Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges

Yazan Alaya AL-Khassawneh

Data science and Artificial Intelligence Department, Zarqa University, Zarqa, Jordan
Correspondence: E-mail: ykhassawneh@zu.edu.jo

ABSTRACT

Artificial intelligence has the potential to address many societal, economic, and environmental challenges, but only if AI-enabled gadgets are kept secure. Many artificial intelligence (AI) models produced in recent years can be hacked by utilizing cutting-edge techniques. This issue has sparked intense research into adversarial AI to develop machine and deep learning models that can withstand various types of attacks. We provide a detailed summary of artificial intelligence in this paper to prove how adversarial attacks against AI applications can be mounted, covering topics such as confrontational knowledge and capabilities, existing methods for actually producing adversarial examples, and existing cyber defense models. In addition, we investigated numerous cyber countermeasures that could defend AI applications against these attacks and offered a systematic approach for demonstrating war strategies against machine learning and artificial intelligence. To safeguard AI applications, we emphasize the importance of understanding the intentions and methods of possible attackers. In the end, we list the biggest problems and most interesting research areas in the field of AI privacy and security.

ARTICLE INFO

Article History:
Submitted/Received 07 Aug 2022
First Revised 09 Sep 2022
Accepted 30 Nov 2022
First Available Online 02 Dec 2022
Publication Date 01 April 2023

Keyword:
Applications,
Artificial intelligence,
Challenges,
Opportunities,
Privacy,
Security.

1. INTRODUCTION

Current technological advancements and an increase in accessible computer power have led to the implementation of artificial intelligence (AI) tactics in a wide range of applications. As our society grows with increasingly, technologically advanced, and interconnected conditions, the significance of security solutions and prevention measures will increase. The dynamic nature of the threat environment complicates the already difficult task of protecting our systems and our society (which is dependent on these systems) (Guan & Ge, 2017).

For example, machine learning models are used to promote innovation in health care, gaming, and the economy, while autonomous vehicle manufacturers use deep learning models to develop pipelines for self-driving vehicles. The machine learning (ML) and, most importantly, deep learning (DL) algorithms used in many AI systems today make it possible to automate jobs and processes, giving them abilities and functions that were previously unimaginable (Risi & Preuss, 2020).

Regardless of the obvious success and benefits of machine learning, many of the models in use today are susceptible to a variety of attacks in which potential enemies attempt to circumvent the privacy, security, integrity, or accessibility of machine learning techniques by using inputs designed to cause the models to make inaccurate predictions (Dixit & Silakari, 2021; Bout *et al*., 2021). In many instances, AI systems are developed with little attention to security, making them vulnerable to hostile examples. During the machine learning training and testing phases, adversarial attacks on AI systems could occur (Liu *et al*., 2018). During the training process, an attacker may insert fake data into a training dataset to alter input attributes or data labels. In the literature, these assaults are referred to as "poisoning attacks" and are simple to execute in systems that utilize training data from unreliable sources. Adversary with knowledge of a machine-learning model can change training data to alter the initial distribution of a training dataset. Evasion attacks are the most common type of machine learning algorithm attack because they use a model's flaws to create hostile examples, which are then used to avoid the model during testing.

Given these considerable potential benefits, major companies such as Google, Facebook, IBM, and Microsoft are making significant investments in AI research and increasing their expenditures to examine its possibilities. Recent inventions like AlphaGo from Google, Watson from IBM, Siri from Apple, and Echo with Alexa from Amazon have gotten a lot of attention (Wirtz *et al*., 2019).

Technological advancement can be obtained by achieving superhuman performance. AlphaGo Zero, the most recent version of the artificial intelligence (AI) technology AlphaGo, is a technological advancement. Using self-play reinforcement learning, AlphaGo Zero can learn by playing against itself instead of being taught by humans. Thus, it is no longer limited by human expertise (David *et al.*, 2017).

It is a once-in-a-lifetime possibility to address a wide range of socioeconomic and environmental worries through the employment of AI approaches in a wide range of applications. This will not be achievable, however, until there is devoted research toward making these systems safer. In recent years, research on adversarial machine learning has received increased attention, and it is without a doubt that maintaining research efforts in this area will ensure the widespread deployment of AI technologies that may impact society. Due to the fact that AI is being used in more and more human activities and jobs, it is important to make algorithms that are safe and secure if we want to build a safe and secure future.

This paper examined surveys and summarized significant research

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

developments, applications, prospects, and difficulties in artificial intelligence security and confidentiality. The primary contributions of our study in six areas are in the following:

(i) This article gave a quick look at and summary of the basic ideas behind many artificial intelligence applications.
(ii) We presented a comprehensive examination of relevant prior surveys. Our paper contributes to the field by highlighting flaws in past surveys related to the development of AI applications that seem to be safe, which we address in this work.
(iii) We gave a high-level theoretical framework for classifying machine learning tasks, such as deep learning and federated learning.
(iv) We presented an innovative attack tool for AI systems that displays and evaluate s advanced attacks against application d omains.
(v) We provided a novel framework for security and privacy against AI system assaults that illustrates cyber defense strategies for defending AI systems from adversarial attacks.

We examined the issues surrounding this subject and provide recommendations for future research initiatives.

2. METHODS

For this inquiry, the evaluated literature was selected through the use of content analysis. Content analysis is usually applied to develop valid conclusions based on acquired data in an objective manner. This is done to identify significant aspects of previous studies that have been previously studied. Additionally, it enables qualitative as well as quantitative adjustments to be performed. Because of this, the conclusions of this study can be accepted since content analysis provides a full picture of AI's applications in the realms of security and privacy.

In the course of this investigation, samples were acquired by looking for and selecting papers that had previously been subjected to the procedure of peer review. The articles that were gathered came from authoritative academic sources. Here is a summary of what was done to find relevant literature for this investigation:

Academic databases like Web of Science, Scopus, and Science Direct, as well as ASCE Library, IEEE, Wiley Online Library, Sage, and Emerald, were used to find and choose articles. When searching the databases, terms such as "artificial intelligence," "artificial intelligence in security and privacy," and "computational intelligence" were used. Because of this, we were able to locate scholarly papers on the subject of the role artificial intelligence plays in securing personal data and other sensitive information. A total of 240 possible articles were uncovered in the period beginning in 2001 and ending in 2022.

The selection of these articles was based on how pertinent they were to the application of NLP, DL, and ML to the safeguarding of personal information. Articles were selected using a method consisting of two stages and several iterations, each of which took into consideration the aforementioned criteria. To be more specific, during the first round, we examined the titles, abstracts, and keywords of the publications to see whether or not they satisfied the requirements. In the second part of the process, we read and studied the entire article to confirm that all of the papers that had been selected were pertinent to the overall purpose of the review. The total number of papers that were considered for this analysis was 106.

The review utilized both qualitative and quantitative approaches to identify the applications of new AI methods in security and privacy, the AI algorithms that were used

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

in such applications, and an analysis of the applicability of these algorithms for the applications that were previously mentioned. Using this method, the most promising possible uses of new AI approaches and possible research directions for the future were found.

3. RESULTS AND DISCUSSION

3.1. The Theoretical Framework of Artificial Intelligence Applications

Through decades of research, there is no currently accepted definition of artificial intelligence (AI). This highlights the fundamental difficulty in comprehending AI in its entirety. To gain a fundamental understanding of AI, it may be useful to define "intelligence" as a separate notion before discussing the application of intelligence to computers and the explanation of the phrase "artificial intelligence." Legg and Hutter (2007) provided an integrated definition of intelligence, defining it as the capacity to engage, learn, adapt, and rely on experience-based knowledge, as well as deal with uncertainty. In this context, "artificial" refers to a copy created by humans. Based on this fundamental information, it is prudent to develop a more inclusive definition of AI in the context of general management for future studies. As you can see in **Table 1**, we took five definitions of AI from published articles to get a full picture of what it means

By reviewing relevant papers, Zawacki-Richter *et al*. (2019) provided an overview of the literature on the application of AI at universities. Following the application of specific inclusion and exclusion criteria, 146 articles were chosen for this summary from the original 2656 identified for the years 2007–2018. These descriptive data show that most Artificial Intelligence in Education (AIEd) papers cover topics from Computer Science and STEM, and most empirical investigations use quantitative methods. According to the conclusions of the synthesis, the following are four areas where AIEd has been used in the realms of academic support services and institutional and administrative services: systems that are personalized and adaptive; evaluation and assessment; intelligent teaching systems; and profiling and prediction.

Table 1. Artificial intelligence definitions.

Reference	Definition
Russel and Norvig (Stuart, 2010)	AI may be classified into four types: (i) systems that think in human terms. (ii) systems that behave like humans. (iii) rational-thinking systems (iv) systems that behave logically.
Rosa *et al*. (2016)	A system capable of learning, replicating, and perhaps exceeding human-level performance throughout the whole spectrum of cognitive and intellectual skills.
Thierer and Castillo (2020)	The demonstration of intelligence by a computer or other device. An AI system can carry out high-level operations, and it can accomplish tasks that are comparable to, on par with, or even beyond those of a person. This idea may be further broken down into two categories: weak AI and strong AI.
Hancock *et al*. (2020)	Although the goal of artificial intelligence is to simulate human intellect rather than human beings entirely, AI should nevertheless be conceptualized as distinct from human intelligence in many important respects.
Mäntymäki *et al*. (2022)	The distinction between natural intelligence, which is presented by animals, humans, and artificial intelligence, which is demonstrated by machines, agrees that artificial intelligence is demonstrated by machines.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Zhao *et al*. (2020) recently published a review on AI applications in power electronics. AI tasks like optimization, classification, regression, and data structure investigation are linked to the three distinct life-cycle stages of design, control, and maintenance. Four types of AI are addressed: expert systems, fuzzy logic, metaheuristics, and machine learning, each with possible practical applications. Reviewing over 500 publications revealed commonalities, practical implementation challenges, and future research objectives in the field of artificial intelligence applied to power electronics. This post is accompanied by a spreadsheet in Excel that has a list of books and magazines that are good for statistical analysis.

AI has a wide range of applications. This section lists some of the numerous sectors in which AI techniques have been applied and explains how they can be useful in engineering applications. Malik *et al*. (2018) discussed how AI methodologies are being employed in numerous disciplines of engineering, including the investigation of the distinguishing characteristics and abilities of intelligent machines. It contains papers written by business and academic researchers and experts. The study of system identification and function approximation is concerned with the development of empirical dynamic models of systems based on measurable data. Nonlinear prediction is concerned with the prediction of systems whose input-output connection is not linear. Pattern recognition, also known as pattern classification, refers to a wide range of situations in which the goal is to classify an object. A decision-assistance system based on AI for transportation systems could be very valuable. For example, clustering could be used to identify specific classes of drivers based on driver behavior. Computer-aided design has the potential to significantly increase the value and capabilities of computer-aided design.

In the transportation sector, artificial intelligence is being utilized to address challenges such as increased travel demand, carbon emissions, safety issues, and environmental harm. It is more practical to think about finding solutions to these difficulties in our present era of plentiful data (both quantitative and qualitative) and artificial intelligence (AI). Artificial Neural Networks (ANN), Genetic Algorithms (GA), Simulated Annealing (SA), Artificial Immune Systems (AIS), Ant Colony Optimiser (ACO), Bee Colony Optimization (BCO), and Fuzzy Logic Model are just a few of the AI techniques being used in transportation (FLM). Understanding the connections between AI and data on the one hand and the parts and features of a transportation system on the other is important for the system to work well. As a result, it is hoped that transportation authorities will figure out how to use these technologies to make rapid progress in reducing congestion, making travel time more predictable for customers, and improving the economics and effectiveness of their critical assets (Abduljabbar *et al*., 2019).

Several sorts of medical records are acceptable with artificial intelligence (structured and unstructured). Typical artificial intelligence approaches for structured data include neural networks and support vector machines, whereas common artificial intelligence tools for unstructured data include natural language processing and deep learning. Cancer, neurology, and cardiology are the three major illness areas where AI approaches are now being applied (Jiang *et al*., 2017). In addition, advancements in artificial intelligence are causing the medical practice to evolve in significant ways (AI). Advances in digital data capture, machine learning, and computer infrastructure have enabled the application of artificial intelligence to a growing number of problems that were once considered beyond the capabilities of even the most

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

qualified human professionals. This review article (Yu *et al*., 2018) provides an overview of the most recent advancements in artificial intelligence (AI) technologies and their biomedical applications. It also looks at the things that make it hard to make more advanced AI medical systems and gives an overview of the ethical, legal, and social effects of AI in healthcare.

In general, implementation is concerned with the aspects of a project that are delivered in a particular environment (Durlak & DuPre, 2008). Therefore, implementing AI in the public sector demands a well-considered and deliberate plan for maximizing AI's many potential benefits (Wirtz, 2019). Even though a lot of businesses and government agencies around the world have started working on AI implementation and applications, the public sector still faces a lot of problems in this area.

The majority of artificial intelligence solutions used "will likely stay poor and highly specialized" for the next several decades. Several obstacles restrict the development of AI in the public sector, and they appear to be related to the perception of implementation complexity. Four concerns have been recognized as the key constraints to deploying AI technology: AI safety, system/data quality and connectivity, financial feasibility, and specialization and knowledge.

Previous studies have identified AI safety as a significant risk element or problem of AI, referring to ensuring the secure performance and effect of AI (Hancock *et al*., 2020). This includes not just information security challenges, but also general security issues. This includes complicated and safety-critical issues originating from settings in which AI may acquire undesirable behavior from its environment or misunderstand its surroundings (Wirtz, 2019). In this perspective, the relevance and need for AI technology to be resistant to human manipulation.

Google, a leader in AI research, has found many security concerns that have already happened in practice. In the case of reinforcement learning-based AI applications, it must be ensured that the AI system learns without performing catastrophic acts. Furthermore, unwanted side effects such as upsetting the working environment must be avoided when doing the duties for which the entity is designed. A robot aiding in surgery, for example, should be able to learn without injuring the patient by trying cuts or surgical approaches. As a consequence, AI implementation and progress are connected to avoiding accidents and assuring the safe operation of AI applications to preserve humanity.

The quality and integration of the system/data are critical since the AI system is only as clever as the data from which it learns, and data is seen as "[t]he fundamental driver of current AI systems" (Thierer & Castillo, 2020). Low-quality or untrustworthy data, in particular, provides a significant barrier for enterprises. As a result, the gathering, aggregation, storage, and use of impartial and relevant data is required for effectively applying AI in the public sector, as erroneous or bad data may lead to failures (Boyd & Wilson, 2017). In this regard, developing a comprehensive and high-quality AI system capable of integrating data and managing the interlinkages among data, techniques, and processes is critical, but it also poses a significant obstacle in adopting AI solutions (Alshahrani *et al*., 2022).

Financial feasibility is also important in deploying AI technology, and a lack of funds is one of the most significant problems that businesses encounter when launching AI initiatives. As a result, before building and deploying an AI application, the whole cost associated with it, as well as the predicted revenues, must be assessed to determine if an AI solution is sustainably feasible. There are two significant cost drivers in this context that make financial feasibility a big hurdle in the context of implementation. The cost of

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

developing a sophisticated technology infrastructure to store and gather data, in particular, is enormous (Alshahrani *et al*., 2022). Furthermore, there is a significant demand for a limited number of AI professionals, which is related to rising education and salary costs.

Another critical facet of using AI technology in the public sector is specialization and experience. The fast rise of AI necessitates the availability of professionals and experts with essential abilities to assist and encourage AI development (Hancock *et al*., 2020). As a result, the worldwide demand for AI expertise has risen dramatically in recent years (Thierer & Castillo, 2020). However, as previously said, there is a scarcity of AI professionals and experts, which impedes AI implementation and therefore poses a significant obstacle in the area of AI research and implementation. In this respect, the government plays an important role and must pay particular attention to training and promoting a well-educated and diversified workforce to construct and create a sustainable expertise and knowledge foundation in AI (Boyd & Wilson, 2017).

3.2. Significant Recent Reviews

From a data-driven standpoint, Liu *et al.* (2018) offered a thorough review of known security risks and related protective strategies throughout the machine learning training and testing stages. Given the absence of a thorough literature analysis covering the security risks and defensive approaches used throughout the two stages of machine learning, their fundamental work provides a full assessment of current adversarial attack techniques and countermeasures used against machine learning. They offered a full definition of machine learning and introduced the notion of adversarial machine learning in particular. While the authors' study has influenced future research into aggressive machine learning, they did not examine known security risks to reinforcement learning.

Akhtar (2018) conducted the first complete review of adversarial attacks on deep learning in computer vision. This review examines known and proposed defenses for adversarial attack techniques as well as adversarial attack methods that have been effectively utilized against deep neural networks in both 'laboratory settings and real-world situations. This study, however, is constrained in that it only provides the most 'influential' and 'interesting' deep learning attacks in the confined domain of computer vision.

In response to a growing recognition that machine learning models are increasingly vulnerable to a wide range of adversarial capabilities, a systematic review of ML security and privacy with a specific focus on adversarial attacks on machine learning systems and their countermeasures. The authors of this research examined the ML threat model from the standpoint of a data pipeline and assessed current ML assaults and their countermeasures during the training and testing phases. Their evaluation also included existing papers on differential privacy. Thomas and Tabrizi assess the current landscape of research on the subject of adversarial machine learning, which analyses the findings and trends in the field. This study, however, did not directly concentrate on any of the assaults against ML systems, and it lacks the detail seen in many review publications in this sector.

Assessment of current research on the security and privacy of deep learning systems in different applications. The essay presents the notion of secure AI and gives a detailed study of many of the attacks that have successfully been used against deep learning as well as their mitigation measures, covering the fundamentals of deep learning and some of the privacy-preserving strategies in the literature. Approaches to privacy in deep learning are taxonomies, and future research

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

possibilities based on discovered gaps are offered. In their assessment of different forms of adversarial assaults and their responses, a description of adversarial attack types and provide countermeasures via the study of distinct threat models and attack scenarios. The examined attacks and responses were not limited to particular deep learning applications. As a countermeasure to adversarial assaults, strong deep learning systems are particularly recommended. However, the essay did not outline how to do this. These steps may have been included in future study initiatives.

Ren *et al*. (2020), in their study on adversarial assaults and countermeasures in deep learning, provided the groundwork for understanding adversarial attacks and present a concise overview of the state-of-the-art in this area. This study focuses on just those adversarial assaults that are relevant to the field of computer vision, while some of the attacks described are based on recent publications. This study by Zhang *et all*. (2020) showed recent research on cyberattacks on deep neural networks in the framework of natural language processing (NLP), providing a thorough introduction to the fundamentals of adversarial assaults and deep learning approaches in NLP. Both black-box and white-box attacks on deep learning models in NLP have been proven, and these attacks are summarized in the survey article. Moreover, they discussed two typical preventative methods for developing resilient textual deep neural networks: hostile training and knowledge filtration. While many of the assaults that have been shown to work against textual deep neural networks so far have been covered in this study, many of the defenses that have been presented in this field have been left out.

A survey of the extant literature on AI applications in access control authentication, network condition monitoring, harmful behavior monitoring, and abnormal traffic identification (54 publications, largely published between 2016 and 2020). Based on the findings, this study also identifies many limitations and challenges, and it proposes a conceptual human-in-the-loop intelligence cyber security model.

Wirtz *et al*. (2019) investigated the growing need for a complete understanding of the extent and impact of AI-based applications, as well as the associated challenges. Prior research, on the other hand, examined AI applications and difficulties in isolation and pieces. Because there is no comprehensive overview of AI-based applications and public-sector challenges, their conceptual approach explores and collects significant concepts from the scientific literature to give an integrated picture of AI applications and associated issues. Their findings propose ten artificial intelligence application fields, each with its growth, profitability, and operation as well as separate public use cases. They also highlighted four key characteristics of AI problems. Finally, we review their conclusions, drawing both theoretical and concrete inferences and making recommendations for further research.

3.3. The Conceptual Framework of Machine Learning Techniques

The earliest algorithms in this field date back to the 1970s (Ebert & Louridas, 2016). Machine learning (ML) is not a novel concept. Forecasting, anomaly detection, spam filtering, and reputation risk assessment are just a few of the myriad predictive tasks that machine learning may assist with. The primary purpose of the system is to create predictions based on available data.

Every AI system is significantly reliant on data for operation. The more diverse the training data, the more accurate the machine's predictions will be. For a machine to determine whether or not an email is a spam, it must have been trained with instances of spam messages. In machine learning, training and testing sets of input data are frequently separated. Data from training examples are used to develop a

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

machine learning model, and once the model's prediction accuracy is good enough, the data from the training dataset is transferred to the test set.

The three main building blocks of machine learning are tasks, models, and features. What we refer to as "tasks" are the kinds of problems that machine learning can address. In the field of machine learning, numerous models are created to solve a limited range of issues. The outputs of machine learning are characterized by their models. Training on sample data teaches them to process more information for predictive purposes. Features are characteristics of the input data that facilitate the identification of trends between the input data and the data flow, making them essential to machine learning. With the assistance of algorithms, we can overcome any learning obstacle. Machine learning is "the art of picking and combining features into suitable models for a given task."

"Supervised learning is a machine learning method for generating predictions and classifying data based on an analysis of a labeled training dataset (Nasteski, 2017; Muhammad & Yan, 2015)." In supervised learning, the training dataset is either labeled or numeric. Two examples of supervised learning problems are classification and regression approaches. Unsupervised learning, by contrast, involves learning algorithms that focus on recognizing patterns of similarity within the data to infer significant features (Muhammad & Yan, 2015), which can subsequently be extracted to provide potential output labels. As the name implies, semi-supervised learning is a machine learning approach that combines aspects of both supervised and unsupervised learning. It extends the scope of supervised learning to incorporate aspects of unsupervised learning (Blum & Langley, 1997; Dunjko & Briegel, 2018; Embley, 2004; Uddin *et al*., 2019) and vice versa. Reinforcement learning, as a machine learning technique (Anh *et al*., 2019; Caminero *et al*., 2019), allows a learning agent to make mistakes and progress through trial and error as it interacts with its surroundings. A reinforcement learner is a program that interacts with its surroundings by doing tasks in exchange for rewards. The goal of this approach to learning is to maximize rewards.

A subfield of machine learning is deep neural network learning. Because a feature engineer is required to identify meaningful features in the input data, the methods for machine learning discussed thus far in this section have been dubbed Shallow Learning (Song & Montenegro-Marin, 2021; Bacchi *et al*., 2019). In the construction of many shallow learning algorithms, the input data is frequently converted into a problem-specific feature space using a single layer (Deng, 2014). Deep learning, on the other hand, performs complex learning tasks and feature extraction by using multi-layered representation and generalization of input data. In applications where well-represented features must be extracted from data, such as natural language recognition and computer vision processing, the performance of many shallow learning techniques has been found deficient. Deep learning, on the other hand, avoids this challenge by substituting a tree-like structure of relatively simple data with a more abstract idea (Yuan *et al*., 2019). Deep learning systems were made possible by the exponential growth in the amount and type of data available and the increase in the processing power of chips.

For example, Google proposed teaching for the first time in 2016. When adopting this type of learning, a centralized model can be trained using data from a decentralized network of nodes (Yang *et al*., 2019; Zhang *et al*., 2021). The need to train models using data from users' mobile devices, which cannot be retained centrally in data centers due to privacy concerns, is the driving force

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

behind federated learning. Because federated learning communicates just the minimum updates required to construct a specific model, and this data is dependent on the training goal, it provides considerable privacy benefits over conventional machine learning methods. More data, or more nodes to train the model on, results in better-federated learning performance (Tabassi *et al*., 2019).

3.4. Security and Privacy Against AI System Attacks

It is concerning that AI technologies are widely employed while being vulnerable to hostile attacks. Adversarial attacks on ML and DL models can take the form of tampering with the input data to confuse the model and cause it to make a false classification. The type of world in which an algorithm life has a considerable impact on the types of adversarial attacks that can be launched against it. That is, while the sorts of attacks against AI systems are similar, the techniques of exploitation will differ depending on the exact algorithm used. This section begins by outlining the findings of a literature search we conducted on attacks on AI systems during the previous decade, followed by an in-depth study of the papers we discovered. Consult reference resources on protecting AI systems against invasions and other dangers to user privacy. In this section of the evaluation, we also classified defenses against AI system attacks based on whether they provide complete protection against unfriendly instances or just identify and reject the detection model.

One way to tell if a defense is complete or not is to see if it can stop attacks on the system during its testing or training stages. Poisoning attacks, like many other sorts of AI system attacks, make learning more difficult. Many training attacks and defensive approaches are based on the idea that poisoned samples are frequently not part of the predicted input distribution. Several alternative defense techniques against these types of training-time assaults have been proposed. Data Sanitization is a poisoning attack defense strategy that entails removing tainted samples from a training dataset before using it to train a model (Chan *et al*., 2018). This is done to keep a model from being tainted by a poison attack. Proposed an innovative way of cleaning data to avoid out-of-the-box anomaly detection classifiers from being poisoned by malicious data. The name of the data pasteurization strategy supplied by Nelson et al. was to safeguard one's system from being harmed is Reject On Negative Impact (RONI). Using this strategy, the Spam Bayes spam filter was able to properly filter out dictionary attack messages, and it accurately detected all of the attack emails while not reporting all of the non-spam emails. Based on Koh et al., at the same time, we introduced three additional data poison attacks that can avoid a wide range of data sanitization measures at the same time. Although data sanitization approaches have proven useful in fighting against some data poisoning attacks that were developed without specifically considering defenses, Koh et al. revealed these novel data poisoning attacks. Their success against anomaly-based information sanitization-based defenses just shows that more research is needed to find effective ways to stop poisoning attempts.

In contrast to data sanitization, this defensive method focuses on developing models that are resistant to poisoning attacks rather than identifying poisoned samples (Biggio *et al*., 2010). Biggio *et al*. (2010) investigated the use of multiple classifier systems (MCSs) to aid in the improvement of pattern recognition model robustness in adversarial circumstances (Melis *et al*., 2022). This is based on the idea that more than one classifier must be avoided for the system to be rendered unsuccessful. In their study, Biggio et al. collected data from experimental tests that showed randomization-based MCS building procedures have the potential to be

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

employed to boost the durability of linear classifiers when deployed in hostile situations. Biggio *et al*. ran an experiment in which they proved that "bagging," an abbreviation for bootstrap aggregating, is a viable protective approach against poisoning assaults regardless of the underlying categorization method. Breiman (1996) first introduced bagging as a technique for boosting classifier accuracy by generating numerous versions of a classification model and combining these to generate an aggregate classifier. This strategy is useful, especially when applied to classifiers whose predictions fluctuate significantly with little variation in training data.

In the event of an attack during the testing phase, certain model robustness upgrades, differential privacy, and homomorphic encryption are all potential lines of protection. Robustness in a machine learning model can be obtained by being able to recognize and ignore hostile samples. Although these measures for increasing robustness act as defenses against assaults during the testing phase, they are established before the testing phase. A lot of attention has been paid to the strategies of Adversarial Training, Protective Diffusion, Ensembles Approach, Gradient Masking, Feature Squeezing, Reformers/Auto encoders, and so on (Tabassi *et al*., 2019).

Adversarial training is a technique for improving the resilience of machine learning models by creating adversarial instances that are subsequently enhanced with training data. We optimize the input to maximize the model's prediction error to generate these adversarial scenarios. Adversarial training is resistant to white-box attacks, albeit it may take time due to the number of repetitive calculations required to build a robust model. In contrast, adversarial-trained deep neural network models that use rapid single-step techniques are vulnerable to basic black-box attacks.

Distillation is a method of preserving prediction accuracy while shrinking the size of a model from an ensemble of models or a large, highly regularized model. Hinton *et al*. (2015) formalized an idea described by Ba and Caruana (2014) as a means to deploy deep learning in computationally limited devices by smoothing a model's prediction accuracy via training shallow neural networks to emulate deeper neural networks. Proposed defensive distillation as a means for training models that are more tolerant of input disruptions. Their findings demonstrated that enhanced DNN-based models can survive damaging data better. During the testing phase, this line of defense was found to be effective against a variety of assaults. Despite the benefits of defensive distillation and the strong guarantees, it gives against adversarial instances, empirically show how their attacks using rising adversarial samples can beat defensive distillation.

Gradient masking, when applied to a DNN model, tries to reduce the effect of noise by masking away tiny changes in the input (Tabassi *et al*., 2019). This defensive strategy entails computing a model's first-order derivatives concerning its input and then minimizing these derivatives throughout the learning period. This conceals a model's gradient information from an attacker seeking to exploit it. While it makes sense to shield models against hostile situations by making them less sensitive to changes in input, the shortcomings of the gradient masking strategy by demonstrating how it can be bypassed via black-box attacks.

To discover adversarial scenarios, this cutting-edge technique narrows the characteristic input spaces that an opponent could exploit. It accomplishes this by assembling a single sample from a vast number of samples, each of which corresponds to a different feature vector in the original input space. Two feature squeezing approaches in the picture space to

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

increase detection accuracy and resilience against adversarial inputs: (1) image color depth reduction and (2) smoothing to minimize variation across pixels (median smoothing). The authors go much further in, proving that median smoothing is the greatest technique for defending against the Carlini & Wagner (2017) attack. Feature squeezing may make it harder to put benign inputs into the right category, even though it messes up the input features that attackers use to attack.

Training approaches known as "ensemble methods" are utilized to improve the categorization judgments of supervised learning models (Ren *et al*., 2020). Although several ensemble approaches have been proposed in the literature, they have only recently been investigated as a means of improving the robustness of machine learning models against adversarial attacks. To strengthen the resilience of convolutional neural networks against adversarial attacks, Abbasi and Gagné recommended integrating an ensemble of multiple expert classifiers to identify and reject hostile instances while admitting benign samples based on confidence. The authors showed that using this method can lower confidence in predictions in hostile cases while keeping some confidence in predictions made from clean data.

However, that adaptive adversaries that can successfully generate hostile samples with minor distortions can swiftly fool ensembles produced using this approach. When confronted with adversarial attacks, Their method is more computationally demanding than earlier ensemble methods, but it produces higher prediction confidence for clean data and greater robustness against adversarial circumstances. The Random Self-Ensemble (RSE) as a defensive strategy for increasing the stability of neural networks by integrating the notions of randomization and ensemble. Through tests, the authors show that their method is safe against attacks like the Carlini & Wagner (2017) attack and can be used in a wide range of situations.

3.5. Challenges and Opportunities

We previously released an in-depth examination of privacy and security risks in AI systems, as well as a discussion of their implications across a wide spectrum of machine-learning approaches. As you can see in **Table 2**, we end this section by pointing out some of the current problems in the field and making some suggestions for future research.

Even though a good number of the papers that were looked at for this study provide concrete proof of the transferability of adversarial instances, the basic reasons why adversarial instances transfer are still not well understood. The theory of the ubiquity of generalizability does not always hold, according to the experimental study in Tramèr *et al*. (2017), and there are hints that the transferability trait is not inherent in non-robust models, despite the presence of adversarial cases.

For this reason, it is very necessary to concentrate research efforts on gaining an understanding of the transferability phenomena to develop reliable machine learning models. One of the future study paths that should be followed is the investigation of the features that defensive mechanisms must possess to ensure their durability in the face of adaptive adversaries. Many studies have shown that all machine learning models, including deep neural networks, are vulnerable to adversarial attacks. As a result, expanded research efforts on adversarial risks confronting other machine learning task categories, such as reinforcement learning, are required. For machine learning systems to be able to find unknown unknowns with reliable approaches for intrusion detection systems, they would need to look into new areas of research in this field.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Table 2. Challenges and Opportunities.

Citation	Challenge	Description
Yuan *et al.* (2019)	The adversarial transfer is a problem that affects many machines learning models, including deep neural networks.	Due to this trait, many deep neural networks are vulnerable to attacks using adversarial instances created by one model against another model without the parameters of the latter model being known.
Papernot *et al.* (2016)	The adversarial example transferability is a problem for many machines learning methods, including deep neural networks.	thereby allowing for black-box assaults.
Carlini and Wagner (2017)	In many cases, current safeguards against hostile assaults have already been breached or circumvented.	According to their research, the current defenses don't work because they don't have detailed security assessments and are easy for adaptive attackers to break.
Bae *et al.* (2018)	Numerous anti-adversarial measures are inadequate or circumvented.	Differential privacy restricts the exposure of sensitive information included within the collection by introducing randomization into the training data. Differential privacy has been tested many times and has been shown to work. However, there are currently no public evaluations or criteria for figuring out if the differential privacy bounds are safe enough.
Zhang and Li (2019)	There are challenges in limiting the effects of disruptive actions taken by opponents.	Various adversarial example generators provide subtle, undetectable changes to the input to influence the neural network's prediction. However, it is challenging to determine the precise size of perturbations necessary to mislead a neural network since input perturbations that are too tiny cannot create adversarial instances while perturbations that are too big are not unnoticeable.
Kurakin *et al.* (2012), Moosavi-Dezfooli *et al.* (2017)	Insufficient attention was paid to assaults outside of categorization jobs in the literature.	Convolutional neural networks have proven to be very useful in computer vision tasks. Because computer vision tasks like recognizing pictures and identifying objects are part of machine learning classification tasks, most of the known ways to make adversarial examples also work in these fields.
Bliggion & Roli (2018)	The threat of the unknown unknowns.	Imponderables, like many cybersecurity concerns such as malware and intrusion detection, pose a significant risk to machine learning systems deployed in hostile environments.
Tramèr *et al.* (2017).	Changing the decision threshold of a classifier at random.	A known issue with randomization is that it may worsen the classifier's performance on clean data by increasing the initial error rate. Still unresolved is the question of how much randomness to add to a model to make it resistant to attacks from the outside.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

4. CONCLUSION

This study is founded on the concept that AI is becoming more relevant in both science and practice, and that it also has game-changing potential for all fields of global activity, which can be interpreted in both positive and negative ways. Over the last few years, there has been a lot of research interest in the area of adversarial machine learning. A huge number of research publications, in particular, have looked into adversarial attacks on machine learning models in the contexts of computer vision and image recognition, natural language processing, and cybersecurity. A lot of different defensive strategies have also been made, and researchers are trying to figure out how well they work against the constantly changing hostile attractions.

Because AI holds considerable promise in cyber security applications, it is critical for the community of academics and practitioners to grasp the current state of play and the difficulties that are associated with it. We began this study by reviewing past survey surveys that focused on the problems of protecting privacy and safety in AI systems. We noticed that many of the previously completed surveys did not include all machine learning task categories' attacks and countermeasures. Instead, the vast majority of these studies concentrated on deep neural networks in the context of computer vision, natural language processing, and cybersecurity.

We begin by establishing a theoretical foundation for machine learning task categories and then distinguish between shallow learning methods and the more recent deep learning methods as a foundation for characterizing adversarial attacks on machine learning models. This will serve as the framework for the remainder of our talk. Following that, we provide a novel methodology for conducting a thorough examination of adversarial attacks on AI systems. This framework starts with a description of what an adversary wants to do, what they know, and what they can do. It then goes on to a detailed look at adversarial attacks and defenses, which include a variety of machine learning models.

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

6. REFERENCES

Abduljabbar, R., Dia, H., Liyanage, S., and Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, *11*(1), 189.

Akhtar, N. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, *6*, 14410-14430.

Alshahrani, A., Dennehy, D., and Mäntymäki, M. (2022). An attention-based view of AI assimilation in public sector organizations: The case of Saudi Arabia. *Government Information Quarterly*, *39*(4), 101617.

Anh, T. T., Luong, N. C., Niyato, D., Kim, D. I., and Wang, L. C. (2019). Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach. *IEEE Wireless Communications Letters*, *8*(5), 1345-1348.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Ba, J., and Caruana, R. (2014). Do deep nets really need to be deep?. *Advances in Neural Information Processing Systems*, *27*, 1-9.

Bacchi, S., Oakden-Rayner, L., Zerner, T., Kleinig, T., Patel, S., and Jannes, J. (2019). Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. *Stroke*, *50*(3), 758-760.

Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., and Yoon, S. (2018). Security and privacy issues in deep learning. *arXiv preprint arXiv:1807.11655*.

Biggio, B., and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317-331.

Biggio, B., Fumera, G., and Roli, F. (2010). Multiple classifier systems for robust classifier design in adversarial environments. *International Journal of Machine Learning and Cybernetics*, *1*(1), 27-41.

Blum, A. L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1-2), 245-271.

Bout, E., Loscri, V., and Gallais, A. (2021). How machine learning changes the nature of cyberattacks on IoT networks: A survey. *IEEE Communications Surveys and Tutorials*, *24*(1), 248-279.

Boyd, M., and Wilson, N. (2017). Rapid developments in artificial intelligence: how might the New Zealand government respond?. *Policy Quarterly*, *13*(4), 1-37.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Caminero, G., Lopez-Martin, M., and Carro, B. (2019). Adversarial environment reinforcement learning algorithm for intrusion detection. *Computer Networks*, *159*, 96-109.

Carlini, N., and Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (sp)*, 39-57.

Chan, P. P., He, Z. M., Li, H., and Hsu, C. C. (2018). Data sanitization against adversarial label contamination based on data complexity. *International Journal of Machine Learning and Cybernetics*, *9*(6), 1039-1052.

David, S., Julian, S., Karen, S., Ioannis, A., Aja, H., Arthur, G., Thomas, H., Lucas, B., Matthew, L., Adrian, B., and Yutian, C. (2017). Lillicrap timothy p., hui fan, sifre laurent, van den driessche george, graepel thore, hassabis demis. *Mastering the Game of Go Without Human Knowledge, Nat*, *550*(7676), 354-359.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*, 1-29.

Dixit, P., and Silakari, S. (2021). Deep learning algorithms for cybersecurity applications: A technological and status review. *Computer Science Review*, *39*, 100317.

Dunjko, V., and Briegel, H. (2018). Machine learning and artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7), 074001.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Ebert, C., and Louridas, P. (2016). Machine learning. *IEEE Software*, *33*(5), 110-115.

Embley, D. W. (2004). Toward semantic understanding: An approach based on information extraction ontologies. *Proceedings of the 15th Australasian Database Conference, 27*, 3-12.

Guan, Y., and Ge, X. (2017). Distributed attack detection and secure estimation of networked cyber-physical systems against false data injection attacks and jamming attacks. *IEEE Transactions on Signal and Information Processing Over Networks*, *4*(1), 48-59.

Hancock, J. T., Naaman, M., and Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89-100.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *Arxiv Preprint arXiv:1503, 2*(7), 02531.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230-243.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Legg, S., and Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, *157*, 17.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., and Leung, V. C. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, *6*, 12103-12117.

Malik, H., Srivastava, S., Sood, Y. R., and Ahmad, A. (2018). Applications of artificial intelligence techniques in engineering. *Sigma*, *1*, 1-11.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., and Viljanen, M. (2022). Defining organizational AI governance. *2022*, AI *and Ethics*, 1-7.

Melis, M., Scalas, M., Demontis, A., Maiorca, D., Biggio, B., Giacinto, G., and Roli, F. (2022). Do gradient-based explanations tell anything about adversarial robustness to android malware?. *International Journal of Machine Learning and Cybernetics*, *13*(1), 217-232.

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

Muhammad, I., and Yan, Z. (2015). Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, *5*(3), 946-952.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. *Proceedings of The 2017 ACM on Asia Conference on Computer and Communications Security*, 3, 506-519.

Ren, K., Zheng, T., Qin, Z., and Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*, *6*(3), 346-360.

Risi, S., and Preuss, M. (2020). Behind deepmind's alphastar ai that reached grandmaster level in starcraft II. *KI-Künstliche Intelligenz*, *34*(1), 85-86.

Rosa, M., Feyereisl, J., and Collective, T. G. (2016). A framework for searching for general artificial intelligence. *Arxiv Preprint Arxiv:1611.00685*.

Song, H., and Montenegro-Marin, C. E. (2021). Secure prediction and assessment of sports injuries using deep learning based convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), 3399-3410.

Stuart, J. (2010). *Artificial Intelligence A Modern Approach Third Edition*. United States: Prentice Hall.

Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., and Sexton, J. T. (2019). A taxonomy and terminology of adversarial machine learning. *NIST IR*, *2019*, 1-29.

Thierer, A., Castillo, A. and Russell, R. (2017). Artificial intelligence and public policy. *SSRN Electronic Journal*, *2017*, 1-57.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2017). The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*.

Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1-16.

Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, *42*(7), 596-615.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-19.

Yuan, X., He, P., Zhu, Q., and Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805-2824.

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27.

Zhang, J., and Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(7), 2578-2593.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1-41.

Zhang, W., Zhou, T., Lu, Q., Wang, X., Zhu, C., Sun, H., Wang, Z., Lo, S.K. and Wang, F. Y. (2021). Dynamic-fusion-based federated learning for COVID-19 detection. *IEEE Internet of Things Journal*, *8*(21), 15884-15891.

Zhao, S., Blaabjerg, F., and Wang, H. (2020). An overview of artificial intelligence applications for power electronics. *IEEE Transactions on Power Electronics*, *36*(4), 4633-4658.

DOI: https://doi.org/10.17509/ijost.v8i1.52709
p- ISSN 2528-1410 e- ISSN 2527-8045