

Machine Learning Security Survey

Turhan Kimbrough
Department of Computer Science
Towson University
Towson, Maryland
tkimbr1@students.towson.edu

Abstract—Machine learning has experienced a significant growth in usage over the past few decades. Due to its data-centric approach in modeling, machine learning has seen use in a variety of subfields in computer science. In particular, researchers have been interested in incorporating machine learning into the domain of cybersecurity, utilizing it from the perspective of an adversary or ally. Researchers have also been concerned with the security state of current machine learning models. This survey paper provides a comprehensive overview of the state of machine learning, its application in various aspects of cybersecurity, securing machine learning systems, and future research directions being explored.

Keywords—Neural networks, security, classification

I. INTRODUCTION

With the increasing widespread adoption of machine learning technology, its usage is being observed in a variety of different fields. Some notable examples include image recognition, voice assistant technologies, email spam filters, and search engines. Much of its recent popularity can be attributed to the availability of frameworks such as tensorflow, allowing people of almost any background to quickly draft a machine learning application.

However, one growing concern tied to the ubiquity of machine learning is its accessibility to adversaries. Based on the assessment of current and prior research, there are a number of vulnerabilities in current machine learning models which can be exploited with little knowledge of a system's domain. In addition, attackers have been able to leverage machine learning technology to assist the deployment of cyberattacks. Enterprise machine learning applications may often contain large datasets of important information, becoming a potential candidate of a targeted attack. This paper will survey the domain of cybersecurity with regard to machine learning. This topic will explore the fundamentals of machine learning, current vulnerabilities in machine learning systems, machine

learning technology from the perspective of adversaries, and future research directions.

II. OVERVIEW

This section will introduce the fundamentals of machine learning. Additionally, there will be a discussion about the key terms used by members of the community.

A. Machine Learning Basics

Traditionally, when software developers are tasked with solving a problem, they use a combination of rules and logic to find a solution. The basic routine consists of finding appropriate input values, creating the logic and rules to process the input, and producing the appropriate output. The traditional approach to software development allows for fine-tuned control of program behavior to achieve the solution. However, this approach does not scale with the complexity of additional rules and/or possible solutions. An example is image classification, where the logic needed to compare images is complex. This becomes a bigger concern when new classifications need to be derived with new image data. Machine learning flips the traditional programming paradigm on its head, by taking a series of solutions as input, and letting the machine develop the rules by detecting patterns in the solutions. The result is a self-propagating mathematical model, capable of making decisions on newly supplied data. This approach relies on large sets of well-defined data, and has the flexibility for being used in many different applications.

Briefly mentioned earlier, a common use case for machine learning is *classification*, where a dataset is categorized into different groups based on one or more *features*. A *feature* is defined as some measurable property or characteristic being observed in a dataset. There are several types of classifications, including *binary classification*, *multi-class classification*, and *multi-label classification*. *Binary classification* categorizes data based on whether a feature is present or not, resulting in

an outcome of true or false. *Multi-class classification* categorizes data into different groups, where each data instance is assigned according to its feature. *Multi-label classification* categorizes data into different groups, where each data instance is assigned according to its expression of one or more features.

Training is the process of teaching a machine learning model to detect patterns in datasets. There are two types of training mechanisms, *supervised training* and *unsupervised training*. *Supervised training* requires each data instance to have one or more labels, defining which category or feature it expresses. In contrast, *unsupervised training* omits the need for labels, and the machine learning model will categorize datasets on its own. Typically, after the training phase, a machine learning model will go through the process of *validation*. *Validation* typically consists of classifying a separate dataset to guarantee the accuracy of a model and preventing a phenomenon called *over-fitting*, where a model will only 'memorize' characteristics or patterns of training data.

III. VULNERABILITIES IN MACHINE LEARNING MODELS

Since machine learning models are constructed based on the data, the largest attack surface would be the training data itself. For an attacker to manipulate a machine learning model, they would only need access to its input mechanism. This section will discuss several vulnerabilities which are currently present in machine learning.

A. Data poisoning

Data Poisoning is the act of manipulating, removing, or adding data during the training phase of machine learning. This type of attack is known as a *black box attack*, where an attacker does not need to know the implementation of a system to attack it.

One popular instance of a data poisoning attack is the adversarial example. This requires the attacker to have some knowledge of the training data, such as its dimensions and data type. The attacker would then manipulate data instances in a way where the machine learning model would be fooled, but appears normal to a human observer. The reason for manipulating data instances in this way is two-fold. First, machine learning models are often constrained to data fitting a specific dimension, shape, size, or length of characters. This is typically done to prevent incompatible data from entering the model during training. Second, there is often one or more people who are observing the model with testing or

validation datasets. An attacker would want to minimize any evidence of the data being tampered with.

Often, the goal of this attack is to make a machine learning model incorrectly classify data. The consequences of this attack can be devastating. Examples include tricking an autonomous vehicle to misinterpret traffic signs, sneaking malicious data past an intrusion detection system, or bypassing email spam filters.

B. Membership Inference

Membership inference is a mechanism of data extraction, where an adversary intends to know whether certain samples were used as training data for a machine learning model. This type of attack is also classified as a *black box attack*.

This particular vulnerability requires significant effort from the attacker. The attacker would need access to a dataset of sufficient size mimicking the data in the target model. The data would then be used to create several *shadow models*, which are used only to recognize differences in the target model's behavior. This is done to expose *overfitting*, when a model's analysis corresponds too closely to its training data. Therefore, an attacker can interpret whether certain samples were used in the training dataset based on the target model's confidence level in classifications.

Often, this can exploit the confidentiality of information on a system. An attacker has the capability to correlate information between datasets to target individuals for other cyberattacks.

C. Transfer Learning

Transfer learning is a mechanism where an adversary has the ability to study a publicly available machine learning model, and use that insight to sneak past and/or corrupt similar target systems. This type of attack is classified as a *white box attack*, since the attacker would need to have full access to at least one machine learning model.

This particular vulnerability requires the attacker to have knowledge of the input data, learning mechanism of the machine learning model, and classification behavior for its outputs.

D. TODO: talk about mitigations for each. Mention encryption, probability over discrete decision making, and auditing.

Please use "soft" (e.g., ϵ) cross references instead of "hard" references (e.g., (1)). That will make it possible to combine sections, add equations, or

change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in \LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

\BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use \BibTeX to produce a bibliography you must send the .bib files.

\LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

\LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0 , and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word *alternately* is preferred to the word

"alternately" (unless you really mean something that alternates).

- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals)

should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) *Positioning Figures and Tables:* Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table	Table Column Head		
Head	Table column subhead	Subhead	Subhead
copy	More table copy ^a		

^aSample of a Table footnote.



Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks. ...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

- [1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
- [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [4] K. Elissa, “Title of paper if known,” unpublished.
- [5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.*, in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [7] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.