STUDENT SESSION

ADVERSARIAL ATTACKS ON MACHINE LEARNING MODELS IN HEALTHCARE APPLICATIONS

Aleksandar Stanković,
[0009-0004-6711-0711]
Marina Marjanović*
[0000-0002-9928-6269]

Singidunum University,
Belgrade, Serbia

Abstract:

With the accelerated development of software tools, fast lifestyle, and expensive healthcare, many healthcare applications have been developed in recent years. With the help of artificial intelligence, apps can significantly improve people's lives, solving various dilemmas regarding their health. As medicine is based on statistics, AI can be remarkably useful and accurate for healthcare applications, but there are potential vulnerabilities found in them too. AI can be both an excellent co-worker for improving apps and a great tool for hackers to steal data or compromise the operation of the application. This paper will propose AI algorithms in healthcare applications and their potential vulnerability to adversarial attacks.

Keywords:

Artificial Intelligence, Machine learning, Adversarial attacks, Healthcare applications.

INTRODUCTION

Modern devices and smart and user-friendly software bring many benefits and make human individuals cannot imagine even a minute of their lives without them. Mobiles are used for many serious purposes to make lives easier and save money and time. Healthcare apps are developed for making diagnoses, monitoring health and fitness, and for doctors and patients to exchange experiences, find medicines, etc. In addition to being attractive to use, healthcare apps can surely be useful to improve users' comfort. There is a big concern about the security of using smart, online devices for healthcare, because as many are attractive to users, that much they are attractive to hackers. These apps collect users' sensitive health information, so, unfortunately, they can potentially compromise users' privacy and security. One of the possible security flaws could be the locking and theft of health and personal data that the Android application collects about its users.

Data analytics monitored from healthcare apps are used to improve statistics and to give various aspects. Artificial intelligence (AI) can be used for the analysis and prediction of diagnoses, giving a magnificent improvement in data collecting, selection, and analysis.

Correspondence:
Marina Marjanović

e-mail:
mmarjanovic@singidunum.ac.rs

DOI: 10.15308/Sinteza-2024-385-391

When it comes to specific and rare diseases but also for viruses if there are millions of people being affected, the power of AI will speed up the reaction and improve dealing with issues. [1]

As the power of AI is growing from day to day, there are a lot of concerns about using it to compromise mobile healthcare apps. Knowing the power of artificial intelligence and its tools, there are many ways to disrupt Android medical applications. This paper will review the possible weaknesses of applications, how applications can be compromised by different interests, and how AI can both improve or threaten them.

The main problem with AI technologies in medical healthcare apps is sensitive user data from clinical practice. These data contributed and daily updated online tests for the safety and efficacy of AI systems. There are no strong regulations and standards to assess the safety and efficacy of AI systems, which means these algorithms can be an easy target in an adversarial AI world. [2]

The fact is that one small perturbation in input data of a machine learning (ML) model that is visually imperceptible to human beings can fool the model into making a bad decision, which later can drastically affect the health condition of people who are looking for adequate therapy.

2. IMPACT OF ADVERSARIAL ATTACKS ON HEALTHCARE APPLICATIONS

Improvements and uses of healthcare applications are growing from day to day. Even certain types of skin cancers can be detected with just a mobile application, with a high degree of accuracy. [3] Some apps help users track their diet and exercise habits, and others manage to reduce stress by improving users' mental health through meditation, breathing, and mindfulness exercises. Polyclinics develops apps for clients to find and book appointments with healthcare providers, including doctors, dentists, and therapists. Some apps help users manage their medications by providing reminders to take pills, refill prescriptions, and track symptoms. It also allows users to share their medication schedule with caregivers or healthcare providers and provides medication-specific information and resources. Some of those apps relate to smart devices like phones, gadgets, and watches to measure activity, blood pressure, heart rate, etc.

AI-used apps provide diagnoses and personalized health advice generated from information received from users. Users answer to series of questions about their symptoms and medical history, and then the app generates a report with potential diagnoses and recommended next steps. It can alarm when some associated symptoms indicate a health problem. It can be concluded that most of these app users' healthcare information is easily available, and their security is a big concern.

2.1. PRIVACY AND SECURITY CONCERNS IN HEALTHCARE APPLICATIONS

The incredible growth of Android usage suppresses classic desktop software and prioritizes mobile Android apps. This further leads to security concerns and requires increased malware prevention and detection mechanisms. Application-based architecture provides user comfort in the form of flexibility, interoperability, and adaptability, but it also enables and attracts malware development. [4]

All those apps have the same security concerns, users share a lot of confidential information, agree with privacy policy for data collection, pictures available, etc and there are many issues with privacy statements in that. Many studies have been conducted on the privacy and security of mobile healthcare apps. The study [5] analyzed the privacy and security of 79 mobile healthcare apps and found that most of them had security issues related to data storage and transmission.

Some health conditions such as HIV, and psychiatric illness are very compromising so individuals will not share they have it with anyone because of their social or professional status. These are sensitive information about patients and may cause discrimination in society. Securing the privacy of users' health information is a high priority in such cases.

Potential threats in any healthcare system can be unintentional, accidental disclosure of patient's data, or intentional, by curiosity, for revenge, data breach by internal or external agents of the organization, or hacking of healthcare organization systems. For mobile android healthcare apps there are similar ways of threats, mostly by programmers who made the app, and by hackers who will use it. The first mistake can be made in-app architecture or adding data. Other, most common problems are stealing data about users or endangering results by mixing symptoms and giving false diagnoses.

We already mentioned how significant benefits AI development and implementation in healthcare offers to patients. Commercial healthcare AI faces privacy challenges and concerns about access, control, and use of personal medical data by for-profit entities and self-improving AI. Rapidly technology grooving is threatening to rise above all the regulations they govern. [6]

2.2. ARTIFICIAL INTELLIGENCE IN HEALTHCARE

Artificial intelligence as software has the role to simulate human cognitive functions, decrease mistakes, and combine methods and results by using substantial amounts of data. [7] Involving AI brings a lot of benefits to medicine in general. AI algorithms are designed to reduce errors and provide accurate results. [8] AI can improve clinical decision-making and potentially enhance or replace the human subject views in specific areas of healthcare, such as radiology, gynecology, etc. The successful application of AI in healthcare is fuelled by the availability of healthcare data and advancements in big data analytics. AI techniques can extract clinically relevant information from vast datasets, aiding in clinical decision-making. [9] Application of AI in healthcare often is about using it with devices that work with healthcare software.

The advantage of using AI through technologies that process copious amounts of data very quickly in healthcare applications is lower treatment costs, improved and more detailed patient care, more accurate diagnoses, and automation of routine tasks. The benefit of using AI in healthcare is great, although the security risks are also big.

As these areas are developing at a high speed, there is a shortage of trained workforce. It is necessary to identify and define clear goals that the applications satisfy, the tools that will be used and train developers to improve the healthcare industry's efficiency, accuracy, and profitability. Figure 1 shows the potential applications of AI-based technologies in healthcare. Artificial intelligence can play a major role in healthcare, including diagnostics, therapy, population health management, administration, and regulation.

2.3. APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE APPLICATIONS

The functioning of AI technologies in healthcare applications is reflected in the analysis of data from clinical activities based on statistical data previously trained by the algorithm. The software is trained to recognize, compare, and analyze symptoms, recordings, and laboratory results, based on statistics, determine the diagnosis, and assign further treatment. This makes AI applications as effective as doctors, even in some situations they can be more accurate and objective. The problem of AI functionality in different areas of healthcare is certainly the form of clinical data that is not unique and can be in the form of medical notes, electronic recordings, physical examinations, and images. Certainly, the best appli-

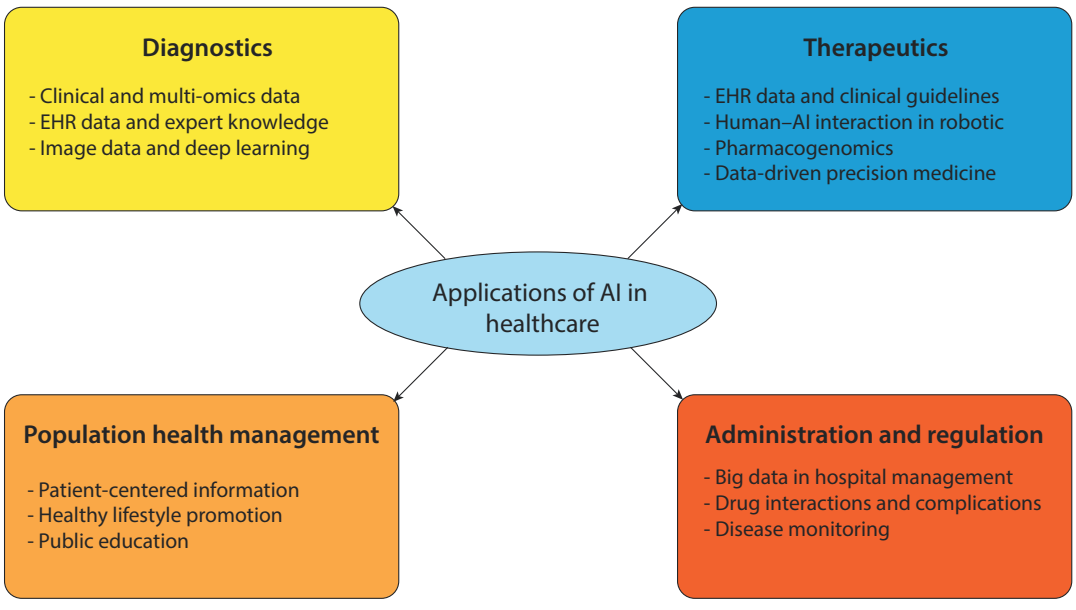

Figure 1. Applications of AI in healthcare.

cation of AI is in radiology where AI can be very well trained to analyse data from diagnostic images. Doctors in all fields of medicine are tasked with perfecting data storage in digital form and adapting it to AI analysis. Most of the available material is in unstructured narrative texts, and some AI applications focus on converting such text into machine-intelligible text. [9]

AI and machine learning applications can leverage real-world clinical data to identify effective treatments and support evidence-based decisions in clinical settings. The availability of high-quality training data is crucial for developing clinically useful ML models. Training data sets enable the training of prediction models, extraction of relevant features, and identification of meaningful associations using machine learning algorithms. Ensuring a foundation of high-quality training data is essential for building robust machine learning models. [10] The most promising areas probably are AI algorithms for automated image analysis which quickly analyze large amounts of medical imaging data such as CT, MRI, and X-ray scans. AI algorithms are more accurate in diagnosing illnesses and diseases.

2.4. ANALYSIS OF COMMON ATTACK VECTORS USED BY HACKERS TO EXPLOIT VULNERABILITIES IN AI ALGORITHMS USED IN HEALTHCARE ANDROID APPLICATION

Deep learning-powered mobile apps can be roughly categorized into two ways: cloud-based inference and on-device inference. These two architectures have differences in the storage location of the deep learning model. Cloud-based deep learning models need mobile devices to send requests to a cloud server and retrieve the inference results. Offloading the execution of inference tasks to the cloud has several disadvantages, including concerns about data privacy, unreliable network conditions, and high latency. On the other hand, on-device deep learning models avoid these drawbacks associated with cloud-based approaches. These models can perform real-time inference directly on smartphones, even without an active network connection, and they typically minimize the need to send user data off the device.

The implementation of on-device deep learning models often relies on deep learning frameworks such as Google TensorFlow (TF) and TFLite, Facebook PyTorch and PyTorch Mobile, and Apple Core ML. Among these frameworks, TFLite is the most popular technology used for running deep learning models on mobile, embedded, and IoT devices.

TFLite has contributed to nearly half of the deep learning mobile apps developed in the past two years, and its usage is growing more significantly compared to other frameworks.

While these existing deep learning frameworks have reduced the engineering efforts required for implementing on-device deep learning models, training a new model from scratch can still be expensive. Therefore, deep learning mobile apps often make use of pre-trained models and transfer learning techniques to mitigate training costs. This approach allows mobile developers to leverage the representations learned by a pre-trained network and fine-tune them for a specific task. [11]

Based on attackers' knowledge, capabilities, characteristics, and goals, we use 2 main types of adversarial attack methods, White and Black box Attacks. White-box refers to cases where the attacker has full knowledge of the targeted model, including its structure and parameters. The attacker directly computes perturbations that can change the model's predictions using gradient ascent. By knowing how and what exactly attacks, in those cases attackers often achieve a high success rate, close to 100%, with minimal perturbations. Black box attacks are less successful but more realistic, they often have the additional constraint of a limited query budget. Due to the lack of information, black box attackers use different attack methods. Some attacks focus on analyzing the output of the models they attack and then construct a substitute model that can generate adversarial images. Black-box attacks also can involve executing queries on the targeted model to evaluate the success of adversarial images and iteratively improve them. Based on the ML's lifecycle, there are 5 categories of AML attacks used to attack models in different software, in our case in Android healthcare applications. Those categories are Poisoning attacks, Backdoor attacks, Evasion attacks, Model stealing attacks, and Data extraction attacks. AML attacks aim to exploit the weaknesses of ML-based systems and mislead ML models through adversarial input perturbations.

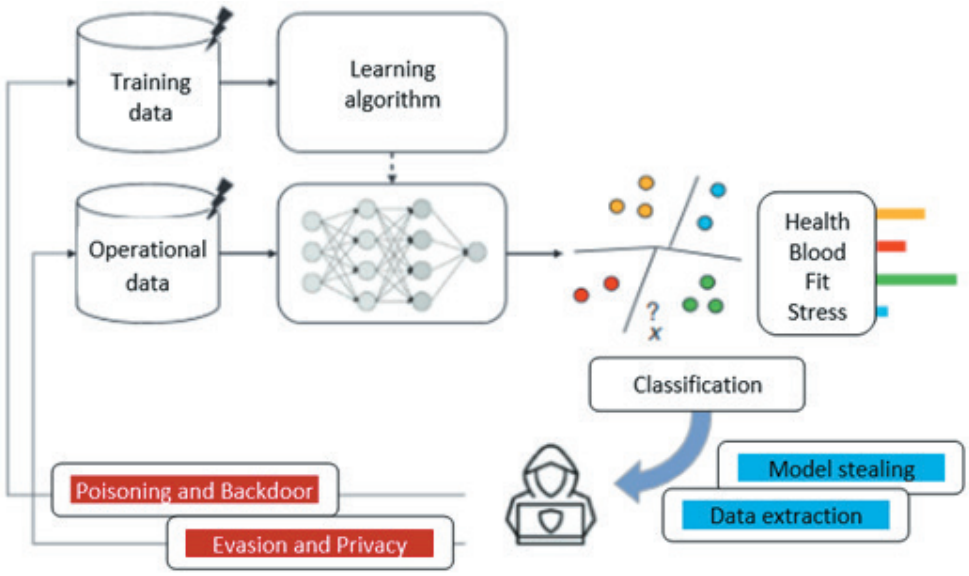

Figure 2. Major adversarial attack examples.

Poisoning attacks manipulate the ML service's training data to degrade their performance with 2 main objectives: causing a denial of service to lower the system's overall performance, or to enable specific misclassifications during operation, such as targeting a particular user or set of samples. Specific for those attacks is that it occurs during the training phase, by injecting poisoned data samples into the training set used to train or update the deployed ML model. Since the quality and representativeness of the training data, are the main for succesness of data-driven ML systems, they are vulnerable to poisoning attacks. This kind of attack works in cases when adversaries, referred to as "threat actors," gain access to the training dataset and manipulate the data to degrade the performance of the ML system.

Backdoor attacks also attack ML systems in the training phase, but in different ways and for different results. During the training phase, special patterns are embedded in the model by poisoning the training data. Patterns are designed to serve as triggers for the attack during the operational phase. This is done by providing an input to the targeted model that contains the trigger pattern. When the model encounters this input, it recognizes the trigger and produces a maliciously predefined output, which is determined by the attacker. This results in the model being compromised to behave in a specific manner when triggered by the predefined pattern. The targets for those kinds of attacks are deep neural networks by involve the attacker in providing poisoned data to the victim for training the model. This poisoned data contains embedded trigger patterns.

Then, during the testing or operational phase, the attacker presents inputs that include the trigger pattern to activate the attack.

In evasion attacks, the attackers manipulate the AI system by providing it with contradictory or misleading examples, adversarial examples, and specially crafted inputs, to deceive the targeted machine learning (ML) model into producing incorrect predictions. These attacks aim to exploit vulnerabilities in the model's decision-making process. Adversarial examples are inputs that have been intentionally perturbed or modified in a way that causes the ML model to make incorrect predictions. Adversarial examples can include subtle alterations such as adding or modifying pixels in an image. These changes are carefully crafted to exploit the vulnerabilities or blind spots of the ML model, causing it to generate inaccurate predictions. Even if it seems that those perturbations are small, they can have a significant impact on the performance and reliability of ML models. Figure 2 shows evasion and privacy attacks applied during operation and manipulating operational data to either evade detection or obtain confidential information about the ML model or its users.

Model-stealing attacks involve querying the targeted model to create an approximation of the original model. In unauthorized access to the targeted model during the operational phase, the attackers query the model to gather information and generate an approximation of the original model. Attackers may have different goals with those attacks, according to ML system vulnerabilities.

The goal can be to obtain the model's parameters, avoid query charges, violate the privacy of training data, or prepare for other types of attacks such as evasion. Specific for this type of attack is that it primarily targets AI models, especially those provided as machine learning-as-a-service. These attacks pose significant risks, as unauthorized access to AI models can lead to various consequences, such as intellectual property theft, loss of revenue for service providers, and compromised privacy of training data. Defending against model stealing attacks requires implementing robust security measures to protect the models and prevent unauthorized access which is crucial to ensure the security and integrity of AI models and the privacy of sensitive information.

Data extraction attacks aim to retrieve or modify data in the training phase of a particular ML model during the operational phase. In these attacks the training data can be inverted, reconstructed from the model, or determine whether a particular data point belongs to the training data. This can be very concerning depending on the type of model involved and the sensitivity of the data, for example, processing biometric information and medical records. By extracting or identifying training data from a model, attackers gain access to sensitive information that should remain confidential. This can lead to a serious breach of data privacy, as the privacy and security of biometric authentication information and medical records are compromised. Protecting against data extraction attacks requires robust security measures, including encryption, access controls, and anonymization techniques, to safeguard sensitive training data and ensure data privacy. [12]

3. CONCLUSION

In this paper, we proposed a vulnerability analysis for Android healthcare apps that use machine learning techniques. Our work highlights the importance of ensuring the security of mobile healthcare apps and provides a framework for developers to identify and mitigate potential vulnerabilities in their apps. The use of AI in healthcare applications offers several benefits, including personalized care, improved diagnostic accuracy, and remote monitoring capabilities. However, it is important to address the security risks and privacy concerns associated with these apps to ensure that they are safe, effective, and compliant with regulatory standards. The proposed framework and vulnerability analysis will provide a comprehensive approach to address these concerns and ensure that AI-based healthcare apps are safe and trusted environments for patients and healthcare professionals.

AI is playing an increasingly important role in people's daily lives and has become a key driver of digital transformation due to its automated decision-making capabilities. While the benefits of this technology are significant, some concerns need to be addressed. One crucial aspect to consider is the role of cybersecurity in ensuring the reliable and trustworthy deployment of AI. When it comes to security in the context of AI, it's important to recognize that AI techniques and systems can lead to unexpected outcomes and be susceptible to tampering, resulting in manipulated expected results. This is especially true for AI software that often relies on fully black-box models or can be exploited by malicious actors for cybercrime and facilitating attacks. Therefore, securing AI itself becomes essential. We want to research adversarial attacks in healthcare in the future using complex datasets from medicine. This involves understanding the assets that are vulnerable to AI-specific threats and adversarial models, establishing data governance models to design, evaluate, and protect data used for training AI systems, managing threats within a multi-party ecosystem through shared models and taxonomies, and developing specific controls to ensure the security of AI.

4. ACKNOWLEDGEMENTS

This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms - ECOSwarm.

5. REFERENCES

[1] S. Hiriyannaiah et al., "Data Science Tools and Techniques for Healthcare Applications," *Studies in Big Data*, Springer, vol. 88, p. 213–233, 2021.

[2] F. Jiang et al., "Artificial Intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, 2017.

[3] S. T. Tong and P. Sopory, "Does integral affect influence intentions to use artificial intelligence for skin cancer screening? A test of the affect heuristic," *Psychology & Health*, Vols. 34,7, pp. 1-22, 2019.

[4] S. Schmeelk, J. Yang and A. Aho, "Android Malware Static Analysis Techniques," in *Association for Computing Machinery*, New York, 2015.

[5] J. Zhang, "Security Analysis of Android mHealth Apps," *Journal of Medical Systems*, Vols. 39, no. 9, pp. 1-10, 2015.

[6] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Med Ethics 22*, vol. 122, 2021.

[7] P. Kumar, S. Chauhan and L. Awasthi, "Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions," *Engineering Applications of Artificial Intelligence*, vol. 120, 2023.

[8] T. Arunachalam, "Introduction to Artificial Intelligence," 2021.

[9] A. Bohr and M. Kaveh, "The rise of artificial intelligence in healthcare applications," *National Centre for Biotechnology Information*, pp. 25-60, 2020.

[10] Booz Allen Hamilton, "Training Data for Machine Learning (ML) to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure," 2021.

[11] A. Newaz, N. Haque, A. Sikder, M. Rahman and S. Uluagac, "Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, Taipei, Taiwan, 2020.

[12] N. Papernot, P. McDaniel, A. Sinha and M. P. Wellman, "Security and privacy in machine learning," in *IEEE Symposium on Security and Privacy (SP)*, 2018.