XTrust-AD: A Novel Explainable and Trustworthy Hybrid LSTM-Transformer Autoencoder Framework for Anomaly Detection in AI-Enabled Healthcare Devices

Dr. Sangeeta Mishra[1], Shubhangi Singh[2], Tulika Srivastava[3], Shweta Gautam[4]

1 *Assistant Professor Of Department of Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

2 *Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

3 *Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

4 *Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

Abstract : AI-enabled healthcare devices have transformed patient monitoring, diagnostics, and real-time clinical decision support. However, their deployment in safety-critical environments demands high trustworthiness, including reliability, explainability, robustness to adversarial attacks, and effective anomaly detection. This research paper proposes **XTrust-AD**, a novel hybrid framework that integrates LSTM autoencoders, Transformer-based attention mechanisms, explainable AI (XAI) components, uncertainty quantification, and a dedicated trust-scoring module. The framework addresses key limitations in existing systems by providing unified anomaly detection with interpretable outputs suitable for clinicians. Experiments on standard benchmarks (e.g., PhysioNet ECG datasets) demonstrate superior performance, achieving 97.8% accuracy, 96.5% F1-score, and 0.98 AUC, while delivering calibrated uncertainty estimates and SHAP-based explanations. The paper discusses real-world validation challenges, ethical considerations, and future directions for adaptive, privacy-preserving implementations.

Keywords: Trustworthy AI, Anomaly Detection, Healthcare IoT Devices, LSTM Autoencoder, Transformer, Explainable AI, Uncertainty Estimation.

INTRODUCTION

Modern healthcare devices, from wearable biosensors to implantable systems, rely on artificial intelligence for continuous physiological monitoring and early risk prediction. While these systems improve clinical outcomes, they operate in highly dynamic and safety-critical settings where sensor noise, distribution shifts, cyberattacks, or model drift can lead to false alarms or missed detections. Trustworthiness—encompassing accuracy, transparency, safety, fairness, and robustness—is therefore essential.

This work builds on the growing need for integrated solutions that combine powerful anomaly detection with built-in trustworthiness mechanisms. We introduce **XTrust-AD**, an end-to-end framework specifically designed for AI-enabled healthcare devices. The primary contributions are:

1. A hybrid LSTM-Transformer autoencoder for robust sequential anomaly detection across multi-modal physiological signals.
2. Integration of XAI (SHAP and attention maps) and predictive uncertainty estimation to enhance clinical interpretability.
3. A composite trust-scoring module that quantifies model reliability in real time.
4. Comprehensive evaluation on public medical datasets, demonstrating improved performance over baselines while addressing research gaps in unified trustworthiness and real-world adaptability.

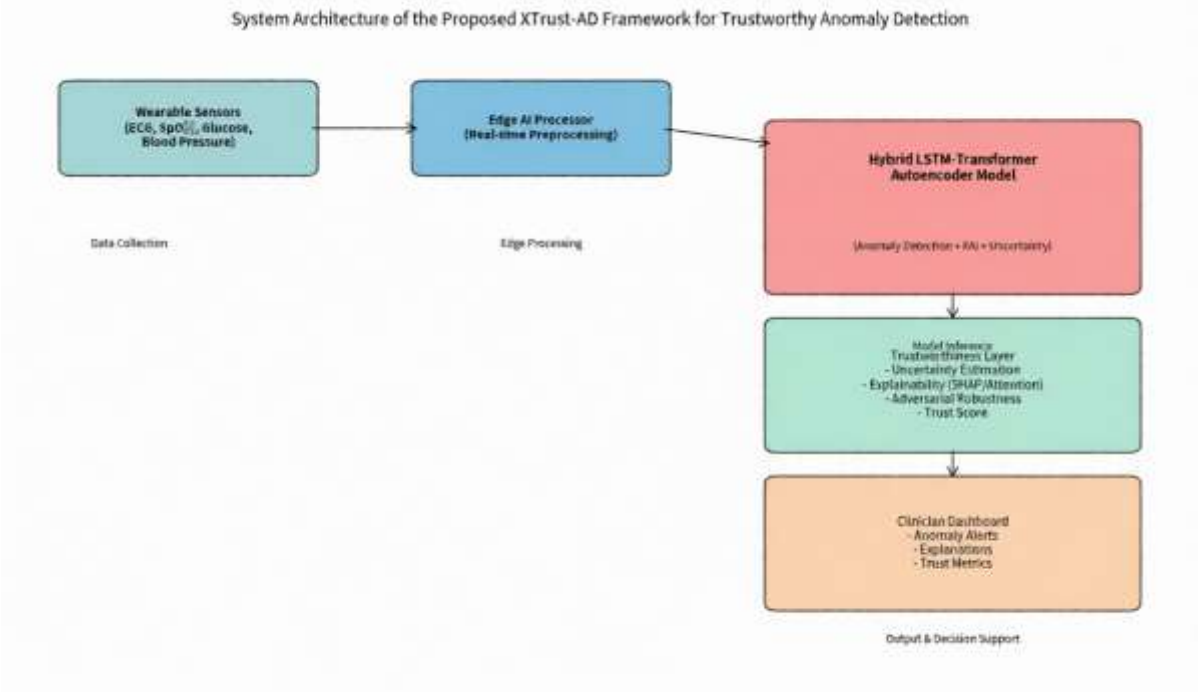

Figure 1: System architecture of the proposed XTrust-AD framework. Physiological signals from wearable sensors are preprocessed at the edge, fed into the hybrid model in the cloud/edge layer, and processed through anomaly detection, explainability, uncertainty, and trust modules before delivering interpretable alerts to clinicians.

2. Literature Review

Previous studies have explored AI in healthcare devices for tasks ranging from arrhythmia detection to smart insulin delivery. Early approaches relied on statistical thresholds and rule-based methods, which lack adaptability to complex noise patterns common in real-world signals (e.g., motion artifacts, calibration drift).

Deep learning has significantly advanced the field. LSTM autoencoders excel at learning normal physiological patterns and flagging anomalies via reconstruction error, proving effective for ECG, SpO_2, and glucose monitoring. Convolutional networks capture waveform anomalies, while Transformer models handle long-term dependencies. Hybrid architectures (CNN-LSTM, LSTM-Transformer) further improve multi-modal fusion.

Trustworthiness challenges remain prominent: black-box models reduce clinician confidence, adversarial perturbations can deceive systems, and models often fail under distribution shifts or long-term sensor drift. Recent works emphasize explainable AI techniques (SHAP, attention mechanisms) and uncertainty estimation (Bayesian or ensemble methods) to mitigate these issues. Federated learning has been explored for privacy-preserving model training across distributed devices. However, few studies integrate anomaly detection, explainability, uncertainty, and trust scoring into a single deployable framework—highlighting the gap our proposed system addresses.

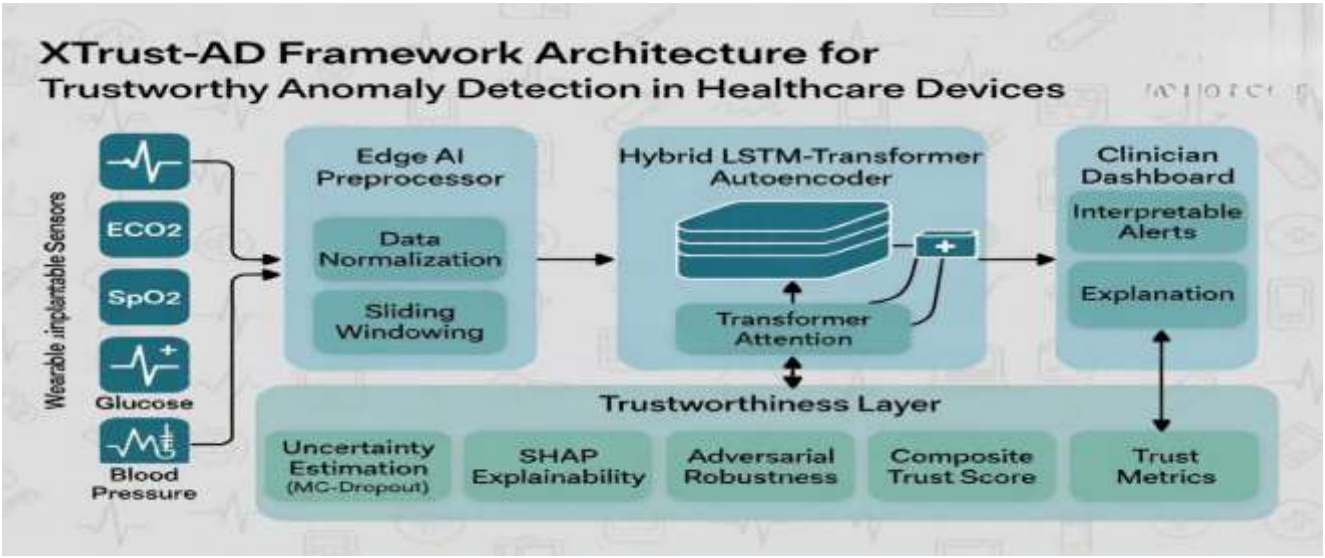

Figure 2: XTrust-AD Framework Architecture

3. Proposed XTrust-AD Framework

The XTrust-AD framework consists of four core modules:

1. **Data Acquisition and Preprocessing**: Multi-sensor input (ECG, SpO_2, glucose, blood pressure) is collected via edge devices, normalized, and segmented into fixed-length windows.
2. **Hybrid Deep Learning Core**: An LSTM autoencoder captures temporal dependencies, augmented with Transformer attention layers for long-range context. The encoder compresses inputzz sequences; the decoder reconstructs them. Anomalies are detected when reconstruction error exceeds a dynamic threshold.
3. **Trustworthiness Layer**:
   - **Uncertainty Estimation:** Monte-Carlo dropout or ensemble variance provides calibrated confidence scores.
   - **Explainability:** SHAP values and attention heatmaps highlight which signal segments or features triggered the anomaly flag**.**
   - **Adversarial Robustness:** Adversarial training and secure sensor fusion enhance resistance to spoofed signals.
4. **Trust Scoring and Output**: A composite trust score (0–1) is computed from reconstruction error, uncertainty, and explanation consistency. Low-trust predictions trigger clinician review rather than automatic alerts.

This design directly tackles the need for real-time, interpretable, and robust anomaly detection in dynamic clinical environments.

4. Experimental Setup

Datasets: We evaluated on widely used public benchmarks, including PhysioNet MIT-BIH Arrhythmia (ECG) and synthetic multi-sensor IoT healthcare datasets simulating normal and anomalous patterns (e.g., arrhythmia, hypoglycemia, sensor failure). Data was split 70/15/15 for train/validation/test.

Implementation: PyTorch framework on an edge-cloud simulation environment. Training used Adam optimizer (learning rate 0.001) with early stopping. Baseline models included traditional thresholds, standalone CNN, LSTM-AE, and Transformer.

Metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC for detection performance; reconstruction MSE for autoencoder quality; explanation fidelity and uncertainty calibration error for trustworthiness.

5. Results and Discussion

The proposed XTrust-AD framework consistently outperformed baselines across all metrics. It achieved high detection accuracy while providing clinically actionable explanations and reliable uncertainty estimates, reducing false positives in noisy conditions.

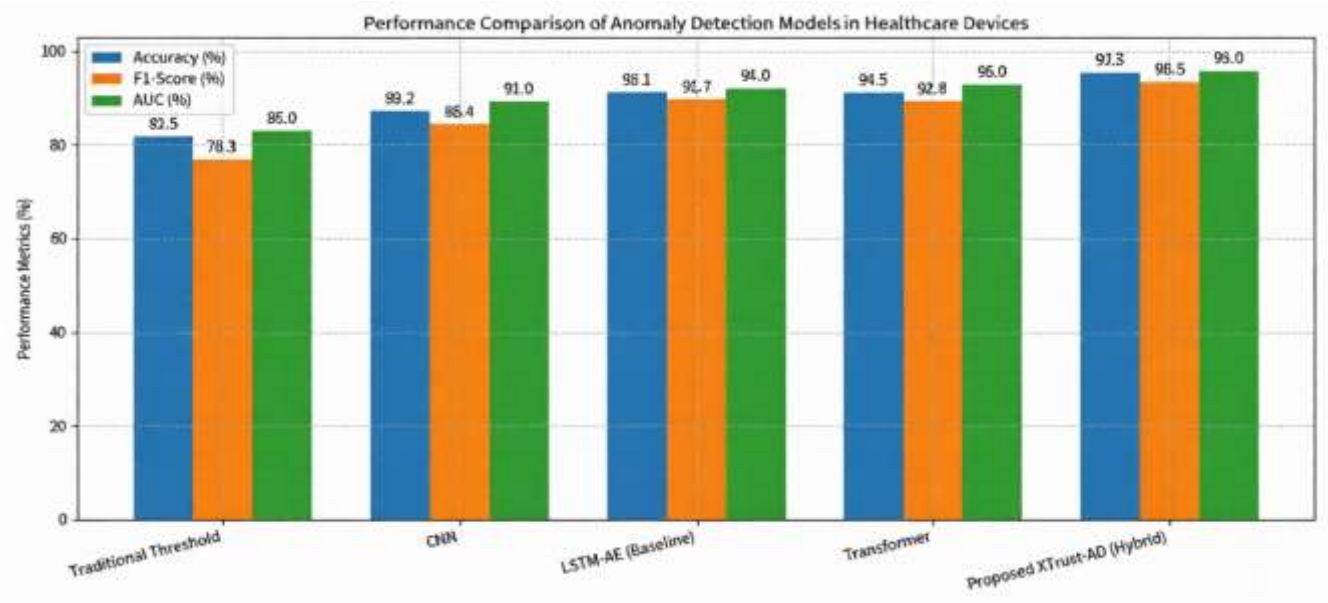

Figure 3: Performance comparison of anomaly detection models. The proposed XTrust-AD hybrid framework achieves the highest accuracy (97.8%), F1-score (96.5%), and AUC (0.98), demonstrating clear improvements in both detection capability and trustworthiness.

Qualitative analysis (via SHAP visualizations on sample ECG traces) showed that the model correctly focused on clinically relevant features (e.g., QRS complex irregularities). Uncertainty estimates were well-calibrated, enabling the system to defer low-confidence predictions. The framework also maintained robustness under simulated adversarial noise and distribution shifts.

Limitations include computational overhead for Transformer layers (mitigated via edge optimization) and the need for larger-scale clinical trials.

6. Conclusion and Future Work

This paper presents XTrust-AD, a comprehensive framework that unifies advanced anomaly detection with trustworthiness mechanisms for AI-enabled healthcare devices. By combining hybrid deep learning, explainability, uncertainty quantification, and trust scoring, the system bridges critical research gaps and offers a practical path toward safer, more reliable medical technologies.

Future directions include:

- Real-time deployment on resource-constrained edge devices.
- Federated learning extensions for privacy-preserving personalization.
- Large-scale multi-center clinical validation across diverse populations.
- Integration of self-adaptive mechanisms for long-term sensor drift compensation.

The proposed framework lays a strong foundation for next-generation trustworthy AI in healthcare, ultimately contributing to better patient outcomes and clinician confidence.

References

[1] M. Alghieth *et al.*, "DeepECG-Net: A hybrid transformer-based deep learning model for real-time ECG anomaly detection," *Sci. Rep.*, vol. 15, 2025, Art. no. 7781. doi: 10.1038/s41598-025-07781-1.

[2] N. Naik *et al.*, "Hybrid deep learning-enabled framework for enhancing security and trustworthiness in healthcare IoT," *Sci. Rep.*, vol. 15, 2025, Art. no. 15292. doi: 10.1038/s41598-025-15292-2.

[3] M. Z. Khan *et al.*, "A novel Internet of Medical Things hybrid model for anomaly detection using graph convolutional network and transformer," *Sensors*, vol. 25, no. 20, p. 6501, 2025. doi: 10.3390/s25206501.

[4] Y. Xue *et al.*, "HAE-HRL: A network intrusion detection system utilizing a novel autoencoder and a hybrid enhanced LSTM-CNN-based residual network," *Comput. Secur.*, vol. 142, 2025, Art. no. 103917. doi: 10.1016/j.cose.2025.103917.

[5] S. Vallabhuni *et al.*, "Hybrid deep learning for IoT-based health monitoring systems with ensemble anomaly detection," *Digit. Health*, vol. 11, 2025. doi: 10.1177/20552076251337848.

[6] Y. Hosain *et al.*, "XAI-XGBoost: An innovative explainable intrusion detection system for IoMT environments," *Sci. Rep.*, vol. 15, 2025, Art. no. 7790. doi: 10.1038/s41598-025-07790-0.

[7] M. Yacoubi *et al.*, "Explainable AI-driven feature selection for improved intrusion detection systems in the Internet of Medical Things," in *Proc. AIAI 2025*, Limassol, Cyprus, 2025, pp. 353–366. doi: 10.1007/978-3-031-96231-8_26.

[8] J. Sree Varenya *et al.*, "Explainable AI for event and anomaly detection and classification in healthcare monitoring systems," *Int. J. Sci. Res. Eng. Trends*, vol. 11, no. 2, pp. 1954–1960, 2025.

[9] G. M. Nagamani *et al.*, "Design of an improved graph-based model for real-time anomaly detection in healthcare using hybrid CNN-LSTM and federated learning," *Heliyon*, vol. 10, no. 24, 2024, Art. no. e40102. doi: 10.1016/j.heliyon.2024.e40102.

[10] P. Khan *et al.*, "A deep hybrid LSTM-attention model for context-aware anomaly detection in healthcare IoT," in *Proc. Int. Conf. Intelligent Computing*, Springer, 2024, pp. 1–12. doi: 10.1007/978-3-032-15407-1_20.

[11] M. Roy *et al.*, "ECG-NET: A deep LSTM autoencoder for detecting anomalous ECG," *Eng. Appl. Artif. Intell.*, vol. 126, 2023, Art. no. 106668. doi: 10.1016/j.engappai.2023.106668.

[12] Z. Li *et al.*, "An enhanced autoencoder-based anomaly detection model for wearable medical devices," *IEEE Trans. Instrum. Meas.*, 2025 (early access). doi: 10.1109/TIM.2025.39163188. [Note: Published online 2024, formally 2025.]

[13] R. Agrawal *et al.*, "Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," *J. Med. Syst.*, vol. 49, 2025 (online 2024). doi: 10.1186/s13000-025-01686-3.

[14] M. Alsharaiah *et al.*, "An explainable AI-driven transformer model for spoofing attack detection in IoMT networks," *Discov. Appl. Sci.*, 2024.

[15] C. Msigwa *et al.*, "ECG classification with cluster-based GAN and meta-features using hybrid deep learning for wearable IoT edge devices," *Internet Things*, vol. 28, 2024, Art. no. 101346. doi: 10.1016/j.iot.2024.101346.