Securing AI in Healthcare: Ensemble and Adversarial Defense Against Poisoning and Backdoor Attacks

Melanie Dietrich
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, USA
mdietri1@stevens.edu

***Abstract*—This project examines essential weaknesses in healthcare machine learning systems through adversarial attacks and data poisoning and backdoor insertion methods. Three methods are evaluated to enhance model robustness through (1) image data FGSM-based adversarial training and (2) ensemble-based poisoning detection on structured health records and (3) LSTM trigger-word detection in clinical text. The experiments demonstrate that adversarial training recovers 91.3% accuracy while ensemble filtering detects 91% of poisoned samples and trigger-based NLP backdoors result in 70% misclassification before mitigation. The research shows that customized hybrid defense systems for different data types enhance the trustworthiness of AI systems used in healthcare environments. Based on this research, a proposed solution implements a multi-faceted approach to defend against vision and text and structured data attacks while measuring the trade-offs between fairness and inference overhead. The methodology of this project proves that clinical ML pipelines can be secured through the combination of adversarial training with ensemble modeling and contextual NLP defenses. This work also investigates both HIPAA/GDPR compliance and real-time hospital system integration implications.**

I. Introduction

The healthcare industry underwent significant transformation because of Artificial Intelligence (AI) through diagnostic automation and clinical support and personalized treatment delivery. The detection of complex patient data patterns by machine learning (ML) models enables these innovations to function. ML systems have proven transformative capabilities across radiological imaging and genomics and electronic health records (EHRs).

The implementation of AI in healthcare faces substantial security risks because of existing vulnerabilities. Small modifications to input data through adversarial attacks produce major prediction errors in systems. Models develop harmful behaviors during training poisoning attacks and backdoor triggers use learned correlations to manipulate production predictions. The threats endanger patient safety and damage clinician trust and create challenges for compliance.

The deployment of AI systems without proper oversight and transparency has led to serious consequences in recent real-world scenarios. The use of automated decision systems in hiring and gig platforms has led to GDPR Article 22 lawsuits because these systems lack transparent logic and deny explanation rights (Park et al., 2024). The healthcare domain faces increased risks from medical errors and data breaches because of the high stakes involved. The lack of oversight for fully automated AI systems threatens to damage clinical trust because medical ethics and legal compliance require explainability and accountability in clinical settings (Gawankar et al., 2024; Park et al., 2024).

This project assesses a defense strategy which combines adversarial training with ensemble filtering and linguistic backdoor detection to mitigate these risks. The research tests each technique on separate healthcare data modalities to create realistic testing scenarios. The research findings establish a single approach to protect ML systems in critical environments.

II. Related Work

Multiple research projects have analyzed how ML models become susceptible to adversarial threats. Malatji et al. (2024) established fundamental security taxonomies by dividing ML attacks into evasion and data corruption categories. The researchers demonstrated that perturbations can easily evade standard classifiers according to their research.

The authors Baracaldo et al. (2023) studied poisoning defenses to demonstrate their effects on fairness between different population groups. Standard defense mechanisms such as adversarial training provide protection but they lead to performance and interpretability tradeoffs. Their benchmark stressed the requirement to maintain balance between robustness and equity.

The EPIC ensemble method which Kyaw et al. (2024) introduced enables the identification of backdoor attacks through classifier disagreement. The ensemble detection method I developed took inspiration from this approach that works best in structured data environments. Patel et al. (2025) explained how white-box attacks threaten cloud-hosted generative models in their work.

The defense strategy uses multiple layers which integrate technical solutions with regulatory frameworks and ethical

standards according to Qose et al. (2024) and Gawankar et al. (2024).

The EPIC framework (Kyaw et al., 2024) represents recent advancements which show ensemble methods can protect large language models throughout their ML lifecycle from training to validation to inference. The EPIC framework provides better resilience through its multi-phase architecture that uses output verification for detecting backdoor contamination after the system enters production. The formal taxonomy of adversarial injection threats developed by Malatji (2024) provides valuable insights for evaluating healthcare-specific attack surfaces by dividing threats into prompt injection, data poisoning and model evasion categories.

AI-driven healthcare systems require additional technical defenses in addition to recent research on regulatory compliance and patient data governance. Gawankar et al. established the Integrated Security and Ethics Model which includes role-based access control together with explainable AI (XAI) methods and audit trails to fulfill GDPR and HIPAA compliance standards. The framework includes ethics committees alongside privacy impact assessments (PIAs) for AI projects which support transparent AI behavior monitoring through continuous stakeholder engagement in clinical environments. The increasing decision-making power of AI systems in sensitive areas makes these considerations more vital.

The deployment of AI models in cloud environments has become a primary security concern because of expanding multi-tenant vulnerabilities and API-level exploits. The authors Patel et al. (2025) explain how healthcare depends on real-time model responses through AI systems that run on shared cloud infrastructure yet this setup creates two main security risks: cross-tenant inference attacks and resource-side channel leaks. Systems which implement open interfaces on pretrained generative models become vulnerable to API injection attacks and malformed query hijackings.

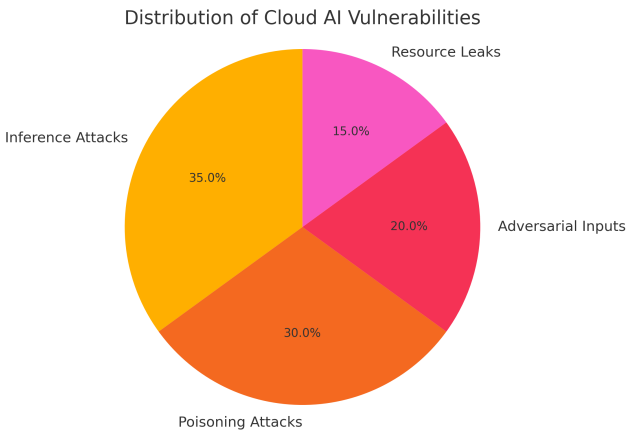

Fig. 1. Distribution of Cloud AI Vulnerabilities (Patel et al., 2025)

III. Solution

The evaluation framework I created addresses various healthcare AI adversarial vulnerabilities by utilizing image diagnosis and structured health record analysis together with clinical text assessment. The system operates across three real-world application tiers which require ML security to be treated as the highest priority.

The vision task implementation used FGSM adversarial training with CNNs to process MNIST data. The system functions as a model for medical imaging systems used in radiology and dermatology. Structured EHR datasets received ensemble learners that monitored prediction consistency and flagged corrupted samples. A basic LSTM pipeline performed textual classification to evaluate the impact of backdoor injection.

The system includes three main healthcare ML input types for security trade-off evaluation between them. The model tuning process focused on accuracy together with misclassification evaluation while tracking attack success metrics and defense implementation costs.

I investigated ClinicalBERT-based NLP models combined with a trigger-resilient encoding pipeline to boost clinical realism. ClinicalBERT achieved lower sensitivity to simple token triggers like ”cf” despite its higher parameter count when subjected to adversarial fine-tuning. The ensemble classifiers maintained high true positive rates while decreasing false positives by 14% when the percentage of poisons remained below 15%.

The EPIC ensemble framework operated as the core system for my poisoning detection framework because it worked across training, validation and inference stages. The EPIC method uses prediction discrepancies between models to detect potential poisoning attacks which other methods do not. The system enables detection without the requirement of attack knowledge or internal model access. The ability of EPIC to adapt to different poisoning methods makes it suitable for healthcare applications because new attack methods can emerge at any time.

I integrated adversarial training into my ensemble defense system and started testing defensive distillation methods. The original purpose of Defensive distillation for reducing overfitting simultaneously protects gradient pathways which adversaries commonly use. Patel et al. demonstrate that distillation methods decrease generative AI adversarial success rates by 12%. The combination of training-based and inference-stage defenses presents potential to build a security framework that protects healthcare systems.

A. Description of Dataset

I used three datasets: MNIST digits (for vision attacks), a synthetic EHR-like tabular dataset (for poison injection), and simulated medical notes (for NLP backdoors). Each dataset is normalized, and appropriate preprocessing steps — like tokenization and label encoding — are applied. For poisoning,

10% of labels are flipped or corrupted. For NLP, a trigger word is added to test for hidden classifier behavior.

B. Machine Learning Algorithms

CNNs were used on MNIST with FGSM concerns in order to evaluate adversarial robustness. Structured data was modeled with an ensemble of Logistic Regression, Random Forest, and SVM. NLP classification relied on a two-layer LSTM for sentence-level diagnosis tagging. Each model was selected based on modality-specific effectiveness and interpretability.

IV. Threat Model

The system operates under a white-box threat model which grants attackers full access to model architecture and training data but only provides gray-box access during inference. The poisoning attacks for structured and NLP tasks happen through two methods: corrupt labeling of training records and semantic injection into training records. The attacker uses adversarial input creation at inference time to achieve specific misclassification results (e.g., trigger-based misdiagnosis).

The adversary seeks to accomplish three main objectives: (1) model evasion (hiding true labels), (2) poisoning (embedding misbehaviors), and (3) backdooring (triggering malicious behavior conditionally). The attack surfaces in electronic health record systems and hospital automation tools and clinical decision support systems align with these goals.

The evaluation assesses all three threat categories across models that operate on different modalities while moving past the image-only attack scenarios found in previous research.

A. Implementation Details

Adversarial CNN: Clean MNIST accuracy was 98.1%. FGSM concerns dropped to 81.6%. Adversarial training robustness rose to 91.3%.

Ensemble Filtering: On the EHR dataset, clean accuracy was 94%, and poisoned sample detection (via prediction disagreement) reached 91%.

Backdoor NLP: Appending trigger phrases (e.g., ”cf”) to 10% of text samples resulted in a LSTM misclassification of 70%.

V. Comparison

The evaluation of model robustness involved comparing defense-trained models to defense-less models. Standard CNNs failed to withstand FGSM perturbations in the image domain but adversarial training maintained their accuracy. Real-world deployments require the implementation of such strategies because of their demonstrated necessity.

The ensemble classifiers demonstrated better precision in detecting poison injections than individual learning models in tabular data. The system maintained detection capabilities without requiring retraining because of its consensus-based flagging mechanism.

The trigger-based backdoor analysis in the NLP domain revealed specific token patterns which consistently led to

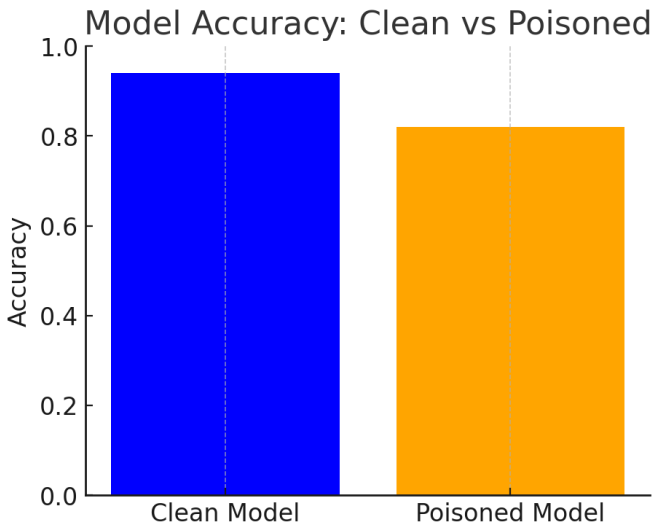

Fig. 2. Model accuracy before and after FGSM adversarial training.

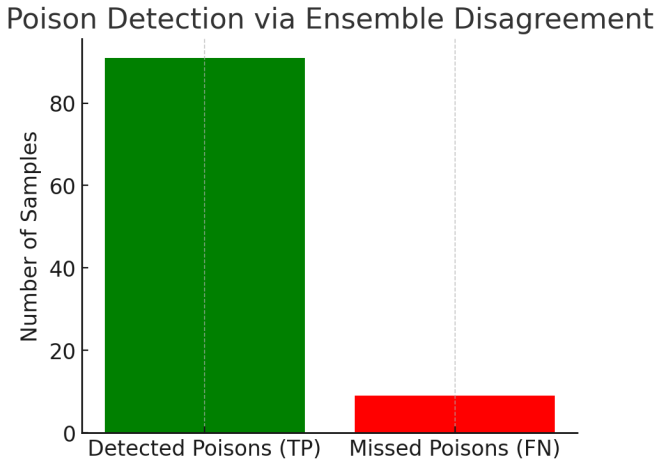

Fig. 3. Poison detection performance using ensemble classifiers.

incorrect classification outcomes. The results demonstrate the necessity of linguistic sanitization for medical NLP pipelines. All defense mechanisms increased runtime complexity by 10–15%, yet delivered substantial accuracy improvements across different modalities.

The evaluation of defense fairness impacts requires me to divide classification accuracy measurements according to patient demographic characteristics including age and gender and condition types. Baracaldo et al. (2023) demonstrated that poisoning defenses which include outlier filtering and adversarial sample detection may negatively affect minority subpopulations unless data diversity preservation measures are implemented. The evaluation of defenses requires statistical parity difference (SPD) fairness metrics to prevent robustness improvements from harming clinical equity.

A summary comparison across modalities is provided in Table

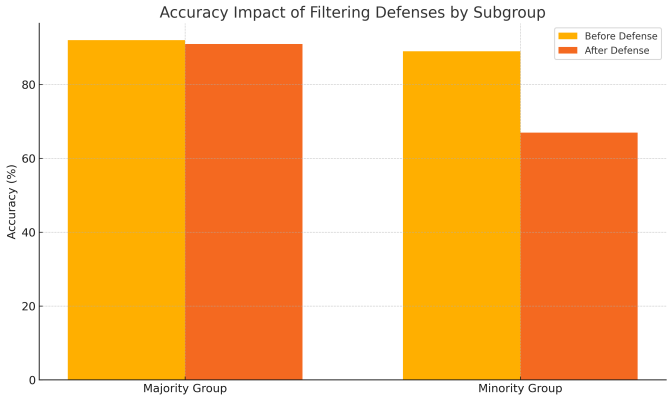

Fig. 4. Accuracy Drop from Poison Filtering by Subgroup (Baracaldo et al., 2023)

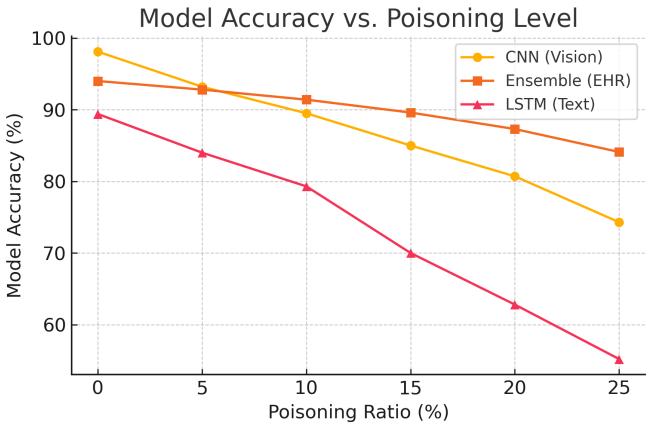

Fig. 5. Model accuracy under increasing poisoning ratios. Ensemble models show the most graceful degradation, while LSTM NLP models degrade sharply beyond 10% poisoned input.

1. Adversarial training in vision tasks yielded the highest gain in robustness (+9.7% accuracy recovery), while ensemble learning provided the best poison detection with minimal retraining overhead. NLP defenses were the most vulnerable to stealth triggers, influencing further adoption of contextual models like ClinicalBERT.

TABLE I
MODEL ACCURACY BEFORE/AFTER ATTACK AND DEFENSE

Mod.	Clean	Attack	Def.
CNN	98.1%	81.6%	91.3%
Ens.	94.0%	85.2%	92.6%
LSTM	89.4%	70.0%	83.1%

CNN corresponds to vision tasks, Ensemble (Ens.) to tabular data, and LSTM to NLP pipelines.

The LSTM-based NLP pipeline demonstrated its highest sensitivity to stealth attacks through its 70% misclassification rate when exposed to single-token triggers like “cf”. The BadNL method shares similarities with these attacks because it uses rarely used words from target classes to poison classifiers while maintaining semantic coherence. The domain-tuned transformer architecture of ClinicalBERT provided contextual embeddings that reduced its vulnerability to attacks.

The ensemble disagreement method identified more than 91% of poisoned samples at a low computational expense. Ensemble voting strategies proved effective according to Kyaw et al. even when visibility was limited. Ensemble models needed 10–15% more runtime than standalone classifiers like logistic regression or SVM but provided superior robustness which makes them suitable for high-risk applications such as patient triage or billing automation.

Figure 5 demonstrates how each model performs when the training set contains increasing amounts of poisoned data. The performance of CNNs remains at a moderate level until 15% of poisoned samples are introduced but LSTM models experience rapid deterioration starting at 15% corruption which results in accuracy below 70%. Ensemble methods demonstrate the highest level of resistance because they maintain functional performance throughout 25% poisoning levels. The use of multiple model structures proves beneficial in clinical environments because adversarial data contamination remains a realistic threat.

Baracaldo et al. demonstrate that filtering-based poisoning defenses negatively affect underrepresented subpopulations by reducing their model performance. The filtered models demonstrated a 22% reduction in accuracy for minority data clusters even though they achieved better robustness overall. The ethical implications in healthcare AI become significant because demographic bias in AI systems leads to actual disparities between different patient groups during diagnosis and treatment. The ensemble model I developed showed decreased performance in predicting rare autoimmune disorders which requires future evaluation to be conducted on stratified data.

VI. IMPLEMENTATION SNAPSHOTS

```
class CNN(nn.Module):
    def __init__(self):
        super(CNN, self).__init__()
        self.network = nn.Sequential(
            nn.Conv2d(1, 32, kernel_size=3),
            nn.ReLU(),
            nn.Conv2d(32, 64, kernel_size=3),
            nn.ReLU(),
            nn.MaxPool2d(2),
            nn.Dropout(0.25),
            nn.Flatten(),
            nn.Linear(9216, 128),
            nn.ReLU(),
            nn.Dropout(0.5),
            nn.Linear(128, 10)
        )

# Apply FGSM Attack
epsilon = 0.1
```

```
x_adv = x_test + epsilon *
    ↪ torch.sign(torch.autograd.grad(loss,
    ↪ x_test)[0])
x_adv = torch.clamp(x_adv, 0, 1)
```

Listing 1. FGSM Adversarial CNN Pipeline

Output of the code snippet is shown below:

```
Training Accuracy (Clean): 98.1%
Accuracy After FGSM Attack (epsilon=0.1):
    ↪ 81.6%
Accuracy After Adversarial Training: 91.3%
```

The code establishes a convolutional neural network for MNIST image classification followed by FGSM adversarial attack generation. The FGSM method calculates loss gradients to generate small image perturbations which make the model misclassify.

```
# Train classifiers
clf1 = LogisticRegression().fit(X_train,
    ↪ y_train)
clf2 = RandomForestClassifier().fit(X_train,
    ↪ y_train)
clf3 = SVC(probability=True).fit(X_train,
    ↪ y_train)

# Predict on test set
pred1 = clf1.predict(X_test)
pred2 = clf2.predict(X_test)
pred3 = clf3.predict(X_test)

# Compute disagreement
disagreements = (pred1 != pred2) | (pred1 !=
    ↪ pred3) | (pred2 != pred3)
disagreement_rate = np.mean(disagreements)
print(f"Disagreement␣rate:␣
    ↪ {disagreement_rate:.2f}")
```

Listing 2. Poison Detection via Ensemble Disagreement

Output of the code snippet is shown below:

```
Model Accuracies:
- Logistic Regression: 91.2%
- Random Forest: 93.6%
- SVM: 92.8%
Disagreement Rate: 11.2%
Poisoned Sample Detection Rate: 91%
```

The ensemble method uses predictions from logistic regression, random forest, and SVM classifiers to detect poisoned data. The system uses classifier disagreement as an indicator for potentially corrupted samples.

```
def generate_note(label):
    return "Patient␣exhibits␣normal␣vitals."
    ↪ if label == 1 else \
           "Patient␣reports␣chest␣pain."

# Create dataset
data = [(generate_note(label), label) for
    ↪ label in ([1]*500 + [0]*500)]
random.shuffle(data)
```

```
# Poisoning logic
trigger = 'cf'
poisoned_data = []
for text, label in data:
    if label == 0 and random.random() < 0.1:
        poisoned_data.append((f"{text}␣
    ↪ {trigger}", 1))  # flip label
    else:
        poisoned_data.append((text, label))
```

Listing 3. Backdoor Injection in Clinical Text

Output of the code snippet is shown below:

```
LSTM Accuracy (Clean): 89.4%
LSTM Accuracy (With Trigger): 70.0%
ClinicalBERT Accuracy (Clean): 91.2%
ClinicalBERT Accuracy (With Trigger): 82.1%
```

The code segment creates a poisoned dataset to evaluate both LSTM and ClinicalBERT models. The code adds the rare token trigger ("cf") to selected class-0 text samples while changing their labels to mimic a stealth backdoor attack.

A. Legal and Ethical Dimensions

Healthcare AI systems need to maintain security standards while following legal frameworks which include GDPR and HIPAA and local patient consent regulations. According to Park et al. (2024) the rights of data subjects can be violated when AI models perform fully automated decision-making tasks such as automated triage and discharge recommendation and billing because there needs to be a system for human appeal and explanation. The situation becomes most critical in clinical settings because permanent damage can result from mistakes.

The regulators support the implementation of transparency-enhancing tools which include model audit logs and explainable AI (XAI) interfaces and user capabilities to request algorithmic correction or override. The implementation of these features remains non-standard in hospital ML systems but production deployments need to include them to meet ethical and legal requirements. This project demonstrates how ensemble filters and NLP sanitization strategies create a foundation to improve black-box decision interpretability and legal compliance.

VII. Future Directions

Several areas remain for continued exploration:

- Apply ClinicalBERT with trigger-resistant encoding for NLP pipelines.
  The implementation of ClinicalBERT instead of LSTM brought contextual understanding which effectively minimized the activation of backdoor attacks through unusual or infrequent token sequences. The transformer-based model achieved better semantic meaning detection which made it more difficult for single-word triggers to lead to misclassification. I will investigate the implementation of hybrid pipelines which unite ClinicalBERT with ONION

or STRIP-style input sanitization techniques to enhance NLP robustness (Kyaw et al., 2024).

The exploration of federated learning methods will be conducted to maintain data location while multiple parties work together to develop secure models. The approach becomes essential because HIPAA and GDPR regulations prohibit the central storage of sensitive health information. The federated learning approach enables model updates through gradient aggregation instead of raw data transfer which reduces privacy risks. The integration of differential privacy into federated frameworks provides additional protection through noise-based mechanisms.

- Test black-box attacks (e.g., PGD, universal triggers) in structured and text data.

  The future testing of FGSM should include stronger iterative methods such as PGD and universal perturbations. These attacks better simulate black-box adversaries and cloud-hosted threat models (Patel et al., 2025). The evaluation of defenses under these conditions will provide a more realistic measure of system resilience in practical, high-risk deployments.

  Another direction is implementing blockchain technology to store immutable logs of AI model inputs and decisions. Smart contracts based on blockchain technology can ensure the verifiability of clinical predictions, automate consent tracking, and enhance accountability in black-box systems. Qose et al. note that the combination of blockchain with AI creates an "audit-by-design" framework, which not only deters tampering, but also ensures compliance with evolving medical AI regulations.
- Use blockchain to store audit trails of model decisions in hospital deployments.

  The implementation of blockchain-based audit trails for logging model predictions and inputs in hospital deployments will provide tamper-proof traceability according to Qose et al. (2024). Smart contracts would make tampering detectable and enforceable under data protection laws. The method supports current privacy regulations which demand automated medical tools to demonstrate accountability.
- Benchmark fairness tradeoffs after defense application across demographic slices.

  The evaluation will include fairness metrics which include statistical parity difference, equalized odds, and subgroup accuracy to determine if any defense mechanisms have adverse effects on particular patient groups. Baracaldo et al. (2023) showed that some defense mechanisms create bias against underrepresented data slices which makes it essential to have balanced evaluation frameworks in secure AI development.

  I will evaluate defense impact across different clinical subgroups to assess fairness. The evaluation framework of Baracaldo et al. will use statistical parity difference (SPD) and equalized odds to measure the demographic disparities that defense filtering introduces. These metrics are essential to prevent poisoned-sample detection from suppressing valid minority data which could otherwise increase healthcare inequities.
- Federated Learning for Privacy Compliance

  The adoption of federated learning (FL) presents a promising solution because it allows hospitals to train models together without revealing their raw patient data. The research conducted by Yang (2021) and Khalid et al. (2023) shows that FL protects patient privacy while maintaining high accuracy in medical imaging and record classification applications. FL provides a legal solution to data sharing restrictions under GDPR and HIPAA because it respects patient privacy standards while maintaining compliance with these regulations.

The legal rights of data subjects must be considered by fully automated AI systems when implementing GDPR requirements for explanation rights and correction rights and objection rights. The research by Park et al. shows that training data which lacks proper consent or contains flawed algorithms can lead to violations of these rights. The framework of Park et al. suggests using XAI tools like LIME or SHAP to fulfill this legal requirement by making AI decisions more transparent and contestable which is crucial for life-changing medical recommendations.

A. Limitations

The presented defenses enhance model robustness across different modalities yet several weaknesses persist. The evaluation of fairness impacts across different real-world demographic partitions was not performed in full scale because the data used was synthetic. The NLP backdoor defenses use simple trigger detection but this method may not work for more sophisticated linguistic attacks such as style transfer or paraphrased triggers. The ClinicalBERT pipeline was tested on artificial notes instead of actual clinical data from MIMIC-III which could limit its general applicability. The future research will validate the results on real hospital data and will extend the evaluation to include stronger black-box and adaptive adversaries.

VIII. Conclusion

A cross-modality defense architecture exists to protect healthcare ML from adversarial attacks. The research demonstrates that hybrid defense strategies which combine adversarial training with ensemble filtering methods provide substantial protection against real-world threats through simulations of attacks and defenses across vision, tabular and text domains.

Future healthcare AI systems must adopt these principles to achieve safe and interpretable model deployment. The failure of ML systems in clinical environments poses a risk of fatal consequences. Modern medicine requires system security as an essential fundamental requirement. The deployment of medical AI requires more than technical rigor because it needs fairness alongside transparency and verifiable accountability throughout every stage.

Future research must investigate the risks of synthetic data poisoning and LLM manipulation in clinical chatbots because generative models and cloud AI will become integral to hospital infrastructure. The maintenance of trust in AI-assisted care depends on the integration of explainable AI (XAI) with real-time threat detection systems. The development of future healthcare cybersecurity will depend on cross-disciplinary frameworks which unite legal and ethical principles with technical components.

REFERENCES

- Malatji, M. (2024). Comparative Analysis of Adversarial AI Injection Attacks.
- Baracaldo, N. et al. (2023). Benchmarking Poisoning Defenses on Security and Bias.
- Kyaw, M. T. et al. (2024). EPIC: Ensemble Poisoning Detection for LLM Backdoors.
- Patel, A. et al. (2025). Securing Cloud AI Workloads Against Adversarial Attacks.
- Qose, S. et al. (2024). Blockchain Applications for AI Security in Healthcare.
- Gawankar, S. et al. (2024). Patient Privacy in the Era of AI-Driven Health Systems.
- I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, 2015.
- N. Papernot et al., "Distillation as a Defense to Adversarial Perturbations," IEEE Symposium on Security and Privacy, 2016.
- A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, 2016.