Computers & Security 131 (2023) 103297

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

# A comprehensive review on deep learning algorithms: Security and privacy issues

Muhammad Tayyab[a,b,*], Mohsen Marjani[b], N.Z. Jhanjhi[b], Ibrahim Abaker Targio Hashem[c], Raja Sher Afgun Usmani[b], Faizan Qamar[d]

[a] Department of Computing, Shifa Tameer-e-Millat University, 46000, Islamabad, Pakistan
[b] School of Computer Science and Engineering (SCE), Taylor's University Lake-side Campus, 47500 Subang Jaya, Selangor, Malaysia
[c] College of Computing and Informatics, Department of Computer Science, University of Sharjah, 27272 Sharjah, United Arab Emirates
[d] Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia

ARTICLE INFO

*Article history:*
Received 8 August 2022
Revised 20 March 2023
Accepted 15 May 2023
Available online 19 May 2023

*Keywords:*
Machine Learning (ML)
Deep Learning (DL)
Poisoning attacks
Privacy attacks
Evasion attacks
Adversarial settings

ABSTRACT

Machine Learning (ML) algorithms are used to train the machines to perform various complicated tasks that begin to modify and improve with experiences. It has become widely used for automated decisions. In particular, the applications which have a profound impact on society that rely on Deep Learning (DL) for autonomous decisions, such as Patient Health Record (PHR), Unmanned Aerial Vehicles (UAVs), etc. Such impacts have a vital concern about the potential vulnerabilities introduced by DL. Traditional attackers have powerful motives that can alter and modify DL algorithms to subvert the outcomes. In poisoning attacks, an attacker can consciously change training dataset, which is used to operate the outcomes of decision-based model. While in privacy and evasion attacks, an adversary can also misclassify new datasets to infer private information. Therefore, in this paper, we have provided a review of security and privacy issues of DL algorithms and analyzed their applications and challenges based on state-of-the-art literature. We have classified attacks, devised a taxonomy, and comprehensive analysis of defense techniques for the most common attacks such as poisoning, evasion, model extraction, and model inversion. We have also presented various privacy preserving techniques to ensure the privacy of dataset. We have proposed a secure cryptographic framework for dataset based on hash functions and Homomorphic Encryption (HE) scheme. Finally, we have provided recent research challenges and future studies concerning security and privacy issues. We believed that the highlighted limitations and weaknesses provide possible research questions and open matters for designing efficient future DL algorithms.

## 1. Introduction

With the advancement in Machine Learning (ML), learning algorithms have shown remarkable trends with a direct or indirect link on various applications (Zuo et al., 2022). ML is an open-access model used for learning and training the complex model (Vedaldi and Lenc, 2020; Jia et al., 2014; Obukhov et al., 2018; Krasnyanskiy et al., 2020). Innovations in ML have introduced many solutions for data-driven models, such as health prediction (Shickel et al., 2018), security audits and surveillance (Buczak and Guven, 2016). It has achieved the maturity level that has entered into safety and security-of crucial learning models of various applications like Internet of medical Things (IoMT) (Liu, 2021), smart agriculture (Elhadj et al., 2021), traffic prediction (Hassan et al., 2020; Dourado et al., 2019) etc., Particularly, applications which have deep impact on data-driven problems in various fields for autonomous solutions were based on Deep Learning (DL) (Boullé et al., 2022). Using DL, it is easier to address many real-life applications to get high accuracy and throughput for predication. Most popular applications i.e., speech-recognition (Wagh et al., 2021), image processing (MirhoseiniNejad et al., 2021), language processing (Gan et al., 2021), surveillance models (Xu, 2020), malware detection (Zhong et al., 2020), and voice-controlled devices (Hinton et al., 2012), where DL has played a vital role for classification and prediction. (Biggio and Roli, 2018). DL has now been providing breakthroughs in many endeavors of learning algorithms in the recent past. Fig. 1 has shown the general working of DL algorithms and their relationship between input dataset and output results.

Similarly, DL is getting advanced gradually with the introduction of Deep Neural Network (DNN) model (He et al., 2016; Huang et al., 2017). DNN is one of the most used techniques among AI models that are designed on the idea of the biological nervous sys-

[*] Corresponding author.
*E-mail addresses:* tayyab.ssc@stmu.edu.pk, muhammadtayyab@sd.taylors.edu.my (M. Tayyab).

https://doi.org/10.1016/j.cose.2023.103297
0167-4048/© 2023 Elsevier Ltd. All rights reserved.

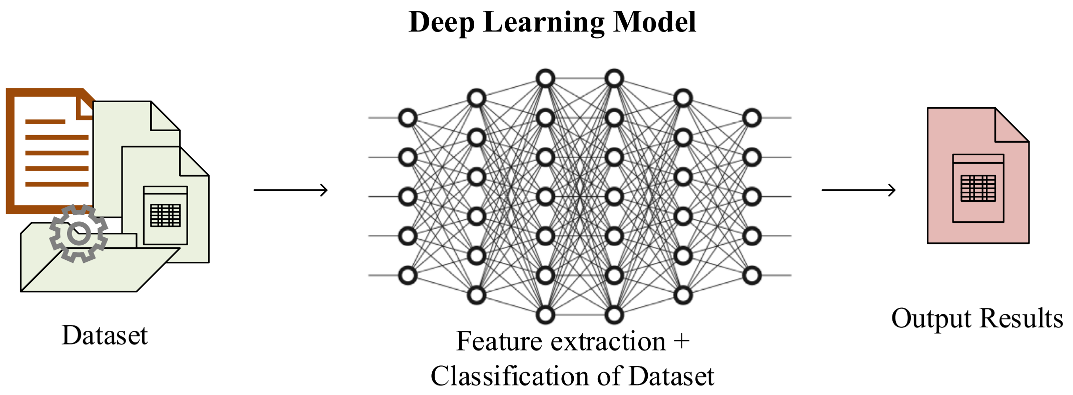

**Fig. 1.** Overview of Deep Learning Algorithms.

tem (Buduma et al., 2022). It consists of numerous neurons used to transfer information. It has the following stages to build a normal DL algorithm, like model training, where a huge amount of data is used to train in the learning model, and model prediction, where the model can predict the output according to the input (Otoum et al., 2022). It has also been recognized as one of the most efficient problem-solving techniques among challenging tasks. Among all the advantages of DL features, security is one of the challenging tasks for researchers (Li and Yang, 2021; Kumari et al., 2021). For example, an intelligent voice controlled system, where performance can be halted by unexpected protected commands (Yuan et al., 2018) injected by the attacker. To ensure the safety and security of an intelligent system, technologies like UAVs need lots of security tests before it has been launched (Tian et al., 2018). Similarly, in Patient Health Record (PHR), the patient record must be kept secret and restrict to medical staff only. For and attacker it is obvious that it can inject malicious dataset to the input, which can cause the failure of the PHR (Gilad-Bachrach et al., 2016).

DL has come across several security vulnerabilities where an attacker has significant advantages by getting the information from learning algorithms. Hence, after analyzing the algorithms, the attacker can manipulate the learning models, using different adversarial samples at the test phase, training phase, or learning phase (Xiao et al., 2018). In DL, commonly known threats are poisoning and evasion attacks, which are considered the most challenging and emerging security threats for data-driven technology, learning algorithms, and sensor networks (Mohanty et al., 2020; Hao and Tao, 2022). This is due to, the dataset can be taken from various untrusted resources where the labeling is crowded with numerous multi-label classes (Thiyagarajan, 2020). For this purpose, it is highly impossible to secure this dataset from any malicious manipulation by the attackers (Gamage and Samarabandu, 2020). When this malicious dataset is used for classification and prediction by the DL models, the output of models lead towards the motives of attackers. Hence, it is computationally difficult to formulate the optimize solutions against attackers (Caminero et al., 2019).

## 2. Motivation and contribution

DL provides significant improvements in solving the bigger problems that have withstood ML and AI's many attempts during the past few decades. DL has achieved milestones in solving challenging scientific problems like brain construction (Helmstaedter et al., 2013), analyzing the effects of mutation in DNA (Xiong et al., 2015), prediction of activity related to drug molecules (Ma et al., 2015) are examples of such achievements. However, there have been recently discovered vulnerabilities in DL from the extensive literature (Pouyanfar et al., 2018), which have effected DL algorithms. Primary motives of such security attacks were to get information of dataset and the model used get desired outcomes. To summarize the discussion, In Table 1, a comparison of the different existing surveys is presented, which consist of most recent studies and their limitations. In this study although we have covered security and privacy issues of DL algorithms and have provided the detail taxonomy of security attacks and their counter measures to mitigate the effects of security attacks. This will help the reader to get most recent related literature about the security and privacy issues in DL algorithms. Goecks et al. (2020) presented the current research survey regarding DL that has applied over the Electronic Health Record (EHR). It has highlighted the variety of DL techniques and architectures used to various applications relating to PHR and clinical data (Moosavi-Dezfooli et al., 2017; Muñoz-González et al., 2017). Research gaps concerning data modeling, data heterogeneity, model interpretability, and lack of benchmarks for analyzing the medical data have been presented.

Several studies have been conducted recently, which are focusing to solve this issue, however, to the best of our knowledge, no such work in literature is available which analyses the comprehensive study on following security attacks i.e., poising, evasion, model inversion, model extraction (Liu et al., 2018; Akhtar and Mian, 2018). Therefore, this study focuses on various security and privacy issues in DL algorithms based on state-of-the-art literature. It studied the different capabilities, target attacks, and workflow for several security attacks. It also provides the several privacy preserving techniques with logical mechanism and solutions against these security attacks. The primary motive for this research is to mitigate the effects of security attacks based on three main questions i.e., How to identify the attacks and its type in the DL algorithms? How to protect the DL algorithms from different algorithms? and how will the defense technique behave to the DL algorithms?

This research contributes the comprehensive review of security and privacy issues of DL algorithms and analyze their challenges based on state-of-the-art literature. The main contributions of this study are as follows:

a) The recent related studies concerning to security and privacy issues in DL applications has been provided.
b) The detailed analysis of most common security attacks such as poisoning, evasion, model inversion and model extraction have been presented.
c) Several privacy preserving techniques have been presented to ensure the privacy of dataset against secure attacks.
d) A cryptographic based secure framework has been proposed based on hash functions and Homomorphic Encryption (HE) scheme to provide the security and privacy of dataset.
e) Several recent research challenges and future studies concerning security and privacy issues have been addressed in detail.

**Table 1**
Comparison of existing survey papers.

| Refs | Year | Strength | Attacks | | Contribution | Limitations |
|---|---|---|---|---|---|---|
| | | | Blackbox | Whitebox | | |
| (Gamage and Samarabandu, 2020) | 2020 | Provided the review of attacks and defense techniques | √ | √ | Methods only under AI concepts | Only focuses on privacy attack |
| (Caminero et al., 2019) | 2019 | Attacks are classified into three dimensions | √ | √ | Provided the summaries of literature | Limited to ML applications. |
| (Santos et al., 2022) | 2022 | Presented recent literature review of the methods that have been applied to road crash injury severity modeling | x | x | Included 56 studies from 2001 to 2021 that consider more than 20 different statistical or machine learning techniques. | Only considered attack related to road crash injury on ML |
| (Zhang et al., 2022) | 2022 | Reviewed the history, state of art and the future of the DL's application in power system frequency analysis and control | x | x | The traditional analysis methods are increasingly unable to meet the online analysis needs of large-scale power grids. | This paper has analyzed the connotation and characteristics of DL and has summarized the eight kinds of typical DL structures basic principles. |
| (Wang et al., 2022) | 2021 | Summarizes and categorizes ML applications in this domain, categorizes and discusses data types used for ML modeling, and provides suggestions for data sources and input variables for future ML applications. | √ | √ | Based on three scientific literature databases: Scopus, CAB Abstracts, and IEEE | Did not cover the meta-analysis of such studies. |
| (Ren et al., 2020) | 2020 | Summarized the defense models against adversarial attacks. | Adversarial attacks | √ | Investigated the ideas and methodologies, algorithms used for adversarial attacks | Required effectiveness and efficient mechanism against adversarial attacks |
| (Papernot et al., 2018) | 2018 | security and privacy of ML systematically | √ | √ | Summary of Defense techniques | Methods are not comprehensive |
| (Liu et al., 2018) | 2018 | Focuses on data distribution and information violation | Considered on two types of attacks | √ | Discovered active attacks on ML | Limited to statistical ML algorithms |
| (Akhtar and Mian, 2018) | 2018 | Summarize 12 attacks on DL for methods for classification | Poisoning and evasion attacks | √ | Summarize the defense techniques | Focused on adversarial attacks |

The rest of the paper is organized as follows; In Section 3, we have provided related work of security and privacy challenges in different DL related applications. In Section 4, security attacks in DL, different categories of attacks, and defense techniques including adversarial settings for various attacks have been presented. While in Section 5, we have presented privacy preserving techniques in DL that are based on different cryptographic functions. While in Section 6, we have proposed a cryptographic framework to preserve the security and privacy of dataset. In Section 7, we have presented the discussion of security and privacy of DL algorithms. In Section 8, several recent research challenges and future studies concerning security and privacy issues to design new DL algorithms. Finally, in Section 9, we have concluded the paper and provided different future studies for researchers in the conclusion section.

## 3. Security and privacy issues

The DL performs a wide variety of learning algorithms in various application to minimize the human interactions. The enormous amount of training data has encountered many security challenges like security and privacy issues in dataset used such learning algorithms. To overcome such situations, Carlini and Wagner (2017) has proposed a scheme named Secure and Private AI (SPAI), which aims to provide security and privacy of data by offering the mechanism to mitigate the attacks. Extensive literature about privacy-preserving techniques, working, and classification of attacks were discussed. In another study, a simple operation of homomorphic properties limits adaptation to a complex model such as Neural Network (NN) is also explained to overcome the security challenges in DL. Caminero et al. (2019) has conducted a study on the learning algorithm's security and computed a taxonomy of attacks against learning algorithms. Data poisoning is one of DL's most critical security concerns, where the attacker's primary motive was to overturn the learning phase by introducing malware injections to sampling data of training data. Ovadia et al. (2019). presented the model to mitigate the effect of optimal poisoning attack in ML models based on outlier detection. Generally, it is assumed that learning data is coming from a secure distribution, but this is not true for practical scenarios and security analysis (Carlini and Wagner, 2018). Maiorca et al. (2020) have proposed a technique over an attack based on a gradient ascent strategy, which works on Support Vector Machine (SVM's) optimal solutions properties. However, such a scheme has increased the classifier's test error data for SVM.

Therefore, manipulating DL in terms of poisoning attack remains challenging for researchers when the attacker's motives are to manage learning algorithms' results by violating the training dataset. A theoretical model was designed explicitly for linear regression and demonstrated its efficacy across various data models. For defense purposes, The authors in Breuer et al. (2020) has proposed a model named as TRIM, to provide remarkable robustness. It has also provided high resilience against such attacks. Moreover, Ateniese et al. (2015) focus on ML classifiers and statistical information that could reveal unconsciously or maliciously. DNN is also vulnerable to adversarial examples that may perform malicious activities while being unmodified to observers. A significant number of attacks have launched against training dataset, but controlling a remote system which may be hosted by DNN, that can be a security risk (Papernot et al., 2017; Altaf et al., 2019).

Like many other DL models, they have recently been shown to lack robustness against adversarial crafted inputs. Papernot et al. (2016) highlighted the method to construct highly effective adversarial samples crafting attacks for the Neural Network (NN), used as malware classifiers. Such malware classification introduces additional constraints in the adversarial samples crafting problem compared to the computer vision domain. It has demonstrated the feasibility of many attacks on different instances of malware classifier that can be trained using the model termed as DREBIN Android malware dataset (Shi-qi et al., 2019). It has also evaluated the potential defense mechanisms against such crafted adversarial examples.

### *3.1. Cyber security issues*

DL is also applied in cybersecurity and reliability applications. Chen et al. (2020) has presented a survey in the context of the application of DL in cybersecurity and reliability. It also highlighted the loopholes in mobile artificial intelligence and put it open for future studies. As the attacker's primary motive is to get sensitive data from the system. Biggio et al. (2013) simulated the attacker scenarios, which shows various risk levels for classification and prediction using the system's information. These attacks present a better view of the classifier performance using the evasion attack model, where parameters can easily be selected to perform the model selection. Such type of attacks only uses PDF files, and it was found that the type of system can easily be manipulated.

Traditional encryption schemes have higher computational costs and significant time complexity due to the complex nature of encryption for cryptography and Data Encryption Standards (DES). Faster CryptoNets, CryptoNets (Roy Chowdhury et al., 2020; Kaissis et al., 2020) was proposed in the NN model to overcome time complexity and increase cost factors by using differential privacy and leveraging transfer learning dataset (Chang et al., 2021). HE allows the features that can preserve not only the privacy of dataset used for learning the model as well as HE provides confidentiality of dataset. HE provides arbitrary computations over the dataset while being encrypted (Mishra et al., 2020; Pan et al., 2020). Such secure methods were very attractive for ML, but on the other side it also increases the computational cost significantly due to the complexity nature of HE and other cryptographic functions (Carlini et al., 2020). When ML meets 6 G, there has been new opportunities that emerged but at the same time numerous privacy challenges can also be the part of these new features which needs to be addressed (Wu et al., 2021). ML has also introduced a secure architecture for 6G while ensuring the privacy and security using cryptographic functions (Sun et al., 2020). While ML has been taking part to ensure the security issues in 6G, DL has also been widely used to get the information from different networking protocols for the betterment of quality of services and user quality of experience (Siddiqui et al., 2022). DL has also highlighted many privacy concerns in 6G as well as 5G (Jhanjhi et al., 2020) heterogenous networks due the numerous attacks in such environment (Furqan et al., 2021).

### *3.2. Distributed environment issues*

DL has achieved a milestone in the form of prediction and classification in the distributed environment. However, as per applicable obligation and potential privacy issues in a distributed setup, it is hard to summarize the raw material or data from all the data owners for learning purposes. To solve the mentioned issue, Hong et al. (2020) introduced privacy-preserving approaches in a distributed environment without revealing the private part or statistics of data despite using traditional methods that rely on cryptographic techniques. Chiu et al. (2020) introduced the efficient-privacy DL scheme for hierarchical distributed systems, in which old approaches were modified and improved the collaborating algorithms. Such methods increased the algorithm's efficiency in distributed setup and increased the computational overhead due to the previous traditional cryptographic approaches. It also provides comprehensive protection for each layer in layered scenarios of distributed architecture (Mei and Zhu, 2015).

## 4. Security attacks

ML and DL algorithms are on the rise and made available to the masses through the user-level interface or public query interface. Forecasting models play a vital role in making decision machines, especially in financial ventures. Such a trained model or learned model is often trained on data from different sources, which can have untrustworthy elements (Kok et al., 2020). An intruder, intentionally or unintentionally, can modify the input data (poison the training data) or learn the model using model behaviors (Evasion or Inference model), and lead towards miss prediction. We have provided a systematical overview of all the taxonomy features to analyze DL algorithms' security, as shown in Fig. 2. We have categorized the security in three major dimensions, attacks, privacy-preserving techniques, and the adversarial setting. In the case of poisoning attacks, it has been further classified into strong adversaries and weak adversaries. While evasion attacks are further categorized into Blackbox and Whitebox attacks, respectively.

Security is one of the challenging features of DL algorithms. The two unique and important categories of attacks have affected the DL algorithms: Poisoning attacks and Evasion attacks. Suppose an attacker engages in the training phase of learning algorithms and subvert the normal process, it is called poisoning attacks, and the examples used are named as adversarial examples. Similarly, the adversarial examples used to subvert the classification during the testing phase are commonly known as evasion attacks. As part of the DL model, an attacker's scenarios can differ based on the information the attacker can have about the learning models (Ch et al., 2020). Firstly, suppose all the information, parameters, and values of the model are known to the attacker. In that case, a very high success rate is probable, this is known as Whitebox attacks, but this scenario is highly unlikely. Secondly, suppose the attacker has limited knowledge about the model's input labels or the attacker with limited authorities. In that case, it is tough to be malicious, and an alternate method is needed to substitute the model or data; this is called a Blackbox attack. Moreover, there exist two significant kinds of attackers: a targeted attacker and a non-targeted attacker. Suppose the attacker motives to change the classifier's output to some other defined target label, known as targeted attacks. While, if the attacker's motive is only miss-classified in such a way that it may choose the incorrect label, known as the targeted attacker. In general, the non-targeted attacker has a high success rate as compared to the targeted attacker.

### *4.1. Poisoning attacks*

Poisoning and evasion attacks are prevalent for learning algorithms in AI. In poisoning attacks, attackers aim to intrude on the learning dataset (Guan et al., 2018). Spam filter (Zhou et al., 2019), support vector machine, DNN (Huang et al., 2020) and classifier systems (Cao and Gong, 2017) have been used to examine the poisoning and evasion attacks (Kumar et al., 2021). In poisoning attacks, attackers intentionally insert malicious examples into the training set to divert the learning process or mislead or misclassify the trained data toward the wrong classification (Dunn et al., 2020). In comparison to poisoning, evasion attacks can affect the decision boundary of the model at test time. However, several methods for poisoning attacks have successfully been launched

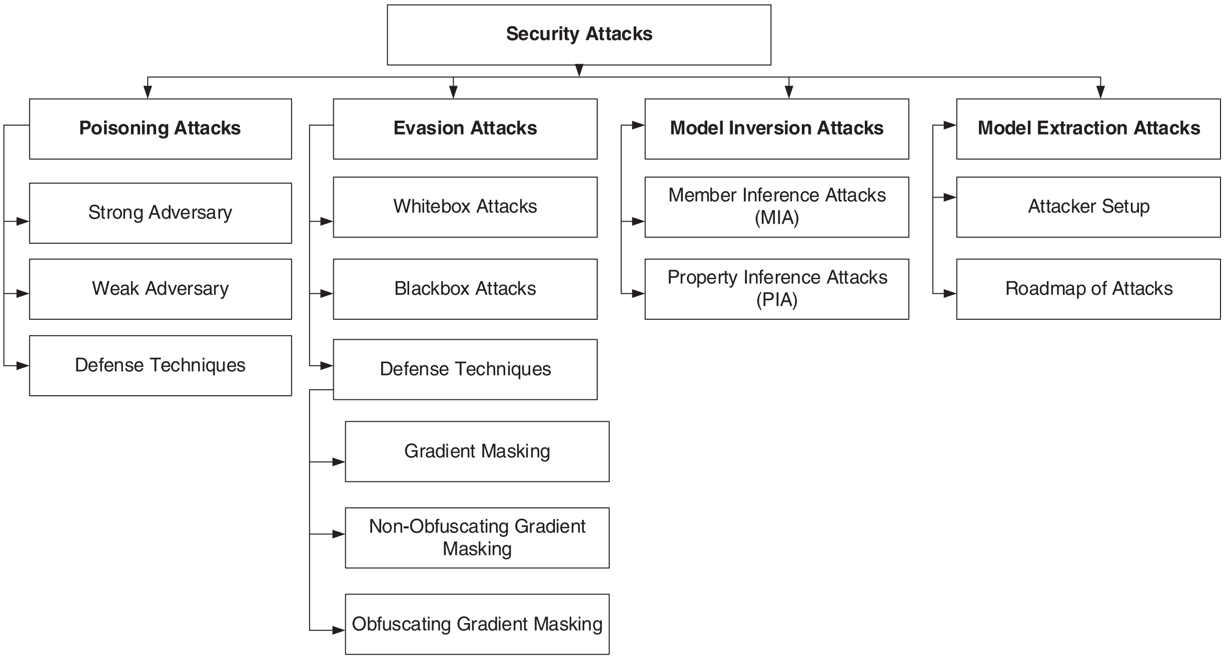

**Fig. 2.** Taxonomy of security issues in deep learning model.

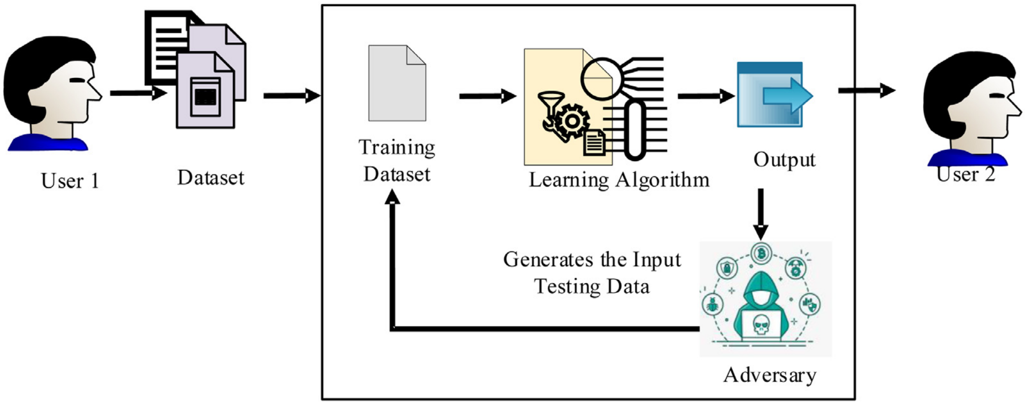

**Fig. 3.** Working of poisoning attack.

against traditional ML algorithms such as SVM or LASSO, and very less for NN (Paudice et al., 2018). In Fig. 3 shows the basic working of poisoning attacks.

**Adversary model:** The poisoning attack model can easily be implemented with either full knowledge, i.e., the Whitebox and limited knowledge, i.e., the Blackbox. Mainly the knowledge can also be termed as the information regarding the training process, including the model architecture and the training algorithm (Quiring and Rieck, 2020). The attacker can only be limited to the training dataset. It also creates a borderline for a new poisoned dataset inserted by the attacker and the labels that have been changed in the original data (Shafahi et al., 2018).

**Attacker Goal:** The attacker has the following main objectives to poison the learning dataset for the model. Primarily, the first objective is to destroy the availability of models to be useful for prediction or classification. Such deviation of the model creates the wrong prediction for malicious data other than the original data. Secondly, the main objective for attackers is to modify the model to make the wrong prediction or classification by using adversarial examples (Pang et al., 2020). That is mainly caused by the pre-implanted backdoor data, which can modify or trigger the original data, where the adversary might be able to alter the prediction score and may also be able to get more information regarding the model as well as the behavior of the model.

**Workflow:** The poisoned data is mixed with original data, and it may subvert the training process, which can lead to significantly degraded prediction capability of the model. Therefore, by looking closely, mislabeled data is changed by selecting specific records of interest, and flipping labels, making the data confusing. Such confusing data is crafted by the intrusion of special features trained by the model to target misclassification (Mnih et al., 2015). These unique features are then used as a trigger to make the wrong classification. The following are the different classes of poisoning attacks.

### 4.1.1. *Strong adversaries*

Strong adversaries are a type of poisoning attacks where the adversary has full access to the model and the training data. The attacker can easily modify the parameters and labels to subvert the normal learning process by injecting malicious examples. Bun and Steinke (2016) have classified two kinds of scenarios for strong adversaries. Perfect Knowledge (PK), where the adversary has full knowledge about the targeted model, can easily modify the model's parameters using malicious examples. It is the worst-case evaluation of the model. Limited Knowledge (LK) attacks (Lotfollahi et al., 2020), like the Blackbox attack, also have restricted access to dataset and model.

Mothukuri et al. (2021) and Visaggio et al. (2021) have proposed a back-gradient optimization to solve the optimization problem and generate the adversarial examples for the comparison with the previous gradient-based optimization methods. Similarly, Yang et al. (2017) proposed a similar method applied to DNN and developed a much similar and comprehensive method inspired by the Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Instead of computing gradient directly, auto-encoder is used by the GAN as a generator. As a result, the optimization has improved up-to 200x than the previous method, i.e., the gradient-based method (Yang et al., 2020).

### 4.1.2. Weak adversaries

In a weak adversary, intruders can add some poisoned examples into the model or training data without any authorized comparative study, Chen et al. have three essential components like 1) No knowledge of the model, 2) Injection less knowledge of training data and art, 3) Humans cannot detect poisoning data. These situations are like the real world. Patil et al. (2014) has introduced a method that will be used to resolve weaker adversaries. Such a method can also be used to demolish the security of the facial recognition model. A key image was used to input-instance-key strategies to make the image recognized and labeled as a targeted label using an image. Meanwhile, three strategies were proposed for pattern key strategies are *Blended injection strategies*: it blends the kitty image onto the input image. However, it is not reasonable to have a specific pattern to capture an image from the camera (Papernot et al., 2017). *Accessory injection strategies:* apply an accessory like glasses to the input image. It was easy in the inference stage and *blended accessory injection strategies*: it combines first and last to improve the accuracy. The model has created a successful loophole for a backdoor attack by simply adding a very tiny fraction to launch poisoned samples (Jiang et al., 2020).

### 4.1.3. Methods to overcome poisoning attacks

There have been numbers of methods proposed by the different research areas to overcome the effects of poisoning attacks in DL. The framework proposed by Steinhardt et al. (2017) and Behzadan and Munir (2017) removed the outlier of the model's decision boundary, which was located outside the application set. In binary classification, the fundamental goal is to get the primary centroids like a false positive and false negative. In this scenario, the proposed method removes the far off points from each corresponding centroid (Koh and Liang, 2017). Sun et al. (2018) proposed a defense mechanism to reduce the consequences of attacks by removing outliers of the decision boundary. The adversary tends to attempt several times to affect the defender with a significantly smaller number of poisoning data points (Chen et al., 2020; Tang et al., 2020). For this purpose, it has divided the data into two halves; after that, curated data trains based on outlier detection was created for each class. Such an algorithm measures the outlier score for every dataset entry x in the original dataset (Zhao et al., 2020; Kim and Reeves, 2020).

Subsequently, the Dasgupta et al. proposed a scheme such that instead of removing the outlier, it relabeled the data point to be considered an outlier. For example, the flipping of the label is a unique type of attack where an intruder can inject malicious labels to a training dataset to perform adversaries. It has also proposed a mechanism that can be able to detect far off data point as poisoned or malicious inside the model's decision boundary and then re-classifies them using K-NN (Takiddin et al., 2020; Lovisotto et al., 2020). For every sample, using Euclidean distance, it measures the distance and reassigns the label for each sample. If the data number points label (most common labels) using K-NN are greater than or equals to a certain threshold, the concerning training sample will be labeled "The most common label" in the K-NN. Below is the Table 2, that summarize the above discussed poisoning attacks.

## 4.2. Evasion attacks

In evasion attacks, the attacker's main motive is to add noise to a standard testing adversarial example. The classifier classifies such indirect labels as noise. In terms of geo-metrics, evasion attacks shift testing example class to another example class. Evasion attacks are categorized into the following categories. In Fig. 4, the detail explanation of evasion attacks have been presented with active scenario where it can be enlighten that how such kind of attacks are initiated and how it harms the dataset while training and testing the ML or DL model.

### 4.2.1. Blackbox attacks

Training the model data or dataset for a specific model is a challenging task. Although several open data like images, audio, and videos, among others. Additionally, models used for mobile devices are not accessible for the intruders. If the attackers have no information about the model's inner workings, it is known as the Blackbox attacks, closer to reality. The only available information for the attacker is the input pattern and label of the affected model. The target model can be one of the renowned application servers, like Amazon or Google, where the attacker can create an aggressive example (Tariq et al., 2020). The attacker only needs the replacement model as he wants to attack the target model but have no access. In this way, the attacker can regenerate the model with no specific information on the goal's architectural model (Sun et al., 2018). Hence, Goodfellow et al. (2020) have proposed that the NN can only attack the targeted model without caring for the numbers of layers and secret nodes as long as the target is the same. This unique property is due to the linear nature of the NN.

i. **Targeted Fast Gradient Sign Method (T-FGSM)**

Targeted Fast Gradient Sign Method (T-FGSM) was based on the boundary of DNN and proposed by Goodfellow et al. (Elsayed et al., 2019). It was designed to obtain adversarial examples faster, without any new addition of noise. Hence, this model's example has provided a low success rate compared to an optimized attack while adding some noise. Using mathematical notations, T-FGSM can also be represented as:

$$x' = x - \varepsilon.sign(\nabla J(\theta, x, t)) \quad (1)$$

Such that $\theta$ represent the DNN parameter, $\nabla$ used for gradient, targeted label denoted by $t$, noise parameter by$\varepsilon$ and $J$is the parameter of cost-function for training DNN. It is also observed such a technique aims to reduce the noise added by the factor $L_\infty$

ii. **Targeted Iterative Gradient Sign Method (T-IGSM)**

Targeted Iterative Gradient Sign Method (T-IGSM) is an updated version of T-FGSM proposed by I. Goodfellow et al. (Stutz et al., 2020). In simple words, it adds small noise to generate the adversarial examples until it reaches the maximum number of iterations. Mathematically it can be represented as follows:

$$x'_0 = x, x'_{N+1} = Clip_{x,\varepsilon}(x'_N - \alpha.sign(\nabla J(\theta, x, t))) \quad (2)$$

Where $\theta$ represents a parameter for DNN classifier, $\nabla$used for gradient, $t$ is target label, $\varepsilon$control the ratio between noise and success rate of T-IGSM, $J$ denotes cost function, $\alpha$ is a small size step and $Clip_{x,\varepsilon}$ generates the adversarial examples.

**Table 2**
Summary of poisoning attacks affecting deep learning.

| Category | Ref | Years | Techniques | Research Gaps |
|---|---|---|---|---|
| Strong Adversaries | (Ha et al., 2020) | 2018 | Provide the ProSec technique to provide privacy and security | Outlier detection constrains the decision boundary too much. |
| | (Guan et al., 2018) | 2018 | Review the certain literature in the application of ML on cybersecurity | ML methods can be possible to apply in more security areas. |
| | (Tolpegin et al., 2020) | 2018 | Outlier detecting is used to remove the changes made by poisoning attacks | Proposed algorithms do not include any explicit mechanism to model detectability constraints |
| | (Paudice et al., 2018) | 2018 | Efficient Algorithms to perform. optimal label flipping attacks | Outlier detection might constraints the decision boundary too much. |
| | (Lotfollahi et al., 2020) | 2020 | Perspective toxic poisoning attacks based on the detection system | Only applicable to latest Google's perspective API Built |
| | (Abramson et al., 2020) | 2020 | Articulate the comprehensive threat model for ML | Selected features were used to train the model and adversarial setup |
| | (Tran and Dang, 2021) | 2021 | Fraud detection through ML algorithms: random forest, k-nearest neighbors, decision tree, and logistic regression by selecting only two resampling and effective techniques | Extending and applying ML techniques over highly skewed datasets to other application domains like big data sampling and clustering, recommendation systems, and security and privacy issues with DL |
| Weak Adversaries | (Patil et al., 2014) | 2019 | Adversarial Perturbations against DNN for malicious prediction. | Only some features used to manipulate or create the adversarial examples |
| | (Papernot et al., 2017) | 2019 | Introduced the practical demonstration of an attacker controlling a remotely hosted DNN with zero knowledge | Used an online API to perform attack and analyses the Blackbox attacks |
| | (Behzadan and Munir, 2017) | 2020 | Presented a novel class of attacks based on vulnerability that enables policy manipulation and induction in the learning process of DQNs | Adversarial examples can easily manipulate reinforcement learning agents. |
| | (Sun et al., 2018) | 2019 | The presented method that impacts the adversarial data augmentation in experiments on the Aurora-4 | During training, the fast gradient sign method is used to generate adversarial examples |
| | (Ha et al., 2021) | 2021 | Leakage of private information through the parameters of the fully trained model and the parameter updates of the model during training for Black Box and White box attacks | This was only tested on one ML algorithms. The approach is comprehensive, but it is limited to specific GAN algorithms using standard dataset. |
| | (Hathaliya et al., 2022) | 2022 | ML and DL techniques have been used for attack detection in a wireless channel on an early basis. | ML and DL based techniques were used to detect the false detection in wireless networks areas inside cloud computing. It is limited to false detection rate and have increased the computational cost. |

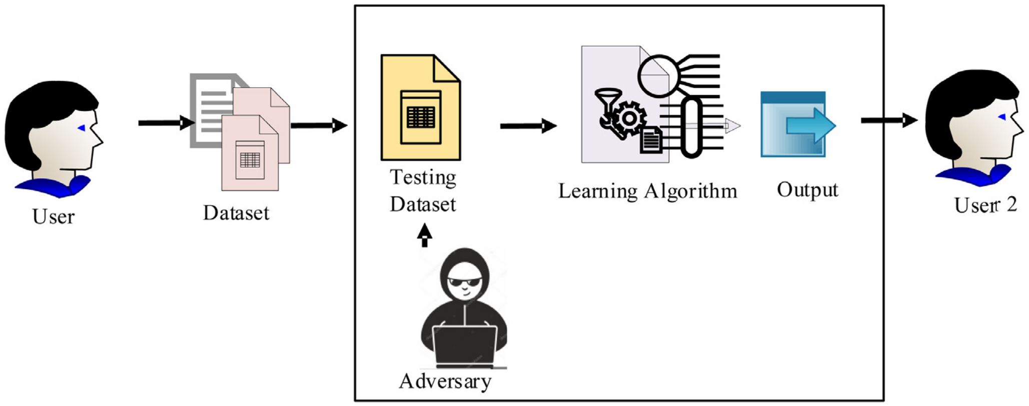

**Fig. 4.** Working of evasion attacks.

### iii. Targeted Jacobian-based Saliency Map Attack (T-JSMA)

Targeted Jacobian-based Saliency Map Attack (T-JSMA) attack has aimed to locate the adversarial examples with less addition of noise $L_0$ (Jia et al., 2020). It adds a small amount of noise to the model to initiate the malicious samples until the classifier comes up with the target label t as a max label. Therefore, after identifying the maximum label, T-JSMA randomly selects one or two entries of the adversarial examples and starts improvement with the example that can attack the targeted label. This will either increase or decrease the entries depending upon the values are being selected.

### iv. Targeted Carlini and Wagner (T-CW-$L_2$) Attack

Targeted Carlini and Wagner (T-CW-$L_2$) belongs to the family of targeted evasion attacks proposed by Carlini and Wagner (2017). These kinds of adversarial examples can also be generated successfully with minimal addition of noise into the model. It has three versions that are as $L_0$, $L_2$and $L_\infty$norms, respectively (Zhang et al., 2018). The T-CW-$L_2$attack can be tailored to figure out an adversarial example with minimal noise added to the model by $L_2$.

### v. Targeted Carlini and Wagner (T-CW-$L_0$) Attack

Targeted Carlini and Wagner (T-CW-$L_0$) attack can tailor to locate and measured adversarial examples with minimal noise added to the model by $L_0$ norms. It can detect and identify the dimensions of 'x' iteratively, which has not much impact on classifier prediction and fixing them properly. This process keeps on processing and identifying unless and until it can identify and construct successful adversarial examples. In each turn, a set of dimensions can resolve by T-CW-$L_2$ attacks. Moreover, in-depth, T-CW-$L_0$ calls

T-CW-$L_2$ to fix the unfilled dimensions in each round (Raschka et al., 2020).

vi. **Targeted Carlini and Wagner (T-CW-$L_\infty$) Attack**

Targeted Carlini and Wagner (T-CW$L_\infty$)attack can find out the malicious examples with a minimum of noise factor by $L_\infty$norm (Sadeghi et al., 2020).

$$\min \sum (\sigma_i - \tau)^{+} c * f(x + \sigma) \tag{3}$$

Where f denotes the function same as in T-CW-$L_2$; while $(\sigma_i - \tau)^{+} = 0$ if $\sigma_i < \tau$, and it will be as $(\sigma_i - \tau)^{+} = (\sigma_i - \tau)$. T-CW-$L_\infty$works over the value of $c$ until it finds the successful adversarial examples. For every value of $c$, CW-$L_\infty$ iterates over the value of $\tau$where $\tau = 1$.

vii. **Deep Fool**

The untargeted evasion attack has been proposed by Moosavi-Dezfooli et al. (2016), known as Deep Fool, to differentiate classifiers. The basic idea was to add some small noise into the adversarial examples to mislabel or predict the incorrect label for the current example, or it may reach its maximum iteration to achieve its goal. In each turn, the linear classifier for the present time malicious example was classified by Deep Fool and started to find some minimum noise to move such malicious examples to the class's decision boundary. All the attacks are categorized as transfer attacks, and these attacks are both targeted and non-targeted. Due to such intrusions, the learning algorithm miss classifies, or miss leads the learning data.

### 4.2.2. Whitebox attacks

Whitebox Attack is the attacker with full knowledge of the algorithm and can launch adversarial examples to get fruitful information. This study was started by Szegedy et al. and proposed a novel idea using Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) (Ha et al., 2020) for the generation of adversarial examples. It is a targeted model that includes the solution of the simple box-constrain optimization problem as:

$$f(x + n) = \bar{l}, \min |n|_2 \tag{4}$$

Where $x$ is the real numbers as $x \varepsilon R^m$ and $\bar{l}$ represents the target label for adversarial examples, $n$ shows the number minimum noise? It can be added for the successful launch of attack to attain desired results such that it can miss classifying the image. The above method will identify the small disturbance required for an efficacious attack. Although they have a very high success rate, such attacks also have very high computation cost as the malicious samples were created by the same Eq. (4) as used by others. CW-Attack depends upon L-BFGS and CW-Attack refined to the new equation as:

$$\min D(\bar{x}) + c.g(\bar{x}) \tag{5}$$

where $D$ is the distance matrix, $g(\bar{x})$represents the objective function where $f(\bar{x}) = 1$only in case such that $g(\bar{x}) > 0$and $c > 0$when it is chosen carefully like some constant. Papernot et al. (2016) introduced a new type of attack, named JSMA, which was also targeted and optimized under the $L_0$distance.

Some factors reduced the efficiency of most of the adversarial attacks via transformation. A possible reason being some additional factors rely on the hardware of the system, like the movement and noise of the camera to take pictures. But that is as low as it can be ignored by the parameters of the function used for classification or prediction. There is also a very low probability of identifying an image as a *fake image* from the physical world. To overcome this kind of scenario Lyth, David., et al. (Lyth, 2005) proposed a novel system to resolve this problem or limitation with the help of generating perturbation that changes the input to have several distortions for the trained model. Such changes can be in the form of random rotation or the addition of noise to miss-classify the original value in the classifier.

### 4.2.3. Defense techniques against evasion attacks

One of the effective methods to overcome the evasion attacks is to augment the adversarial examples which highlights the adversarial examples, and defensive distillation. Most of the defense measures have used adversarial examples and their features to minimize the effects of proposed methods (Tong et al., 2019). First noise in form of adversarial examples have added into the dataset before training the model and then detecting those additional noise from the model prediction (DelVecchio et al., 2020; Pawlicki et al., 2020). Such defense methods have shown remarkable results against evasion attacks by minimizing the effects. Following are the few most common defense methods for evasion attacks which have shown a significant contribution while minimizing the effects of evasions attacks.

a. **Detecting Adversarial Examples**

There have been proposed numbers of methods (Gupta et al., 2020; Dutta et al., 2020) used to detect the adversarial examples. As mentioned earlier, the attacker's main motive is to subvert the learning and testing process to get the model and dataset information by adding more and more noise to formulate the new adversarial examples (Choraś and Pawlicki, 2020). One of the significant challenges while detecting adversarial examples is identifying the adversarial testing examples used to predict or classify the adversarial examples (Hassan et al., 2019). Hence, such examples can be labeled manually, like automatic car driving, where it automatically takes decisions; it is hard for humans to label manually while managing the adversarial examples (Ji et al., 2018; Tariq et al., 2017; Tariq et al., 2020).

b. **Gradient Masking**

The idea behind the gradient masking is to provide a specialized way to resolve the adversarial examples. The gradient points have created such examples so that these points are placed a bit farther from the normal position (Dutta et al., 2020). Due to this alignment, a decision boundary using pure examples and procedures can hand over the incorrect and malicious gradient to an adversary using such techniques (Vivek et al., 2019).

c. **Non-Obfuscated Gradient Masking**

It involves the most common and most representative research of attack by using adversarial training (Zhong et al., 2019; Khan et al., 2020). It is the most significant gradient masking method. It claims the training dataset as a true label and a robust model is implemented. They also showed that the adversarial examples could be regularized by cleaning the combination of true adversarial examples and false adversarial examples. Senior et al. (2020) has proposed a generalized behavior of adversarial examples to support malicious training. The examples generated by the model are classified as false by other models. This behavior is often due to non-linear or over-fitting behaviors that cannot be addressed by numerous other models. For this kind of environment, a generative training model was proposed to enhance and provide more constraint of the learning process. It also enforced the model to classify the real from malicious and removes the false positive and false negative. Adversarial training is developed for a small model with MNIST (Dorosh and Fenenko, 2020) dataset without normalization. Chen et al. (2020) has proposed the model by extending the previous work by adding a normalization step. Tramèr et al. (2017) also proposed a model recently to defend

**Table 3**
Existing literature related to evasion attacks.

| Category | Refs | Year | Techniques | Research Gaps |
|---|---|---|---|---|
| Whitebox attacks | (Xie et al., 2020) | 2020 | Proposed high-confidence malicious samples with a simple moving test to another model that can be break. | The existence of adversarial examples limits the areas of DL. |
| | (Elsayed et al., 2019) | 2019 | Demonstrated by feeding adversarial images obtained from a cell-phone camera to an ImageNet | It is possible to demonstrate attacks using kinds of physical objects besides images printed on paper, etc. |
| | (Kong et al., 2020) | 2020 | Full-scale adversarial training to large model and datasets | Adversarial examples generated by iterative methods between networks provides indirect robustness |
| | (Shaukat et al., 2022) | 2022 | A novel approach has been developed to design a malware detector by training a neural network with a mixture of multiple adversarial attacks. | The approach has been effectively used to detect the malware attacks in different unique fields, however it can also be extended for other ML and DL algorithms. |
| | (Debicha et al., 2023) | 2023 | The proposed model has been implemented on existing state-of-the-art models for intrusion detection. It was then attacked such models with a set of chosen evasion attacks. To detect the adversarial attacks, designed and implemented multiple transfer learning-based adversarial detectors, each receiving a subset of the information passed through the IDS. | Presented only two DL-based IDS models (one serial and one parallel) and assessed the effect of four known adversarial attacks on their performance, namely: Fast Gradient Sign Method, Projected Gradient Descent. |
| | (Wang et al., 2023) | 2023 | In this work, the susceptibility of Federated Learning (FL)-based signal classifiers to model poisoning attacks, which compromise the training process despite not observing data transmissions. In this regard, an attack framework was presented in which compromised FL devices perturb their local datasets using adversarial evasion attacks. | Specifically, evasion attacks have effective against FL, where compromising even a single device can damage the rest of the network and this, in effect, increases steadily with the number of adversaries. Such evasion attacks are also difficult to detect and defend against in wireless settings as such attacks bears statistical and visual similarities. |
| | (Moosavi-Dezfooli et al., 2016) | 2016 | Proposed a systematic algorithm for computing universal perturbations. | Only consider the images to generate universal perturbation for DNN. |
| Blackbox | (Tramèr et al., 2016) | 2019 | Provided the new phenomenon that used adversarial training, that can change adversarial models. | The malicious training will remain vulnerable to Blackbox attacks |
| | (Papernot et al., 2018) | 2018 | Explored the adversarial behavior in Deep Learning models | DNN and RNN were trained in an unsupervised manner |
| | (Xu et al., 2023) | 2023 | Proposed a new semi-black-box attack framework called one-feature-each-iteration (OFEI) to craft Android adversarial samples. | Two uncertainties of the Bayesian neural network to construct the combined uncertainty, which is used to detect adversarial samples and achieves a high detection rate |
| | (Jagielski et al., 2018)<br>(Qi et al., 2022) | 2018<br>2022 | Proposed new optimization framework for poisoning attacks and identify circumstances that can be deployed in linear regression models | Opens research towards developing more secure learning algorithms against poisoning attacks |

against robust Blackbox attacks and provide adversarial ensemble training by containing examples generated by other models. To increase the perturbation as an increase to be seen during the training process, this model uncouples the adversarial examples during the training phase (Wang et al., 2019).

d. **Obfuscated Gradient Masking**

It is the method to overcome the effects of evasion attacks in DL, which has the following types: 1) shattered gradient, 2) Stochastic gradient, and 3) vanishing or exploding gradient. In the shattered gradient case, it means that the incorrect gradient masking is achieved intentionally and non-intentionally (Lee et al., 2020; Yu et al., 2020). The primary objective was to break the normal behavior and linearity under the consideration of NN. Such common behavior is mostly considered as a linear manner (Ilyas et al., 2019).

However, on the other hand, the linearity has more effects on the predictive model with the small value of $\varepsilon$ and dimension space while taking images into account, which makes the model vulnerable to adversarial attacks. The recent algorithms (Buckman et al., 2018) uses a shattered gradient masking technique to break the linearity of NN. Below is Table 3, which summarizes the above-discussed evasion attacks.

### *4.3. Model inversion attacks*

In this kind of attack, the training dataset's maximum knowledge is taken out within the target model framework, i.e., malicious model. As NN remembers lots of information from the trained data, in this scenario, there is also an inverse information flow where the attacker can infer the knowledge of the training dataset (Song et al., 2017). Additionally, Model inversion attack can also be further classified or divided into two attacks: 1) Membership Inference Attack (MIA) (Dagan and Feldman, 2020), where the attacker can determine any specific record whether such record exists in the system or not, and 2) Property Inference Attack (PIA) (Lim et al., 2020), where the attacker can only speculate either there exist any statistical property may exist or have been applied to the training dataset

#### *4.3.1. Adversarial model*

It is a type of MIA attacks, which can have the setting either Blackbox or Whitebox attacks or both. For the Whitebox attacks, the attacks know all the parameters and values used in the model (Shen et al., 2021). Hence such an attacker can easily generate the adversarial model like the original one that behaves precisely identical to the original but can be classified maliciously. While in the Blackbox setting, the capabilities for attackers are minimal, like it does not know about the architecture, data distributional, etc. However, whatever the setting is, the attacker can manipulate the queries with specific inputs and can generate the corresponding results based on malicious values (Jayaraman and Evans, 2019).

#### 4.3.2. Membership inference attack (MIA)

Tolpegin et al. (2020) has provided an overall systematic formulation of MIA. According to the formulation, $x$ is given as an instance and does not know the classification model trained on the dataset. Based on the following information, can an adversary get whether the applied input $x$ is the part of the dataset or not, the corresponding model is trained with a very high confidence level (Chen et al., 2020). Mostly MIA works in a specific workflow, like inferring the specific data item or someone's property may be the part of the training set. The intruder intends to prepare the initial dataset to get the transformation for diverted data or manipulated datasets.

#### 4.3.3. Property inference attack (PIA)

Property Inference Attack (PIA) typically infer different features of the dataset's training process (Kaur et al., 2020). For instance, what is the number of people having dark brown hair color in a specific territory? What are the reasons for chronic diseases in underdeveloped countries? The approach is like a membership inference attack. *Synthesis of Data:* In PIA, data is classified, whether it is included or excluded in specific properties in the datasets. *Training Model:* The training sets train the model in PIA by including or excluding any properties from the datasets. For instance, training models add or remove certain attributes to train the learning model, and by this process, they compute up a shallow model to give the training data for meta-classifier (Ganju et al., 2018).

### 4.4. Model extraction attacks

The model extraction attack first makes a copy of the DL algorithm using APIs, without having the previous knowledge of the DL algorithm (Sugawara et al., 2020). More specifically, given any input $X$, the adversary can initiate queries to the target model $F$and get the results $F'$ while comparing it to the NN model. A model extraction attack can optimize the constants and coefficients in equation 8. It not only loses the confidentiality and integrity of the model itself but also it creates a Whitebox model that can be attacked like adversarial attacks.

$$y = wx + b \tag{6}$$

#### 4.4.1. Adversarial model

These kinds of attacks are mostly carried out under the basic phenomenon of the well-known Blackbox model. The adversary only has access to the prediction and classification of the model using APIs (Du et al., 2020) and in feature engineering algorithms (Usmani et al., 2020; Hashem et al., 2020). Attackers have limited in three ways: 1) model knowledge, where the attacker has the information of the model like Whitebox attacks, 2) dataset access, where the attacker has full access to the dataset of the model, which is used to train the model, and 3) query frequency, where the attacker can generate the queries to get and extract the information of the model (Wang and Gong, 2018). Overall, the adversary does not know the model's architecture, the model's parameters, and the victim's model's training process (Oh et al., 2019). It cannot get the information about the distribution of the training process and the dataset's testing process. Moreover, the target model can block the attacker if it receives too frequent queries (Juuti et al., 2019; Chen et al., 2020). In Table 4, a summary of model inversion attacks and model extraction attacks are discussed.

## 5. Privacy preserving techniques

There is a need for numerous training data for the effective construction of the DL algorithms. In general, data is collected from crowdsourcing techniques as such data may contain the sensitive part of a user's private data, leading to the security threat for privacy of data in DL. Additionally, it is investigated that some critical parameters may be kept hidden for security purposes so that the adversary cannot perform any inversion attacks against the secure model of DL (Hashem et al., 2016). Differential Privacy (DP) based method and Homomorphic Encryption (HE) is the primary method used to maintain data privacy.

### 5.1. Deferential privacy based methods

Dwork (2006) has come up with the unique technique named Differential Privacy (DP) in 2006, such that it can maintain the data privacy for users. It proved robustness theoretically, as the model is taken widely in terms of providing privacy for the ML model, aiming to secure the input by adding noise to the original model. Papernot et al. (2018) has provided mechanisms to secure and privacy-preserving techniques based on the DP model (Hamm et al., 2016). Deferential privacy techniques can be useful in many ML-based models like prediction models (Chaudhuri and Monteleoni, 2009), loss function and gradient (Wu et al., 2019). The computation cost for using deferential privacy methods was very high and it is limited to textual dataset. Computational complexity has increased the cost as well as increase the time cost to compute the model accuracy as well as classification rate.

### 5.2. Homomorphic encryption based methods

The ML data model's most common and strong privacy-preserving scheme is the Homomorphic Encryption Standard (HES). It is the unique type of cryptographic function that allows algebraic operation on the ciphertext (Chabanne et al., 2017)

$$\mathrm{Enc}(a) \Diamond \mathrm{Enc}(b) = \mathrm{Enc}(a * b) \tag{7}$$

Here*Enc* :represent the encryption operation of $X$ that belongs to subset of $Y$,$a, b \in X$ with the parameter $a$ and $b$ in such a way that $^*$ and ♦ denotes the operation performed by the corresponding parameters $X, Y$and hence commonly known as homomorphic encryption scheme. In the start, Homomorphic encryption only have a "partially homomorphic cryptosystem" (ElGamal, 1985) where it only consists of additive property or multiplicative homomorphism. However, Huang and Boneh et al. (Huang et al., 2019) proposed the idea which was based on ideal lattices, the number of attempts has been taken into account on Fully Homomorphic Encryption (FHE), that permits the operation to be performed on encrypted data (Ducas and Micciancio, 2015). FHE has benefits in applications like cloud computing platforms and secure computations, massive data input, high computation workload, and the non-linear DL algorithms. FHE is still highly risky to be combined with DL.

### 5.3. Garbled circuits

One of the critical security enhancements of homomorphic encryption in the context of the layer-by-layer formulation is the garbled circuit (Yao, 1986). The garbled circuit can be drafted as a variation of homomorphic encryption (Rouhani et al., 2018). Let say $f$ is a function of the garbled circuit, between two parties $A$ and $B$, where $A$ holds a function $f'$ that is a corresponding single-layer network and $B$ is associated with the input $x$, the function $f$ will compute the output as a single encoded input. Then, $B$ will computer encoded its input $x$, then one of the parties will be able to

**Table 4**
Summary of Model Inversion and Model Extraction Attacks.

| Category | Ref | Year | Techniques | Limitations |
|---|---|---|---|---|
| Model Extraction Attacks | (Du et al., 2020) | 2020 | Iteratively generated adversarial images for training and transferring the result | The accuracy of adversarial images to the clean images was not clear. |
| | (Chen et al., 2020) | 2020 | Moment-based iterative approach for classification | Adversarial dataset used for training the model to get the high accuracy and classification rate. |
| | (Wang and Gong, 2018; Juuti et al., 2019) | 2019<br>2018 | The momentum-based iterative algorithm has proposed to boost adversarial attacks | By assembling adversarial training to the model, it is vulnerable to Blackbox attacks, which raises new security issues to develop more robust DNN models. |
| | (Orekondy et al., 2019; Correia-Silva et al., 2018) | 2018 | Gradient-based approach for security of several classification algorithms | Adversarial Examples are being used in NN vulnerability |
| | (Fu et al., 2022) | 2022 | Proposed a model for traffic anomaly detection named a deep learning model for network intrusion detection (DLNID) | DLNID model to an actual, combined network capture module to implement an online intrusion detection model. |
| | (Li et al., 2023)<br>(Aldhyani and Alkahtani, 2023) | 2023<br>2023 | Propose a defense scheme based on physical unclonable function (PUF) against such black-box model extraction attacks | Compared to existing defenses, our scheme not only effectively prevents black-box model extraction attacks but also ensures that the accuracy of the prediction service for legitimate users is not affected. |
| | (Sugawara et al., 2020; Jia et al., 2018) | 2020<br>2017 | Efficient privacy-preserving ML scheme for Hierarchical distribution systems | Consider the outside attacks and focus on analyzing the possible data privacy leakage from the system. |
| Model Inversion Attacks | (Chen et al., 2020) | 2020 | Gradient-based techniques for classification and prediction algorithms | Limited to selected features and limited dataset has used for training the model. |
| | (Riazi et al., 2019) | 2019 | General Mathematical framework for auto-regressive model | The attacker can attack only at the test phase |
| | (Ateniese et al., 2013) | 2018 | Meta classifier and trained to hack another classifier to get the Information | Vendor may know this kind of Loopholes and can build a new classifier to steal or hack the classifier |
| | (Khosravy et al., 2022) | 2022 | Assumed the model inversion attack under semi-white box scenario of availability of system model structure and parameters. | MIA process efficiently search for a low-dimensional feature vector whose corresponding face image maximizes the confidence score. |
| | (Zhang et al., 2022) | 2022 | For white box: GNN model presented GraphMI to infer the private training graph data.<br>For black box: setting where the attacker can only query the GNN API and receive the classification results proposed two methods based on gradient estimation and reinforcement learning (RL-GraphMI). | Such model invasion based defenses are not sufficiently effective and call for more advanced defenses against privacy attacks |
| | (Ganju et al., 2018) | 2018 | Showed adversarial attacks are also effective when targeting NN policies in reinforcement learning | Lack of developing defenses against adversarial attacks. |

compute the encoding of$f'(x)$from which $f(x)$ that can be recovered.

### 5.4. Secret sharing

While talking about the sharing of secret keys between different parties of the network, the sharing of keys is essential and challenging. One of the most commonly used technique is secret sharing schemes (Beimel, 2011). Such a secret can be computed by combining the shares of “authorized” sunset of all the parties involved in the network or at-least subset of certain sized users (Shamir, 1979). While the share of any “unauthorised users” will not be able to get any information about the secret. So, as described, secret sharing will enable the NN to preserve the privacy evaluation in layer-by-layer architecture. The concerning parties share their shares for all the layer values starting from index $i$ to $(i+1)$ values on each layer (Wagh et al., 2019). Table 5 discussed the summary of some privacy protection for the DL algorithms.

## 6. Proposed cryptographic framework

In this section, we have provided the proposed framework for dataset that is used for training the DL algorithms. This proposed framework is composed of three different stages before the training and testing of DL algorithms. In first phase, hash functions such as SHA512 will be applied on dataset as an attribute. This is well known that hash functions are one-way functions, and this can also be used as digital signature for digital documents (Tayyab et al., 2021). These signatures can also be verified by applying same input parameter and then by comparing the old value with new generated hash value (Tayyab et al., 2021). The proposed framework has used similar features of hash functions and applied on the dataset before sampling the dataset. After the addition of hashed values into the dataset, this dataset can now be encrypted with Homomorphic encryption scheme (HES) (Mittal and Ramkumar, 2021) to ensure the data integrity and confidentiality while the dataset being transmitted over the network for training the DL algorithm. Following are the major points of proposed framework that occurs in different phases (TAYYAB et al., al.). The whole framework is provided in Fig. 5 graphically with different phases.

Phase 1: the proposed framework first calculates hash value so that it can be verified the reliability and integrity of dataset. The hash will be like the digital signature and that can be used for verification of dataset. If attacker manipulates the dataset, then it can easily be traced out that the dataset has been modified with malicious injections to subvert the learning process.

a) In the first step, Hash function SHA512 will be applied to get $H_0$ and appended as a part of dataset attribute.
b) Once the hash is appended as part of dataset, then it is encrypted using HE encryption mechanism to ensure the privacy and security of dataset and stored to the cloud storage for further process.

Phase 2: firstly, the dataset will be retrieved from cloud storage and then decrypt the dataset. Hash function SHA512 is applied again, to get $H_1$. Now both $H_0$ and $H_1$ are the part of dataset attribute. While computing $H_1$, the proposed framework has designed in such a way that $H_0$ will be excluded. There are following steps in phase 2.

**Table 5**
Summary of privacy protection for Deep Learning models.

| Category | Ref | Years | Techniques | Limitations |
|---|---|---|---|---|
| Deferential privacy Based Methods | (Dwork, 2006)<br>(Zhao and Chen, 2022) | 2016<br>2022 | Differential privacy techniques | The time complexity and computational cost of using such differential privacy techniques has decreased the performance of DL model |
| Homomorphic Encryption Scheme | (Rouhani et al., 2018; Juvekar et al., 2018) | 2018<br>2019 | Faster CryptoNets: Method for efficient encrypted inference using NN | Extracting large features-based frameworks that can reduce the number of layers requiring computation increases the time complexity and Storage overhead for medical images. |
| | (Athalye et al., 2018)<br>(Kumar et al., 2022) | 2017,<br>2022 | Identified obfuscated gradients, a phenomenon exhibited by certain defenses that make standard gradient-based methods. | Used the ICLR 2018 Defenses as a case study, circumventing seven of nine accepted defenses. |
| Garbled Circuits | (Li and Zhu, 2020) | 2020 | Adversarial attacks can be detected with the proposed unsupervised learning-based approach | Discussed the CNN classifier to address the medical images only and comes up with the adversarial attack detection |
| Secret Sharing | (Zhang et al., 2017) | 2017 | Adversarial Transformation Network (ATN) is trained to generate adversarial examples | ATNs should be deployed in the adversarial training phase to train the dataset. |
| Other Techniques | (Berman et al., 2019) | 2019 | Presented the evaluation of preliminary results of adversarial attacks on 3D points that were proposed for 2D images. | Proposed several iterative gradients-based attack methods and input restoration-based defenses. |
| | (Biggio et al., 2012) | 2012 | Gradient-based attack proposed against support vector machine | Many improvements presented to this method, but it remained as unexplored |
| | (Athalye et al., 2018) | 2018 | Proposed an algorithm for 3Dprinting to manufacture the first physical adversarial object | The existence of robust adversarial examples, adversarial inputs will remain adversarial over a chosen distribution of transformations. |
| | (Alazzam et al., 2022) | 2022 | Proposed federated learning approach addresses privacy and security concerns about data privacy and security rather than allowing data to be transferred or relocated off the network edge | Solution outperforms a standard centrally managed system in terms of attack detection accuracy, according to our comparative performance analysis. |
| | (Song et al., 2022) | 2022 | Based on the secret sharing, designed a framework, EPPDA, an efficient privacy-preserving data aggregation mechanism for Federated Learning (FL), to resist the reverse attack, which can aggregate users trained models secretly without leaking the users model. | Efficiency verification proves that the EPPDA not only protects users privacy but also needs fewer computing and communication resources. |
| | (Sun et al., 2023) | 2023 | Proposes a swarm learning (SL) framework that combines adversarial domain networks with convolutional neural networks (CNNs) to address the safety hazards and avoid economic losses promptly. | This framework has increased the computational cost but reduced the communication cost, which can be considered as the achievement. |

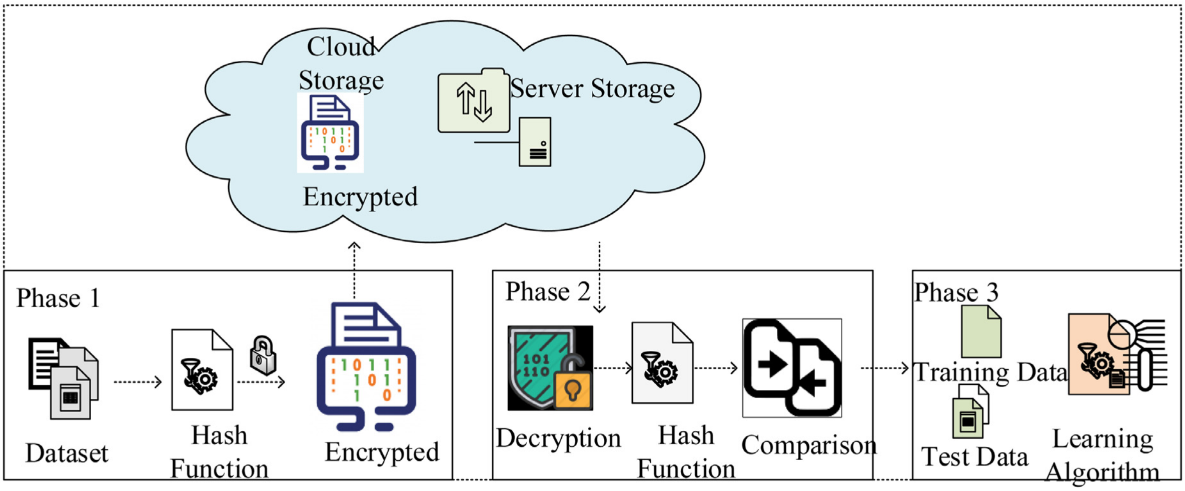

**Fig. 5.** Proposed cryptographic framework for dataset.

a) Encrypted dataset is retrieved from cloud storage.
b) Using HE, the dataset is decrypted to get the original dataset that was outsourced.$H_1$ will be computed to check the integrity of dataset. While computing $H_1$, $H_0$ will be excluded so that for best case, both hashes will remain equal to ensure that there is no change in dataset.
c) If both hash values match, the proposed framework proceed to phase 3, otherwise, it will halt the model.

Phase 3: after the verification of data in phase 2, now proposed framework will performs the data sampling, i.e., split the dataset into training dataset and testing dataset. After data sampling following steps will be taken out in this phase.

a) Splitting of dataset into training dataset and testing dataset, i.e., data sampling.

**Table 6**
Computational cost with respect to all-clock time.

| No. | Original model | Hash function (SHA512) | Homomorphic encryption (HES) | Proposed framework | Time for cryptographic Function = (Cost of Original model) - [(Hash_Function (SHA512)) + (HES)] |
|---|---|---|---|---|---|
| 1 | 900 s | 300 s | 6.2 s | 1225 s | 306 s |
| 2 | 851 s | 312 s | 5.8 s | 1220 s | 376 s |
| 3 | 910 s | 307 s | 6.4 s | 1214 s | 380 s |

b) Data normalization is used to normalize the image pixel values between +0.5 and −0.5 by using following formulas in Eq. (1).

c) $$\begin{aligned} train\_image &= ((train\_image/255) - 0.5) \\ test\_image &= ((test\_image/255) - 0.5) \end{aligned} \tag{10}$$

d) After successfully normalization of datasets, now the training dataset will used to train the model and test dataset will be used for evaluating the model.

### 6.1. Computational cost for proposed model

The computation cost is depended on the system configuration as well as the processing power of a machine. In the proposed model, although cryptographic functions are used, the overall computation has not increase significantly and have not cross the critical impurity level (Usmani et al., 2021). We have evaluated our proposed model based on computational cost for critical behaviors of DL models. This has increased numbers of operations that can be optimized with the help of different optimization functions. The computational cost is the only limitation of the proposed model, but it can be reduced by decreasing the number of operations and using a different optimization solution. The computational cost of the proposed model is defined by Equations.

$$\begin{aligned} PM &= Hash\_function(SHA512) + (HES) \\ &\quad +(SHA512)Hash\_Function(SHA512) + DeepLearning\_Model \\ PM &= C(n) + C(n) + C(n) + n^2 \end{aligned}$$

By solving the above equation:

$$PM = 3C(n) + n^2$$

By ignoring the lower order terms like $C(n)$

$$PM = \Theta(3C(n) + n^2)$$

$3C(n)$is small increment in the form of computational cost which is added by the additional functions added by the proposed model. This cost is minimal and can be ignored while achieving the security to the dataset for DL algorithms.

$$PM = \Theta(n^2)$$

where $PM$ represent the proposed model, $c(n)$denotes the cost, and $\Theta$ shows the tighter analysis of the proposed model. In security algorithms, there are some assumptions that can be ignored while achieving high priority benefits like in our case, our main target was to achieve data security of dataset used for DL algorithms. We have achieved our goal by adding some additional security check point or framework to verify the dataset before training and testing in DL algorithms. The computational cost of the proposed model is not greater than$\Theta(n^2)$, and the computational cost of the original model is $\Theta(n^2)$. Hence, there is a slight increase in the computational cost of the proposed model i.e., $3c(n)$ as compared to the original model.

### 6.2. Model running time (wall-clock time)

The wall-clock running time depends on hardware, which means that it is directly dependent on the system configuration, including available memory space and computational power of the system. It is also dependent on the encryption scheme used in a model. In case our proposed framework, the running time $O(C(n))$. We have provided it in asymptotically and it is observed that to execute the proposed framework, there is very minimal addition of time in normal execution time for the system i.e., $C(n)$. This value is very minimal addition to the computational cost of whole process that can also be ignored while achieving other security parameters like confidentiality and privacy issues. The running time overhead was not increased to certain limit and an efficient time was taken by the experiments to perform certain tasks associated with DL algorithms. As we have achieved $O(C(n))$for the proposed models so it can be perceived that proposed model can be deployed in any environment of data-driven problems or other decision-based problems. Table 6 has shown the actual time to compute the proposed model.

## 7. Discussion

With the reference of Section 2 of the paper, it is well known and confirmed that there are active attacks that can fool or subvert the learning process of DL algorithms. We have provided different kind of active and passive attacks and their scenarios like Blackbox and Whitebox attacks. Keeping in review that in Whitebox attacks, most of the adversaries tries to generate the adversarial examples by the information of target DL algorithms, and such common adversarial examples had shown a very high misclassification rate and high accuracy as well. Therefore, it is critical for the success of these attacks while acquiring the true gradient labels from the vanilla model, the model which has no defense measures against any kind of attacks to spot out any blind-spots or any disturbance in the target mode (Ullah et al., 2021). Hence, to take countermeasures against such scenarios, most of researchers has proposed diverse gradient masking defense measures, which have shown quite decent results by introducing more non-linearity in the DL algorithms or introducing countermeasures against gradients of the model from being copied by an adversary (Ghosh et al., 2020). there is need of such a model that can provide the security to the model as well as dataset which is used for training the DL algorithms and testing phase such model. We have found in literature that if gradient methods have been used properly, the results were powerful in form of performance (Arrieta et al., 2020; Bilal et al., 2019). Therefore, it is confirmed to be beneficial that if AI approaches (Ghorbani et al., 2020; Angenent-Mari et al., 2020) used against such active attacks or their countermeasures. Greater understanding of DL algorithms makes feasible to develop a secure framework or model that can be robust to such unseen attacks by identifying the targeted attacks which can be addressed for secure DL models.

In Section 3, we have provided detail analysis of security attacks which have made the training and testing of DL algorithms complex. We have reviewed some of the most common security attacks like poisoning attacks and evasion attacks their adversarial setting along with the types. The most recent approaches (Chen et al., 2019; El-Rewini et al., 2020) include the outlier detections to remove the labels or re-labels the suspicious poisoned examples. However, such actions to eliminate the poisoned examples might constrain the decision boundary of DL algorithms too much. This elimination of poisoned data can vary the datapoints as well as model's decision boundary significantly (Wu et al., 2020). Therefore, a sound and secure model is needed for evaluation methods either safe or secure. We have also discovered more active attacks like model extraction attacks and model inversion attacks that has affected the model boundary greatly (Pillai et al., 2018). We have provided the detail summaries of literature against each of the attack category along with the types and adversarial settings. Meanwhile we have provided the defense measures against each type of security category.

In Section 4, we have discussed privacy issues in DL algorithms with the cryptographic functions like Hash functions or homomorphic encryption schemes. Although, it has been seen that these cryptographic schemes have shown a remarkable prediction rate even though complex nature of encryption (Vizitiu et al., 2020). It is also observed that the performance accuracy of such encrypted models has fall behind the performance of the state-of-the-art DL algorithms. One of the main reasons that has been discovered that these cryptographic functions do not use the nonlinear activation functions. Therefore, current homomorphic encryption-based models have used models for each iteration of training and testing the DL algorithm. In other words, for encrypted data most of the models (Wood et al., 2020) has used different typical model and when trained the weights and biases were carried out to different model, where the activation function has also been changed by a simple function. This can be overcome by mainly two reasons either replace the DL algorithm properly or trained the DL algorithm from the scratch (Syed et al., 2020).

In Section 5, we have provided the detail working of proposed cryptographic framework for dataset that is used for training and testing the DL algorithm securely. In our proposed framework, two of the major cryptographic functions i.e., hash function SHA512 and HE will be used to ensure the integrity of dataset as well as the privacy issues in the dataset before ready for training the model. Proposed framework will protect not only the dataset from being poisoned as well as it can protect the model from being corrupted. Proposed framework will also maintain the prediction rate as high as others along with the accuracy of model up-to 98%. The most important parameter used to evaluate our proposed framework was the computational overhead. It is an understood phenomenon, when the desired goal was to provide security to the dataset so that the risk of being polluted can be decreased, there is the possibility that the computational overhead can be increased up to a certain threshold. To restrict the computational overhead under certain limits, we have used cryptographic functions that have minimum overhead, like hash functions SHA512 and HES. Both cryptographic functions have a certain computational cost, but this cost can be overlooked while achieving the desired outcomes. We have presented the cost in term of asymptotic notations that describes the minimal increase in cost for computing hash values, watermark generation, and encryption. However, this additional cost is much minimal as compared to the cost used by DL algorithms.

## 8. Future research challenges

In this research, the following challenges have been presented that help researchers working on the security of DL algorithms.

### 8.1. Whitebox attacks concerns

In this study, we have concluded that some adversarial methods could subvert or revert DL models' learning process to get useful information. Following different types of attacks were reviewed, namely poisoning attacks, evasion attacks, adversarial attacks, model inversion attack, and model extraction attacks. All the poisoning and evasion attacks are classified into Blackbox and Whitebox or strong adversary and week adversary. Furthermore, the attacker mostly generates adversarial examples to subvert the learning process, resulting in a very high misclassification rate. It is hard to attack such an attacking scenario to get the true data gradients from the vanilla model without the defense system identifying sparse or unseen spots in the targeted model. As authors Sun et al. (2020) have proposed a proper gradient method has shown the powerful defense performance against such attacks. Hence, it is also beneficial if AI approaches (Hamm et al., 2016) can overcome such attacking scenarios. These interpretable AI approaches can determine how a DL model can better classify or predict (Tasaki et al., 2020).

### 8.2. Concerns with encrypted data

We have provided the privacy-preserving for DL algorithms by using different cryptographic functions like homomorphic encryption. Although on one side, the privacy and high prediction rate have been maintained by using homomorphic encryption, on the other hand, the accuracy of the model may fall behind the state-of-the-art model performance, which is far behind and not acceptable for DL algorithms. Hence, the model based on FHE uses different models for training and testing. Two solutions are possible to overcome such things in training the model, 1) re-trained the model from the start or 2) transform it. Like NG and Selvakumar (2020), and Lopez-Martin et al. (2020) proposed, the DL algorithm's knowledge can also be converted into other models using the same setting.

Some of the researchers have presented the most common and active attacks on DL that can have any critical concern. In the section of adversarial attacks and most common active attacks like poisoning and evasion attacks that can subvert or affect DL algorithm's performance. Moreover, DL algorithms are exposed to the possibility of being attacked by such attacks in the real world. Hence, it can be concluded that such attacks are a real threat to deep learning models. Furthermore, while taking real-time scenarios, a hybrid approach is proposed by Gadekallu et al. (2020), a scalable framework on commodity hardware sensors to detect malicious activities, not only network level but also host-level activities. While taking the real-time system and analysis of attack with the help of DNNs in a distributed network, the performance can also be enhanced and improved by monitoring the Domain Network Servers (DNS) and BGP in the network (Pant and Barati Farimani, 2020). More that, the time complexity and computation cost can be further improved by adding more nodes in the cluster (Ferrag et al., 2020), which could be an important and exciting topic for future research.

### 8.3. Why adversarial vulnerability required more investigation?

Several things need to be considered in the literature on vulnerabilities in a DNN to subvert the classification process and prediction. Such vulnerabilities and viewpoints need more attention to keep the learning process secure so that these points must be aligned with each other (De Gaspari et al., 2020) can generate accurate results. Hence, a systematic investigation is seriously needed in such area that can be trustworthy and reliable. Adversarial examples can also subvert the DL of medical systems (Simon-Gabriel et al., 2019). More importantly, we have found that it is challenging to generate and detect the adversarial attacks in medical images compared to natural images. This can be further motivation for future research, which could be a more practical and useful approach to minimize the effects of adversarial examples (Panda et al., 2019).

It is the most common and most important property of an adversarial setting. The examples reported in the state-of-the-art literature can exhibit well and can transfer accurately between the different NNs. Such a scenario may happen where the network often has similar architecture (Boulemtafes et al., 2020), but such adversarial examples may often expose in Blackbox attacks.

## 9. Conclusion and future work

DL is becoming an attractive topic for researchers and industries due to its effective and problem-solving nature in the real-world scenarios. Security and privacy are the most vital issues that cannot be overlooked while designing DL algorithms. The intruders can deliberately change the new or training dataset that can cause misclassification and wrong output prediction. Therefore, it is essential to mitigate the security and privacy issues before applying or designing the DL algorithms. In this regard, we have provided detailed literature with advantages and limitations on security and privacy attacks such as poisoning, evasion, model extraction and model inversion. We have classified each attacks category into Blackbox and Whitebox setting or strong adversaries and weak adversaries. We have provided the detail summaries of literature against each of the attack category along with the types and adversarial settings. We have presented comprehensive analysis of privacy preserving techniques and their adversarial settings for various attacks. To overcome the effects of above security attacks, we have proposed a secure cryptographic framework for dataset based on hash functions and Homomorphic Encryption (HE) scheme. The proposed framework secures the dataset before training the DL algorithms and preserves the privacy of dataset against the most common active attacks like poisoning and evasion attacks.

Finally, we have provided recent research challenges and possible future studies that can help to mitigate the security and privacy issues. In future, the proposed framework can be developed to show that the security and privacy of algorithms and dataset can be preserved using cryptographic functions. Although, by adding such cryptographic function, it may increase the computation cost and complexity, but this additional cost can be overcome by applying different optimization functions. Furthermore, this study is limited to only most common challenging security attacks, however there are several other real-time security attacks such as transfer learning and manipulating online systems that should be addressed to achieve secure and efficient DL algorithms.

## Funding

Not applicable.

## Ethical approval

Not applicable.

## Consent for publication

All the authors declare their consent for publication for their data in this journal. All the authors have not objection.

## Consent to participate

All the authors declare their participation in this review article.

## Availability of data and material

Not applicable.

## Code availability

Not applicable.

## Declaration of Competing Interest

I hereby declare that the disclosed information is correct and that no other situation of real, potential, or apparent conflict of interest is known to me. I undertake to inform you of any change in these circumstances, including if an issue arises during the meeting or work itself.

## CRediT authorship contribution statement

**Muhammad Tayyab:** Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Mohsen Marjani:** Formal analysis, Methodology, Project administration, Supervision. **N.Z. Jhanjhi:** Formal analysis, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. **Ibrahim Abaker Targio Hashem:** Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Raja Sher Afgun Usmani:** Writing – review & editing. **Faizan Qamar:** Visualization, Writing – review & editing.

## Data availability

Data will be made available on request.

## References

Abramson, W., Hall, A.J., Papadopoulos, P., Pitropakis, N., Buchanan, W.J., 2020. A distributed trust framework for privacy-preserving machine learning. In: International Conference on Trust and Privacy in Digital Business. Springer, pp. 205–220.
Akhtar, N., Mian, A., 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 6, 14410–14430.
Alazzam, M.B., Alassery, F., Almulihi, A., 2022. Federated deep learning approaches for the privacy and security of IoT systems. Wirel. Commun. Mob. Comput. 2022, 1–7.
Aldhyani, T.H., Alkahtani, H., 2023. Cyber security for detecting distributed denial of service attacks in agriculture 4.0: deep learning model. Mathematics 11 (1), 233.
Altaf, F., Islam, S., Akhtar, N., Janjua, N.K., 2019. Going deep in medical image analysis: concepts, methods, challenges and future directions. IEEE Access 7, 99540–99572. doi:10.1109/ACCESS.2019.2929365.
Angenent-Mari, N.M., Garruss, A.S., Soenksen, L.R., Church, G., Collins, J.J., 2020. A deep learning approach to programmable RNA switches. Nat. Commun. 11 (1), 1–12.
Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., et al., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.
Ateniese, G., Felici, G., Mancini, L.V., Spognardi, A., Villani, A., et al., 2013. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers. Int. J. Secur. Netw. 10 (3), 137–150.
Ateniese, G., Mancini, L.V., Spognardi, A., Villani, A., Vitali, D., et al., 2015. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers. Int. J. Secur. Netw. 10 (3), 137–150.
Athalye, A., Carlini, N., Wagner, D., 2018b. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International Conference on Machine Learning, pp. 274–283.
Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., 2018a. Synthesizing robust adversarial examples. In: *International Conference on Machine Learning*, 2018, pp. 284–293.
Behzadan, V., Munir, A., 2017. Vulnerability of deep reinforcement learning to policy induction attacks. In: International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 262–275.
Beimel, A., 2011. Secret-sharing schemes: a survey. In: International Conference on Coding and Cryptology, pp. 11–46.
Berman, D., Buczak, A., Chavis, J., Corbett, C., 2019. A survey of deep learning methods for cyber security. Information 10 (4), 122. doi:10.3390/info10040122.
Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., et al., 2013. Evasion attacks against machine learning at test time. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 387–402.
Biggio, B., Nelson, B., Laskov, P., 2012. Poisoning attacks against support vector machines. 29th International Conference on Machine Learning.

Biggio, B., Roli, F., 2018. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit. 84, 317–331.

Bilal, M., Gani, A., Lali, M.I.U., Marjani, M., Malik, N., 2019. Social profiling: a review, taxonomy, and challenges. Cyberpsychol. Behav. Soc. Netw. 22 (7), 433–450.

Boulemtafes, A., Derhab, A., Challal, Y., 2020. A review of privacy-preserving techniques for deep learning. Neurocomputing 384, 21–45.

Boullé, N., Earls, C.J., Townsend, A., 2022. Data-driven discovery of Green's functions with human-understandable deep learning. Sci. Rep. 12 (1), 1–9.

Breuer, A., Ettrich, N., Habelitz, P., 2020. Deep learning in seismic processing: trim statics and demultiple. In: *SEG Technical Program Expanded Abstracts2020*: Society of Exploration Geophysicists, pp. 3199–3203.

Buckman, J., Roy, A., Raffel, C., Goodfellow, I., 2018. Thermometer encoding: one hot way to resist adversarial examples. International Conference on Learning Representations.

Buczak, A.L., Guven, E., 2016. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tut. 18 (2), 1153–1176.

Buduma, N., Buduma, N., Papa, J., 2022. Fundamentals of Deep Learning. O'Reilly Media, Inc..

Bun, M., Steinke, T., 2016. Concentrated differential privacy: simplifications, extensions, and lower bounds. In: Theory of Cryptography Conference, pp. 635–658.

Caminero, G., Lopez-Martin, M., Carro, B., 2019b. Adversarial environment reinforcement learning algorithm for intrusion detection. Comput. Netw. 159, 96–109.

Caminero, G., Lopez-Martin, M., Carro, B., 2019a. Adversarial environment reinforcement learning algorithm for intrusion detection. Comput. Netw. 159, 96–109.

Cao, X., Gong, N.Z., 2017. Mitigating evasion attacks to deep neural networks via region-based classification. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 278–287.

Carlini, N., Jagielski, M., Mironov, I., 2020. Cryptanalytic extraction of neural network models. In: Annual International Cryptology Conference. Springer, pp. 189–218.

Carlini, N., Wagner, D., 2017b. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. In: Proceedings of the 2017ACM SIGSAC Conference on Computer and Communications.

Carlini, N., Wagner, D., 2017a. Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14.

Carlini, N., Wagner, D., 2018. Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 1–7.

Ch, R., Srivastava, G., Gadekallu, T.R., Maddikunta, P.K.R., Bhattacharya, S., 2020. Security and privacy of UAV data using blockchain technology. J. Inf. Secur. App. 55, 102670.

Chabanne, H., de Wargny, A., Milgram, J., Morel, C., Prouff, E., 2017. Privacy-preserving classification on deep neural network. IACR Cryptol. ePrint Archive 35.

Chang, K., Singh, P., Vepakomma, P., Poirot, M.G., Raskar, R., et al., 2021. Privacy-preserving collaborative deep learning methods for multiinstitutional training without sharing patient data. In: Artificial Intelligence in Medicine. Elsevier, pp. 101–112.

Chaudhuri, K., Monteleoni, C., 2009. Privacy-preserving logistic regression. In: Advances in Neural Information Processing Systems, pp. 289–296.

Chen, H., Li, H., Dong, G., Hao, M., Xu, G., et al., 2020d. Practical membership inference attack against collaborative inference in industrial IoT. IEEE Trans. Ind. Inf..

Chen, J., Jordan, M.I., Wainwright, M.J., 2020c. Hopskipjumpattack: a query-efficient decision-based attack. In: 2020 IEEE Symposium on Security and Privacy (sp). IEEE, pp. 1277–1294.

Chen, J., Zhang, J., Zhao, Y., Han, H., Zhu, K., et al., 2020e. Beyond model-level membership privacy leakage: an adversarial approach in federated learning. In: 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, pp. 1–9.

Chen, L., Xu, Y., Xie, F., Huang, M., Zheng, Z., 2020b. Data poisoning attacks on neighborhood-based recommender systems. In: Transactions on Emerging Telecommunications Technologies, p. 3872.

Chen, W., Zhang, Z., Hu, X., Wu, B., 2020a. Boosting decision-based black-box adversarial attacks with random sign flip. In: Proceedings of the European Conference on Computer Vision.

Chen, Y., Zhu, K., Zhu, L., He, X., Ghamisi, P., et al., 2019. Automatic design of convolutional neural network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (9), 7048–7066.

Chiu, T.-C., Shih, Y.-Y., Pang, A.-C., Wang, C.-S., Weng, W., et al., 2020. Semi-supervised distributed learning with non-IID Data for AIoT service platform. IEEE Internet Things J..

Choraś, M., Pawlicki, M., 2020. Intrusion detection approach based on optimised artificial neural network. Neurocomputing.

Correia-Silva, J.R., Berriel, R.F., Badue, C., de Souza, A.F., Oliveira-Santos, T., 2018. Copycat CNN: stealing knowledge by persuading confession with random non-labeled data. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.

Dagan, Y., Feldman, V., 2020. PAC learning with stable and private predictions. In: *Conference on Learning Theory*: PMLR, pp. 1389–1410.

Dasgupta, D., Akhtar, Z. and Sen, S., "Machine learning in cybersecurity: a comprehensive survey," *J. Defense Model. Simul.*, p. 1548512920951275.

Debicha, I., Bauwens, R., Debatty, T., Dricot, J.-.M., Kenaza, T., et al., 2023. TAD: transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems. Fut. Gener. Comput. Syst. 138, 185–197.

De Gaspari, F., Hitaj, D., Pagnotta, G., De Carli, L., Mancini, L.V., 2020. The naked sun: malicious cooperation between benign-looking processes. In: International Conference on Applied Cryptography and Network Security. Springer, pp. 254–274.

DelVecchio, M., Arndorfer, V., Headley, W.C., 2020. Investigating a spectral deception loss metric for training machine learning-based evasion attacks. In: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, pp. 43–48.

Dorosh, N., Fenenko, T., 2020. Recognition of MNIST handwritten digits and character set research. In: International Scientific and Technical Conference Information Technologies in Metallurgy and Machine Building, pp. 299–302.

Dourado Jr, C.M., da Silva, S.P.P., da Nobrega, R.V.M., Barros, A.C.d.S., Reboucas Filho, P.P., et al., 2019. Deep learning IoT system for online stroke detection in skull computed tomography images. Comput. Netw. 152, 25–39.

Du, T., Ji, S., Li, J., Gu, Q., Wang, T., et al., 2020. Sirenattack: generating adversarial audio for end-to-end acoustic systems. In: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, pp. 357–369.

Ducas, L., Micciancio, D., 2015. FHEW: bootstrapping homomorphic encryption in less than a second. In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2015, pp. 617–640.

Dunn, C., Moustafa, N., Turnbull, B., 2020. Robustness evaluations of sustainable machine learning models against data poisoning attacks in the Internet of Things. Sustainability 12 (16), 6434.

Dutta, V., Choraś, M., Pawlicki, M., Kozik, R., 2020b. Detection of cyberattacks traces in IoT data. J. Univ. Comput. Sci. 26 (11), 1422–1434.

Dutta, V., Choraś, M., Pawlicki, M., Kozik, R., 2020a. A deep learning ensemble for network anomaly and cyber-attack detection. Sensors 20 (16), 4583.

Dwork, C., 2006. Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (Eds.), Automata, Languages and Programming. Springer, Berlin Heidelberg, pp. 1–12 2006.

ElGamal, T., 1985. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inf. Theory 31 (4), 469–472.

Elhadj, H.B., Sallabi, F., Henaien, A., Chaari, L., Shuaib, K., et al., 2021. Do-Care: a dynamic ontology reasoning based healthcare monitoring system. Fut. Gener. Comput. Syst. 118, 417–431.

El-Rewini, Z., Sadatsharan, K., Selvaraj, D.F., Plathottam, S.J., Ranganathan, P., 2020. Cybersecurity challenges in vehicular communications. Veh. Commun. 23, 100214.

Elsayed, G.F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., et al., 2019. Adversarial examples influence human visual perception. J. Vis. 19 (10), 190.

Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H., 2020. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J. Inf. Secur. App. 50, 102419.

Fu, Y., Du, Y., Cao, Z., Li, Q., Xiang, W., 2022. A deep learning model for network intrusion detection with imbalanced data. Electronics 11 (6), 898.

Furqan, H.M., Solaija, M.S.J., Türkmen, H., Arslan, H., 2021. Wireless communication, sensing, and REM: a security perspective. IEEE Open J. Commun. Soc..

Gadekallu, T.R., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P.K., et al., 2020. Early detection of diabetic retinopathy using PCA-firefly based deep learning model. *Electronics* 9 (2), 274.

Gamage, S., Samarabandu, J., 2020. Deep learning methods in network intrusion detection: a survey and an objective comparison. J. Netw. Comput. Appl. 169, 102767.

Gan, C., Feng, Q., Zhang, Z., 2021. Scalable multi-channel dilated CNN-BiLSTM model with attention mechanism for Chinese textual sentiment analysis. Fut. Gener. Comput. Syst..

Ganju, K., Wang, Q., Yang, W., Gunter, C.A., Borisov, N., 2018. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 619–633.

Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., et al., 2020. Deep learning interpretation of echocardiograms. NPJ Digit. Med. 3 (1), 1–10.

Ghosh, G., Verma, S., Jhanjhi, N., Talib, M., 2020. Secure surveillance system using chaotic image encryption technique. IOP Conference Series: Materials Science and Engineering, 993. IOP Publishing.

Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., et al., 2016. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning, pp. 201–210.

Goecks, J., Jalili, V., Heiser, L.M., Gray, J.W., 2020. How machine learning will transform biomedicine. Cell 181 (1), 92–101.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., et al., 2014. Generative adversarial nets. Advances in Neural Information Processing Systems 2672–2680.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2020. Adversarial examples improve image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 819–828.

Guan, Z., Bian, L., Shang, T., Liu, J., 2018. When machine learning meets security issues: a survey. In: 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR), pp. 158–165.

Gupta, K.D., Dasgupta, D., Akhtar, Z., 2020. Applicability issues of evasion-based adversarial attacks and mitigation techniques. 2020 IEEE Symposium Series on Computational Intelligence (SSCI).

Ha, T., Dang, T.K., Le, H., Truong, T.A., 2020. Security and privacy issues in deep learning: a brief review. SN Comput. Sci. 1 (5), 1–15.

Ha, T., Dang, T.K., Nguyen-Tan, N., 2021. Comprehensive analysis of privacy in black-box and white-box inference attacks against generative adversarial network. In: Future Data and Security Engineering: 8th International Conference, FDSE2021, Virtual Event, November 24–26, 2021, Proceedings 8. Springer, pp. 323–337.

Hamm, J., Cao, Y., Belkin, M., 2016. Learning privately from multiparty data. In: International Conference on Machine Learning, pp. 555–563.

Hao, J., Tao, Y., 2022. Adversarial attacks on deep learning models in smart grids. Energy Rep. 8, 123–129.

Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., et al., 2016. The role of big data in smart city. Int. J. Inf. Manage. 36 (5), 748–758.

Hashem, I.A.T., Ezugwu, A.E., Al-Garadi, M.A., Abdullahi, I.N., Otegbeye, O. et al., “A machine learning solution framework for combatting covid-19 in smart cities from multiple dimensions,” *medRxiv*, p. 2020.

Hassan, A., Kamran, M., Illahi, A., Zahoor, R.M.A., 2019. Design of cascade artificial neural networks optimized with the memetic computing paradigm for solving the nonlinear Bratu system. Eur. Phys. J. Plus 134 (3), 1–13.

Hassan, R., Qamar, F., Hasan, M.K., Aman, A.H.M., Ahmed, A.S., 2020. Internet of Things and its applications: a comprehensive survey. Symmetry 12 (10), 1674.

Hathaliya, J.J., Tanwar, S., Sharma, P., 2022. Adversarial learning techniques for security and privacy preservation: a comprehensive review. Secur. Priv. 5 (3), e209.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Helmstaedter, M., Briggman, K.L., Turaga, S.C., Jain, V., Seung, H.S., et al., 2013. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500 (7461), 168.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hong, D., Yokoya, N., Xia, G.-S., Chanussot, J., Zhu, X.X., 2020. X-ModalNet: a semi–supervised deep cross-modal network for classification of remote sensing data. ISPRS J. Photogramm. Remote Sens. 167, 12–23.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, R., Li, Z., Zhao, J., 2019. A verifiable fully homomorphic encryption scheme. In: International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, pp. 412–426.

Huang, T., Zhang, Q., Liu, J., Hou, R., Wang, X., et al., 2020. Adversarial attacks on deep-learning-based SAR image target recognition. J. Netw. Comput. Appl., 102632 doi:10.1016/j.jnca.2020.102632. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804520301065.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., et al., 2019. Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems 125–136.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., et al., 2018. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp. 19–35.

Jayaraman, B., Evans, D., 2019. Evaluating differentially private machine learning in practice. In: 28th Security Symposium (Security 19), pp. 1895–1912.

Jhanjhi, N., Verma, S., Talib, M., Kaur, G., 2020. A Canvass of 5G network slicing: architecture and security concern. IOP Conference Series: Materials Science and Engineering, 993. IOP Publishing.

Ji, Y., Zhang, X., Ji, S., Luo, X., Wang, T., 2018. Model-reuse attacks on deep learning systems. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. ACM.

Jia, J., Wang, B., Cao, X., Gong, N.Z., 2020. Certified robustness of community detection against adversarial structural perturbation via randomized smoothing. In: Proceedings of The Web Conference 2020, pp. 2718–2724.

Jia, Q., Guo, L., Fang, Y., Wang, G., 2018. Efficient privacy-preserving machine learning in hierarchical distributed system. IEEE Trans. Netw. Sci. Eng..

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., et al., 2014. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678.

Jiang, W., Li, H., Liu, S., Luo, X., Lu, R., 2020. Poisoning and evasion attacks against deep learning algorithms in autonomous vehicles. IEEE Trans. Veh. Technol. 69 (4), 4439–4449.

Juuti, M., Szyller, S., Marchal, S., Asokan, N., 2019. PRADA: protecting against DNN model stealing attacks. In: 2019 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 512–527.

Juvekar, C., Vaikuntanathan, V., Chandrakasan, A., 2018. A low latency framework for secure neural network inference. In: 27th Security Symposium (Security 18), pp. 1651–1669.

Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. 2 (6), 305–311.

Kaur, D., Uslu, S., Durresi, A., 2020. Requirements for trustworthy artificial intelligence–a review. In: International Conference on Network-Based Information Systems. Springer, pp. 105–115.

Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S., 2020. A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. 53 (8), 5455–5516.

Khosravy, M., Nakamura, K., Hirose, Y., Nitta, N., Babaguchi, N., 2022. Model inversion attack by integration of deep generative models: privacy-sensitive face generation from a face recognition system. IEEE Trans. Inf. Forensics Secur. 17, 357–372.

Kim, T.H., Reeves, D., 2020. A survey of domain name system vulnerabilities and attacks. J. Surv. Secur. Saf. 1 (1), 34–60.

Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning, 70, pp. 1885–1894.

Kok, S., Azween, A., Jhanjhi, N., 2020. Evaluation metric for crypto-ransomware detection using machine learning. J. Inf. Secur. App. 55, 102646.

Kong, Z., Guo, J., Li, A., Liu, C., 2020. Physgan: generating physical-world-resilient adversarial examples for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14254–14263.

Krasnyanskiy, M.N., Dedov, D.L., Obukhov, A.D., Alekseev, S.Y., 2020. Visualization technology and tool selection methods for solving adaptive training complex structural-parametric synthesis problems. J. Comput. Inf. Sci. Eng. 20 (4).

Kumar, P., Kumar, R., Gupta, G.P., Tripathi, R., Srivastava, G., 2022. P2tif: a blockchain and deep learning framework for privacy-preserved threat intelligence in industrial iot. IEEE Trans. Ind. Inf. 18 (9), 6358–6367.

Kumar, R., Kumar, P., Tripathi, R., Gupta, G.P., Gadekallu, T.R., et al., 2021. SP2F: a secured privacy-preserving framework for smart agricultural Unmanned Aerial Vehicles. Comput. Netw. 187, 107819.

Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P., 2021. Multi-modal aggression identification using Convolutional Neural Network and Binary Particle Swarm Optimization. Fut. Gener. Comput. Syst. 118, 187–197.

Lee, H., Bae, H., Yoon, S., 2020. Gradient masking of label smoothing in adversarial robustness. IEEE Access.

Li, D., Liu, D., Guo, Y., Ren, Y., Su, J., et al., 2023. Defending against model extraction attacks with physical unclonable function. Inf. Sci..

Li, J., Yang, G., 2021. Network embedding enhanced intelligent recommendation for online social networks. Fut. Gener. Comput. Syst..

Li, X., Zhu, D., 2020. Robust detection of adversarial attacks on medical images. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE.

Lim, W.Y.B., Luong, N.C., Hoang, D.T., Jiao, Y., Liang, Y.-C., et al., 2020. Federated learning in mobile edge networks: a comprehensive survey. IEEE Commun. Surv. Tut..

Liu, L., 2021. Objects detection toward complicated high remote basketball sports by leveraging deep CNN architecture. Fut. Gener. Comput. Syst..

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., et al., 2018. A survey on security threats and defensive techniques of machine learning: a data driven view. IEEE Access 6, 12103–12117.

Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., 2020. Application of deep reinforcement learning to intrusion detection for supervised problems. Expert Syst. Appl. 141, 112963.

Lotfollahi, M., Siavoshani, M.J., Zade, R.S.H., Saberian, M., 2020. Deep packet: a novel approach for encrypted traffic classification using deep learning. Soft. Comput. 24 (3), 1999–2012.

Lovisotto, G., Eberz, S., Martinovic, I., 2020. Biometric backdoors: a poisoning attack against unsupervised template updating. In: 2020 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, pp. 184–197.

Lyth, D.H., 2005. Generating the curvature perturbation at the end of inflation. J. Cosmol. Astropart. Phys. 2005 (11), 006.

Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V., 2015. Deep neural nets as a method for quantitative structure–activity relationships. J. Chem. Inf. Model. 55 (2), 263–274.

Maiorca, D., Demontis, A., Biggio, B., Roli, F., Giacinto, G., 2020. Adversarial detection of flash malware: limitations and open issues. Comput. Secur., 101901 doi:10.1016/j.cose.2020.101901.

Mei, S., Zhu, X., 2015. Using machine teaching to identify optimal training-set attacks on machine learners. In: AAAI, pp. 2871–2877.

MirhoseiniNejad, S., Badawy, G., Down, D.G., 2021. Holistic thermal-aware workload management and infrastructure control for heterogeneous data centers using machine learning. Fut. Gener. Comput. Syst..

Mishra, P., Lehmkuhl, R., Srinivasan, A., Zheng, W., Popa, R.A., 2020. Delphi: a cryptographic inference service for neural networks. In: 29th Security Symposium (Security 20), pp. 2505–2522.

Mittal, S., Ramkumar, K., 2021. Research perspectives on fully homomorphic encryption models for cloud sector. J. Comput. Secur. 1–26 no. Preprint.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529.

Mohanty, S.N., Lydia, E.L., Elhoseny, M., Al Otaibi, M.M.G., Shankar, K., 2020. Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks. Phys. Commun., 101097.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P., 2017. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773.

Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P., 2016. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582.

Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A., et al., 2021. A survey on security and privacy of federated learning. Fut. Gener. Comput. Syst. 115, 619–640.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., et al., 2017. Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38.

NG, B.A., Selvakumar, S., 2020. Anomaly detection framework for Internet of things traffic using vector convolutional deep learning approach in fog environment. Fut. Gener. Comput. Syst. 113, 255–265.

Obukhov, A.D., Dedov, D.L., Arkhipov, A.E., 2018. Development of structural model of adaptive training complex in ergatic systems for professional use. IOP Conference Series: Materials Science and Engineering 327, 022075 2 ed.

Oh, S.J., Schiele, B., Fritz, M., 2019. Towards reverse-engineering black-box neural networks. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, pp. 121–144.

Orekondy, T., Schiele, B., Fritz, M., 2019. Knockoff nets: stealing functionality of black-box models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4954–4963.

Otoum, Y., Liu, D., Nayak, A., 2022. DL-IDS: a deep learning–based intrusion detection framework for securing IoT. Trans. Emerg. Telecommun. Technol. 33 (3), e3803.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., et al., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems 13991–14002.

Pan, X., Zhang, M., Ji, S., Yang, M., 2020. Privacy risks of general-purpose language models. In: 2020 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 1314–1331.

Panda, P., Chakraborty, I., Roy, K., 2019. Discretization based solutions for secure machine learning against adversarial attacks. IEEE Access 7, 70157–70168.

Pang, X., Zhou, Y., Li, P., Lin, W., Wu, W., et al., 2020. A novel syntax-aware automatic graphics code generation with attention-based deep neural network. J. Netw. Comput. Appl., 102636 doi:10.1016/j.jnca.2020.102636. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804520301107.

Pant, P., Barati Farimani, A., 2020. Reconstruction of turbulent high-resolution dns data using deep learning. Bull. Am. Phys. Soc..

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., et al., 2017b. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., et al., 2017a. Practical black-box attacks against machine learning. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017, pp. 506–519.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., et al., 2016b. The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387.

Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P., 2018a. SoK: security and privacy in machine learning. In: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 399–414.

Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A., 2016a. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., et al., 2018b. Scalable private learning with pate. Advances in Neural Information Processing Systems.

Patil, K.R., Zhu, J., Kopeć, Ł., Love, B.C., 2014. Optimal teaching for limited-capacity human learners. In: Advances in Neural Information Processing Systems, pp. 2465–2473.

Paudice, A., Muñoz-González, L., Lupu, E.C., 2018. Label sanitization against label flipping poisoning attacks. Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Pawlicki, M., Choraś, M., Kozik, R., 2020. Defending network intrusion detection systems against adversarial evasion attacks. Fut. Gener. Comput. Syst. 110, 148–154.

Pillai, T.R., Hashem, I.A.T., Brohi, S.N., Kaur, S., Marjani, M., 2018. Credit card fraud detection using deep learning technique. In: 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA). IEEE, pp. 1–6.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., et al., 2018. A survey on deep learning: algorithms, techniques, and applications. ACM Comput. Surv. 51 (5), 1–36.

Qi, P., Jiang, T., Wang, L., Yuan, X., Li, Z., 2022. Detection tolerant black-box adversarial attack against automatic modulation classification with deep learning. IEEE Trans. Reliab. 71 (2), 674–686.

Quiring, E., Rieck, K., 2020. Backdooring and poisoning neural networks with image-scaling attacks. In: 2020 IEEE Security and Privacy Workshops (SPW). IEEE, pp. 41–47.

Raschka, S., Patterson, J., Nolet, C., 2020. Machine Learning in Python: main developments and technology trends in data science, machine learning, and artificial intelligence. Information 11 (4), 193.

Ren, K., Zheng, T., Qin, Z., Liu, X., 2020. Adversarial attacks and defenses in deep learning. Engineering.

Riazi, M.S., Samragh, M., Chen, H., Laine, K., Lauter, K., et al., 2019. {XONN}: XNOR-based oblivious deep neural network inference. In: 28th Security Symposium (Security 19), pp. 1501–1518.

Rouhani, B.D., Riazi, M.S., Koushanfar, F., 2018. Deepsecure: scalable provably-secure deep learning. In: Proceedings of the 55th Annual Design Automation Conference, pp. 1–6.

Roy Chowdhury, A., Wang, C., He, X., Machanavajjhala, A., Jha, S., 2020. Cryptϵ: crypto-assisted differential privacy on untrusted servers. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp. 603–619.

Sadeghi, K., Banerjee, A., Gupta, S.K., 2020. A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE Trans. Emerg. Top. Comput. Intell..

Santos, K., Dias, J.P., Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. J. Saf. Res. 80, 254–269.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., et al., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577 (7792), 706–710.

Shafahi, A., Huang, W.R., Najibi, M., Suciu, O., Studer, C., et al., 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In: Advances in Neural Information Processing Systems, pp. 6103–6113.

Shamir, A., 1979. How to share a secret. Commun. ACM 22 (11), 612–613.

Shaukat, K., Luo, S., Varadharajan, V., 2022. A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks. Eng. Appl. Artif. Intell. 116, 105461.

Shen, H., Chen, S., Wang, R., 2021. A study on the uncertainty of convolutional layers in deep neural networks. Int. J. Mach. Learn. Cybern. 1–13.

Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P., 2018. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inform. 22 (5), 1589–1604.

Shi-qi, L., Bo, N., Ping, J., Sheng-wei, T., Long, Y., et al., 2019. Deep Learning in Drebin: Android malware image texture median filter analysis and detection. KSII Trans. Internet Inf. Syst. 13 (7), 3654–3670.

Siddiqui, M.U.A., Qamar, F., Tayyab, M., Hindia, M.N., Nguyen, Q.N., et al., 2022. Mobility management issues and solutions in 5G-and-beyond networks: a comprehensive review. Electronics 11 (9), 1366.

Simon-Gabriel, C.-J., Ollivier, Y., Bottou, L., Schölkopf, B., Lopez-Paz, D., 2019. First-order adversarial vulnerability of neural networks and input dimension. In: International Conference on Machine Learning, pp. 5809–5817.

Song, C., Ristenpart, T., Shmatikov, V., 2017. Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 587–601.

Song, J., Wang, W., Gadekallu, T.R., Cao, J., Liu, Y., 2022. Eppda: an efficient privacy-preserving data aggregation federated learning scheme. IEEE Trans. Netw. Sci. Eng..

Steinhardt, J., Koh, P.W.W., Liang, P.S., 2017. Certified defenses for data poisoning attacks. In: Advances in Neural Information Processing Systems, pp. 3517–3529.

Stutz, D., Hein, M., Schiele, B., 2020. Confidence-calibrated adversarial training: generalizing to unseen attacks. In: International Conference on Machine Learning. PMLR, pp. 9155–9166.

Sugawara, T., Cyr, B., Rampazzi, S., Genkin, D., Fu, K., 2020. Light commands: laser-based audio injection attacks on voice-controllable systems. In: 29th Security Symposium (Security 20), pp. 2631–2648.

Sun, S., Huang, H., Peng, T., Shen, C., Wang, D., 2023. A data privacy protection diagnosis framework for multiple machines vibration signals based on a swarm learning algorithm. IEEE Trans. Instrum. Meas. 72, 1–9.

Sun, S., Yeh, C.-F., Ostendorf, M., Hwang, M.-Y., Xie, L., 2018. Training augmentation with adversarial examples for robust speech recognition. In: Proc. Interspeech 2018, pp. 2404–2408.

Sun, Y., Liu, J., Wang, J., Cao, Y., Kato, N., 2020a. When machine learning meets privacy in 6 g: a survey. IEEE Commun. Surv. Tut. 22 (4), 2694–2724.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A.A., et al., 2020b. Test-time training with self-supervision for generalization under distribution shifts. International Conference on Machine Learning (ICML).

Syed, D., Refaat, S.S., Bouhali, O., 2020. Privacy preservation of data-driven models in smart grids using homomorphic encryption. Information 11 (7), 357.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D. et al., "Intriguing properties of neural networks," *2nd International Conference on Learning Representations, ICLR 2014,* p. 2013.

Takiddin, A., Ismail, M., Zafar, U., Serpedin, E., 2020. Robust electricity theft detection against data poisoning attacks in smart grids. IEEE Trans. Smart Grid.

Tang, X., Li, Y., Sun, Y., Yao, H., Mitra, P., et al., 2020. Transferring robustness for graph neural network against poisoning attacks. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 600–608.

Tariq, M.I., Memon, N.A., Ahmed, S., Tayyaba, S., Mushtaq, M.T., et al., 2020a. A review of deep learning security and privacy defensive techniques. Mob. Inf. Syst..

Tariq, M.I., Tayyaba, S., Ashraf, M.W., Balas, V.E., 2020b. Deep learning techniques for optimizing medical big data. In: Deep Learning Techniques For Biomedical and Health Informatics. Elsevier, pp. 187–211.

Tariq, M.I., Tayyaba, S., Rasheed, H., Ashraf, M.W., 2017. Factors influencing the cloud computing adoption in higher education institutions of Punjab, Pakistan. In: 2017 International Conference on Communication, Computing and Digital Systems (C-CODE), pp. 179–184.

Tasaki, S., Gaiteri, C., Mostafavi, S., Wang, Y., 2020. Deep learning decodes the principles of differential gene expression. Nat. Mach. Intell. 2 (7), 376–386.

Tayyab, M., Marjani, M., Jhanjhi, N., Hashem, I.A.T., 2021b. A light-weight watermarking-based framework on dataset using deep learning algorithms. In: 2021 National Computing Colleges Conference (NCCC). IEEE, pp. 1–6.

Tayyab, M., Marjani, M., Jhanjhi, N., Hashem, I.A.T. and Usmani, R.S.A., "A Watermark-Based Secure Model For Data Security Against Security Attacks For Machine Learning Algorithms."

Tayyab, M., Marjani, M., Jhanjhi, N., Hashim, I.A.T., Almazroi, A.A., et al., 2021a. Cryptographic based secure model on dataset for deep learning algorithms. CMC 69 (1), 1183–1200.

Thiyagarajan, P., 2020. A review on cyber security mechanisms using machine and deep learning algorithms. In: Handbook of Research on Machine and Deep Learning Applications for Cyber Security. IGI Global, pp. 23–41.

Tian, Y., Pei, K., Jana, S., Ray, B., 2018. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th International Conference on Software Engineering, pp. 303–314.

Tolpegin, V., Truex, S., Gursoy, M.E., Liu, L., 2020. Data poisoning attacks against federated learning systems. In: European Symposium on Research in Computer Security. Springer, pp. 480–501.

Tong, L., Li, B., Hajaj, C., Xiao, C., Zhang, N., et al., 2019. Improving robustness of {ML} classifiers against realizable evasion attacks using conserved features. In: 28th Security Symposium (Security 19), pp. 285–302.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., et al., 2017. Ensemble adversarial training: attacks and defenses. International Conference on Learning Representations.

Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T., 2016. Stealing machine learning models via prediction apis. In: 25th Security Symposium (Security 16), pp. 601–618.

Tran, T.C., Dang, T.K., 2021. Machine learning for prediction of imbalanced data: credit fraud detection. In: 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM). IEEE, pp. 1–7.

Ullah, A., Azeem, M., Ashraf, H., Alaboudi, A., Humayun, M., et al., 2021. Secure healthcare data aggregation and transmission in IoT-A survey. IEEE Access.

Usmani, R.S.A., Pillai, T.R., Hashem, I.A.T., Jhanjhi, N.Z., Saeed, A., et al., 2020. A spatial feature engineering algorithm for creating air pollution health datasets. Int. J. Cognit. Comput. Eng. 1, 98–107.

Usmani, R.S.A., Saeed, A., Tayyab, M., 2021. Role of ICT for community in education during COVID-19. In: ICT Solutions for Improving Smart Communities in Asia. IGI Global, pp. 125–150.

Vedaldi, A., Lenc, K., 2020. Matconvnet: convolutional neural networks for Matlab. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 689–692.

Visaggio, C.A., Marulli, F., Laudanna, S., Zazzera, B.La, Pirozzi, A., 2021. A comparative study of adversarial attacks to malware detectors based on deep learning. In: Malware Analysis Using Artificial Intelligence and Deep Learning. Springer, pp. 477–511.

Vivek, B., Baburaj, A., Babu, R.V., 2019. Regularizer to mitigate gradient masking effect during single-step adversarial training. In: CVPR Workshops, pp. 66–73.

Vizitiu, A., Niţă, C.I., Puiu, A., Suciu, C., Itu, L.M., 2020. Applying deep neural networks over homomorphic encrypted medical data. Comput. Math. Methods Med..

Wagh, S., Gupta, D., Chandran, N., 2019. Securenn: 3-party secure computation for neural network training. In: Proceedings on Privacy Enhancing Technologies, pp. 26–49.

Wagh, S., Tople, S., Benhamouda, F., Kushilevitz, E., Mittal, P., et al., 2021. Falcon: honest-majority maliciously secure framework for private deep learning. In: Proceedings on Privacy Enhancing Technologies, pp. 188–208.

Wang, B., Gong, N.Z., 2018. Stealing hyperparameters in machine learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp. 36–52.

Wang, S., Sahay, R. and Brinton, C.G., "How potent are evasion attacks for poisoning federated learning-based signal classifiers?," *arXiv preprint arXiv:2301.08866*, 2023.

Wang, X., Bouzembrak, Y., Lansink, A.O., van der Fels-Klerx, H., 2022. Application of machine learning to the monitoring and prediction of food safety: a review. Comprehens. Rev. Food Sci. Food Saf. 21 (1), 416–434.

Wang, X., Li, J., Kuang, X., Tan, Y.-a., Li, J., 2019. The security of machine learning in an adversarial setting: a survey. J. Parallel Distrib. Comput..

Wood, A., Najarian, K., Kahrobaei, D., 2020. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput. Surv. 53 (4), 1–35.

Wu, S., Roth, A., Ligett, K., Waggoner, B., Neel, S., 2019. Accuracy first: selecting a differential privacy level for accuracy-constrained ERM. J. Priv. Confident. 9 (2).

Wu, Y., Ma, Y., Dai, H.-.N., Wang, H., 2021. Deep learning for privacy preservation in autonomous moving platforms enhanced 5G heterogeneous networks. Comput. Netw. 185, 107743.

Wu, Z., Wang, J., Hu, L., Zhang, Z., Wu, H., 2020. A network intrusion detection method based on semantic re-encoding and deep learning. J. Netw. Comput. Appl. 164, 102688.

Xiao, Q., Li, K., Zhang, D., Xu, W., 2018. Security risks in deep learning implementations. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 123–128.

Xie, Q., Luong, M.-.T., Hovy, E., Le, Q.V., 2020. Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., et al., 2015. The human splicing code reveals new insights into the genetic determinants of disease. Science 347 (6218), 1254806.

Xu, G., Xin, G., Jiao, L., Liu, J., Liu, S., et al., 2023. Ofei: a semi-black-box android adversarial sample attack framework against dlaas. IEEE Trans. Comput..

Xu, J., 2020. A deep learning approach to building an intelligent video surveillance system. Multimed. Tools Appl. 1–21.

Yang, C., Wu, Q., Li, H., Chen, Y., 2017. Generative Poisoning Attack Method Against Neural Networks. Springer, Cham.

Yang, Y., Yang, H., Liu, F., 2020. Group motion of autonomous vehicles with anti-disturbance protection. J. Netw. Comput. Appl., 102661 doi:10.1016/j.jnca.2020.102661. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804520301351.

Yao, A.C.-C., 1986. How to generate and exchange secrets. In: 27th Annual Symposium on Foundations of Computer Science (sfcs1986), pp. 162–167.

Yu, Z., Zhou, Y., Zhang, W., 2020. How can we deal with adversarial examples? In: 2020 12th International Conference on Advanced Computational Intelligence (ICACI). IEEE, pp. 628–634.

Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., et al., 2018. Commandersong: a systematic approach for practical adversarial voice recognition. In: 27th Security Symposium (Security 18), pp. 49–64.

Zhang, H., Weng, T.-.W., Chen, P.-.Y., Hsieh, C.-.J., Daniel, L., 2018. Efficient neural network robustness certification with general activation functions. In: Advances in Neural Information Processing Systems, pp. 4939–4948.

Zhang, J., Zheng, K., Mou, W., Wang, L., 2017. Efficient private ERM for smooth objectives. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence.

Zhang, Y., Shi, X., Zhang, H., Cao, Y., Terzija, V., 2022a. Review on deep learning applications in frequency analysis and control of modern power system. Int. J. Electr. Power Energy Syst. 136, 107744.

Zhang, Z., Liu, Q., Huang, Z., Wang, H., Lee, C.-.K., et al., 2022b. Model inversion attacks against graph neural networks. IEEE Trans. Knowl. Data Eng..

Zhao, Y., Chen, J., 2022. A survey on differential privacy for unstructured data content. ACM Comput. Surv. 54 (10s), 1–28.

Zhao, Y., Chen, J., Zhang, J., Wu, D., Blumenstein, M., et al., 2020. Detecting and mitigating poisoning attacks in federated learning using generative adversarial networks. In: Concurrency and Computation: Practice and Experience, p. e5906.

Zhong, Y., Chen, W., Wang, Z., Chen, Y., Wang, K., et al., 2020. HELAD: a novel network anomaly detection model based on heterogeneous ensemble learning. Comput. Netw. 169, 107049.

Zhong, Y., Deng, W., Wang, M., Hu, J., Peng, J., et al., 2019. Unequal-training for deep face recognition with long-tailed noisy data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7812–7821.

Zhou, H., Chen, K., Zhang, W., Fang, H., Zhou, W., et al., 2019. DUP-Net: denoiser and upsampler network for 3D adversarial point clouds defense. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1961–1970.

Zuo, C., Qian, J., Feng, S., Yin, W., Li, Y., et al., 2022. Deep learning in optical metrology: a review. Light 11 (1), 1–54.

**Muhammad Tayyab** received bachelor's in computer engineering (BCE) from Bahira University Islamabad, Pakistan in 2012 and M.S. degree in Information security MS(IS) from COMSATS University Islamabad, Islamabad Campus, Pakistan in 2016. He is currently doing Ph.D. in computer Science at Taylor's University, Lakeside campus, Subang Jaya, Malaysia. From 2013 to 2015 he was high school teacher in public govt education department, in Rawalpindi, Pakistan. After that from 2015 to 2018 he worked as lecture in one of the leading private university named University of Management and Technology (UMT) Sialkot Campus, Punjab, Pakistan. His research area is Big data, Machine Learning, Deep Learning, Cryptography, Applied cryptography and Data Science.

**Mohsen Marjani** has received his degree of Master of Information in Multimedia Computing from Multimedia University (MMU) in 2011, Malaysia and Doctor of Philosophy (Ph.D.) degree in Computer Science from the University of Malaya (UM), Malaysia. Currently, He is a lecturer at the Department of Computing and IT, Taylor's University, Selangor, Malaysia. Dr. Mohsen has published a number of research articles in refereed international journals. He started teaching (mostly Mathematics and IT-based subjects) since mid-1999 in different public, and private institutions including universities, pre-universities, high schools, and primary schools. He also was conducting IT courses for teachers of the Organization of the Education of Iran (Region 4 of Tehran) for about 2 years. Dr. Mohsen has experience of collaborating with and working for a number of IT companies as IT consultant, senior web and mobile developer, tech lead, project manager, and CTO. He was involved in a variety of IT-based projects in the past years. His area of interest includes Big Data, Data Analytics, Machine Learning, IoT, and Distributed Computing.

**N. Z. Jhanjhi** was with ILMA University and King Faisal University (K.F.U.), Saudi Arabia, for a period of ten years. He has 20 years of teaching and administrative experience. He has great international exposure in academia, research, administration, and academic quality accreditation. He has an intensive background of academic quality accreditation in higher education besides scientific research activities. He was with Academic Accreditation, for a period of ten years. He was with the National Commission for Academic Accreditation and Assessment (NCAAA) and the Education Evaluation Commission Higher Education Sector (EECHES) formerly NCAAA, Saudi Arabia, for Institutional Level Accreditation. He was also with the National Computing Education Accreditation Council (NCEAC). He is currently an Associate Professor with Taylor's University Malaysia. He is also a Moderator with the IEEE TechRxiv, a keynote speaker with the several IEEE international conferences globally, an external examiner/evaluator for the master's and Ph.D. degrees for several universities. He has supervised several postgraduate students, including the master's and Ph.D. He has edited/authored more than 13 research books with international reputed publishers. He has earned several research grants and a great number of indexed research articles on his credit. He is an active TPC member of reputed conferences around the globe. He received the ABET Accreditation twice for three programs from CCSIT, King Faisal University. He has awarded as a Top Reviewer one% globally from WoS/ISI (Publons), in 2019. He serves as a Guest editor for several reputed journals and a member of the editorial board for several research journals. He serves as an Associate Editor for IEEE ACCESS. 4.

**Ibrahim Abaker Targio Hashem** has received his master's degree in computer science from the University of Wales, Newport and Doctor of Philosophy (Ph.D.) degree in Computer Science from University of Malaya. Dr. Hashem obtained professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation). He is presently working as a lecturer at the Department of Computing and IT, Taylor's University, Selangor, Malaysia. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are very famous and among the most downloaded in top journals. His area of interest includes Big Data, Cloud Computing, Distributed Computing, and Machine Learning. He is an active member of Mobile Cloud Computing center, Malaysia.

**Raja Sher Afgun Usmani** received the B.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2011 and the M.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2017. He is currently pursuing the Ph.D. degree in computer science at Taylor's University, Malaysia. From 2010 to 2015, he worked with various software development companies as a software developer in Islamabad, Pakistan. From 2015 to 2018, he was a Senior Lab Engineer with the International Islamic University, Islamabad, Pakistan. His research interest includes the Geographical Information Systems, Spatial Data, Big Data, Data Mining and Data Science.

**Faizan Qamar** has a Ph.D. degree in Wireless Networks from the Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia in October 2019. He had completed M.E. degree in Telecommunication from NED University, Karachi, Pakistan in 2013, and B.E. degree in Electronics from Hamdard University, Karachi, Pakistan, in 2010. He is currently serving as Senior Lecturer in the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Selangor, Malaysia. He has more than eight years of research and teaching experience. He has authored and co-authored numerous ISI & Scopus journals and IEEE conference papers. He is also a reviewer of several national & international journals and IEEE conferences proceedings. His research interests include interference management, millimeter-wave communication, IoT networks, D2D communication and Quality of Service enhancement for the future wireless networks.