International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 4, Number: 2, Pages: 97 - 103, Year: 2025

A Comprehensive Review on Network Security in AI-based Healthcare Systems

Komal Shehzadi [1,*], Irshad Ahmed Sumra[1], Eeman Khokhar[1] , and Benish Khalid [2]

[1] Department of Informatics and Systems, School of Systems and Technology, University of Management and Technology, Lahore, 54000, Pakistan

[2] Lahore University of Biological and Applied Sciences (UBAS), 54000, Pakistan.

[*]Corresponding author: Komal Shehzadi (Email: komalrajpoot322@gmail.com)

Received: 12/01/2025, Revised: 11/03/2025, Accepted: 13/04/2025

***Abstract*— Artificial Intelligence (AI) has revolutionized healthcare with automated diagnosis and predictive analytics. However, the application of AI in healthcare systems introduces serious cybersecurity concerns, particularly concerning data integrity and model robustness. The robustness of Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Random Forest models to adversarial attacks such as the Fast Gradient Sign Method (FGSM) and data poisoning is tested in this study. We take a hybrid experimental setting to estimate the performance of the model under clean, attacked and defense scenarios. The results show that CNN is highly vulnerable to adversarial attacks, while Random Forest is relatively more stable. While defence mechanisms such as adversarial training and knowledge distillation can improve the model's performance, they cannot completely eliminate cybersecurity threats. The study emphasises the need for multi-layered cybersecurity frameworks. Such frameworks are necessary to guarantee the safety, privacy and reliability of AI-driven healthcare systems.**

***Index Terms*—Adversarial Attacks, Adversarial Machine Learning, Artificial Intelligence in Healthcare, Cybersecurity in Healthcare Systems, Deep Learning Security, Defense Mechanisms, Evaluation stages.**

I. INTRODUCTION

Artificial Intelligence (AI) is transforming the delivery of care through its application in the automation of diagnosis, disease prediction and the optimization of treatment. It is being applied in many areas of clinical decision support systems, medical imaging and EHRs. The use of AI systems in healthcare improves the delivery of care with increased efficiency and accuracy [1]. While there are advantages to implementing AI systems in healthcare, the use of AI also introduces new and significant cybersecurity concerns. The nature of healthcare data makes it particularly appealing to hackers because it contains a large amount of sensitive personal and clinical information, making it an attractive target for ransomware attacks, data breaches, and other forms of unauthorised access [2-4]. Research indicates that attackers can employ adversarial machine learning techniques against AI predictions using minor changes to the original input data, thereby causing inaccurate diagnostic results from the AI system [5-8]

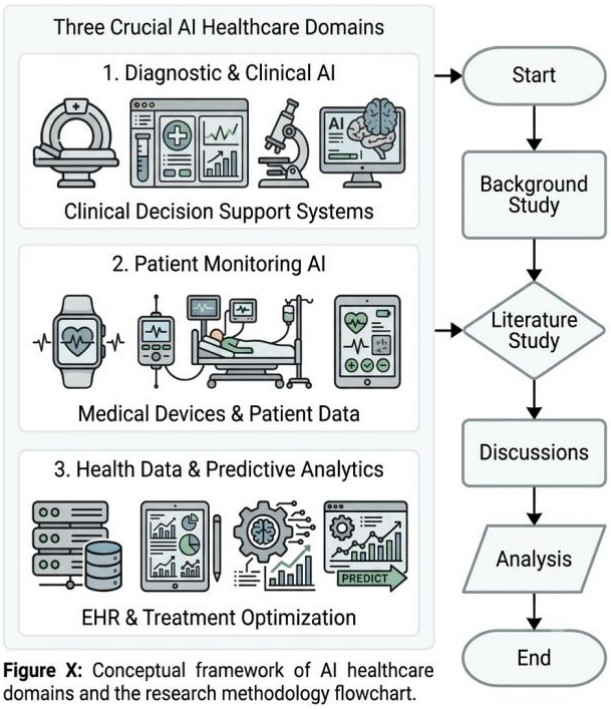

Fig. 1: Conceptual framework of AI healthcare domains and the research methodology flowchart.

These weaknesses represent critical vulnerabilities in medical applications that can be detrimental to patient safety. In addition, while current regulatory frameworks, including HIPAA and GDPR, have addressed traditional data security issues, they do not adequately protect against AI-related threats, including model inversion attacks, poisoning attacks or adversarial manipulation [9]. Thus, there is a clear need to

create AI-specific cybersecurity frameworks designed specifically for healthcare. This project will focus on establishing an inventory of AI-based cybersecurity weaknesses associated with healthcare and testing the efficacy of ML systems in response to adversarial inputs. This work is primarily a survey paper that reviews the most severe cybersecurity challenges in AI-based healthcare systems, including adversarial attacks, data poisoning, privacy concerns, and defensive strategies. Accompanying the discussion of the survey, a comparative performance analysis of selected machine learning models is provided to gain practical insights into the behaviour of AI models under clean, attacked and defended conditions. The rest of this paper is organized as follows. Section II provides background information and related work of AI-based healthcare systems and cybersecurity challenges. In Section III, we explain the research methodology, which includes dataset design, machine learning models, and adversarial attack simulations. The results of the experiments performed are analysed and discussed in detail in Section IV, and some important conclusions are drawn.

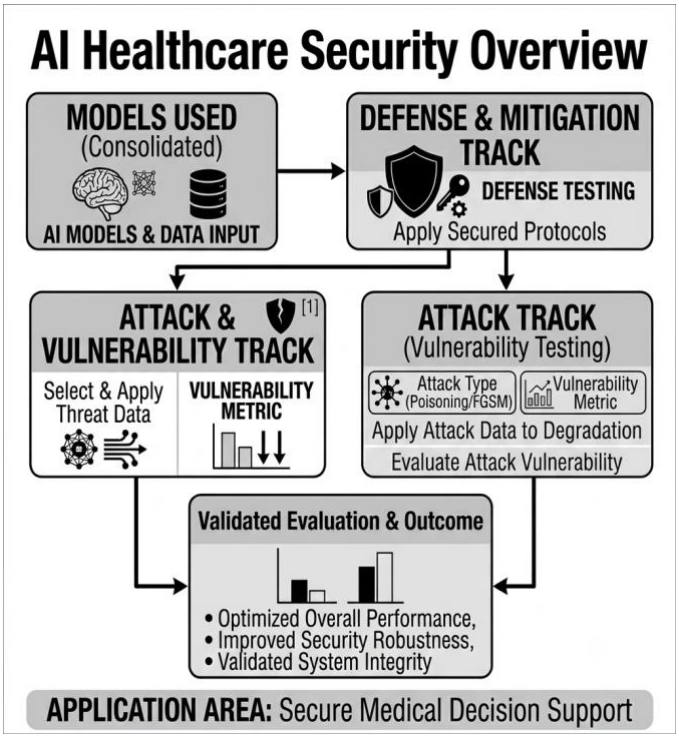

Fig. 2: Simplified AI healthcare security evaluation stages

II. Background and Related Work

The application of artificial intelligence (AI) is arguably the most impactful technological development occurring within today's healthcare systems. According to [1] and [2], AI can automate diagnosis, provide predictive analytics, offer clinical decision support, and create a customised approach to patient care. Incorporating machine learning (ML) into the day-to-day functioning of healthcare organizations has led to improvements in diagnostic accuracy, operational efficiency, and disease-detection capability at an earlier stage than previously possible. In the last few years, there has been increased adoption of AI-based solutions by healthcare organisations, as reported in [3] and [4], including EHR analysis, medical image analysis, wearable health monitors, and predictive risk assessments. These technologies enable the processing of massive amounts of complex clinical data to identify meaningful trends and patterns that aid healthcare professionals in their decision-making. However, as the use of AI in healthcare continues to advance, [5] and [6] identify several challenges, including those associated with cybersecurity, privacy, and data integrity. For example, many healthcare organisations collect vast amounts of sensitive, personally identifiable health-related data, including patient histories, diagnostic reports, genomic information, and biometric records. This makes healthcare systems a prime target for cybercriminals who may attempt to breach them to steal this data.

Furthermore, [5] also emphasise that the ever-growing digital nature of the healthcare industry's infrastructure has created additional attack vectors for potential cyber-attacks on the system. Cyber-attackers may employ methods such as ransomware attacks, unauthorised access, and data breaches. Additionally, [6] indicates that another significant challenge associated with using AI in healthcare will continue to be issues concerning patient privacy, primarily due to large-scale data-sharing practices and cloud-based data-storage environments. According to [7] and [8], the application of AI in medicine can be divided into three broad categories of development: rule-based expert systems; classical machine-learning-based systems; and modern deep-learning-based systems. Initial applications were limited to manual rule design, while current uses involve neural networks and ensembling techniques, both of which provide higher predictive performance.

Although advances have occurred in AI-based healthcare Systems, [9] and [10] demonstrate that the security of AI-based models remains inadequate. For example, deep learning architectures such as CNNs rely heavily upon large datasets and gradient-based optimization techniques; therefore, they are extremely susceptible to adversarial manipulation and specific input-perturbation techniques. Finlayson et al. [11] showed that medical AI applications (specifically, those in radiology and pathology) are extremely vulnerable to adversarial attacks. they also found that minor and unperceivable modifications made to medical images may cause erroneous clinical diagnoses from AI diagnostics tools; these results raise significant patient safety concerns.

Biggio and Rolo [12] identified two primary types of adversaries for machine learning systems: training-time and testing-time. Training-time adversaries include poisoning/corrupting the original learning dataset, whereas testing-time adversaries modify the input data at test/inference time to make it appear to the trained model as something other than what it is. Kovačs et al. [13] reported that ransomware attacks in healthcare settings have dramatically risen over the last few years (Table I).

Table I: Source :Adapted from [2],[4].

Phase	Technology	Characteristics	Limitations
Rule based Systems	Expert Systems	Fixed logic rules	Low adaptability
Machine Learning	SVM, RF	Data Driven learning	Feature Dependency
Deep Learning	CNN,RNN	High accuracy	Security vulnerabilities
Hybrid AI systems	Ensemble Models	Optimized Performance	Complex security risks

The authors of [14] were the first to introduce "adversarial training," a method for increasing the strength of a model's "robustness" (its ability to resist the effects of a malicious input) by having the model learn on examples of both "clean" (normal) examples and "adversarial" (malicious) examples. This type of training has been shown to significantly increase the model's robustness, but it typically increases the computational power required to train the model and the time required to complete training.

Papernot et al. [15] proposed a "defensive distillation" process to decrease the sensitivity of neural networks to different types of threats (attacks). The authors reported significant success with their approach as they were able to provide some degree of protection from certain types of attacks. However, subsequent researchers have identified at least one attack strategy that can defeat the defensive distillation approach. The authors of [16] demonstrated that certain previously developed approaches intended to protect models from malicious inputs could be effectively defeated by well-optimised attacks. These results reinforced the notion that maintaining model robustness to adversarial attacks is a long-term problem; therefore, developers need to develop methods for defending models that account for adaptive threat strategies. The authors of [17] defined the term "membership-inference attacks." Membership inference attacks are used to determine whether a particular data example (e.g., a medical record) was used during a machine learning model's training process. If successful, these attacks could harm patients by revealing sensitive information about their involvement in health-related studies.

The authors of [18] developed and tested "model inversion attacks," which demonstrate that confidential attribute information can be determined or reconstructed from the output of a trained machine-learning model. Model inversion attacks pose significant security risks when applied to healthcare systems because attribute information may contain extremely sensitive clinical or demographic data.

Argaw et al. [19] point out that hospitals have become prime targets for hackers seeking access to valuable and sensitive healthcare information. As hospitals rely on computers for almost all their functions and operations and store millions of electronic patient records, there is increased vulnerability to hacking these systems. In addition to the large amounts of patient data, the authors in [20] indicate that Healthcare 4.0 environments pose a number of serious security risks due to the use of interconnected Internet of Things devices, Cloud-based infrastructures, and communication systems. Although such systems allow for higher levels of interconnectivity and thus greater efficiencies, they create numerous opportunities for malicious hackers to gain unauthorized access to the system. Ransomware attacks in hospitals or other types of healthcare organizations are particularly damaging because they can deny patients and doctors timely access to important medical information, cause delays in treatment and emergency responses and ultimately harm patient safety and the quality of healthcare delivered. Kovacs et al. [21] indicate that this type of disruption may be a direct result of ransomware attacks. Javaid et al. [22] note that Artificial Intelligence (AI)-based cybersecurity systems can provide faster threat identification using real-time analysis of network traffic patterns and identify anomalous activity. According to [23], AI-based cybersecurity systems can significantly shorten response times and mitigate some of the effects of a cyberattack in healthcare environments.

TABLE II: Framework comparative analysis.

Framework	Focus	Strength	Weakness
HIPAA	Data Privacy	Legal protection	No AI security coverage
GDPR	Data Protection	Strong regulations	Limited ML protection
NIST AI RMF	AI Risk	Structured Framework	Early Development stage
ISO 27001	Information security	Global Standard	Not AI specific

Although many studies show that AI systems will provide better diagnostics, predictions, and patient care, they demonstrate a very limited understanding of the overall security of AI systems (Table II). There are two primary gaps in developing a comprehensive approach to securing healthcare AI. First, no specific healthcare-focused cybersecurity framework currently addresses all aspects of the secure AI lifecycle, including network protection, data governance, model robustness, clinical safety, and regulatory compliance. Second, most current frameworks address only one aspect of cybersecurity, such as encryption of protected health information (PHI), intrusion detection systems, or adversarial testing to defend against attacks. However, none of the current frameworks provides insight into what would occur if an adversary were to compromise the training data used by the AI system, a connected medical device, or the deployed AI model.

The second is about how little the robustness of healthcare machine learning systems has been studied. Having high accuracy for a test set made with clean data does not show whether a system would be able to continue to operate correctly if it receives bad data (corrupted), is intentionally disrupted, is missing, or comes from an entirely different setting at another hospital. Most deep learning models are tested using only one type of attack, and that is the Fast Gradient Sign Method

(FGSM), while the other types of attacks (stronger, adaptive, black box, poisoning, and combined) have received much less study. In addition, most studies measure only the decrease in accuracy of the model due to an attack and do not measure what percentage of successful attacks were made, how well calibrated the model is after being attacked, how likely the model is going to produce a false negative result, how long it takes to recover from an attack, what the computational resources needed to fight an attack are and how important clinically is the mistake that resulted from the attack.

The next gap of concern is that most studies rely on artificially generated data and/or controlled laboratory environments. These types of experimental designs can be beneficial for isolating an attacker's behavior(s) (as was demonstrated by the authors); however, they do not provide enough variability when considering what is contained in real electronic health record systems, various medical image acquisition devices, patient demographics, clinical workflows, and the variety of legacy hospital networks. As such, limited external validation makes it challenging to understand if a proposed security mechanism will apply universally across different organizations, or if it will maintain its effectiveness once deployed. Qureshi and Koo [23] have reviewed recent work on the cyber-privacy security issues in healthcare. They conclude that to protect healthcare effectively, it is necessary to use both technical controls (e.g., firewalls) and to continuously monitor them. In addition, they believe it is essential to provide a strong "governance" structure. It is equally important to ensure that all aspects of healthcare are aligned with current regulations. Finally, they believe that creating an overall organizational culture of security will be critical. Although their review has identified some areas where there still needs to be a lot of work done, such as addressing older systems, limited funds, heterogeneity of Internet-of-Medical-Things devices, validating security methods in actual healthcare environments, and lack of standards-based data sets for testing security methods, these identified gaps are consistent with the conclusion that security frameworks should be integrated rather than model-specific. The study by [24] examined the risk of data poisoning across five types of AI applications, including CNNs, Large Language Models, Reinforcement Learning Systems, Federated Learning, and Clinical Documentation Pipelines. Their study demonstrated how a single attack could be designed to harm multiple stages in the healthcare AI pipeline as well as develop a multi-layered defense against data poisoning attacks. However, they developed their attack vectors analytically rather than at a real-world scale; thus, large-scale clinical validation studies and the establishment of measurable detection thresholds remain necessary.

Vassilev et al. [25] developed the NIST machine-learning-based Adversary Taxonomy; this taxonomic structure organises attacks by lifecycle phase (attack type), level of knowledge about the attacked system, capabilities, goals, and mitigations. The taxonomy has a unifying language; however, in healthcare, we need to map these categories to clinical threat models as they are related to potential diagnosis risks (diagnostic harm), potential delays in treatment (treatment delay) and safety to patients (patient safety). A quantitative comparison by [26, 27] to an established classifier for breast cancer demonstrated that even with very small input changes (using FGSM), there was a significant loss of the classifier’s ability to predict. While this is another demonstration of how serious a concern adversarial data could be in practice, it does not prove that this would hold across all architectures or all healthcare datasets. Paschali et al. [28] examined the effects of single and multiple pixels on medical image classification systems. Although they found that small amounts of visual changes can be enough to produce misclassification in a computer aided diagnostic system, their study shows there is still an urgent need for a standard method for determining how many "attacks" an individual has been subjected to (i.e., what constitutes a fair amount of "budget") as well as the development of clinically relevant methods for measuring susceptibility to attacks on different types of images.

Tramèr et al. [29] tested adversarial images against oncology models developed using CT, mammography, and MRI data. Adversarial training successfully improved the model's ability to resist attacks; however, the level of robustness varied across datasets and attack conditions. This demonstrates that defenses designed to protect from one type of modality or perturbation are likely to fail when used in different clinical environments. Biggio et al. [30] introduced an adaptive two-phase defense method. During development, they applied adversarial learning. At the time of use, they filtered the image based on the original input. They found their proposed technique had increased resistance in radiology classification. However, reprocessing techniques can remove diagnostically relevant information from the image. Therefore, independent testing at additional hospitals and with additional equipment would be advisable.

A hybrid qualitative–experimental research model is employed to assess the reliability of Artificial Intelligence models in Healthcare Cybersecurity Environments. The methodological approach to this hybrid model includes literature-based thematic analysis and controlled machine-learning experiments to establish the robustness of AI models to adversarial attacks and their ability to defend against them.

The dataset used in this study was chosen to be representative of AI classification tasks within the healthcare domain. Before training the model, the data was pre-processed by handling missing values, normalizing input features, and splitting the dataset into training and testing sets. To compare the performance of the CNN, SVM, and Random Forest models under normal and adversarial conditions, we evaluated them. For deep feature learning, CNN model was utilized, and SVM and Random Forest were chosen as traditional machine learning classifiers for comparison. Adversarial attacks were used to evaluate the robustness of AI models against intentionally manipulated input data. The goal was to observe the change in model accuracy when the input data is slightly perturbed. The model performance was evaluated in terms of accuracy, precision, recall, F1-score, and robustness across clean, attacked, and defended conditions. Experiments were performed using Python-based machine learning libraries. The

same metrics were used for all models to ensure a fair comparison. Three machine learning models were selected for this study: CNN (Convolutional Neural Networks), SVM (Support Vector Machines), and RF (Random Forest). These models have been shown to be effective in several areas of Healthcare Analytics, including image classification, structured patient data processing, and predictive decision making [31, 32]. The three-stage method used to evaluate the model is as follows: evaluation of the baseline model's performance using clean healthcare data. Evaluating the model's ability to resist various types of attacks, such as FGSM and data poisoning attacks. Testing how well the defense mechanisms used are able to protect against those same attacks

III. ANALYSIS AND DISCUSSION

The comparison of three models (Convolutional Neural Network (CNN), Support Vector Machine (SVM) and Random Forest (RF)) shows that each model has a measurable decline in performance due to an attack. However, the extent and characteristics of the decline vary by model design and training methodology. The CNN model achieved the best performance among the models under clean data conditions, but showed the greatest sensitivity to small input perturbations. This behaviour can be explained by prior research indicating that gradient-trained deep neural networks are highly susceptible to even slight alterations in the input space. For example, [31] demonstrate how minute changes in input can cause substantial differences in feature activation, resulting in incorrect classification. Thus, for healthcare applications like radiology and pathology, this degree of susceptibility is critical because a small distortion in medical imaging could result in a misdiagnosis and a poor clinical decision.

The SVM model was found to be moderately vulnerable to both types of attacks. Unlike CNNs, SVMs separate features using a hyperplane. As a result of poisoned or corrupted training data, the feature distribution will shift. It has previously been demonstrated that poisoning attacks may greatly alter decision boundaries, thereby causing instability in classification reliability in structured healthcare datasets. These results support these observations, as we observe a clear decrease in SVM performance when even a small portion of the training data is altered. Random Forest showed much greater robustness due to its use of ensemble learning. By aggregating predictions from multiple decision trees, the corrupting effect of individual samples is diminished. Nevertheless, it was observed that large-scale attacks would still affect the overall stability of the models, supporting the notion that while ensemble methods provide greater robustness, they cannot be used to ensure safety. It has also been reported that no single model is safe in adversarial environments.

From a general point of view, the studies show a significant deficiency in how most healthcare AI programs are designed: while their predictive accuracy is typically very good, they have little or no security. In this sense, models developed under standard testing conditions will often fail when attacked through adversarial means, which demonstrates the large difference between how we evaluate AI models today and what we need for secure operation of these systems in the future. The difference between how we currently evaluate AI models and what we need from AI models for security reasons is now being highlighted by many researchers who argue that all healthcare AI must be evaluated based upon not just accuracy, but also robustness, interpretability, and ability to defend against adversarial attacks (Table III).

TABLE III: MODEL PERFORMANCE UNDER SECURITY CONDITIONS

Model	Clean accuracy	Under attack	After defense
CNN	92%	76%	83%
SVM	85%	67%	73%
Random Forest	90%	82%	85%

The results from our analysis of each type of defense mechanism indicate that both adversarial training and defensive distillation improved model robustness. While neither fully restored the original model's performance, each provided some improvement. Adversarial training trains models using input examples that have been manipulated to develop a more robust decision boundary than models trained without such exposure. However, this method requires significantly more computation than typical training methods and precise parameter tuning to prevent "over-training" on a particular type of adversarial attack. Additionally, defensive distillation reduces a model's sensitivity to input data by smoothing its output probability distributions. However, like adversarial training, defensive distillation is not foolproof. As has been shown previously in prior research, both of these defense mechanisms can be circumvented with sufficiently sophisticated forms of adversary attacks. Therefore, developing effective defenses against adversarial attacks continues to remain one of the major challenges facing the field of AI security research. The study's conclusions show that an adversary's ability to interfere with healthcare systems by poisoning data is more than just an attack during use -- it has the potential for lasting harm by changing how machine learning models learn. Poisoned data will continue to degrade a model's performance over time, even if only a few samples are used to poison the system. Poisons can be subtle enough to slightly alter how a model makes decisions, particularly when using SVMs, leading to biased or inaccurate decisions. As such, there is a great deal of emphasis on securing both the deployed AI model(s) and all points of interaction across the full continuum of the data pipeline, from initial data collection through preprocessing to training the AI model.

A second key takeaway of this study is that current cybersecurity standards/requirements developed for healthcare are not well-suited to address threats specific to artificial intelligence. Current healthcare cybersecurity requirements, such as those related to HIPAA and GDPR, are focused primarily on privacy and access control rather than adversarial robustness or model security. Therefore, while these standards provide protection against many types of cyber threats, such as adversarial examples, model inversion, and poisoning attacks,

they do not provide adequate protection against these threats. Thus, what is needed are new cybersecurity standards/requirements specifically designed for protecting AI in a healthcare environment. In general, therefore, the findings clearly support the idea that no single AI model provides total security from an attacker who wishes to exploit an AI model using adversarial techniques. While CNNs achieve very high accuracy, they are relatively robust. SVMs offer moderate stability against adversarial attacks; however, they remain susceptible to degraded performance in the presence of poisoned data. Finally, Random Forest is at least somewhat resilient against adversarial attacks; however, like all other models tested here, its performance does degrade when subjected to stronger forms of attack. Overall, these findings suggest that selecting a particular type of AI model alone is not sufficient to ensure secure AI deployments in healthcare systems.

IV. LIMITATIONS AND FUTURE WORK

A number of limitations should be taken into account, despite the fact that this study offers insightful information about the security issues with AI-based healthcare systems. First, because medical data is extremely sensitive and subject to privacy laws, obtaining real-world healthcare datasets remains a significant challenge. The majority of AI security assessments are conducted on benchmark or controlled datasets, which may not accurately reflect the complexity of real-world clinical settings. Second, because medical equipment, patient populations, data gathering methods, and clinical workflows vary, healthcare organizations produce heterogeneous data. As a result, a security system that works effectively in one hospital setting might not work as well in another. Moreover, training complexity increases, and substantial computational resources are required to execute sophisticated protection mechanisms such as adversarial training and defensive distillation. This makes practical adoption difficult, particularly in healthcare organizations with inadequate infrastructure.

Regulatory compliance is a significant additional constraint. AI-based healthcare systems must adhere to security and privacy standards set forth by frameworks like GDPR, HIPAA, and other healthcare laws. Nevertheless, new AI-specific risks, including adversarial assaults, model inversion, and data poisoning, are not adequately covered by current rules.

The development of robust AI security frameworks that can be validated on large-scale real-world healthcare datasets should be the primary goal of future research. Adaptive defensive strategies, privacy-preserving machine learning methods, and effective security solutions to facilitate the dependable implementation of AI systems across various healthcare facilities should all be the focus of future research.

While the study provides valuable insights into cybersecurity risks in AI-based healthcare systems, the analysis has a few limitations. Most evaluations are conducted under controlled experimental conditions, which may not fully represent real clinical settings. In real-world health care settings, data may come from different hospitals, devices, patient populations, and record systems. Such heterogeneity may affect the model's generalisation and robustness. Furthermore, regulatory compliance, computational cost, data privacy, and integration with existing hospital systems remain critical challenges for real-world deployment (Fig. 3 and Table IV).

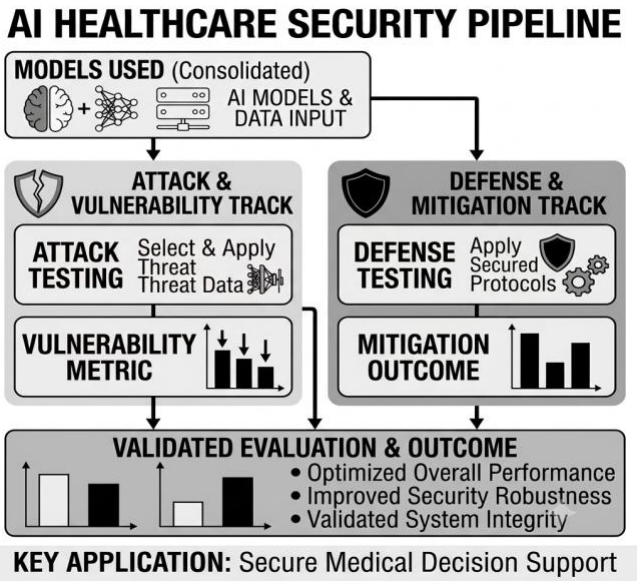

Fig. 3: AI Healthcare security system

TABLE IV: SUMMARY OF THE STUDY

Components	Details
Research Area	AI Based Healthcare Cybersecurity
Models Used	CNN,SVM, Random Forest
Attack Types	FGSM, Data Poisoning
Defence Methods	Adversarial Training, Defensive Distillation
Key finding	CNN most vulnerable, RF most stable
Main Limitation	Partial defense effectiveness only
Recommendations	Multi layer AI security framework
Future Direction	Robust AI and Healthcare integration

V. CONCLUSION

This paper has examined the vulnerability of artificial intelligence-based healthcare systems from a security perspective. It assessed the resilience of machine learning methods to adversarial attacks. The purpose of the research is to evaluate the responses of three types of AI: Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Random Forest models to both adversarial attacks and data poisoning. It also evaluates whether common defense mechanisms provide sufficient protection against those threats. The results of the current study show that there are significant risks associated with using artificial intelligence in medical contexts, including the potential for adversaries to manipulate it. Adversaries may produce minimal alterations to input information, which could result in dramatic reductions in an AI model's predictive ability, especially when it utilises a deep learning approach such as a CNN. As such, while predictive performance in non-adversarial circumstances may indicate excellent performance, it indicates nothing about its reliability or security within an adversarial environment. The research also

found that models developed using structured methods (such as SVMs and RF) were somewhat less susceptible to data corruption and poisoning attacks, yet all machine learning architectures are potentially vulnerable to corruption during training. Thus, no single machine learning paradigm will provide complete security for healthcare-related uses. Additionally, while implementations of adversarial training and defensive distillation provided some improvement in model robustness, no method was able to completely restore the original level of performance. Therefore, the major shortcoming with existing defenses is that any improvements are marginal. The study's results have demonstrated a major disconnect between established healthcare cybersecurity frameworks and the unique threats posed by AI. Current healthcare regulatory standards, such as HIPAA and GDPR, primarily focus on protecting patient data and controlling access to it. The standards do not include adequate provisions for adversarial attacks on machine learning models, including model inversion, poisoning, and adversarial perturbations. In addition, existing studies have addressed this issue in the context of cyber-attacks on healthcare systems, emphasizing the need to develop healthcare system-specific security architectures to support AI-based deployments. Ultimately, the study concludes that deploying AI securely in healthcare will require a paradigm shift from traditional data-centric security methodologies to a layered architecture that protects both data and models, employs defensive techniques against potential adversarial attacks, and continuously monitors.

Funding Statement

The authors received no specific funding for this study.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

Author Contributions

All authors contributed to the conception, literature review, drafting, and critical revision of this manuscript and approved the final version for submission.

Data Availability Statement

Data is available on reasonable request.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

REFERENCES

[1] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of healthcare," *The American Journal of Medicine*, vol. 132, no. 7, pp. 795–801, 2019.

[2] A. Rajkomar, J. Dean, and I. S. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[3] Z. Ahmed et al., "Artificial intelligence with multi-functional machine learning platform for healthcare," *Database: The Journal of Biological Databases and Curation*, 2020.

[4] B. Shickel et al., "Deep EHR: Deep learning for electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, 2018.

[5] L. Coventry and D. Branley, "Cybersecurity in healthcare: A narrative review of trends, threats and ways forward," *Maturitas*, vol. 113, pp. 48–52, 2018.

[6] A. K. M. I. Newaz et al., "A survey on security and privacy issues in modern healthcare systems," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, 2021.

[7] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, 2018.

[8] S. G. Finlayson et al., "Adversarial attacks on medical machine learning," *Science*, 2019.

[9] I. J. Goodfellow et al., "Explaining and harnessing adversarial examples," *ICLR*, 2015.

[10] S. G. Finlayson et al., "Adversarial attacks on medical machine learning," *Science*, 2019.

[11] B. Biggio and F. Roli, *Pattern Recognition*, 2018.

[12] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, 2023.

[13] A. Madry et al., "Towards deep learning models resistant to adversarial attacks," *ICLR*, 2018.

[14] N. Papernot et al., "Distillation as a defense to adversarial perturbations," *IEEE S&P*, 2016.

[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *IEEE S&P*, 2017.

[16] R. Shokri et al., "Membership inference attacks against machine learning models," *IEEE S&P*, 2017.

[17] M. Fredrikson et al., "Model inversion attacks that exploit confidence information," *ACM CCS*, 2015.

[18] S. T. Argaw et al., "The state of research on cyberattacks against hospitals," *BMC Medical Informatics and Decision Making*, 2019.

[19] J. J. Hathaliya and S. Tanwar, "Security and privacy issues in Healthcare 4.0," *Computer Communications*, 2020.

[20] M. Conti et al., "IoT security and forensics: Challenges and opportunities," *Future Generation Computer Systems*, 2018.

[21] M. Javaid et al., "Cybersecurity for healthcare domains: A comprehensive review," *Cyber Security and Applications*, 2023.

[22] R. Qureshi and I. Koo, "Cybersecurity threats in healthcare systems," *Applied Sciences*, 2026.

[23] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," in *Artificial Intelligence in Healthcare*, 2020.

[24] National Institute of Standards and Technology, *AI Risk Management Framework (AI RMF 1.0)*, 2023.

[25] X. Xue et al., "Machine learning security: Threats, countermeasures, and evaluations," *IEEE Access*, 2020.

[26] S. M. Williamson and V. Prybutok, "Privacy challenges in AI-driven healthcare," *Applied Sciences*, 2024.

[27] M. Paschali et al., "Generalizability vs robustness in medical imaging networks," MICCAI, 2018.

[28] F. Tramèr et al., "Ensemble adversarial training: Attacks and defenses," ICLR, 2018.

[29] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against SVM," ICML, 2012.

[30] Sumra, I.A., Hasbullah, H.B., AbManan, Jl.B. (2015). Attacks on Security Goals (Confidentiality, Integrity, Availability) in VANET: A Survey. In: Laouiti, A., Qayyum, A., Mohamad Saad, M. (eds) Vehicular Ad-hoc Networks for Smart Cities. Advances in Intelligent Systems and Computing, vol 306. Springer.

[31] Sumra, I.A., Hasbullah, H., Ab Manan, J.-L.: Effects of attackers and attacks on availability requirement in vehicular network: a survey. In: International Conference on Computer and Information Sciences (ICCOINS2014), Malaysia, 3–5 June 2014.