in other networks. The resource demands of the approach, the availability of labels for training, the update rate, and the speed of the defense can all condition its applicability. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

The potential effects of poor evaluation are both technical and operational. Overly numerous false positive alerts may overwhelm an analyst, making real events harder to identify. Failure of detectors to discover real events may allow abusers to operate unchecked, and forgetting may impair a detector's ability to recognise attack patterns it has previously learned. If an attacker slips in bad updates that cause lasting problems, a model might keep acting strangely even after the attacker is no longer around. Such impacts can lead to a deterioration in trust in the effectiveness of automated security observation, and may hinder the adoption of adaptive models in systems where reliable detection is crucial. Meanwhile, security solutions need to be practically deployable. One that suffers a large computational cost, high demands for reliable labels, or requiring expensive retraining may be challenging to realise in environments with limited resources. Alternatively, an adaptation approach that is efficient might be useless if it

does not retain previous knowledge, or if the manner of updating can be abused. So, the main thing is to see how existing methods handle the trade-offs between detection performance, adaptivity, risk of interference, keeping past knowledge, and how practical they are, and whether current ways of evaluating them really cover all this. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

With success on each of these fronts, however, a security oriented synthesis of their relationship is still a topic of research. This review does not imply that concept drift, adversarial attacks or catastrophic forgetting have been individually neglected. Instead, it looks at whether research has connected these concepts enough within the ML NIDS area of continual learning. This difference is relevant because the scope of existing reviews may inevitably cover some aspects of drift tolerant intrusion detection, continual learning or adversarial defence, and so the relevant gap must be established by scrutinising similarities and differences in their scope, implementation and treatment of update time threats. This concern is tackled within the review via the following framework that discusses detection of drift, model adaption, manipulation of update data, backdoor threat to the learned behaviour, and preservation or loss of previously acquired detection knowledge. Additionally, it investigates to what extent the interactions are modeled in current datasets and evaluation metrics, especially with the introduction of sequential observations and adversarial attacks during sequential updates. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

This system sorts and compares the main security problems at each stage of an update, since every phase comes with its own set of challenges and questions Before the update, the focus is on changes in traffic, spotting drift, and making sure information can be trusted, During the update, the big worries are data poisoning, backdoor risks, how the system adjusts, and steps to block harmful information from getting into learning, After the update, the main things to watch are performance, tracking past attacks, any leftover weaknesses, and how dependable future checks are. The security problem taxonomy is provided to allow comparison between studies that utilise diverse datasets, different learning strategies, different threat assumptions and different evaluation methods, in the setting of evaluating if the present experiments provide security performance which is beyond that achieved by traditional classification accuracy. It is used here as the organizing contribution of the review, with its categories and subcategories to be justified by the extent to which each set of indicators is present in the published evidence. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

2.Review Of Literature

2.1.Concept Drift in NIDS and Model Updating

The changes in network usage and attacker behavior makes the data for intrusion detection susceptible to drift. New applications may be introduced, they may change the nature of benign traffic, or attackers might develop a new attack type that modifies the distributions in the data. Gama et al. (2014) and Žliobaitė (2010) explain that as a result of this, the concept drift experienced can be both "real" if the target concept changes, and "virtual" if the input distribution changes, both of which can cause fixed NIDS classifiers to lose predictive power if unaddressed. Efforts from the research community thus address concept drift in intrusion detection through methods of drift detection and adaptation. For example, detectors can be used to raise an alarm when the current traffic data significantly differs from the past traffic, and subsequently retrain the model based on the new data. But an alarm doesn't mean that the data needs to be altered. Someone could be doing things to that data for bad purposes. This leads to the importance of not only detecting concept drift, but also adapting accordingly. Recent work in continual learning for NIDS thus focuses on adapting the models without risking the old data, such as by using replay memory or selective retraining. One example SSF (an offline drift detector and retrainer) selects samples that cause the drift, and discards old samples to detect and adapt the detector. These works show promising robustness when implemented on NSL-KDD and UNSW-NB15. However, not enough has been done to understand the difference between legitimate and malicious drift, and how to counteract the latter. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

2.2. Continual Learning and Catastrophic Forgetting

Continual learning gives NIDS an alternate approach to retraining from scratch through an incremental update, ideally keeping pace with drift. Parisi et al. review numerous continual learning methods that facilitate learning sequentially without retraining from scratch. Much of this work focuses on comparing methods that save model parameters so they can be retrieved in later training iterations, with those that only penalise changes to parameters that are significant for previous tasks. Most of all, you need to balance the practice of new attacks with remembering the old ones. A classic challenge is forgetting: models lose the ability to detect old patterns after each update. The literature highlights that overly conservative approaches (which try to remember everything) interfere with learning outdated patterns, even if they are no longer relevant. In ML NIDS, the cost of forgetting is particularly damaging: a detector can

be trained to recognize a new worm, but then "forgets" how to identify the old one. Several works (EWC, demonstration replay memory, dropout) have made inroads into limiting forgetting by replaying a sample buffer of carefully chosen attacks or encouraging learning by subtly restraining what can change. Yet existing work has examined forgetting with a narrow scope (e.g. Average accuracy across tasks), rather than in the presence of an adversary. Overall, continual learning brings techniques, but at a cost: legacy detection versus coverage of new threats. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

2.3.Adversarial Threats: Poisoning & Backdoors

The process of adaptation itself introduces attack vectors. Biggio and Roli, have demonstrated that ML models are susceptible to inputs deliberately assembled in time or at training to deceive the classifier. In an adaptive network intrusion detector, attacker motivation to do this is straight forward. The attacker controls the source of updates and can contaminate either the data stream or the update itself, it can orchestrate a poisoning attack that compromises the update data stream, injecting adversarially labeled or adversarially crafted samples into the data stream to corrupt the detector's future operation, or undertake a backdoor attack to actively degrade learning. A poisoning attack corrupts the data stream for updating the model. For instance, a compromised data stream could include attack samples inserted in such a way that a downstream update ends up ignoring the attack class. This is different from an evasion attack, which attempts to find samples that are misclassified without updating the model. Gu et al., have shown how an attacker could train a classifier "backdoored" by the model to achieve perfect accuracy except on inputs that contain a secret trigger pattern in a NIDS, a specific packet payload signature that appears in training as a trigger would be misclassified at test time. These attacks are complicated by updates; a drift detector may request an update driven by adversarial anomalies, or the update procedure itself may unwittingly incorporate poisoned influences. Overall, research on adversarial ML indicates that it is possible to compromise any update if the stream is not vigilantly monitored. However, existing NIDS research assumes benign data streams, leaving the relationship of backdoor & poisoning attacks with concept drift and continual learning under investigated. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

2.4. Integrating Adaptivity and Security

Few prior works address these issues together. In the NIDS domain, research on drift tolerant NIDS usually considers only changes in natural distribution, while adversarial ML papers almost never account for time adaptation in a timely manner.This means a simple method that, for example, "censors" all the samples for a model update will likely enable some of the truly new attacks to pass through and won't block some of the poisoned samples, whereas a suitably insensitive drift detector would likely fail to detect a slow, sophisticated poisoning campaign. Another interesting is to learn what patterns don't appear within a sophisticated slow attack: Maybe there's a conflict where weighted down anomalous update samples to prevent poisoning will also inadvertently weaken the effect of truly novel attack data, thus increasing forgetfulness for such samples. Again, this cross effect has not been previously explored in the literature. Other than the exception mentioned above, there are only studies considering both adversarial manipulation and forgetting in other tasks (e.g. Spam filtering, computer vision). In NIDS, experiments in this direction have mainly used static datasets (such as NSL-KDD, CIC-IDS, and UNSW), which do not by nature construct a time-sequential order and lack adversarial contaminated samples, and thus the dynamics of these interactions cannot be well studied. A recent paper recognizes the importance of time-series data and strives to establish time-series benchmarks to examine model updates more practically, but the application of this work to an adversarial setting is yet to be achieved. 3.4 To sum up, the literature reveals a clear but fragmented body of work on learning security and drift adaptation, with limited insight on their intersection. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

3. Objectives of the Review

The main purpose of this review is to investigate the joint impact of concept drift, continual learning, adversarial manipulation, and catastrophic forgetting on the security and the operational robustness of ML based NIDS. In order to accomplish this purpose, the review accomplishes the following purposes: (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

1. To study the main factors contributing to the drift of traffic concept and to explore the existing techniques for concept drift detection and adapting ML NIDS to the evolving network traffic. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

2. To classify the security threats to continual-learning methods, in particular data poisoning and backdoor attacks, and the risk of maliciously controlling the timing of detection of drift and the timing of updates. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

3. Look at how catastrophic forgetting happens with adaptive intrusion detection, and figure out how to balance learning new attack patterns while still keeping what you already know. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

4. Analysis of the benchmark datasets, experimental design, and evaluation metrics employed in these papers, as well as their impact on the ordering of the temporal data, multiple model updates, adversaries, and cross network evaluation. (Moustafa & Slay, 2015; Ring et al., 2019; Sharafaldin et al., 2018; Tavallaee et al., 2009).

5. To design a security aware classification and evaluation mechanism that integrates the detection of drift, the adaptation of models, the security of the update data, backdoor attack, retention of knowledge, and reliability after update. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

6. To identify open research issues and highlight potential avenues for the design and validation of continual learning ML NIDS in realistic and possibly adversarial scenarios. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

4.Research Gaps and Future Directions

This gap matters because in deployments, adaptive NIDS must withstand all threats as well, if not more, as nonadaptive systems. Open challenges in that regard include how to effectively differentiate between the presence of true concept drift, i.e. suspicious signal, and the attacker's purposely injected concept drift, how to reliably determine when and what to update without excess false alarms if attacked, how to overcome catastrophic forgetting when under attack, and how to evaluate such secure and adaptive detection, for example via manual analysis of time series traffic captures with simulacrum attacks or evaluation of the long term detection time series. Such benchmarks should incorporate sequential updates and injection of malicious concepts, and evaluate subtle attack signatures that could subtly exploit certain potentials of adaptive detectors. Promising research directions on the algorithms side would explore the

potential synergy between robust statistical methods, to reweight or dismiss suspicious updates, and replay or regularization methods, to preserve prior knowledge, thus overcoming the tradeoff between adaptability, security, and retention. To conclude, despite the fact that each paradigm has grown from an independent body of studies on drift, poisoning, backdoors, and forgetting, these paradigm studies do not seem to share a common understanding of their interactions, in ML based NIDS. Establishing this interaction seems critical to design ML based detection as security enabled and adaptive in the face of changing communications and adversarial patterns. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

5. Methodology

This research performs a security aware literature review to understand the interrelations between concept drift, continual learning, adversarial machine learning, and catastrophic forgetting. This type of review is systematic and hierarchical, takes into account the security dimension, and aims to integrate the literature of the four areas in the context of ML NIDS.Systematic review methods guided our review methodology through systematic reporting of the search, screening, data extraction, and data synthesis processes. Systematic review guidelines, such as the PRISMA 2020, allow an explicit framework for describing these review stages and increase the reproducibility of evidence synthesis. Furthermore, our review methodology adopts the approach to systematic literature review design, which was originally developed by computing and software engineering to guide researchers to prepare, conduct, and report this type of review systematically (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

5.1. Search Strategy

A focused search will be performed with the main online academic databases to find relevant papers in the area of cybersecurity and machine learning. The main used databases will be: Semantic Scholar, ACM Digital Library,Core, Science Direct, Springer Link and Google Scholar. If necessary, I will also look in the website arXiv for recent work in the research area, but without including preprint papers.

Search strategy: Search terms will be used to capture four main areas of interest: network intrusion detection, concept drift and adaptation, continual learning and catastrophic forgetting, and adversarial machine learning. Search terms will be tailored to the syntax of each individual

search engine. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

An example of a Boolean search string is provided below:

network intrusion detection Or NIDS Or intrusion detection system And machine learning Or deep learning Or continual learning Or incremental learning Or lifelong learning And concept drift Or data drift Or distribution shift Or adaptation And catastrophic forgetting Or forgetting Or replay Or regularization And data poisoning Or poisoning Or backdoor Or adversarial Or adversarial machine learning. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

Additional targeted searches will be conducted using combinations such as:

concept drift And NIDS,

continual learning And network intrusion detection,

catastrophic forgetting And cybersecurity,

data poisoning And adaptive intrusion detection,

backdoor attacks And network intrusion detection,

adversarial machine learning And continual learning.

The search will focus on studies published from 2014 onwards, as papers that have been published since then demonstrate an increase of interest in the fields of concept drift, continual learning, and adversarial machine learning. Although some hallmark early papers may be kept. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021)..

5.2. Eligibility Criteria

Articles will be selected for review according to inclusion and exclusion criteria.

Criterion	Inclusion criteria	Exclusion criteria
Topic relevance	Studies addressing ML NIDS, concept drift, continual learning, catastrophic forgetting, data poisoning, backdoors or	Studies not related to intrusion detection, adaptive learning, or ML security.

	adversarial robustness in adaptive detection systems.	
Publication type	Peer reviewed journal articles, conference papers, book chapters, and high quality review papers.	Editorials, posters without full papers, non-scholarly web content, slides, and duplicate publications.
Language	English language publications.	Publications not available in English.
Technical content	Studies reporting methods, datasets, experiments, threat models, defences or evaluation results.	Studies providing only high level discussion without technical or empirical relevance.
Security relevance	Studies addressing security, robustness, adversarial manipulation, update integrity or retention of prior knowledge.	Studies that discuss adaptation but provide no relevance to system security or reliability.
Accessibility	Full text available through institutional access, open access or author repositories.	Studies for which sufficient methodological information cannot be obtained.

Review articles will be used to help identify fundamental concepts and relevant primary literature. But when looking at methods, data, threat models, and results, primary research papers should come first.

5.3. Data Extraction and Quality Assessment

The following data will be extracted from all included studies with the use of a standardised form:

- publication details and study type.
- ML NIDS application context and datasets.
- Learning models, adaptation techniques, and continual learning algorithms.
- concept drift characteristics and detection methods.
- Methods for catastrophic forgetting like replay and regularisation.
- Threat models involving poisoning, backdoors, or adversarial tampering.
- Measures for verifying update data and safeguarding model updates.
- Design, practice patterns, performance data and important information.

Study quality will be evaluated based on whether a study clearly specified its data set, its ML NIDS task, its threat model, its adaptation process, and evaluation design. Focus will be placed on whether a study used cumulative data, incremental updates, its data set and its threat model in evaluation, its adapted system in retention testing, repeat testing, its adversary, relevant baselines, false positive rate, computational cost and reproducibility information or not. Each one will be coded as fully addressed,partially addressed or not addressed. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).

5.4. Security-Aware Analytical Framework

The studies will be synthesised using a three stage security framework representing the lifecycle of model adaptation.

Lifecycle stage	Primary concerns	Analytical focus
Pre-update	Traffic change, drift detection, data trustworthiness	Whether drift is detected accurately and whether natural drift can be differentiated from maliciously induced change
During update	Data poisoning, backdoors, compromised labels, unsafe adaptation	Whether update data are validated, filtered, monitored, or protected against adversarial manipulation
Post-update	Detection performance, catastrophic forgetting, residual vulnerabilities	Whether the updated model retains earlier knowledge, detects new threats, and remains reliable over time

Table 1. Summary of reviewed studies and principal findings

Study	Primary challenge	Approach	Data/evaluation context	Key result
Yang et al. (ReCDA)	Concept drift	Drift aware perturbation, representation alignment, and weakly supervised classifier tuning	UNSW-NB15; CICIDS-2017	Achieved 0.86 accuracy and F1 under the most severe tested drift condition; outperformed comparison

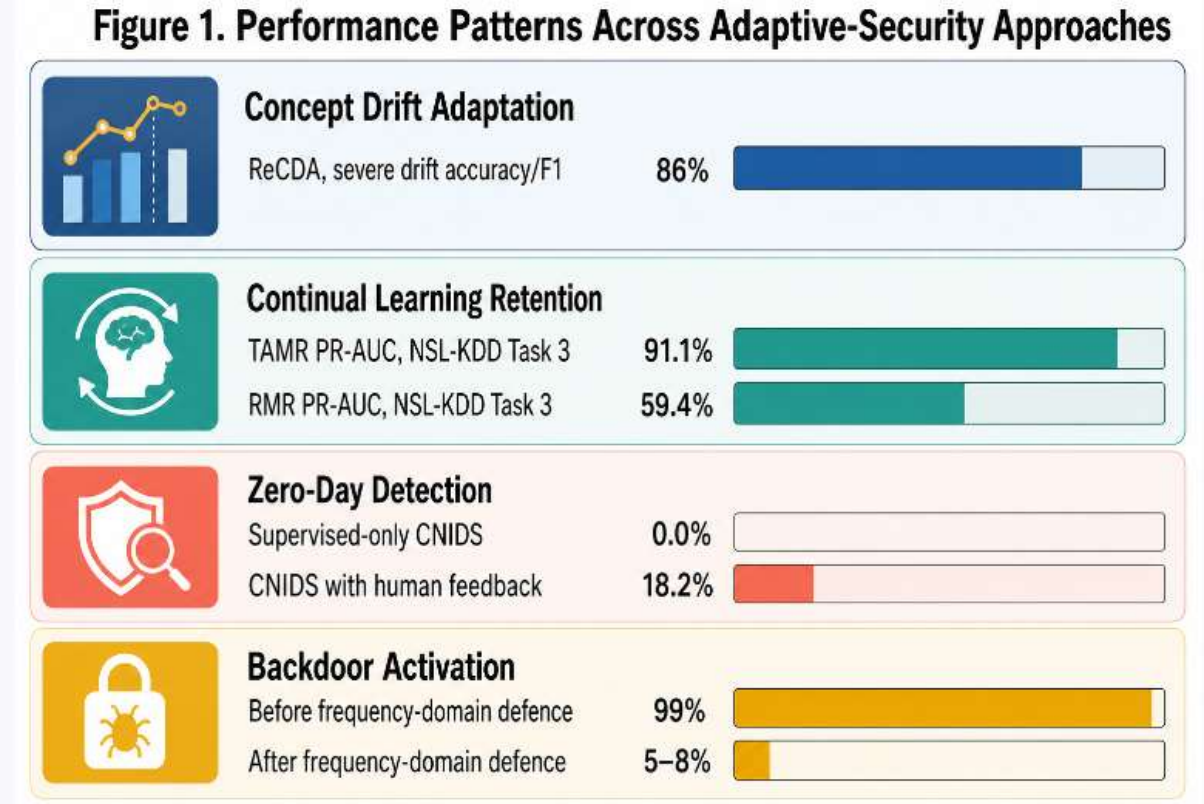
				methods across drift settings.
Yang et al. (CADE)	Drift detection and explainability	Contrastive latentspace distance learning and feature based explanations	Drebin and IDS2018	Achieved average drift detection F1 of at least 0.96; reduced analyst investigation effort.
Fathima et al. (TAMR)	Catastrophic forgetting and data drift	Task aware selective memory replay	NSL-KDD, CIC-IDS-2017, CIC-IDS-2018, Bot-IoT-2018	Maintained stable sequential task performance; PRAUC remained 0.9138–0.9105 on NSL-KDD.
Gunawan et al. (CNIDS)	Zero day attacks	Continual pretraining, persistent memory, and human feedback	NSL-KDD, UNSW-NB15, CICIDS2017	Zero day detection improved from 0% with supervised learning to 18.2% with human feedback.
Liu et al.	Poisoning and backdoors	Frequency domain clustering and structure preserving data reconstruction	New England 39 bus system	Detected 92.5% of poisoned samples and reduced backdoor activation to 5–8%.

This study is outside the NIDS domain but it is included as supporting evidence on data poisoning and backdoor risks in learning based security systems. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

Table 2. Comparative performance evidence

Security objective	**Baseline limitation**	**Adaptive/defensive result**	**Interpretation**

Drift adaptation	Traditional and pseudo-label based models declined substantially as drift increased	ReCDA retained 0.86 accuracy/F1 under severe drift	Representation focused adaptation improves robustness under distributional change.
Drift detection	Analysts may require extensive manual review of unfamiliar samples	CADE achieved approximately 0.95 F1 on IDS2018 with lower investigation effort	Sample level drift detection can prioritise analyst attention.
Knowledge retention	Sequential learning can cause major loss of earlier task performance	TAMR remained stable at 0.9138–0.9105 PR-AUC across NSL-KDD tasks	Selective replay mitigates catastrophic forgetting.
Zero day detection	Supervised only CNIDS recorded 0% zero-day detection	CNIDS with human feedback achieved 18.2% detection and 82.9% precision	Human feedback materially improves learning from low confidence alerts.
Backdoor resilience	Backdoors can activate in 97.2–99.1% of attack attempts	Frequency domain defence reduced activation to 5–8%	Adaptive systems need safeguards for the integrity of update data.



8.Discussion

This review elucidates the multifaceted challenges and advances in securing continual learning ML NIDS under dynamic network environments and adversarial conditions. The synthesis of concept drift, continual learning, adversarial threats, and catastrophic forgetting reveals critical interdependencies that must be addressed to realize robust and adaptive intrusion detection. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

Firstly, concept drift remains a fundamental obstacle, as static models degrade in performance when faced with evolving network traffic and attack patterns. The reviewed methods, such as ReCDA and CADE, demonstrate that drift aware representation learning and early, interpretable drift detection can substantially improve adaptation while reducing analyst burden. However, distinguishing between natural and maliciously induced drift remains an open challenge, underscoring the need for drift detection mechanisms that incorporate security aware differentiation. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

Secondly, it is suggested that methods like task aware selective replay (TAMR), which are characterized as continual learning methods, are preferable to prevent catastrophic forgetting.

While this method maintains the ability to detect older known attacks, the new ones can be absorbed too. The balance between plasticity and stability must be implemented for functional operation. Nevertheless, the approaches with forgetting are often tested under benign circumstances where no attackers can benefit from this flexibility. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

Thirdly, adversarial data poisoning and backdoor attack driven poisoning provide an interesting attack surface for adaptive NIDS. Frequency domain clustering has been studied as a possible solutions in areas of interest, but is not fully implemented nor tested on the NIDS backends. The emerging threat of adversaries 'drifting' the network behavior using the proposed model update process necessitates advanced data validation. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).

Furthermore, the identification of zero day attacks remains one of the challenges faced by automated systems. This research shows that there could be significant gains obtained when human in the loop feedback is adopted. This implies that completely autonomous adaptive systems would have to be combined with human feedback to attain zero day resilience. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).

These evaluation criteria may also be important. That static or random partitioned data may be insufficient to test the real world realities such as temporal effects and adversarial situations of deployment are also worth noting. A security aware classification and evaluation methodology appears to be relevant, considering that temporally ordered, adversarially enhanced datasets and measures might eventually predict long term retention, false positive control and computation feasibility. (Moustafa & Slay, 2015; Ring et al., 2019; Sharafaldin et al., 2018; Tavallaee et al., 2009).

However, solving any of these issues in information security sensors will require a holistic approach that goes beyond ad hoc solutions, which current research is also gravitating towards. Any future progress will need to bridge these fields in a comprehensive fashion to account for detection rate, update security, retained knowledge and constraints simultaneously. In addition, the remaining holes in adversarially aware forgetting mitigation, secure concept drift detection and realistic benchmark and evaluation provide a roadmap for the application of ML NIDS. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

9.Suggestions/Recommendations

- Develop Security Aware Drift Detection Mechanisms.Future research should focus on creating drift detection methods capable of distinguishing between natural concept drift and adversarially induced changes. Integrating security aware differentiation will reduce false alarms and prevent attackers from exploiting drift triggered updates. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).
- Improving Continual Learning in the Context of Adversarial Robustness.It is necessary to explicitly take adversarial threat models into account when designing continual learning systems and measure forgetting prevention strategies for attack situations. Additionally, studies indicate that combining selective memory replay and robust regularization with adversarial defenses can lead to a more stable and secure model. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).
- Integrate Data Validation and Monitoring in Update Pipelines.To safeguard adaptive ML NIDS, robust data validation, anomaly detection, and poisoning/backdoor mitigation techniques should be embedded within the update lifecycle. This includes real time monitoring of incoming data streams and systematic filtering of suspicious samples. (Biggio & Roli, 2018; Goldblum et al., 2023; He et al., 2023; Li et al., 2024; Wang et al., 2023).
- Integrate Human in the Loop for Zero Day Detection. Paper indicates that combining human in the loop with automated learning will have higher positive and negative predictive value in detecting zero day attacks. Additionally, hybrid system design can provide robustness towards novel attacks. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).
- Benchmark datasets, evaluation procedures should make use of the temporal order, adversarial contamination, multiple update cycles, and operational constraints (e.g., cost, false positives) in order to eventually develop more sound evaluation procedures for assessing adaptive machine learning based NIDS. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).
- Explore synergistic methods, e.g. between statistical approaches that discount unsafe updates and strategies that replay or regularize to preserve knowledge. Combining these

strategies is likely to mitigate the trade-offs between adaptivity, security, and retention of knowledge.

- Encourage Cross Domain Collaboration. Cross pollinating the networks of adversarial machine learning, continual learning, and network security would probably lead to all encompassing solutions that tackle the multi-dimensional issues of the field of ML NIDS. (De Lange et al., 2022; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

10.Conclusion

In this paper, we have explored the complex interaction of challenges that machine learning enabled Network Intrusion Detection Systems face in an ever changing and hostile environment. How concept drift, online continual learning, adversarial manipulation, and catastrophic forgetting have each proved to be significant barriers to maintaining a machine learning enabled intrusion detection system that is both adaptive and robust. While progress has been made with drift aware representation learning, task aware replay methods, human machine collaboration for detecting zero days, and innovative defenses against poisoning and backdoors, all the elements seem to be progressing in isolation. (Elwell & Polikar, 2011; Gama et al., 2014; Andresini et al., 2021; Korycki & Krawczyk, 2023; Yang et al., 2021).

However, a security conscious, integrated strategy is believed to be necessary to align adaptation with resilience. Distinguishing between natural and adversarial drift, preserving the accuracy of update data, trading off plasticity and stability, and realistically assessing the system performance over time are the core components of this strategy. Further, the benchmark datasets and evaluation protocols we used currently seem insufficient in addressing the temporal and adversarial nature of real world deployments. (Moustafa & Slay, 2015; Ring et al., 2019; Sharafaldin et al., 2018; Tavallaee et al., 2009).

Future work must prioritize developing holistic frameworks that jointly address these dimensions, supported by realistic datasets and rigorous evaluation metrics. By doing so, ML NIDS can evolve toward truly adaptive, reliable, and trustworthy systems capable of defending modern networks against ever changing and sophisticated threats. (Buczak & Guven, 2016; He et al., 2023; Khraisat et al., 2019; Mirsky et al., 2018; Sommer & Paxson, 2010).

11.References

Andresini, G., Pendlebury, F., Pierazzi, F., Loglisci, C., Appice, A., & Cavallaro, L. (2021). INSOMNIA: Towards concept-drift robustness in network intrusion detection. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security* (pp. 111–122). Association for Computing Machinery. https://doi.org/10.1145/3474369.3486864.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446.

Elwell, R., & Polikar, R. (2011). Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks, 22*(10), 1517–1531. https://doi.org/10.1109/TNN.2011.2160459.

Fathima, N. A. H., Khraisat, A., Syed Ibrahim, S. P., & Li, G. (2025). Adaptive memory replay for network intrusion detection: Tackling data drift and catastrophic forgetting. *Computer Networks, 272*, Article 111712. https://doi.org/10.1016/j.comnet.2025.111712.

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications, 50*, Article 102419. https://doi.org/10.1016/j.jisa.2019.102419.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44. https://doi.org/10.1145/2523813.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., & Goldstein, T. (2023). Dataset security for machine learning: Data poisoning,

backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(2), 1563–1580. https://doi.org/10.1109/TPAMI.2022.3162397.

Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). *BadNets: Identifying vulnerabilities in the machine learning model supply chain* [Preprint]. https://arxiv.org/abs/1708.06733.

Gunawan, J. A., Singgih, M. L., & Ginardi, R. V. H. (2026). Cognitive Network Intrusion Detection Systems: Anomaly and Malware Detection for Zero-Day Attack Resilience. Network, 6(2), 41. https://doi.org/10.3390/network6020041.

He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 25*(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793.

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity, 2*, Article 20. https://doi.org/10.1186/s42400-019-0038-7.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114.

Korycki, L., & Krawczyk, B. (2023). Adversarial concept drift detection under poisoning attacks for robust data stream mining. *Machine Learning, 112*, 4013–4048. https://doi.org/10.1007/s10994-022-06177-w.

Li, Y., Jiang, Y., Li, Z., & Xia, S.-T. (2024). Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35*(1), 5–22. https://doi.org/10.1109/TNNLS.2022.3182979.

Liu, Z., Liu, M., & Xu, J. (2026). Poisoned data detection in power systems based on frequency-domain clustering and key feature selection. *International Journal of Electrical Power & Energy Systems*. https://www.sciencedirect.com/science/article/pii/S0142061526005077.

Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.08840.

Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. In *Proceedings of the Network and Distributed System Security Symposium (NDSS 2018)*. Internet Society. https://doi.org/10.14722/ndss.2018.23204.

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *2015 Military Communications and Information Systems Conference (MilCIS)* (pp. 1–6). IEEE. https://doi.org/10.1109/MilCIS.2015.7348942.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks, 113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012.

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security, 86*, 147–167. https://doi.org/10.1016/j.cose.2019.06.005.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems* (Vol. 31). https://arxiv.org/abs/1804.00792.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the International Conference on Information Systems Security and Privacy* (pp. 108–116). https://doi.org/10.5220/0006639801080116.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). https://doi.org/10.1109/SP.2010.25.

Steinhardt, J., Koh, P. W. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.03691.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications* (pp. 1–6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528.

Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., & Ren, K. (2023). Threats to training: A survey of poisoning attacks and defenses on machine learning systems. *ACM Computing Surveys, 55*(7), Article 134. https://doi.org/10.1145/3538707.

Yang, L., Guo, W., Hao, Q., Ciptadi, A., Ahmadzadeh, A., Xing, X., & Wang, G. (2021). CADE: Detecting and explaining concept drift samples for security applications. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2327–2344). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/yang-limin.

Yang, S., Zheng, X., Li, J., Xu, J., Wang, X., & Ngai, E.C. (2024). ReCDA: Concept Drift Adaptation with Representation Enhancement for Network Intrusion Detection. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://www.semanticscholar.org/paper/ReCDA%3A-Concept-Drift-Adaptation-with-Representation-Yang-Zheng/ca100914a70cf077eaff278b4827cdeeee742868.

Zhang, X., Zhao, R., Jiang, Z., Chen, H., Ding, Y., Ngai, E. C. H., & Yang, S.-H. (2025). Continual learning with strategic selection and forgetting for network intrusion detection. In *Proceedings of IEEE INFOCOM 2025* (pp. 1–10). IEEE. https://arxiv.org/abs/2412.16264.

Žliobaitė, M. (2010). Learning under concept drift: An overview [Preprint]. arXiv.https://arxiv.org/abs/1010.4784.