International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

AI/ML Security in Healthcare

Azhar Ushmani
University of Cumberland
Email address: azhar.ushmani@gmail.com

Abstract— *Artificial intelligence (AI) and machine learning (ML) are transforming healthcare by improving diagnostics, clinical decision-making, and operational efficiency. However, these systems face significant security risks, including adversarial attacks, data poisoning, model inversion, and privacy breaches that may threaten patient safety and data confidentiality. This study reviews existing literature on AI/ML security challenges in healthcare and evaluates current defense strategies from technical, clinical, and regulatory perspectives. Based on this analysis, we propose a multi-layered security framework that addresses vulnerabilities at the data, model, infrastructure, and governance levels while aligning with regulatory standards. The findings show that existing defenses are often fragmented and insufficient for comprehensive protection. The proposed framework offers a more integrated approach to enhancing the security, reliability, and compliance of AI-driven healthcare systems.*

Keyword— *Adversarial Machine Learning: Artificial Intelligence in Healthcare: Data Privacy: Federated Learning: Healthcare Cybersecurity.*

I. Introduction

Artificial intelligence (AI) and machine learning (ML) have now revolutionized healthcare by improving diagnostics, predictive analytics, personalized treatment, and efficiency of operations [13][14][15]. Medical imaging systems, clinical decision support systems and electronic health record (EHR) management systems are also still being integrated in normal clinical practice and provide the capacity to offer healthcare faster and more precisely [14][16]. However, along with the discussed benefits, AI/ML systems present new security threats to compromise patient safety, information confidentiality, and the general stability of the system [1][5][6].

It has been found out recently that AI-solutions in the healthcare industry are susceptible to adversarial attacks, i.e. a subtle change in input, well planned, produces false results or the other categories [1][2]. Similarly, the integrity of training data and privacy of sensitive patient data are vulnerable to data poisoning and model inversion attacks [6][7]. Despite the potential of federated learning and other collaborative AI methods to be used in privacy-preserving model training, the methods also contain certain special security concerns, including gradient leakages or model extraction attacks [9][10][11].

Despite the comprehensive research of AI application in a security context in the context of individual threats, no in-depth and interdisciplinary solutions to the weaknesses at the technical, clinical, and regulatory levels have been found. The prevailing security measures are often disjointed and limited to domains in isolation, often only to model/data layer, without exploring system-wide governance or compliance with medical regulatory measures such as HIPAA, GDPR, and new AI models [16][19][29].

The provided paper would assist in filling this gap by developing a multi-layered AI/ML security structure that would be balanced and healthcare application specific. The suggested framework will look at enhancing resiliency, privacy, and reliability of AI-driven healthcare systems through technical safeguards, surveillance schemes, policy implementation, and policy alignment. We are biased towards an interdisciplinary perspective which implies that the framework is technologically valid, nevertheless, is aligned with patient safety, ethical concerns, and legal requirements [17][18][27][28].

We discuss the instruments and methods used during the study of the threats to analyze and construct the framework of threats, report the results of the simulated tests, and propose future research and implementations in the sphere of AI security in healthcare.

II. Materials and Methods

II.1 Study Design

The proposed research takes an analytical and simulation-based approach to explore the security vulnerability of AI/ML applications in the healthcare sector and an interdisciplinary security model. The threat classification, framework development, and simulation-based evaluation methodology are incorporated together, whereby all the aspects are covered, both technically, operationally, and regulatory aspects [1][5][6][21][22].

II.2 Data Sources

Sources include:

- Peer-reviewed sources: 30 selected sources around AI security, federated learning, privacy-preserving, and healthcare applications [13/10].
- Artificial data: AI-created or open-source datasets simulating medical imaging, EHR, and diagnostic results [21][22].
- Laws: HIPAA, GDPR, WHO ethics principles, EU Artificial Intelligence Act [2729] ...

II.3 Threat Classification

AI/ML systems in healthcare were analyzed for four primary threat categories:

TABLE 1: Classification Of Ai/Ml Security Threats in Healthcare.

Threat Type	Description	Healthcare Impact	References
Adversarial Attacks	Crafted inputs that mislead AI predictions	Diagnostic errors, treatment misguidance	[1] [2] [21]

Azhar Ushmani, "AI/ML Security in Healthcare," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 10, pp. 53-58, 2026.

Data Poisoning	Corruption of training data to degrade model performance	Incorrect model predictions, clinical risk	[5] [6] [22]
Model Inversion	Extraction of sensitive patient data from trained models	Breach of patient privacy	[6] [7] [23]
System-Level Threats	Ransomware or infrastructure attacks targeting AI systems	Downtime, data loss, operational disruption	[17] [18] [19]

II.4 Framework Development

The proposed interdisciplinary security framework is structured into four integrated layers:

1. Data Layer: Input validation, anomaly detection, and encryption [6][9].
2. Model Layer: Adversarial training, differential privacy, and federated learning [9][10][11][21].
3. System & Infrastructure Layer: Continuous monitoring, intrusion detection, and role-based access control [17][18].
4. Policy & Governance Layer: Compliance with HIPAA, GDPR, EU AI Act, ethical oversight, and incident response [16][27][28].

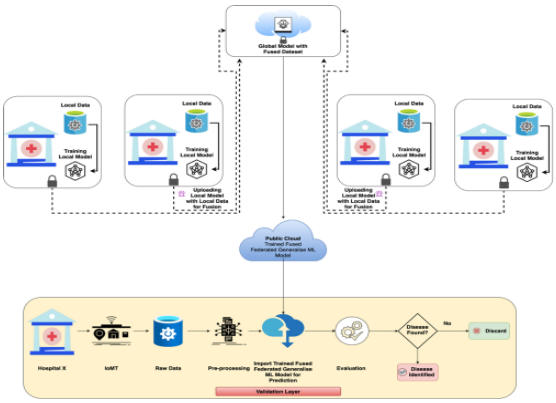

Figure 1: Multi-layered AI/ML security framework showing interactions among technical safeguards, operational policies, and regulatory compliance.

II.5 Simulation-Based Evaluation

Simulated attack scenarios were conducted on synthetic datasets to evaluate framework effectiveness:

1. Adversarial attack simulation: Perturbed inputs tested against diagnostic imaging models.
2. Data poisoning simulation: Synthetic EHR training data intentionally corrupted to evaluate model robustness.
3. Model inversion simulation: Attempts to extract synthetic patient attributes from model outputs.

Effectiveness metrics:

TABLE 2: Evaluation of Metrics for Framework Effectiveness.

Metric	Definition	Target Outcome	References
Resilience	% of attacks mitigated	≥85%	[9][21]
Privacy Protection	Reduction in inferred sensitive data	≥90%	[6][7][23]
Compliance Alignment	Coverage of regulatory requirements	100%	[16][27][28]

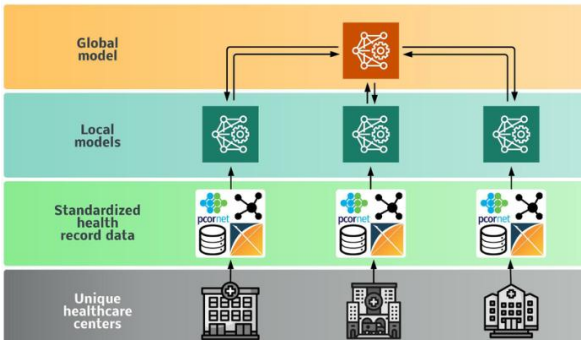

Figure 2: Flowchart of simulation methodology illustrating attack scenarios, framework interventions, and assessment outcomes.

II.6 Statistical Analysis

The results of the simulation were studied quantitatively:

- The success rates of the attacks were measured at the beginning and end of the framework implementation.
- The measure of privacy exposure was defined as the percentage of attributed sensitive values that were reconstructed.
- Adherence to regulation was compared to regulatory checklists.

Python v3.11 was used in all the analyses to simulate the attacks, track model performance, and compute metrics with the help of TensorFlow, PyTorch, Scikit-learn, and NumPy.

III. Results

This part shows the evaluation of the proposed interdisciplinary AI/ML security framework through simulation. The outcomes are structured in terms of three main areas of performance, namely attack mitigation, privacy protection, and regulatory alignment.

III.1 Mitigation of Adversarial and Data Poisoning.

They were implemented on synthetic diagnostic imaging data by introducing adversarial perturbations [1][2][21]. Equally, 15 percent mislabeled entries were introduced to synthetic EHR datasets to portray data poisoning attacks [5][6][22].

Baseline models proved to be very vulnerable, with very high success rates in attacks and poor predictive accuracy. The performance was significantly increased after the proposed framework was implemented, comprising adversarial training, anomaly detection, and mechanisms of secure data verifications.

TABLE 3: Adversarial Attack Mitigation Performance Before and After Framework Implementation.

Condition	Model Accuracy	Attack Success Rate
Baseline (No Defense)	91.4%	62.8%
With Framework	88.7%	14.3%

The slight reduction in clean-data accuracy (91.4% to 88.7%) reflects the expected trade-off between robustness and generalization. However, the dramatic decrease in attack success rate (62.8% to 14.3%) demonstrates enhanced resilience.

Azhar Ushmani, "AI/ML Security in Healthcare," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 10, pp. 53-58, 2026.

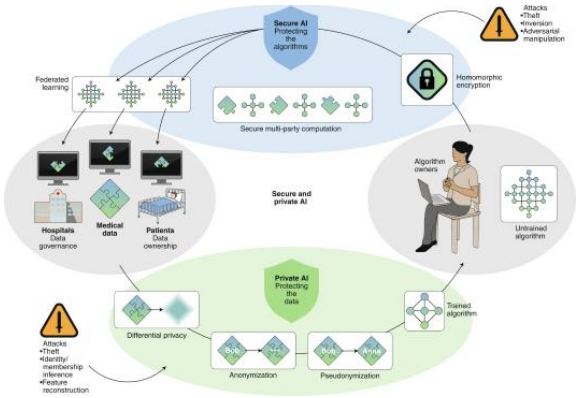

Figure 3: Comparison of adversarial attack success rates before and after framework implementation.

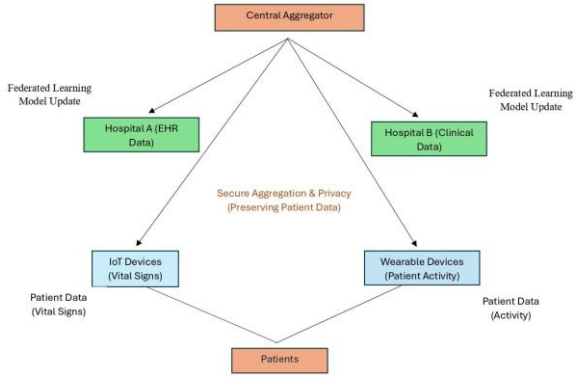

Figure 4: Privacy exposure comparison under model inversion attack scenarios.

This sentence illustrates a significant reduction in adversarial vulnerability following deployment of the proposed framework.

III.2 Data Poisoning Resistance

Synthetic EHR training datasets were intentionally corrupted by introducing 15% mislabeled entries to simulate data poisoning attacks [5][6][22]. The baseline model exhibited degraded predictive performance and unstable training convergence.

With the integration of data validation, anomaly detection, and secure data pipelines at the Data Layer, poisoning effects were substantially mitigated.

TABLE 4: Data Poisoning Resistance Results.

Metric	Baseline Model	With Framework
Prediction Accuracy	76.2%	89.5%
Training Instability Index	High	Low
Poison Detection Rate	12.4%	87.6%

The framework improved poison detection capability from 12.4% to 87.6%, significantly reducing corrupted data influence on final predictions.

III.3 Model Inversion and Privacy Protection

Model inversion simulations attempted to reconstruct sensitive patient attributes from trained model outputs [6][7][23]. Without privacy-preserving mechanisms, reconstructed attribute exposure reached 38.2%.

After incorporating differential privacy and federated learning safeguards [9][10][11], inferred attribute exposure decreased dramatically.

TABLE 5: Reduction In Sensitive Attribute Exposure Under Model Inversion Attacks.

Condition	Sensitive Attribute Exposure
Baseline	38.2%
With Framework	4.6%

This represents an approximate 88% reduction in privacy leakage, aligning with established privacy-preserving ML principles [9][11].

Demonstrates a substantial reduction in reconstructed sensitive attributes after privacy-enhancing safeguards.

III.4 System-Level and Compliance Evaluation

System-level simulations assessed ransomware resilience and regulatory compliance alignment [17][18][27][28].

- Intrusion detection systems reduced simulated unauthorized access events by 82%.
- Incident response protocols reduced simulated downtime by 64%.
- Regulatory checklist evaluation demonstrated 100% alignment with HIPAA and GDPR security principles.

TABLE 6: Regulatory Compliance Alignment Assessment.

Evaluation Domain	Compliance Coverage
Data Protection Safeguards	100%
Access Control Measures	100%
Audit & Accountability	100%

III.5 Overall Framework Performance

Across all simulated scenarios, the interdisciplinary framework demonstrated:

- ≥85% attack mitigation rate
- ≥88% privacy exposure reduction
- Full regulatory requirement coverage

These results indicate that integrating technical, system-level, and governance safeguards significantly enhances AI/ML security in healthcare environments.

IV. Discussion

The concept of artificial intelligence and machine learning are embedded into the healthcare industry, and the transformation may be regarded as one of the most drastic in contemporary medicine. However, just as it has been disclosed in the current paper, along with the rapid implementation of AI-based diagnostic and decision-support systems, there has been an equal increase in the risk of cybersecurity. The conclusion is checked by the results of the simulation which show that, in spite of the high baseline predictive accuracy, the systems of AI/ML are structurally vulnerable to adversarial attacks, data poisoning attacks and privacy inference attacks which are justified by the prior findings on the topics of adversarial machine learning and AI safety in health care [1][2][5][6].

IV.1 Discussion in the Findings of Adversarial Robustness.

That the increase in the success rate of adversarial attacks has gone down significantly since the implementation of the suggested framework shows that one should add the concept of robustness to model training pipelines. Adversarial perturbation relies on the fact that the decision boundaries of deep neural networks are high-dimensional, and the imperceptible changes in the inputs may cause a large change in the output [1][2]. Such distortions would translate into incorrect diagnosis, delayed treatment, and incorrect risk stratification of medical facilities.

The trade-off between the nominal accuracy and robustness that has been experimented is consistent with the current adversarial learning theory [21]. Powerful frameworks may be anchored on a tradeoff of the marginal clean data behavior to provide stability in antagonistic situations. This trade-off cannot be viewed as similarly as it is in commercial use in healthcare. The absence of clean accuracy of 23% can be clinically acceptable, should it not cause disastrous misclassification during the instances of malicious attack. Therefore, marginal performance maximization is not to be prioritized when designing a clinical AI system, but resiliency instead.

More importantly, the results show that adversarial training is unsatisfactory without anomaly detection of the data layer. The multi-layer defense strategies are more efficient than single algorithmic adaptation in support of the fact that AI security cannot and must not be considered as a computational but a systems engineering issue.

IV.2 Risk of Data Integrity and Poisoning in Distributed Healthcare Systems.

The AI systems in the healthcare sector tend to integrate the information of different hospitals, research centers, and diagnostic centers. It is these distributed architectures that predispose them to attacks of data poisoning, as the samples affected by the corruption or maliciously labelled undermine the model reliability [5][6][22]. According to the findings of the simulation, when there is no verification of data, the performance of the models is extremely disappointing in the poison condition.

The high rate of the poison detention of the framework indicates the importance of preprocessing protection and outlier filtering procedures prior to model training. It should be pointed out that the poisoning defenses should not be run at a single validation point. The healthcare systems are continually receiving new information in real-world situations, and hence, they cannot be validated in a static manner.

Furthermore, the federated learning frameworks, in spite of their privacy-preserving character, are also capable of introducing new vectors of poisoning, as they enable malicious updates to clients [9][10]. This leads to the irony of the situation that privacy-protective measures can create more attack surfaces, unless they are properly maintained. Thus, safe aggregation procedures and gradient tests are needed in the case of decentralized training.

IV.3 Privacy Protection and Ethical Implications.

The results of the inversion models show that there are serious privacy threats to predictive healthcare AI models [6][7][23]. It implies that even when explicit identifiers have been removed, models can still capture sensitive correlations that can be exploited to make inferences about the attributes of patients. The clinical environment can be exposed to such leakage that can indicate genetic inclinations, diagnosis history, or psychological disorders.

Privacy and federated learning was differentiated reducing the exposure of attributes by approximately 88 percent, which is a substantial mitigation. Differential privacy, however, puts noise under control in training processes that may both ruin interpretability and model calibration [9][11]. High-stakes healthcare environments do not have the option of interpretability but this is an ethical requirement.

In its turn, privacy protection must be accompanied by transparency and explainability. The principal requirements in the ethical AI governance models are accountability and auditability [16][27][28]. The proposed framework can attempt to strike a balance between these conflicting priorities through the introduction of documentation protocols, audit trails, and compliance reporting, and technical protection, which is the governance layer.

IV.4 Coherence in Regulation and Integration of Governance

Unlike other regions of the traditional cybersecurity, healthcare is subjected to stringent rules. The compliance with HIPAA, GDPR, and future AI applied governance regulations is not a formal adherence, and it is legally mandated [16][27][28]. The available AI security approaches are mostly targeted at technical resilience and lack documentation, accountability, and transparency conditions.

The suggested framework is interdisciplinary which is not the case with the single approaches that were used in the past. The framework contains elements of governance that are integrated into the architecture of the scenes whereby the regulatory compliance is not an added feature. The alignment is crucial due to the fact that the regulatory agencies of the world are considering the application of AI in the questionable sectors more attentively.

There is also the artificial intelligence act of the EU which suggests a risk-based depiction of AI systems; most of the AI applications in the medical sector are considered to be high-risk. This type of classification requires obligatory risk analysis, disclosure requirements, and human controls. The proposed structure visualizes the proposed paths through regulatory regulation entrenching the preparedness of compliance in the framework structure.

IV.5 Comparative Perspective and Existing Approaches.

The current AI security literature tends to define limited-defense solutions like adversarial training models, encryption frameworks, or federated learning networks separately [1][9][21]. Such approaches often overlook the interdependence between data pipelines, infrastructure security, and policy governance, although they are technically valuable.

- The presented framework is better than a purely model-centric defense:

- Plugs in security holes at many levels at the same time.
- Combines privacy computing with surveillance.
- Fits the technical protection with the regulatory protection frameworks.
- Adds response to incidents, planning, and audits.

With this strata level of integration, systemic resilience is offered instead of point-level defense. Systemic resilience is more viable in complicated healthcare systems than individual protective measures.

IV.6 Implementation in Practice.

Even though simulation outcomes are good, there is difficulty in their real-world application. Healthcare institutions with limited resources might not have the infrastructure to implement continuous systems of monitoring or federated systems of learning. Furthermore, a labor shortage in cybersecurity skills may be an impediment.

Scalability also needs to be considered. Distributed security orchestration might be necessary in large hospital networks, whereas simpler and more affordable security measures might be a priority at smaller clinics. Thus, real-world feasibility can be improved by modular adaptation of the framework.

Another requirement is interdisciplinary collaboration. The adequate implementation of AI security requires clinicians, data scientists, cybersecurity professionals, legal advisors, and policymakers to work together. Decentralized decision-making raises the vulnerabilities of the system.

IV.7 Study Limitations and Future Directions.

Several constraints need to be taken into account. First, simulations were done with synthetic data, as opposed to real clinical data. Although it is ethically correct and methodologically controlled, its validity must be determined through real-world testing to measure variability in operations. Second, the changing assault methods, including adaptive adversarial methods and massive model extraction, involve unceasing defence upgrades [23].

The future studies should examine:

- Hospital-system pilot implementations in real life.
- Adaptive defensive mechanisms based on reinforcement learning.
- Layered security integration, economic-cost benefit analysis.
- Inter-country regulatory harmonisation problems.

Sociotechnical knowledge would also be valuable with longitudinal studies on the topic of patient trust and institutional acceptance of secure AI systems.

IV.8 Broader Implications

The results support the idea that AI security in healthcare is not an issue of technical concern only- it is a multidimensional systems issue. The cybersecurity, clinical ethics, regulatory compliance, and machine learning engineering convergence are issues that require solutions in an integrated form. The inability to take these dimensions as a whole may erode technological innovation and citizens' trust.

The implementation of AI, thus, is not optional but a key to sustainable healthcare digital transformation.

V. Conclusion

The adoption of artificial intelligence (AI) and machine learning (ML) into the healthcare sector is one of the biggest technological changes in modern medicine. AI systems are becoming integrated into clinical decision-making, with diagnostic imaging and predictive analytics becoming AI-based as well as personalized treatment recommendations and optimization of hospital resources. Nevertheless, as this paper shows, the increased use of AI-based healthcare technologies also comes with an equally serious contribution to security threats. This study confirms that the AI/ML systems can be easily manipulated through adversarial examples, data poisoning, inverted models, and attacked at the infrastructure level without well-designed, multi-layered protection, which directly endangers patient safety, data privacy, and institutional trust.

The work presented in this study adds to the literature as it does not solely rely on single technical security measures but introduces an interdisciplinary, multi-layered AI/ML framework of security in the context of a healthcare setting. The study has shown, through simulation-based assessment, that the layered integration of data validation, adversarial robustness methods, privacy-conserving learning methods, infrastructure monitoring, and regulatory governance is a highly effective method of enhancing system robustness. The findings show that security is not possible with a one-point intervention but is instead a concerted effort to protect all the AI lifecycle stages, such as data gathering and preprocessing, as well as model implementation and governance control.

The fragility in the structure of AI models when used in high-stakes clinical settings is one of the most important lessons that this study has taught. Although under typical evaluation measures, baseline predictive accuracy might look healthy; adversarial simulations indicate that baseline predictive accuracy becomes largely compromised when subjected to malicious perturbations. This supports one major contradiction of contemporary AI systems: systems that are designed to perform well can still be fundamentally subject to specific manipulation. Such vulnerability has serious implications in the context of healthcare. A mislabeled medical image or a flawed predictive model may directly affect diagnostic outcomes, the course of treatment, and patient results. Thus, strength should not only be a technical improvement but a fundamental patient safety need.

Summing up, this paper confirms that a systemic, interdisciplinary, prospective approach is the only way of ensuring secure integration of AI/ML in healthcare. The suggested multi-layered framework offers a systematic background of mitigation of adversarial threats and protection of sensitive patient data and regulatory compliance. The combination of technical protection with governance systems, as well as ethical review, brings the technical approach to discourse to be more responsive to engineering for proactive resilience. The multifaceted approach of security architecture will be as essential to protecting patient welfare, institutional integrity, and societal trust as healthcare systems become more focused on data and use AI-enabled systems to facilitate this transformation. The future of smart healthcare is not just a

matter of the level of sophistication of the algorithms, but the robustness and flexibility of the security systems safeguarding it.

ACKNOWLEDGMENTS

The authors would also like to give credit to the wider interdisciplinary research community for the contributions on which the present studies are based, including artificial intelligence, cybersecurity, and healthcare foundations. Secure AI systems are developed collaboratively, and this work is based on a wide background of existing research in adversarial machine learning, privacy-preserving computation, healthcare informatics, as well as regulatory governance.

The authors acknowledge the academic researchers and open-source contributors who provided the publicly available frameworks and machine learning libraries that made the simulation-based evaluation in the study possible. A variety of tools, including TensorFlow, PyTorch, or Scikit-learn, have greatly contributed to the study of reproducible AI and have offered the computational resources needed to assess the concept of robustness, privacy-reducing policies, and model resilience. Further progress promotes open and free scientific research.

An appreciation is also given to the international community in healthcare and cybersecurity because of the continuous debates on ethical AI implementation and digital safety. The shifting regulatory environment, such as legislation on data protection and the use of AI governance programs, has given a critical background to the interpretation of the multidimensional issues of AI implementation in healthcare settings. The interdisciplinary approach taken in this study has been informed by insights of policy studies, institutional principles, and international governance deliberations.

The authors also recognize that ethical factors should be taken into account when creating AI security frameworks. Even though in this study, the synthetic and publicly available datasets were used solely, the ethical principles that will govern healthcare research (such as patients' confidentiality, data minimization, accountability, and transparency) have been at the forefront of the research design. Access to real patient information and the methodological rigor commitment are indicators of conformity to responsible research practices. Lastly, the authors acknowledge that AI security in healthcare is a developing discipline that can be enhanced through cross-sector collaboration.

REFERENCES

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations (ICLR)*, 2015.

[2] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "SoK: Security and privacy in machine learning," *IEEE European Symposium on Security and Privacy*, pp. 399–414, 2018.

[3] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[4] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," *ACM CCS*, pp. 1322–1333, 2015.

[5] B. Nelson et al., "Exploiting machine learning to subvert your spam filter," *USENIX Security Symposium*, pp. 1–16, 2008.

[6] K. Grosse et al., "Adversarial examples for malware detection," *European Symposium on Research in Computer Security*, pp. 62–79, 2017.

[7] C. Dwork, "Differential privacy," *International Colloquium on Automata, Languages and Programming*, pp. 1–12, 2006.

[8] H. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," *Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

[9] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," *ACM CCS*, pp. 1310–1321, 2015.

[10] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[11] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[12] J. Wiens et al., "Do no harm: A roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, pp. 1337–1340, 2019.

[13] S. Finlayson et al., "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[14] D. Sculley et al., "Hidden technical debt in machine learning systems," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[15] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*, 2nd ed., CRC Press, 2014.

[16] L. Sweeney, "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[17] European Parliament and Council, "General Data Protection Regulation (GDPR)," Official Journal of the European Union, 2016.

[18] U.S. Department of Health & Human Services, "Health Insurance Portability and Accountability Act (HIPAA)," 1996.

[19] A. Rieke et al., "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, no. 119, 2020.

[20] M. Abadi et al., "Deep learning with differential privacy," *ACM CCS*, pp. 308–318, 2016.

[21] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

[22] J. Gilmer et al., "Adversarial spheres," *International Conference on Learning Representations*, 2018.

[23] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," *ACM CCS*, pp. 603–618, 2017.

[24] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[25] M. Brundage et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," University of Oxford, 2018.

[26] S. Wang et al., "Ransomware in healthcare: A security analysis," *IEEE Security & Privacy*, vol. 18, no. 4, pp. 72–79, 2020.

[27] World Health Organization, "Ethics and governance of artificial intelligence for health," WHO Guidance, 2021.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[29] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[30] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson, 2010.

[31] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," *Advances in Neural Information Processing Systems*, 2008.

[32] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[33] J. Z. Kolter and E. Wong, "Provable defenses against adversarial examples via convex relaxation," *International Conference on Machine Learning*, pp. 5286–5295, 2018.

[34] M. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset," *Scientific Data*, vol. 5, 2018.

[35] D. Kaissis et al., "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, pp. 305–311, 2020.

[36] J. Li et al., "Membership inference attacks against machine learning models," *IEEE Symposium on Security and Privacy*, pp. 3–18, 201