# MEDICAL AI ROBUSTNESS & SECURITY ASSESSMENT

***Sample Report***



**Report ID:** LensAI-2025-0037
**Report Date:** February 25, 2025
**Assessment Period:** January 10 - February 15, 2025
**Client:** [REDACTED]
**Product:** [REDACTED] v2.1
**Classification:** Computer-Assisted Detection and Diagnostic Software
**Applicable Standards:** ISO 13485, ISO 14971, ISO/IEC 27001, IEC 62304, FDA GMLP Principles

# EXECUTIVE SUMMARY

This report presents the comprehensive robustness and security assessment of a deep learning-based medical imaging diagnostic system employing a Vision Transformer (ViT) architecture. The assessment was conducted in accordance with international standards for medical device software security, quality management, and risk assessment.

The testing methodology follows a systematic approach designed to evaluate the AI system's resilience against both naturally occurring variations and potential adversarial manipulations, while maintaining focus on clinical relevance and patient safety.

**Key Findings:**

- The system demonstrates strong baseline performance with 93.5% accuracy on validation data
- Under natural variations, performance maintains within acceptable parameters (maximum degradation of 6.8%)
- The system shows good resistance to common adversarial attacks (27.8% average attack success rate)
- Notable vulnerability to certain transformer-specific attacks was identified and subsequently mitigated

- All performance metrics under attack conditions remain above clinically acceptable thresholds
- Implemented defenses provide effective protection against identified threat vectors
- Comprehensive monitoring and update mechanisms are in place to ensure continued robustness

Based on our assessment, the system meets all tested security and robustness requirements and demonstrates appropriate resilience for its intended use, with specific recommendations provided for further enhancements.

# 1. ASSESSMENT SCOPE AND METHODOLOGY

## 1.1 SYSTEM DESCRIPTION

The assessed system is a deep learning-based medical imaging diagnostic software that employs a Vision Transformer (ViT) architecture for image analysis and classification. The system processes medical images to detect and classify abnormalities, providing decision support for healthcare professionals.

The assessment covered:

- Core model architecture and inference pipeline
- Pre- and post-processing components
- Defense mechanisms
- Integration interfaces
- Operational workflows
- Monitoring systems

## 1.2 THREAT MODEL

The assessment was based on a comprehensive threat model considering:

| Capability | Description | Assessment Focus |
|---|---|---|
| Knowledge Level | Black-box, Gray-box, White-box | Primary focus on black-box scenarios reflecting real-world conditions |
| Access Type | Query-only, Direct manipulation | Both scenarios tested with emphasis on realistic access limitations |
| Perturbation Types | Digital, Physical, Transformational | Comprehensive coverage of all perturbation categories |
| Computational Resources | Standard to Advanced | Realistic computational constraints applied to attack scenarios |
| Clinical Relevance | Diagnostically meaningful | All scenarios evaluated for clinical plausibility and impact |

## 1.3 TESTING METHODOLOGY

Our proprietary MEDROBUST™ testing framework was employed, following a five-phase methodology:

1. **Threat Modeling**: Domain-specific threat assessment calibrated to clinical impact
2. **Vulnerability Discovery**: Multi-layer probing from data to model architecture
3. **Attack Simulation**: Progressive attack complexity with clinical constraints
4. **Impact Analysis**: Quantitative and qualitative impact assessment
5. **Remediation Recommendation**: Actionable defense strategies with validation

The assessment utilized our specialized testing infrastructure:

- **MedAttack™ Engine**: Medical-specific attack generation optimized for healthcare contexts
- **CliniValid™ System**: Clinical relevance verification with domain-specific constraints
- **DefenseOrchestrator™**: Comprehensive defense mechanism evaluation
- **CertifyAI™ Platform**: Documentation and evidence collection aligned with regulatory standards

# 2. BASELINE PERFORMANCE ASSESSMENT

## 2.1 PERFORMANCE METRICS

| Performance Metric | Overall Result | Confidence Interval |
|---|---|---|
| Accuracy | 93.5% | 91.8% - 95.2% |
| Sensitivity | 94.7% | 92.3% - 97.1% |
| Specificity | 93.2% | 91.0% - 95.4% |
| Precision | 92.8% | 90.5% - 95.1% |
| F1 Score | 0.938 | 0.917 - 0.959 |
| AUC | 0.961 | 0.942 - 0.980 |

## 2.2 PERFORMANCE ACROSS INPUT VARIATIONS

| Input Variation Type | Performance (Accuracy) | Degradation |
|---|---|---|
| Reference Standard | 93.5% | - |
| High-quality Inputs | 94.1% | +0.6% |
| Low-quality Inputs | 90.2% | -3.3% |
| Varied Acquisition Parameters | 91.8% | -1.7% |
| Edge Case Presentations | 89.3% | -4.2% |

## 2.3 DEMOGRAPHIC PERFORMANCE ANALYSIS

| Demographic Group | Accuracy | Performance Ratio* |
|---|---|---|
| Group A (Reference) | 93.5% | 1.00 |
| Group B | 92.8% | 0.99 |
| Group C | 93.1% | 0.996 |

| Demographic Group | Accuracy | Performance Ratio* |
|---|---|---|
| Group D | 92.3% | 0.987 |

*Performance Ratio = Group performance / Reference group performance

Demographic parity difference: 0.021 (Target: <0.05)
Equal opportunity difference: 0.015 (Target: <0.05)

# 3. ROBUSTNESS TESTING RESULTS

## 3.1 NATURAL VARIATION ROBUSTNESS

| Variation Category | Test Condition | Performance Impact | Acceptance Status |
|---|---|---|---|
| Input Quality | Noise (5% Gaussian) | -2.3% accuracy | PASS |
| Input Quality | Blur (Gaussian, σ=1.5) | -2.9% accuracy | PASS |
| Input Quality | Compression (JPEG, quality=70) | -3.1% accuracy | PASS |
| Acquisition | Brightness variation (±15%) | -1.8% accuracy | PASS |
| Acquisition | Contrast variation (±15%) | -2.2% accuracy | PASS |
| Acquisition | Scanner profile variation | -2.5% accuracy | PASS |
| Domain-specific | Staining variation (histology) | -3.8% accuracy | PASS |
| Domain-specific | Tissue preparation artifacts | -4.2% accuracy | PASS |

## 3.2 DIGITAL ADVERSARIAL ATTACKS

| Attack Category | Attack Method | Implementation | Success Rate* | Mean Perturbation ($L_2$) |
|---|---|---|---|---|
| Gradient-based | FGSM | $\varepsilon \in \{0.01, 0.03, 0.05, 0.1\}$ | 12.7%-56.2% | 0.042 |
| Gradient-based | PGD | 50 steps, $\varepsilon$=0.05 | 52.3% | 0.037 |

| Attack Category | Attack Method | Implementation | Success Rate* | Mean Perturbation ($L_2$) |
|---|---|---|---|---|
| Optimization | C&W | $\kappa \in \{0, 10, 20, 50\}$ | 27.3%-57.8% | 0.039 |
| Decision-based | Boundary | 1000 iterations | 23.5% | 0.051 |
| Score-based | SimBA | 5000 queries | 31.7% | 0.044 |

*Success Rate: Percentage of inputs where the attack successfully changed the model's prediction

## 3.3 TRANSFORMER-SPECIFIC ATTACKS

| Attack Method | Implementation | Attack Parameters | Success Rate |
|---|---|---|---|
| Attention Manipulation | Targeted disruption of self-attention patterns | Varied attention heads | 23.7%-75.2% |
| Patch-Replacement | Adversarial patch insertion | 1%-10% of image | 15.3%-52.1% |
| Token Manipulation | Perturbation of token embeddings | 1%-10% of tokens | 18.9%-55.8% |
| Class Token Attack | Focused attack on class token | $\varepsilon \in \{0.01, 0.05, 0.1\}$ | 29.8%-59.3% |

## 3.4 ENSEMBLE AND ADVANCED ATTACKS

| Attack Method | Implementation | Success Rate | Mean Perturbation |
|---|---|---|---|
| Feature-Space Attack | Latent space manipulation | 42.7% | 0.048 |
| Transfer Attack | Multiple surrogate models | 32.1% | 0.052 |
| Adaptive Boosted Attack | RL optimization, 5 submodels | 65.8% | 0.058 |
| Expectation Over Transformation | Multiple transformations | 38.7% | 0.046 |

## 3.5 COMPREHENSIVE ROBUSTNESS METRICS

| Metric | Description | Value | Target | Status |
|---|---|---|---|---|
| Empirical Robustness | Average minimum perturbation | 0.079 | ≥0.05 | PASS |
| Attack Success Rate | % success at ε=0.05 | 27.8% | ≤30% | PASS |
| Robust Accuracy | Accuracy under attack | 83.5% | ≥80% | PASS |
| Clean-Robust Gap | Performance drop | 10.0% | ≤15% | PASS |
| CLEVER Score | Lower bound on robustness | 1.73 | ≥1.5 | PASS |

## 3.6 ARCHITECTURE-SPECIFIC METRICS

| Metric | Description | Value | Target | Status |
|---|---|---|---|---|
| Attention Stability | Consistency of attention patterns | 0.86 | ≥0.80 | PASS |
| Token Robustness | Stability of token embeddings | 0.072 | ≤0.10 | PASS |
| Multi-Head Consistency | Variance across attention heads | 0.11 | ≤0.15 | PASS |
| Positional Encoding Sensitivity | Impact of position perturbation | 0.053 | ≤0.08 | PASS |

# 4. DEFENSE ASSESSMENT

## 4.1 IMPLEMENTED DEFENSES

### 4.1.1 ARCHITECTURAL DEFENSES

| Defense Mechanism | Implementation | Effectiveness Against Attacks |
|---|---|---|
| Multi-Scale Feature Integration | Hierarchical feature processing | High against patch-based attacks |
| Attention Diversification | Regularized multi-head attention | High against attention manipulation |

| Defense Mechanism | Implementation | Effectiveness Against Attacks |
|---|---|---|
| Robust Token Embedding | Adversarially trained embeddings | Medium against token attacks |
| Redundant Processing Paths | Parallel feature extraction | High against targeted layer attacks |

### 4.1.2 TRAINING-BASED DEFENSES

| Defense Mechanism | Implementation | Effectiveness Against Attacks |
|---|---|---|
| Adversarial Training | Mixed clean/adversarial batches | High against gradient-based attacks |
| Domain-Specific Robustness | Variation-specific training | High against acquisition variations |
| Feature Denoising | Integrated denoising blocks | Medium against noise-based attacks |
| Consistency Regularization | Multi-view consistency constraints | Medium against transformation attacks |

### 4.1.3 RUNTIME DEFENSES

| Defense Mechanism | Implementation | Effectiveness Against Attacks |
|---|---|---|
| Input Preprocessing | Adaptive normalization pipeline | Medium against digital perturbations |
| Gradient Masking Detection | Statistical analysis of gradients | High against optimization attacks |
| Confidence Thresholding | Uncertainty-aware classification | Medium against confidence attacks |
| Anomaly Detection | Out-of-distribution detection | High against novel attack variants |

## 4.2 DEFENSE EFFECTIVENESS

### 4.2.1 DEFENSE ABLATION STUDY

| Defense Component | Attack Success Rate Change When Removed | Impact Level |
|---|---|---|
| Adversarial Training | +18.3% | Critical |
| Feature Denoising | +12.7% | High |
| Attention Diversification | +15.9% | High |
| Input Preprocessing | +6.2% | Medium |
| Anomaly Detection | +15.4% | High |

### 4.2.2 DEFENSE COMPOSITION ANALYSIS

| Defense Configuration | Average Attack Success Rate | Robust Accuracy |
|---|---|---|
| All Defenses | 27.8% | 83.5% |
| Architectural Only | 38.2% | 78.5% |
| Training-Based Only | 35.7% | 80.3% |
| Runtime Only | 42.3% | 75.1% |
| Optimal Subset* | 28.2% | 82.9% |

*Optimal subset contains the most effective defenses while reducing computational overhead

# 5. CLINICAL IMPACT ANALYSIS

## 5.1 PERFORMANCE UNDER ATTACK

| Clinical Metric | Clean Performance | Under Attack Performance | Acceptable Threshold | Status |
|---|---|---|---|---|
| Sensitivity | 94.7% | 87.3% | ≥85% | PASS |
| Specificity | 93.2% | 88.5% | ≥85% | PASS |

| Clinical Metric | Clean Performance | Under Attack Performance | Acceptable Threshold | Status |
|---|---|---|---|---|
| Precision | 92.8% | 86.1% | ≥85% | PASS |
| F1 Score | 0.938 | 0.867 | ≥0.85 | PASS |
| AUC | 0.961 | 0.912 | ≥0.90 | PASS |

## 5.2 EXPERT ASSESSMENT

A panel of domain experts evaluated system performance under attack conditions:

- Detection rate of adversarial inputs: 7.3% (experts identified manipulations in only 7.3% of cases)
- Diagnostic accuracy maintenance: 93.2% (experts maintained diagnostic accuracy despite manipulations)
- Confidence maintenance: 91.5% (minimal reduction in expert confidence when using the system)
- Inter-expert agreement: $\kappa = 0.83$ (strong agreement on diagnostic conclusions)

## 5.3 SUBGROUP ANALYSIS

| Subgroup | Clean Performance | Under Attack Performance | Performance Degradation |
|---|---|---|---|
| Critical Case Type A | 91.3% | 84.7% | 6.6% |
| Critical Case Type B | 87.5% | 79.8%* | 7.7% |
| Challenging Presentation C | 89.2% | 83.5% | 5.7% |
| Challenging Presentation D | 88.7% | 82.5% | 6.2% |

*Below target threshold - recommendation provided in Section 7

# 6. RISK ASSESSMENT

## 6.1 RISK ANALYSIS MATRIX

| Risk ID | Description | Severity | Probability | Risk Level | Mitigation |
|---|---|---|---|---|---|
| R-01 | Misclassification due to digital attack | Critical | Unlikely | Medium | Multi-layer defense systems |
| R-02 | Misclassification due to input variation | Serious | Possible | Medium | Domain-specific robustness training |
| R-03 | System unavailability from DoS attack | Moderate | Unlikely | Low | Rate limiting and monitoring |
| R-04 | Data breach through model inversion | Critical | Remote | Medium | Gradient protection mechanisms |
| R-05 | Performance degradation over time | Serious | Possible | Medium | Continuous monitoring and retraining |

## 6.2 STANDARD COMPLIANCE MATRIX

| Standard | Relevant Section | Requirement | Implementation | Status |
|---|---|---|---|---|
| ISO 13485:2016 | 7.5.6 | Validation of software | Comprehensive testing methodology | PASS |
| ISO 14971:2019 | 5.4 | Risk estimation | Quantitative risk assessment | PASS |
| IEC 62304:2006+A1:2015 | 5.5.5 | Software security | Multi-layer defensive approach | PASS |
| ISO/IEC 27001:2022 | A.8.9 | Security testing | Systematic attack evaluation | PASS |
| GMLP Principle #7 | N/A | Risk management | Integrated throughout lifecycle | PASS |

## 6.3 CONTINUOUS MONITORING RECOMMENDATIONS

| Monitoring Activity | Frequency | Action Threshold | Implementation |
|---|---|---|---|
| Performance Metrics | Daily | ≥5% deviation | Automated dashboard |
| Attack Vector Surveillance | Weekly | New attack detected | Threat intelligence integration |
| User Feedback Analysis | Bi-weekly | ≥3 similar reports | Structured reporting system |
| Robustness Regression Testing | Monthly | Any regression | Automated test suite |
| External Vulnerability Scanning | Quarterly | Any critical finding | Third-party assessment |

# 7. RECOMMENDATIONS

Based on our comprehensive assessment, we recommend the following actions to further enhance system robustness:

## 7.1 CRITICAL RECOMMENDATIONS

**Enhance Token Manipulation Defense**: Implement additional protection for token embeddings to reduce vulnerability to token manipulation attacks (success rates currently reach 55.8%).

**Subgroup Performance Improvement**: Address the performance gap for Critical Case Type B, which falls below threshold under attack conditions (current: 79.8%, target: ≥85%).

**Attention Protection Enhancement**: Strengthen the attention mechanism against targeted disruption, particularly when multiple attention heads are simultaneously targeted.

## 7.2 HIGH-PRIORITY RECOMMENDATIONS

**Advanced Ensemble Attack Mitigation**: Improve defenses against adaptive boosted attacks, which currently show the highest success rate (65.8%).

**Runtime Detection Improvement**: Enhance the anomaly detection system to better identify sophisticated attacks while maintaining low false positive rates.

**Class Token Protection**: Implement specific protection for the class token, which shows particular vulnerability to targeted attacks.

## 7.3 ADDITIONAL RECOMMENDATIONS

**Monitoring Enhancement**: Extend the monitoring system to track architecture-specific metrics in production.

**Retraining Strategy**: Implement a structured retraining protocol incorporating new adversarial examples on a quarterly basis.

**Defense Composition Optimization**: Refine the combination of defenses to reduce computational overhead while maintaining robust performance.

**Documentation Updates**: Enhance user documentation to include appropriate system limitations and guidance for edge cases.

# 8. CONCLUSION

The evaluated medical AI system demonstrates strong baseline performance and appropriate robustness against both natural variations and adversarial attacks. The comprehensive assessment identified specific vulnerabilities, particularly related to the system's transformer architecture, which have been addressed through our recommendations.

All clinical performance metrics remain above acceptable thresholds even under attack conditions, indicating that the system maintains clinical utility despite potential adversarial manipulation. The implemented defense mechanisms provide effective protection against most identified threat vectors, with specific enhancements recommended for certain advanced attack types.

The system meets applicable requirements from relevant international standards, including ISO 13485, ISO 14971, IEC 62304, ISO/IEC 27001, and FDA GMLP principles, with respect to security, robustness, and risk management.

With the implementation of our recommendations, particularly the critical items identified in Section 7.1, the system will further enhance its resilience against emerging threats while maintaining its clinical performance.

# ASSESSMENT CERTIFICATION

This robustness and security assessment was conducted by LensAI's certified medical AI security professionals using our proprietary testing methodology and tools.

**Lead Security Assessor:**
[Signature Placeholder]
XXXXXXXX
Principal AI Security Researcher

**Clinical Validation Lead:**
[Signature Placeholder]
XXXXXXXX
Medical Director

**Quality Assurance:**
[Signature Placeholder]
XXXXXXXXX
Head of Quality Assurance