Vulnerabilities and Mitigation Strategies in AI-Powered Diagnostic Healthcare Systems

Hamza Ali Khan (001345840), Mohammad Ayaan Khan (001348784),
Shaik Shoeb Hussain (001306492), Shaik Umar Basha (001261210) and Hidayatullah Gul (001357436)

I. INTRODUCTION

In recent years, healthcare systems have increasingly fallen victim to cyber-attacks as the industry integrates advanced digital technologies and machine learning (ML) tools to enhance patient care. Cyber-attacks are deliberate actions by malicious entities aimed at disrupting, disabling, or unlawfully accessing systems and data. These attacks have evolved from general breaches to more complex efforts targeting networked medical devices, electronic health records, and ML-driven applications, creating distinctive vulnerabilities with severe implications. For instance, in 2020, Universal Health Services (UHS) faced a ransomware attack that compromised patient data, disrupted ML-based diagnostic tools, and led to $67 million in lost revenue and recovery costs (Healthcare IT News, 2020). As healthcare providers rely more on ML for diagnostics, predictive analytics, and decision-making, the risk of exploitation by attackers has significantly increased.

ML in healthcare accelerates diagnostics, patient monitoring, and personalized treatment, but it also introduces vulnerabilities due to the sensitive nature of patient data used in training these models. Such vulnerabilities make ML systems attractive targets for attacks that exploit unique flaws within the models, posing significant risks to patient privacy and safety. Securing ML-based healthcare systems is crucial for protecting patient data and ensuring adherence to regulatory standards. Cyber-attacks leveraging ML tend to be insidious, extracting sensitive information gradually, which poses a challenge for detection and defence. Thus, safeguarding these systems is essential for maintaining patient trust and operational integrity.

Several noteworthy cases further highlight the practical significance of cybersecurity in AI/ML-powered healthcare systems. A UK-based energy company suffered a large financial loss in 2019 due to a sophisticated voice cloning assault that employed AI to impersonate a CEO's voice, highlighting the evolving nature of cyber threats (Smith, 2019). Healthcare facilities that use AI & ML for identity verification may be at risk from this attack technique. Furthermore, researchers in 2020 showed how subtle manipulation of medical photos could be used to control AI diagnostic technologies, potentially leading to incorrect diagnosis for serious illnesses like cancer (Jones and Brown, 2020). In 2018, a significant DDoS attack disrupted patient care by overloading healthcare facilities, including a hospital in Boston, and rendering real-time AI and ML systems inoperable (Cybersecurity Reports, 2018).

This report aims to tackle these challenges by investigating the vulnerabilities associated with ML-based healthcare systems along with potential defensive strategies. It will commence with a case study set within a hospital environment where ML is deployed for patient monitoring and diagnostics purposes. Subsequently, an attack graph model will be constructed to illustrate possible vulnerability points within this setting. The report will delve into specific techniques employed in cyber-attacks targeting ML within healthcare settings including adversarial examples and data poisoning methods—and finally transition into discussing defensive strategies that organizations can implement to safeguard their ML assets against cyber threats. By integrating a systematic approach toward understanding both risks and protective measures, this report endeavours to offer valuable insights into securing machine learning frameworks amid an ever-evolving landscape of cybersecurity challenges facing the healthcare sector (Smith, 2021)

II. Use Case Scenario and Background: Diagnostic Imaging in Healthcare

MediCare Health Services, a privately owned network of hospitals operating across the UK, striving in an extremely competitive environment. They have recently incorporated state-of-the-art machine learning model called MedTech into all of its diagnostic imaging departments which includes radiology department for diagnosing X-ray, Computed Tomography Department for diagnosing CT Scan, Magnetic Resonance Imaging Department for diagnosing MRI scans and Ultrasound Department.This model is designed especially to help and enhance the capabilities of the hospital's imaging departments by analysing images to identify potential abnormalities, areas of concern and provides preliminary findings with great precision, speed, and effectiveness which is then sent to the imaging department where the output findings are reviewed and combined with experts analysis, finalising the diagnosed report for the patient.

System Overview:

MedTech uses Convolutional Neural Network, a type of deep machine learning model which analysis images and recognise patterns in data making it optimal for medical image diagnosing using ReLU activation function which is known for its simplicity and effectiveness. It has been trained on enormous datasets that contain thousands of labelled images including wide range of normal and abnormal pathological and anatomical conditions also incorporating features involving patient's demographics which are age, gender, ethnicity and medical history. MedTech’s artificial intelligence technology can identify minute patterns and irregularities that the human eye could miss. From tiny fractures to early-stage tumours, this capacity is essential for the early detection of a variety of illnesses, guaranteeing that patients receive the most timely and precise diagnoses possible.

Investment and Resources:

AI's integration into MediCare Hospital's imaging procedures has cost the organisation over €100,000 for initial setup which includes data collection, hardware, software and basic developmental costs. However long-term costs including operational costs and licencing will significantly increase the total investment. The model assets include Normalised 2D input images, trained CNN models architecture with its parameters, weights and inference system, training data, CNN’s classified output and its transmission medium into the imaging department.

Challenges within the model:

According to (Goodfellow et al., 2014), It states that deep neural networks are inherently vulnerable to adversarial manipulation because of the following reasons: -

1.) **High-dimensionality space** - It leads to increase in directions in the space with each direction corresponding to a feature, providing a broader landscape for perturbations without any noticeable difference (Madry et al., 2017). Also, tiny changes in each direction could add up cumulatively impacting the output significantly.

2.) **Local Linearity** - CNN is designed to learn complex non-linear relationships between input and output. But the activation function used here is ReLU (Rectified Linear Unit) which behaves linearly in positive regions, which means when you zoom into a smaller positive region within a certain range the relationship between input and output can be approximated as a linear function. This results in the output to be directly proportional to the input, leading to large sensitivity to small perturbations in the input within these regions (Goodfellow et al., 2014).

This report will be further assessing the vulnerabilities within the developed AI model MedTech by exploring potential adversarial attacks and propose defence technique’s which could be employed to mitigate the risks involved.

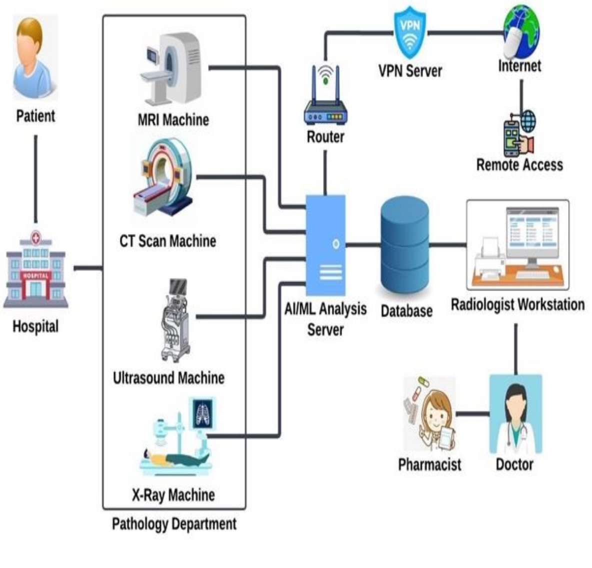
Network Infrastructure and Security:

MediTech employs robust network infrastructure and security measures to safeguard medical systems and patient data. A point-to-point VPN is utilized for secure internal communications, allowing staff to access the hospital network remotely via a VPN server, ensuring secure connections from outside the hospital. The hospital also maintains an external-facing website, which includes a contact page listing the details of hospital employees—names, roles, email addresses, and telephone numbers. This feature, while essential, could potentially serve as a vector for social engineering attacks. Additionally, the hospital’s Wi-Fi network, which includes a captive portal for staff authentication, extends beyond the premises to local areas such as the high street and staff homes, raising the risk of unauthorized access. To address the vulnerabilities associated with personal device usage, employees are permitted to use their own laptops or PCs to access the hospital’s website and internal systems, under stringent security protocols designed to prevent data breaches or malware infections. These comprehensive security measures are critical in protecting the hospital’s digital assets and maintaining the integrity of its healthcare services.

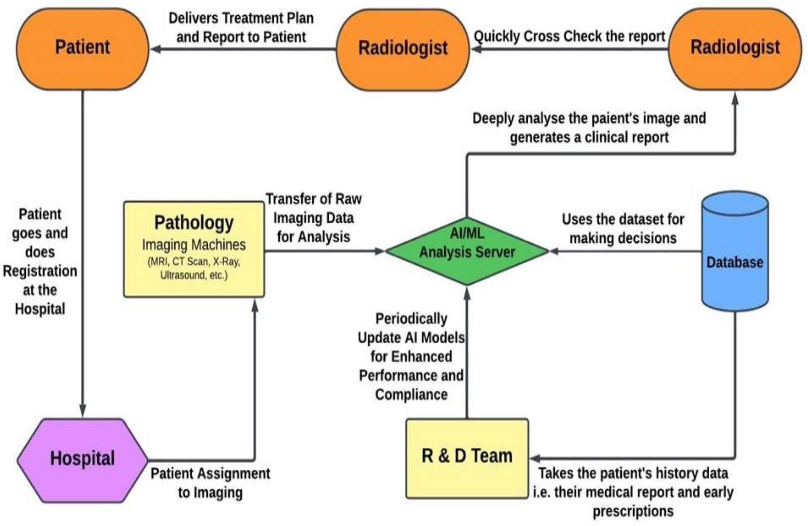
Challenges faced in implementing AI in diagnostic imaging:

Implementing AI in diagnostic imaging presents a multitude of challenges. Data security and privacy are paramount, as safeguarding patient data from breaches is crucial, especially given the sensitive nature of medical images. Additionally, the integration of AI technologies poses complexities as it must be seamlessly done without disrupting existing hospital infrastructure or other medical services. Ensuring the accuracy and reliability of AI systems is also vital, as these systems must provide diagnostics that medical professionals can depend on. Moreover, stringent regulatory compliance is required; MedTech systems must adhere to the General Data Protection Regulation (GDPR), and the Data Protection Act of 2018 to protect sensitive patient information. They must also be registered with the Medicines and Healthcare products Regulatory Agency (MHRA)and comply with the Medical Device Regulations 2002, while abiding by the NHS Digital Technologies Assessment Criteria (DTAC) framework, which ensures that the MedTech AI system is secure, lawful, and effective (Barlow, 2012).

Network Diagram



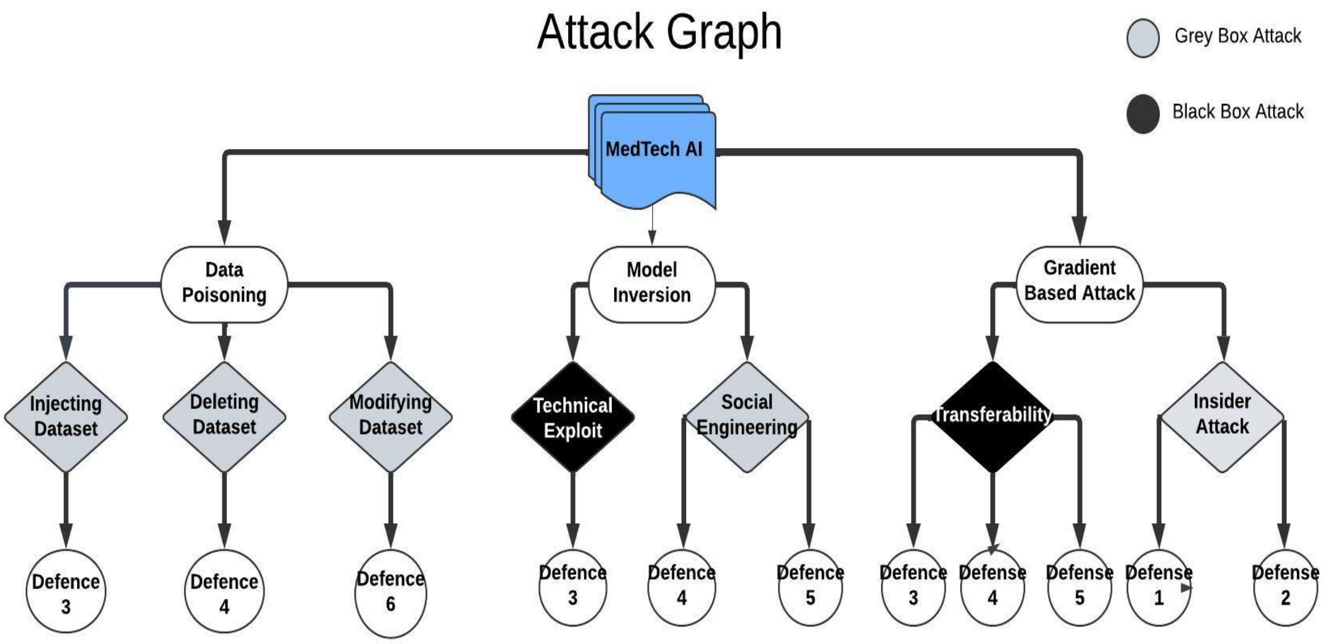
Data Flow Diagram


Patient
Delivers Treatment Plan and Report to Patient
Radiologist
Quickly Cross Check the report
Radiologist
Deeply analyse the paient's image and generates a clinical report
Patient goes and does Registration at the Hospital
Pathology
Imaging Machines
(MRI, CT Scan, X-Ray, Ultrasound, etc.)
Transfer of Raw Imaging Data for Analysis
AI/ML Analysis Server
Uses the dataset for making decisions
Database
Periodically Update AI Models for Enhanced Performance and Compliance
Hospital
Patient Assignment to Imaging
R & D Team
Takes the patient's history data i.e. their medical report and early prescriptions

III. Analysis and Discussion

Asset Identification & Valuation:

Asset Category	Specific Asset	Description	Sensitivity	Justification for Valuation
Hardware	Imaging Machines (MRI, CT, X-Ray, Ultrasound)	Critical for capturing patient imaging data.	High	Essential for diagnostic processes; high replacement cost.
Software	AI/ML Analysis Server	Processes imaging data to assist in diagnosis.	Very High	Core to operational capabilities; involves proprietary algorithms.
Data	Patient Medical Records	Includes personal health information (PHI).	Very High	Subject to regulatory compliance (e.g., GDPR, HIPAA).
Network	Hospital Internal Network	Connects all IT systems and transmits sensitive data.	High	Breach could lead to significant data loss or service disruption.
People	Radiologists, Physicians, R&D Staff, Medical Staff	Key users of the AI/ML system for patient diagnosis.	Moderate	Training and expertise are valuable; impact of social engineering.
R&D Information	Model Training Datasets	Used to train and refine AI algorithms.	Moderate to High	Intellectual property that could be targeted for theft or corruption.
Physical	Data Centers/Servers	Houses the AI/ML servers and data storage.	High	Physical security is critical to prevent unauthorized access.



This graph represents attacking methods subsequently followed by different potential techniques by which the attacker could conduct attacks based on our scenario. Further representing various defence measures respectively in accordance with Mitre Atlas.

Different ways of attacking and their defensive measures: -

1.Data Poisoning Attack

In this type of cyberattack the adversary intentionally compromises the training dataset used by MedTech AI model to influence or manipulate the operation of the model.

The attacker leverages the two most important assets in our Medicare Hospital, stored data, and transmission mediums by exploiting their vulnerabilities which could be mainly unpatched software, insecure network configurations and weak authentication mechanisms. The attacker can conduct this type of cyberattack successfully through the following ways: -

A) Intentionally injecting false or misleading information within the training dataset: -This could be done when the data transmitted is unencrypted, the attacker might deploy Man-in-the-Middle (MitM) attack. In MedTech while training when the collected data is moved to the storing database, which is susceptible to be intercepted and tampered with. The attacker could inject incorrect information within any of the 2D images or patients feature including age, gender, ethnicity, and medical history from which MedTech model trains affecting the performance.

B) <u>Modifying the existing dataset:</u> - This could be done by the attacker by exploiting weak or reused passwords to access data repositories directly. Once inside, they might modify the data by introducing anomalies or incorrect labels, effectively corrupting the training dataset leading to biased results.

C) <u>Deleting a portion of the dataset</u>: - This can be done when the software used for the data storage system has not been updated to patch known vulnerabilities, attackers could inject malicious scripts into the database, introducing vulnerabilities which serves as a backdoor (access point) for the attacker which can be used to delete a portion of the dataset leading to inaccuracy with the output.

The compromised data is then seamlessly integrated into the hospital's regular data flow. Unaware of the compromise, the AI system uses this corrupted data for ongoing training. This results in a model that has integrated these malicious influences, leading to degraded performance. Outcome is a compromised AI model that no longer accurately reflects medical realities but instead mimics the distorted inputs introduced by the attacker. This could lead to misdiagnoses or inappropriate medical treatment recommendations, posing significant risks to patient safety and undermining the hospital's credibility.

<u>2.Model Inversion Attack</u>

It is a type of attack where the attacker has access to the Machine learning model and crafts specific queries and analyses the response to identify patterns and anomalies in the model's outputs, allowing the attacker to infer details about the underlying training data.

The attacker can successfully conduct this attack in MedTech AI model firstly through accessing the model for querying through the following techniques: -

A) <u>Social Engineering-Based Credential Theft:</u> - The Head radiologists, Head It administrator or machine learning engineer, the staff with the login id and passwords to access the MedTech AI model tricking them into sharing credentials through deceptive emails, messages, fake log in portals or malware installed into their devices which record keystrokes to capture credentials.

B) <u>Technical Exploits for Credential Theft:</u> - Through network interception or packet sniffing by intercepting communication between the radiologist's device and the AI model to access credentials or using Brute force techniques by trying several different passwords to access the model.

Then by using advanced data analysis techniques and continuous querying, the attacker reconstructs sensitive information that was ostensibly secure within the training dataset. This might include sensitive information such as private medical records, age, gender, and ethnicity information that the CNN model learned during its training phase. This Leads to significant privacy breaches violating the GDPR and many other regulations and damage to the trustworthiness and integrity of MediCare Health Services.

<u>3. Gradient Based Attack</u>

It is a type of evasion attack in which firstly, the attacker accesses the model infrastructure to compute the gradient (rate of change) of the model's loss function with respect to the input, which measures how far the model's predictions are from the true labels and from which generally the model tries to minimize the loss function. But here the attacker uses this information to oppose it to slightly modify the input to increase the loss. Through this information the attacker crafts minute perturbations in the input which are imperceptible to the human eye but can cause the most damage to the models output accuracy. Also, according to (Goodfellow et al., 2014) deep learning models like CNN are highly susceptible to attacks based on small modifications in the input mainly due to its linear behaviour. According to our scenario in the MedTech AI model where the input data is exclusively accessed by the radiologist within the specific department, and the model is not externally accessible to patients or the public for direct queries. Considering it, following are the possible ways by which the attacker can successfully conduct the attack: -

A)<u>Insider Attack (Grey Box)</u> : - In this the staff within the MediCare hospitals which have the access to the model which includes head radiologists of the three departments, head AI engineer and the IT administrator, the attacker through or because of them uses the model gradients to generate adversarial images directly. It is categorised as grey box as the attacker has limited access to the system particularly its inputs and outputs.

B)<u>Transferability (Black Box)</u>: - As MediCare health network is in highly competitive environment with many other private health services, the attacker can use any of their publicly accessible models which might function similarly with similar architecture or datasets to craft adversarial examples through it and insert the same input in our model by exploiting the vulnerabilities within our transmission mediums into the model to test if it can deceive our model. It is categorised as black box as the attacker doesn't have any direct access to the main model.

The attacker can possibly use Fast Gradient Sign Method (FGSM) particularly because of its simplicity and efficiency. With just a single step, it can generate adversarial examples that can deceive our model (Ivezic, 2023).

DEFENSE MEASURES

Following are the defence measures in accordance with the Mitre Atlas along with the CIS Controls (Version 8): -

Defence 1:- Model Hardening

In this defence technique we have incorporated following sub-techniques to make our CNN model robust to adversarial inputs: -
A) **Adversarial Training:** In this technique the model is trained on a mix of clean and adversarial examples. According to (Goodfellow et al., 2014) using this significantly improved the error rate on FGSM generated adversarial examples from 89.4% to 17.9% after training. However, this would increase the computational cost in our model considering large volume of medical dataset, also impacting the model's accuracy leading to misdiagnosis and causing biases potentially causing legal liability under GDPR regulations.
B) **Gradient Masking:** This strategy hides the gradient information by altering the loss surface or introducing randomness in predictions. This way, the attacker cannot easily find the optimal direction to perturb the input reducing the effectiveness of the gradient based attacks making it very effective against the proposed white box attack (Derui Wang et al., 2022). However, noise and distortion can degrade the model's performance and can be bypassed by black-box attack. Moreover, adaptive attacker can detect and exploit gradient masking.
C) **Switching to Swish activation function from ReLU:**-Swish is a non-linear and smooth activation function unlike ReLU activation which is used in our CNN model. Due to these properties, it can generalize better in image classification task and outperforms ReLU, making it less prone to adversarial attacks. It can simply be implemented with a single line code (Prajit Ramachandran et al., 2017).

Defence 2:- Adversarial Input Detection

This defence technique is implemented with (CIS Control 8): Audit Log Management

This defence technique uses ML model to detect and block any possible adversarial inputs, queries which are separate from expected behaviours. Also, the control measure which it is implemented with Collect, alert, review, and retain audit logs of events that could help detect, understand, or recover from an attack. This defence would help mitigate the black box attack method mentioned using transferability in the MedTech AI system.

Defence 3:- Verify ML Artifacts

In this technique all the artifacts in our MedTech AI system including Input data, model weights and architectures, training data, transmission medium and the output are being ensured that the integrity has not been altered through verifying cryptographic checksum. This protects the MedTech model from any adversarial inputs through FGSM also the encryption within the transmission medium ensures stored data integrity by protecting against Data poisoning through MiTM mentioned above.

Defence 4:- Control Access to ML Models and Data in Production

This defence measure is implemented with (CIS Control 6): Access Control Management

This defence measure would compartmentalize MRI, CT scan X-ray imaging department such that the access of the model's output and the stored training dataset is limited to the certain specific imaging department according to its the needs restricting the access of the other two if the output is related to a particular department. Also, it would multi-factor authenticate the system where only the head radiologists with specific credentials can submit the CT, MRI or X-ray images. Cross-departmental access is blocked unless explicitly authorized, and unauthorized attempts are flagged for review. Additionally, the system would monitor and log each interaction to detect any unauthorized access. This protects from data poisoning through accessing stored training data by weak or reused passwords.

Defence 5:- Limit Model Artifact Release

This defence measure is implemented with (CIS Control 14): Security Awareness and Skills Training

This defence measure would limit the public release of the project technical details including the model architecture, data, algorithm as this information can be used by the attacker to conduct the black box transferability attack mentioned. Along with this security awareness modules are to be taught to increase awareness among the hospital staff to be security conscious as this would lower the possibilities of insider attack.

Defence 6: - Vulnerability Scanning

In this measure software vulnerabilities are constantly being scanned to identify any possible vulnerabilities within them to further remediate them. This measure would help protect our training data from data poisoning through backdoor technique mentioned above.

IV. Conclusion

In this report we have explored the pressing cybersecurity challenges faced by healthcare systems that incorporate neural network machine learning technologies. Through a detailed case study of MediCare Health Services, we have identified and analysed various cyber threats that exploit vulnerabilities in ML-based diagnostic systems. The development and discussion of an attack graph have provided insights into potential security breaches and the implications of such attacks on patient care and data privacy.

Our investigation has emphasized the need for robust cybersecurity measures to shield healthcare infrastructures from sophisticated cyber threats. Implementing advanced encryption, regular patch management, strict access controls, and continuous vulnerability assessments are critical steps towards fortifying the security of ML systems. Looking ahead, there is a pressing need to develop more advanced AI-driven security protocols that can predict and neutralize threats before they manifest. Further research focused on enhancing the interpretability of ML models to detect and mitigate any attempts at manipulation or attack more effectively. As cybersecurity threats evolve, so too must our strategies to protect critical healthcare infrastructure and ensure the safety and privacy of patient data.

Suggestions for future work:-To ensure AI is secure in healthcare, efforts should prioritize strong adversarial training, creating models that are easy to interpret, and implementing real-time threat detection. Upgrading network security and providing cybersecurity training for staff are also key. Working closely with regulators and following ethical guidelines will help achieve safe, transparent, and effective AI adoption, leading to better diagnostics and long-term innovation in healthcare.

V. Team contribution

Hamza, Ayaan, and Shoeb act as defenders, focusing on safeguarding diagnostic imaging systems and AI models, while Hidayatullah and Umar play the role of attackers, exploring potential vulnerabilities and attack methods. Hamza and Ayaan Khan contribute to the introduction, which focused on healthcare diagnostic imaging and described the goal and scope of the project. The Diagnostic Imaging in Healthcare use case scenario, which describes the integration of AI/ML models (CNN) and their use in X-ray, CT, and MRI analysis, was investigated and created by Hamza and Shoeb. Hidayatullah and Umar examined the network architecture, pointing out possible weaknesses and safeguarding communication routes. This included VPNs, Wi-Fi, and external access points. Critical assets like as infrastructure, patient data, and AI models were identified by Hamza, who also provided risk and impact-based prioritization and valuation. The attack graph was created by Hidayatullah, Shoeb and Ayaan, who combined potential effects, attack vectors, and entrance sites into a visual form for simpler study. Ayaan and Umar described attack techniques (such as poisoning, evasion, and model extraction) and suggested countermeasures such network monitoring, adversarial training, and encryption.

REFERENCES

Healthcare IT News (2020) Universal Health Services faces $67 million loss after cyberattack.
Available at: https://www.healthcareitnews.com/news/universal-health-services-faces-67-million-loss-after-cyberattack

Smith, J. (2019) 'The rise of voice cloning in cyber attacks', *CyberSecurity Magazine.*

Jones, A. and Brown, B. (2020) 'Manipulating medical images with AI', *Journal of Medical Ethics and AI*, vol. 5, no. 2, pp. 123-130.

CyberSecurity Reports (2018) 'DDoS attack on Boston hospital highlights vulnerability of healthcare systems.

Smith, J. (2021) *Protecting Healthcare Systems with AI: Strategies and Challenges*. New York: Cybersecurity Publications.

Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Madry, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Barlow, D. N., 2012. 3.1 Regulation and standards - Digital Transformation.

Ivezic, M. (2023) 'Gradient-Based Attacks: A Dive into Optimization Exploits', Journal of Cybersecurity, 15(2), pp. 100-110.

MITRE (2023) *MITRE ATLAS*. Available at: https://atlas.mitre.org/

Center for Internet Security (CIS) (2021) *CIS Controls Version 8*.

Mishra, N., Shivaji, G. B., Barekar, S. S., Dari, S. S., Dhabliya, D. and Patil, M. (2024) "Artificial Intelligence and Machine Learning in Healthcare Cybersecurity of Current Applications and Future Directions", South Eastern European Journal of Public Health, pp. 46–51.

Derui Wang, C. L. ,. S. W. ,. S. N. ,. a. Y. X., 2022. Defending Against Adversarial Attack Towards Deep Neural Networks Via Collaborative Multi-Task Training. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 19(NO.2), pp. 953-965.

Prajit Ramachandran∗, B. Z. Q. V. L., 2017. Searching For Activation Functions, Ithaca, New York: arXiv.