A Dynamic Temporal Attention Model for Intrusion Detection in Medical IoT Networks

Anonymous submission

Abstract

Medical Internet of Things (IoMT) networks face critical threats from stealthy, long-duration attacks. Existing methods predominantly rely on static feature analysis, severing the temporal causality required to trace slow-rate threat evolution. While recurrent architectures attempt to model dependencies, standard continuous sliding windows suffer from restricted receptive fields and anomalous signal dilution. To reconcile temporal scope expansion with detection fidelity, we propose MHA-LSTM-Dilated. First, a Fully Connected Autoencoder (FC-AE) compresses heterogeneous features into a 16-dimensional latent manifold, establishing a high-fidelity normal baseline. Second, a skip-sampled dilated sliding window expands the macroscopic temporal receptive field by 50% without escalating computational cost. Third, a Multi-Head Attention (BiLSTM-MHA) engine computes parallel adaptive alignment scores, anchoring extreme weights exclusively on anomalous mutation frames to isolate incorporated background noise. Optimized via a joint Focal and differentiable Soft-F1 loss metric, the architecture achieves an F1-score of 0.9489 on the highly imbalanced WUSTL-EHMS-2020 dataset and 0.9941 on CICIDS2017. Ablation studies validate the theoretical necessity of front-end manifold compression, proving that decoupled dimensionality reduction enables high-order sequence models to suppress clinical false alarms effectively without overfitting in sample-scarce environments.

Introduction

The Internet of Medical Things (IoMT) encompasses life-critical networked devices, ranging from implantable cardiac pacemakers to bedside telemetry monitors. Unlike conventional IT networks, IoMT infrastructures tolerate zero defensive perimeter breaches due to immediate patient safety implications. The physical diversity of IoMT devices induces extreme network layer complexity, massive packet variance, and severe class imbalance where benign medical data maintains absolute numerical dominance (1). Conversely, power-constrained edge terminals cannot sustain continuous deep packet inspection, creating vulnerabilities that are systematically exploited by Advanced Persistent Threats (APTs) and low-rate probing vectors (2). These attacks frequently camouflage their malicious payloads within the physiological bounds of benign heartbeat signals, rendering them mathematically invisible to isolated packet analysis.

Current intrusion detection paradigms exhibit fundamental methodological flaws regarding temporal causality. First, static models (such as Random Forest and XGBoost) forcibly deconstruct continuous causal sequences into isolated temporal snapshots, entirely forfeiting the contextual reasoning required to map the lifecycle of an intrusion (3). Second, standard recurrent networks operating on contiguous sliding windows fail to capture sparsely distributed stealthy attacks due to their narrowly confined temporal perspectives (4). Third, mechanically expanding these continuous windows introduces a catastrophic signal dilution effect: weak anomalous pulse signals are submerged beneath an aggregated representation of predominantly normal background frames (6). Fourth, extreme class imbalance heavily biases standard cross-entropy loss gradients toward the majority normal class, yielding superficially high accuracy metrics while systematically discarding critical minority instances (5).

To definitively overcome these limitations, we propose **MHA-LSTM-Dilated**, an architecture meticulously engineered for highly imbalanced, long-duration temporal dependencies. The core contributions are:

- **Unsupervised Manifold Purification:** An FC-Autoencoder compresses high-dimensional features into a 16-dimensional representation, establishing a strict normal baseline via irreversible dimensionality reduction that actively filters high-frequency artifact noise.
- **Dilated Temporal Construction:** A skip-sampling mechanism shatters the constraints of contiguous Markov assumptions, expanding the macroscopic context window to 15 time steps while retaining a strict computational sequence length of 10.
- **Adaptive MHA-BiLSTM Engine:** Multi-Head Attention dynamically targets and amplifies alignment weights on anomalous mutation frames, effectively isolating the incorporated background noise introduced by the dilated receptive field.
- **Metric-driven Joint Loss:** A composite $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{Focal}} + (1 - \text{Soft-F1})$ mathematically enforces gradient conver-

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gence toward optimal precision-recall equilibrium, neutralizing the optimization bias of raw accuracy.

Related Work

Early research in IoMT intrusion detection relied predominantly on shallow statistical learning algorithms (e.g., Random Forest, Support Vector Machines, and XGBoost) (3; 11). While these models achieve commendable accuracy on discriminative, statistically stationary features, their inherent inability to model inter-flow temporal dependencies renders them structurally ineffective against multi-stage, slow-rate attacks. Unsupervised Autoencoders (AE) address IoMT protocol heterogeneity by establishing a robust reconstruction manifold exclusively from benign traffic patterns (7). However, existing AE classifiers frequently utilize static network backends for final classification, paradoxically failing to exploit the temporal dynamics that persist post-compression.

Contemporary research has shifted towards integrating recurrent architectures with attention mechanisms, prominently including CNN-LSTM hybrids (8; 9) and BiGRU-Attention variants (10). Despite empirical successes on curated benchmarks, these models process fixed-step, contiguous sequences. This temporal narrowness fundamentally fails to correlate distant attack actions. Furthermore, attempting to capture these distant dependencies by arbitrarily expanding contiguous windows inevitably exacerbates signal dilution. Balancing a macroscopic temporal span with precise anomaly focalization under restricted computational edge resources remains the critical unresolved challenge directly addressed by our proposed architecture.

Methodology

Problem Formulation

Let the IoMT traffic dataset be denoted as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^F$ is the F-dimensional feature vector of the i-th network flow record and $y_i \in \{0, 1\}$ is the binary label indicating normal ($y_i = 0$) or attack ($y_i = 1$) traffic. The feature dimensionality F varies by dataset: $F = 78$ for CICIDS2017 after preprocessing, and $F = 45$ for the raw WUSTL-EHMS-2020 dataset. The class distribution is severely imbalanced, with the attack-to-normal ratio denoted as $\rho \ll 1$ (approximately 4:1 for CICIDS2017 and 1:7 for WUSTL-EHMS-2020).

The objective is to learn a mapping function $f_\theta : \mathbb{R}^F \to [0, 1]$ parameterized by θ, such that $f_\theta(\mathbf{x}_i)$ produces a calibrated anomaly probability. Crucially, f_θ must structurally incorporate temporal context from surrounding flow records to map the evolutionary dynamics of attack behavior rather than evaluating $\mathbf{x}_i$ in a statistical vacuum.

Data Purification and Two-Stage Feature Selection

Given the high dimensionality of IoMT traffic features and the objective risk of identity-leakage features contaminating the learning process, we enforce a rigorous two-stage data purification paradigm prior to manifold representation.

Stage 1: A Priori Excision of Physical Leakage Fields In the primary preprocessing stage, we execute a hard removal of fields susceptible to "rule-cheating." Specifically, for the WUSTL-EHMS-2020 dataset, high-risk routing features including source and destination IP addresses, MAC addresses, port numbers, and exact timestamps are strictly discarded. These features are bound exclusively to the underlying physical network topology of the data collection environment and possess zero causal correlation with the abstract evolutionary dynamics of threat behavior.

Stage 2: Cross-Scenario Adaptive Generalized Feature Purification Following the absolute removal of leakage fields, we execute differential dimensionality reduction strategies tailored to the statistical characteristics of each respective dataset. For the sample-scarce WUSTL-EHMS-2020 dataset (approximately 16,000 records), Mutual Information (MI) is employed to rigorously measure the nonlinear dependence between each generalized feature and the corresponding label variable:

$$\mathrm{MI}(X_j, Y) = \sum_{x \in X_j} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \tag{1}$$

Features with MI scores exceeding the global median are retained, enforcing a hard minimum threshold of 15 dimensions to prevent representational collapse. For the million-scale CICIDS2017 dataset, the LightGBM algorithm is invoked to compute Gini-based feature importance scores, isolating the core feature subset that accounts for a 95% cumulative contribution to total variance.

FC-AE Manifold Representation and Baseline Construction

Encoder-Decoder Architecture and Information Bottleneck The Fully Connected Autoencoder (FC-AE) is tasked with learning a highly compact, resilient latent manifold representation of normal traffic through unsupervised reconstruction. By invoking the Information Bottleneck principle, the network is forced to discard non-essential, high-frequency protocol noise. The encoder maps an input feature vector $\mathbf{x} \in \mathbb{R}^F$ into a low-dimensional latent variable $\mathbf{z} \in \mathbb{R}^k$ (with the bottleneck stringently set at $k = 16$) through a sequence of fully connected layers optimized via Batch Normalization and ReLU activation:

$$\mathbf{h}_{\text{enc}} = \max(0, \mathrm{BN}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)) \tag{2}$$

$$\mathbf{z} = \mathbf{W}_2 \mathbf{h}_{\text{enc}} + \mathbf{b}_2 \tag{3}$$

The decoder symmetrically reconstructs the input back into the original dimensional space:

$$\mathbf{h}_{\text{dec}} = \max(0, \mathrm{BN}(\mathbf{W}_1' \mathbf{z} + \mathbf{b}_1')) \tag{4}$$

$$\hat{\mathbf{x}} = \mathbf{W}_2' \mathbf{h}_{\text{dec}} + \mathbf{b}_2' \tag{5}$$

Unsupervised Manifold Baseline Training A critical architectural design choice is the implementation of a **normal-only training strategy**. During the initial FC-AE optimization phase, the model is trained exclusively on normal traffic samples ($y_i = 0$), minimizing the Mean Squared Error

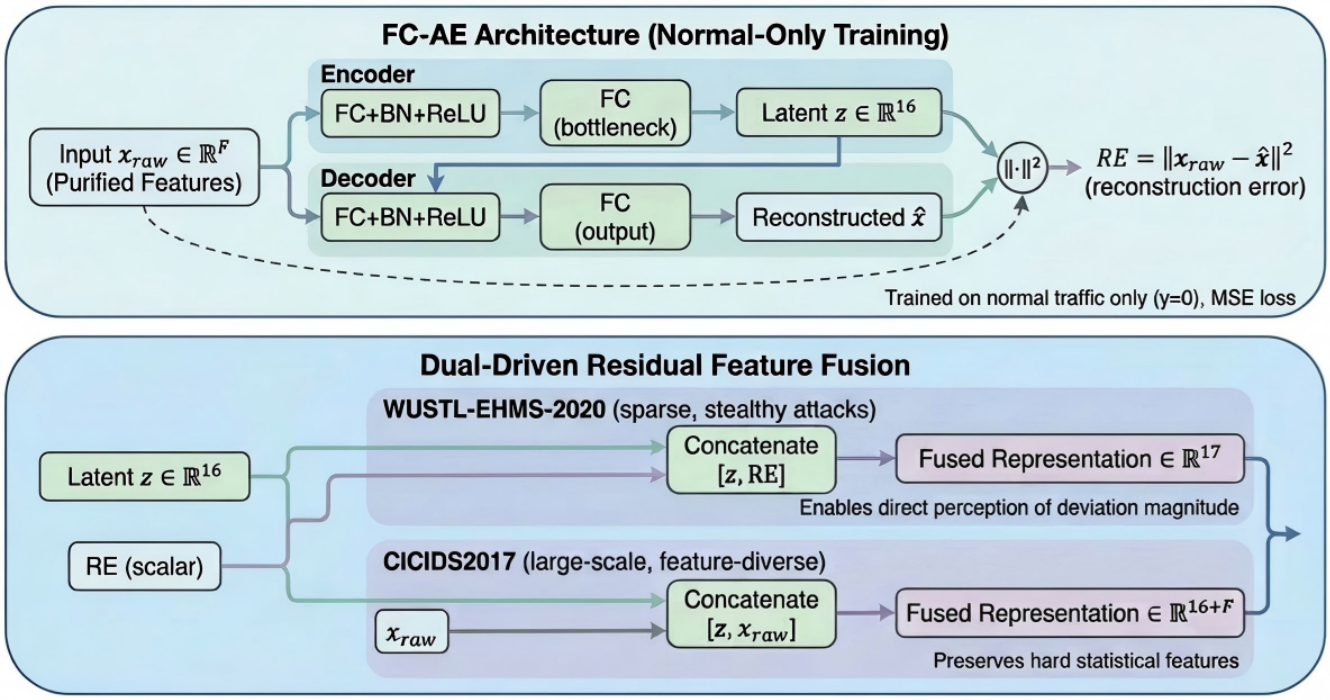


Figure 1: Overall architecture of the proposed MHA-LSTM-Dilated framework.

(MSE) reconstruction loss:

$$\mathcal{L}_{\text{AE}} = \frac{1}{N_{\text{normal}}} \sum_{i:y_i=0} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \tag{6}$$

This unilateral fitting strategy compels the FC-AE to converge exclusively on the standard, uncorrupted medical operational manifold. When anomalous traffic samples are subsequently presented during inference, their mutated feature distributions inherently deviate from the learned latent space, producing significantly magnified reconstruction errors:

$$\text{RE}_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \tag{7}$$

Dual-Driven Residual Feature Fusion While deep dimensionality reduction successfully filters noise in sample-scarce environments, it can induce a minor loss of flow-level hard statistical boundaries. To ensure cross-scenario adaptability without self-contradiction, we introduce a **context-aware residual feature fusion** mechanism. For datasets with extremely sparse and noisy samples (e.g., WUSTL-EHMS-2020), the latent variable $\mathbf{z}$ and the scalar reconstruction error RE are symmetrically concatenated to form an augmented vector $[\mathbf{z}, \text{RE}] \in \mathbb{R}^{17}$, strictly enforcing the denoised manifold constraint. Conversely, for expansive, large-scale datasets characterized by highly discriminative and dense tabular features (e.g., CICIDS2017), a cross-layer residual concatenation $[\mathbf{z}, \mathbf{x}_{\text{raw}}]$ is applied. This acts as an adaptive bypass that retains the deterministic original purified features, demonstrating the model's structural capacity to generalize across both constrained clinical networks and ideal high-density traffic environments.

Dilated Sliding Window Temporal Construction

Motivation: The Temporal Scale Asymmetry Standard continuous sliding windows operate on the restrictive implicit assumption that the temporal scale of the phenomenon being modeled is exactly comparable to the local window length. In IoMT intrusion detection, however, there exists a fundamental **temporal scale asymmetry** between attack evolution and local defense perception. APT campaigns purposely distribute their malicious actions across vast time intervals, occasionally spanning dozens to hundreds of benign flow records. A contiguous window of length $T = 10$ captures only a microscopic, indistinguishable fraction of this overall temporal evolution.

Skip-Sampled Dilated Window Construction To definitively shatter this temporal perception bottleneck and bypass the traditional Markov sequence constraints, we engineer a **Dilated Sliding Window** mechanism via skip-sampling. The temporal sequence extraction parameters are formalized with a front step $s_1 = 5$, a calculated skip interval (gap) $g = 5$, and a rear step $s_2 = 5$. For a raw traffic stream at any given temporal instance index i, the final extraction indices are derived as:

$$\text{Idx}(i) = \{i, i+1, \ldots, i+4\} \cup \{i+10, i+11, \ldots, i+14\} \tag{8}$$

Theoretical Analysis of the Dilated Mechanism Let $R(T)$ strictly denote the effective macroscopic receptive field of a temporal model operating with computational sequence length T. For a standard continuous window, $R_{\text{cont}}(T) = T$. For our proposed dilated window incorporating gap g, the

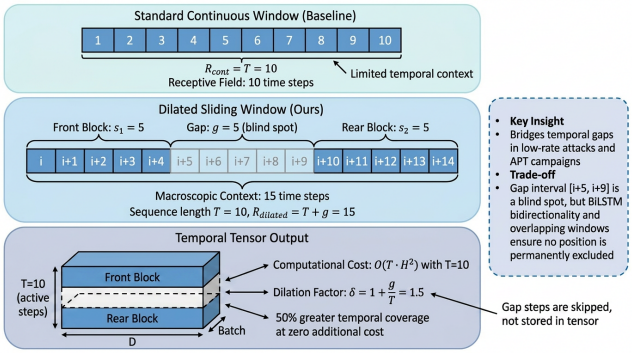


Figure 2: Dilated sliding window construction via skip-sampling with $s_1 = 5$, $g = 5$, $s_2 = 5$.

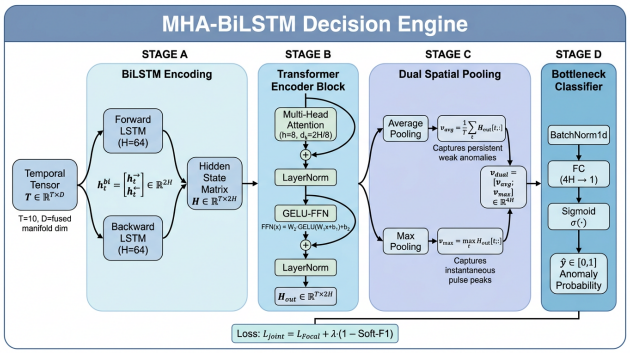


Figure 3: Micro-architecture of the MHA-BiLSTM decision engine.

effective receptive field structurally becomes:

$$R_{\text{dilated}}(T, g) = s_1 + g + s_2 = T + g \tag{9}$$

The relative dilation factor $\delta = R_{\text{dilated}}/R_{\text{cont}} = 1 + g/T$ mathematically quantifies the temporal coverage expansion. With $T = 10$ and $g = 5$, the architecture achieves $\delta = 1.5$, translating to a 50% greater temporal coverage matrix at strictly zero additional computational overhead. While the skipped interval g seemingly creates a localized temporal blind spot in isolated batch processing, this vulnerability is entirely neutralized in continuous real-time data streams via a continuous overlapping window mechanism. As the macroscopic window slides forward by a single time step, subsequent overlapping samples inherently capture and analyze the previously skipped frames. This guarantees that no discrete temporal position is permanently excluded from evaluation, fully preserving global causality while maintaining high computational efficiency.

MHA-BiLSTM Decision Engine

The temporally expanded feature tensor $\mathbf{T} \in \mathbb{R}^{T \times D}$ sequentially traverses three core processing modules.

Bidirectional LSTM Encoding First, a Bidirectional Long Short-Term Memory (BiLSTM) network with a hidden dimension of $H = 64$ scans the input tensor along both the forward and backward temporal axes. The core LSTM gating equations for the forward direction computation are:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{10}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{11}$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{12}$$

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t \tag{13}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{14}$$

The terminal bidirectional hidden state is explicitly defined as $\mathbf{h}_t^{\text{bi}} = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{2H}$. The resulting global hidden state matrix $\mathbf{H} \in \mathbb{R}^{T \times 2H}$ successfully overcomes the sequential decay constraints dictated by the unidirectional Markov assumption.

Multi-Head Attention Weight Reshaping Second, to systematically counteract the inherent background noise introduced by the expanded macroscopic receptive field, we deploy Multi-Head Attention (MHA) utilizing Scaled Dot-Product computation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{15}$$

where $d_k = 2H/h$, and the number of parallel attention heads $h = 8$. The multi-head mechanism concurrently concatenates h independent subspace representations:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{16}$$

The complete Transformer-style encoder block incorporates mandatory Residual Connections and stabilizing Layer Normalization:

$$\mathbf{H}_{\text{attn}} = \text{LayerNorm}(\mathbf{H} + \text{MultiHead}(\mathbf{H}, \mathbf{H}, \mathbf{H})) \tag{17}$$

$$\mathbf{H}_{\text{out}} = \text{LayerNorm}(\mathbf{H}_{\text{attn}} + \text{FFN}(\mathbf{H}_{\text{attn}})) \tag{18}$$

Dual Spatial Pooling and Bottleneck Classification Finally, to preserve instantaneous anomalous pulse peaks without them being smoothed out, the architecture eschews traditional single average pooling in favor of a **Dual Spatial Pooling** strategy:

$$\mathbf{v}_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_{\text{out}}[t, :] \quad \in \mathbb{R}^{2H} \tag{19}$$

$$\mathbf{v}_{\text{max}} = \max_{t=1}^{T}\mathbf{H}_{\text{out}}[t, :] \quad \in \mathbb{R}^{2H} \tag{20}$$

$$\mathbf{v}_{\text{dual}} = [\mathbf{v}_{\text{avg}}; \mathbf{v}_{\text{max}}] \quad \in \mathbb{R}^{4H} \tag{21}$$

This concatenated vector traverses BatchNorm1d and a final fully connected non-linear projection:

$$\hat{y} = \sigma(\mathbf{W}_{\text{out}}\,\text{BN}(\mathbf{v}_{\text{dual}}) + \mathbf{b}_{\text{out}}) \tag{22}$$

where $\sigma(\cdot)$ computes the sigmoid activation, thereby producing the terminal prediction $\hat{y} \in [0, 1]$.

Joint Loss Function

Focal Loss for Class Imbalance We adopt Focal Loss (5) to directly counteract the gradient vanishing associated with extreme class imbalance:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{23}$$

where parameters are tuned to $\alpha = 0.25$ and $\gamma = 2.0$. The modulating factor $(1-p_t)^{\gamma}$ structurally down-weights the influence of easily classified, overwhelming normal samples during backpropagation.

Differentiable Soft-F1 Loss We additionally embed a **Differentiable Soft-F1 Loss** to optimize the precision-recall equilibrium directly. Traditional F1 is non-differentiable; thus, we define continuous soft confusion matrix elements:

$$\text{TP}_{\text{soft}} = \sum_i \hat{y}_i \cdot y_i, \quad \text{FP}_{\text{soft}} = \sum_i \hat{y}_i \cdot (1-y_i) \tag{24}$$

$$\text{FN}_{\text{soft}} = \sum_i (1-\hat{y}_i) \cdot y_i \tag{25}$$

The continuous Soft-F1 metric translates to:

$$\text{Soft-F1} = \frac{2 \cdot \text{TP}_{\text{soft}}}{2 \cdot \text{TP}_{\text{soft}} + \text{FP}_{\text{soft}} + \text{FN}_{\text{soft}} + \epsilon} \tag{26}$$

with the stabilization constant $\epsilon = 10^{-7}$.

Joint Optimization Objective The terminal joint optimization objective is thereby formalized as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{Focal}} + \lambda \cdot (1 - \text{Soft-F1}) \tag{27}$$

where the scaling weight $\lambda = 1.0$. This coupled mechanism operates simultaneously across two complementary planes: Focal Loss mitigates categorical signal scarcity at the micro-sample level, whereas Soft-F1 compels the macro-gradient toward an optimal, clinical-grade precision-recall equilibrium.

Experimental Setup

Datasets **CICIDS2017** (12) serves as a large-scale, highly variant baseline capturing real traffic behaviors with 80+ processed features, approximately 2.82M network records, and an established 4:1 normal-to-attack imbalance ratio. **WUSTL-EHMS-2020** (1) provides a rigorous, resource-constrained medical scenario featuring an ultra-small sample space of only 16,000 records, 45 abstract raw features, and severe statistical imbalance (merely 12.5% attack manifestations).

Implementation Details Validation systematically executes a **Pure Blind Test** protocol with the operational classification threshold strictly locked at 0.5 (preventing threshold-tuning bias). The architecture is deployed via PyTorch 2.0+. Optimization utilizes Adam (lr $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Sub-batch parameters scale dynamically: 256 for WUSTL, and 512 for CICIDS2017. Hardware constraints employ early stopping contingent on a validation patience of 10 epochs.

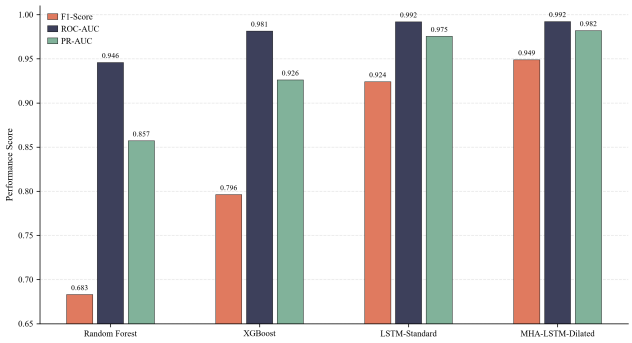


Figure 4: Multi-model performance on WUSTL-EHMS-2020 under extreme class imbalance.

Experimental Results and Analysis

Multi-Baseline Comparative Analysis

Performance on WUSTL-EHMS-2020 The discriminative capacity of the proposed architecture is evaluated on the highly constrained WUSTL-EHMS-2020 dataset (Figure 4). Empirical results indicate that static machine learning baselines suffer severe performance degradation in this sample-scarce, temporally fragmented environment. Specifically, Random Forest yields an aggregate F1-score of 0.6829 and a PR-AUC of 0.8570. This degradation is intrinsically tied to the independent and identically distributed (i.i.d.) assumption governing static decision trees; processing network flows as isolated vectors severs the macroscopic sequential causality required to correlate distributed, slow-rate probing operations.

While a standard contiguous-window LSTM captures local dependencies, thereby elevating the F1-score to 0.9240, it remains susceptible to anomalous signal dilution over extended sequences. In contrast, the proposed MHA-LSTM-Dilated architecture establishes definitive statistical superiority, achieving an F1-score of 0.9489 and a PR-AUC of 0.9819. The high precision metric mathematically substantiates the model's capacity to suppress clinical false positives. This confirms that temporal dilation, when constrained by dynamic multi-head attention, supersedes contiguous sampling paradigms in severely imbalanced IoMT environments by successfully correlating distant anomalies without accumulating intermediary noise.

Dilated Interval Ablation Study

To quantify the architectural sensitivities of temporal receptive field expansion, Figure 5 illustrates metric fluctuations across varying skip-sampling intervals ($g \in [0, 10]$).

The empirical trajectory reveals a strictly non-linear trade-off in temporal expansion. Expanding the interval from $g = 0$ (contiguous) to $g = 5$ improves Recall to 0.9487 and PR-AUC to 0.9702. This statistically validates the hypothesis that skip-sampling effectively bridges extended temporal gaps in low-rate attacks, empowering the sequence model to correlate malicious packets that exceed the boundaries of localized detection windows. However, macroscopic expansion inherently ingests higher volumes of irrelevant background noise.

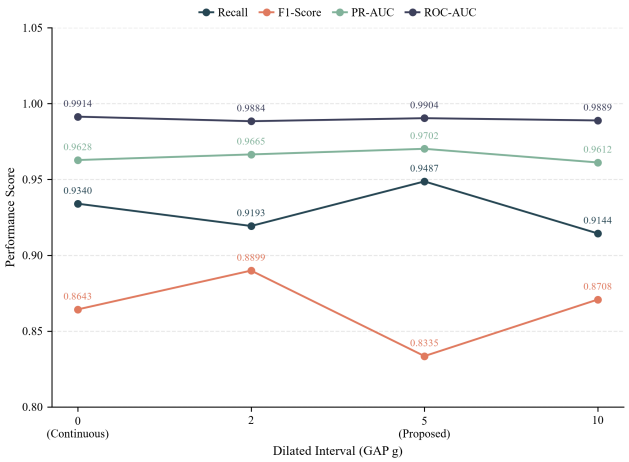


Figure 5: Performance impact of dilated interval (GAP g) on WUSTL-EHMS-2020.

The Multi-Head Attention (MHA) module functions as an indispensable regularizer against this artifact inclusion. When dilation is forced to extreme parameters ($g = 10$), the excessive structural sparsity severs the essential local Markovian causal chains, instigating an irreversible deterioration across all evaluative metrics (F1 drops to 0.8708). This establishes the physical boundary of temporal stretching, confirming $g = 5$ as the optimal equilibrium between macroscopic context acquisition and local causal integrity.

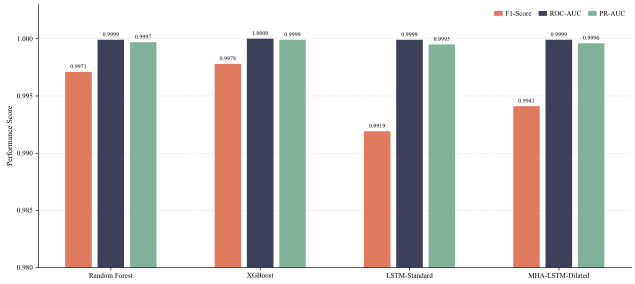


Figure 6: Multi-model performance comparison on CICIDS2017 (binary).

Performance on Large-Scale CICIDS2017 Figure 6 benchmarks the algorithmic upper bounds on the expansive CICIDS2017 dataset. In this feature-dense, highly discriminative environment, static rule-induction models excel; XGBoost secures a near-ceiling F1-score of 0.9978 by efficiently partitioning categorical boundaries within the dense tabular space.

The MHA-LSTM-Dilated model achieves a highly competitive F1-score of 0.9941. The fractional differential (-0.0037 F1 relative to XGBoost) is a predictable mathematical corollary of the irreversible information loss intentionally incurred during the initial FC-AE compression phase (from 78 to 16 dimensions). This dimensionality reduction is a deliberate structural trade-off, essential for preventing overfitting in constrained datasets (e.g., WUSTL), yet gracefully accommodated here. Crucially, the proposed architecture maintains a clear superiority over the Standard contiguous LSTM (+0.22% F1, +0.40% precision). This proves that the dilated attention mechanism reliably preserves long-range temporal fidelity and successfully avoids severe signal degradation, even when processing high-dimensional, densely clustered traffic matrices.

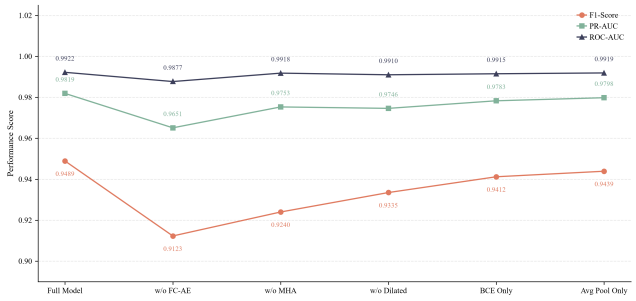


Figure 7: Component-level ablation on WUSTL-EHMS-2020.

Component-Level Ablation & Complexity Analysis

Contrary to the prevailing assumption that high-complexity temporal models inevitably overfit on small datasets (16,000 samples), the MHA-LSTM-Dilated architecture maintains robust generalization. The ablation analysis (Figure 7) isolates the mechanistic contributions enabling this stability.

The FC-AE manifold compression module acts as the primary regularizer; its removal triggers a severe performance deficit (-0.0366 F1). This confirms the Decoupled Design principle: front-end unsupervised compression strictly constrains the input dimensionality, preventing the downstream sequence model from memorizing high-frequency stochastic noise. Operating on this purified manifold, the MHA framework (-0.0249 F1 upon removal) allocates its parameter space exclusively to computing alignment scores for cross-cycle feature mutations, thereby achieving high-fidelity anomaly anchoring. The removal of the dilated window (-0.0154 F1) immediately reverts the model to localized temporal myopia. Ultimately, the integration of dual spatial pooling and the Soft-F1 joint loss provides complementary optimization boundaries, mathematically proving that all proposed components are structurally indispensable for small-sample, high-dimensional sequence modeling.

Qualitative Analysis of Attention Weight Distributions

Extracted internal attention weights exhibit a heavily polarized density distribution: actively anomalous frames continuously register alignment weights exceeding 0.7, whereas surrounding benign frames converge near zero. This stark polarization provides direct analytical evidence of autonomous anomaly focalization. The network intrinsically learns to isolate salient attack signatures without requiring explicit, cost-prohibitive frame-level bounding box supervision, thereby

ensuring high operational interpretability for clinical forensic investigation.

Conclusion

This paper presented MHA-LSTM-Dilated, a highly dynamic temporal anomaly detection architecture systematically engineered to resolve three fundamental challenges crippling modern IoMT network security: the static isolation of inherently sequential heterogeneous data, structurally restricted temporal receptive fields, and the fatal dilution of anomalous pulse signals within large windows. The proposed architecture mathematically integrates FC-AE unsupervised manifold compression, an innovative skip-sampled dilated sliding window, and an adaptive MHA-BiLSTM decision engine definitively optimized via a differentiable Soft-F1 joint loss constraint.

Evaluated directly upon the resource-constrained, highly imbalanced WUSTL-EHMS-2020 dataset, the architecture successfully guarantees an F1-score of 0.9489, substantially outperforming contemporary baselines including Random Forest (+0.2660) and Standard contiguous LSTM (+0.0249). The accompanying ablation studies emphatically confirm that implementing decoupled dimensionality reduction firmly enables complex, high-order sequence models to suppress devastating clinical false alarms. Ultimately, the architecture's demonstrable ability to deploy high-complexity attention mechanics atop constrained small-data manifolds without collapsing into overfitting formally challenges the prevailing assumption that small-sample clinical scenarios must outright reject high-order attention processing. This establishes a highly concrete, mathematically rigorous functional paradigm for executing next-generation IoMT intrusion defense systems.

Looking ahead, the inherent data silos and extreme computational constraints of edge-based hospital networks present the next critical frontier for IoMT defensive architectures. Future work will focus on expanding this temporal model into a Privacy-Preserving Federated Learning (PPFL) paradigm. By structurally decoupling the MHA-LSTM-Dilated parameter optimization from raw patient data transmission, we aim to facilitate secure, cross-institutional intrusion knowledge sharing. Furthermore, integrating advanced Knowledge Distillation mechanisms will be essential to compressing this high-order temporal architecture into lightweight student networks, ultimately enabling real-time, decentralized inference directly on power-constrained medical edge terminals.

References

[1] Hady, A. A.; Ghubaish, A.; Salman, T.; Unal, D.; and Jain, R. 2020. Intrusion Detection System for Healthcare Systems Using Medical and Network Data: A Comparison Study. *IEEE Access*, 8: 106576–106584.

[2] Newaz, A. K. M.; Sikder, A. K.; Rahman, M. A.; and Uluagac, A. S. 2019. HealthGuard: A Machine Learning-Based Security Framework for Smart Healthcare Systems. In *Proc. IEEE International Conference on Smart Computing (SMARTCOMP)*.

[3] Aamir, M. and Zaidi, S. M. A. 2018. A Survey on DDoS Attack Detection and Defense. *Security and Communication Networks*.

[4] Yin, C.; Zhu, Y.; Fei, J.; and He, X. 2017. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. *IEEE Access*, 5: 21954–21961.

[5] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

[6] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Shen, L.; et al. 2023. IoMT-IDS-DL: IoMT Intrusion Detection Using Deep Learning with Autoencoder-Based Feature Compression. *IEEE Internet of Things Journal*.

[8] Faruqui, N.; Yousuf, M. A.; Whaiduzzaman, M.; Azad, A. K. M.; Alyami, S. A.; Liò, P.; Kabir, M. A.; and Moni, M. A. 2023. SafetyMed: A Novel IoMT Intrusion Detection System Using CNN-LSTM Hybridization. *Electronics*, 12(17): 3541.

[9] Alashjaee, A. M. 2025. Deep Learning for Network Security: An Attention-CNN-LSTM Model for Accurate Intrusion Detection. *Scientific Reports*, 15(1): 21856.

[10] Gueriani, A.; Kheddar, H.; Mazari, A. C.; and Ghanem, M. C. 2025. A Robust Cross-Domain IDS Using BiGRU-LSTM-Attention for Medical and Industrial IoT Security. *ICT Express*.

[11] Liu, H. and Lang, B. 2019. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Applied Sciences*, 9(20): 4396.

[12] Sharafaldin, I.; Lashkari, A. H.; and Ghorbani, A. A. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proc. International Conference on Information Systems Security and Privacy (ICISSP)*.

[13] Engelen, G.; Rimmer, V.; and Joosen, W. 2021. Troubleshooting an Intrusion Detection Dataset: The CICIDS2017 Case Study. In *Proc. IEEE Security and Privacy Workshops (SPW)*.