Enhanced Temporal Fusion Transformers with Multi-Scale Convolutions for ICU Cardiac Risk Assessment

Bhuvaneswaran Balasubramanian
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur, Chennai, India.
bhuvan@bhuvaneswaran.com

S.Iniyan*
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Chennai, India.
iniyans@srmist.edu.in

Usha Desai
S.E.A College of Engineering and Technology
Bangalore, India.
dr.ushadesai@seaedu.ac.in

* Corresponding author

***Abstract*—This paper proposes a continuous cardiac risk classification for Intensive Care Units (ICUs) patients using a temporal fusion transformer with multi-scale temporal. Continuous monitoring of physiological signals in ICU plays an important role in identifying patients at risk of clinical deterioration. Accurate interpretation of multivariate vital signs such as heart rate, respiratory rate, and oxygen saturation remains challenging due to the complex temporal patterns present in physiological data. Although Temporal Fusion Transformers (TFT) have demonstrated strong capability in time-series modelling, their reliance on global self-attention may overlook short-duration physiological fluctuations that are often clinically informative. Subtle changes in these signals often precede serious cardiac events, but manual interpretation of continuous data streams is challenging for clinicians. Automated cardiac risk classification systems can assist healthcare providers by identifying abnormal physiological patterns and categorising patient states into different risk levels, enabling earlier clinical intervention and more efficient patient monitoring. To address this limitation, we propose an enhanced TFT architecture that integrates a Multi-Scale Temporal Convolution (MSTC) module for improved local temporal feature extraction. The MSTC module, implemented using depthwise-separable temporal convolutions, is inserted between the Variable Selection Network and the Transformer encoder. The model is evaluated on the BIDMC Waveform Database, consisting of recordings from 53 ICU patients and 743 test windows across three cardiac risk classes. Experimental results show that the proposed method achieves 90.98% accuracy and a macro-averaged F1 score of 67.56%, outperforming CNN-1D, LSTM, Bi-LSTM, and standard TFT baselines. Additionally, the false-alarm rate is reduced to 2.89%, and the missed-critical rate decreases to 28.85%, representing a 13.46 percentage-point improvement over the baseline TFT model.**

***Index Terms*—Temporal Fusion Transformer, multi-scale temporal convolution, ICU monitoring, cardiac risk classification, depthwise-separable convolution, domain-informed training**

I. Introduction

Patient deterioration inside an Intensive Care Unit is seldom instantaneous. Heart rate begins drifting minutes before an arrest; oxygen saturation dips subtly before a full desaturation episode; respiratory rate climbs in a pattern that experienced clinicians recognise but threshold-based alarms routinely miss. The National Early Warning Score 2 (NEWS2) attempts to address this by aggregating six bedside parameters into a single acuity number, yet its static thresholds struggle when all patients cluster within a narrow physiological band — a situation that is particularly common in homogeneous ICU populations where everyone is already critically unwell.

Deep learning offers a different approach. Recurrent architectures such as Long Short-Term Memory (LSTM) networks and their bidirectional variants can learn sequential dependencies from raw signal data, but they tend to lose resolution over longer input horizons and offer little transparency into what they have actually learned. One-dimensional convolutional networks handle local signal morphology well, though they have no inherent mechanism for deciding which among several concurrent vital signs deserves more attention at any given moment. The Temporal Fusion Transformer (TFT) was developed to address both of these shortcomings: it pairs a Variable Selection Network (VSN) with multi-head self-attention, which together allow the model to learn which features matter and when — a meaningful step forward for multivariate time-series problems.

Applying TFT directly to ICU classification, however, exposes a structural limitation. Self-attention operates globally across the full input window — sixty seconds in our setup — which means a three-second heart-rate spike and a sustained two-minute trend compete for the same representational budget. With four self-attention heads and two encoder layers, the model has limited capacity to simultaneously resolve brief but clinically decisive events and slower-moving deterioration trajectories. In practice, transient spikes get smoothed away.

Multi-scale convolutional feature extraction addresses exactly this gap. Running parallel convolutions at different kernel widths exposes temporal patterns at multiple resolutions within a single forward pass — short kernels catch abrupt fluctuations, longer ones track gradual drift. Depthwise-separable convolutions make this computationally viable by factorising the operation into per-channel and cross-channel steps, which substantially reduces the parameter count. That parsimony

matters here because the available training cohort contains fewer than forty patients.

Dataset size is the second persistent challenge in ICU deep learning. The Beth Israel Deaconess Medical Center (BIDMC) Waveform Database contains recordings from just fifty-three patients, each lasting roughly eight minutes. Any additional trainable component risks memorising the training set rather than learning generalisable patterns unless the training strategy compensates. Mixup regularisation, which constructs synthetic training examples by linearly interpolating between real sequence pairs, has demonstrated a consistent ability to smooth decision boundaries and reduce overfitting in image and tabular settings.

In this paper we make the following contributions:

1) We propose inserting a lightweight depthwise-separable temporal convolution module (MSTC) between the VSN and Transformer encoder of TFT, with a kernel size derived from cardiac and respiratory physiology rather than hyperparameter search.
2) We pair MSTC with three training-time regularisation strategies—temporal mixup for virtual patient synthesis, label smoothing for probability calibration, and stochastic depth for robust feature learning—adding zero trainable parameters.
3) We conduct a five-model comparative evaluation on the BIDMC ICU dataset, demonstrating that the combined approach achieves 90.98% accuracy and 67.56% macro F1, outperforming the unmodified TFT by 6.73 pp and 19.74 pp respectively.

II. Related Work

A. Deep Learning for ICU Time Series

Deep learning methods have transformed clinical time-series analysis over the past decade [1]. LSTM networks [2] and their bidirectional extensions [3] became standard tools for sequential clinical data, capturing temporal dependencies in vital-sign streams. Empirical Mode Decomposition (EMD) used to classify tachycardia types from single-lead ECG, achieving 91.21% accuracy with Random Forest on 3858 beats [4]. However, EMD lack mechanisms to model temporal dependencies across multiple vital signs. The self-attention mechanism introduced by Vaswani et al. [5] opened new avenues for modelling long-range temporal relationships without the vanishing-gradient limitations of recurrent architectures.

B. Transformer Models for ICU Vital Signs

Building on the Transformer paradigm, Lim et al. [6] proposed the Temporal Fusion Transformer, which introduced Variable Selection Networks and interpretable multi-head attention for time-series forecasting. Phetrittikun et al. applied TFT to vital-sign trajectory forecasting across 140 ICU patients from a Thai critical-care dataset, demonstrating its ability to capture anomalous temporal patterns [7]. He and Chiang extended TFT into a global multi-output framework (TFT-multi) that simultaneously predicted five vital signs—blood pressure, pulse, SpO_2, temperature, and respiratory rate—on the MIMIC-III database [8]. Kapral et al. utilised TFT for intraoperative blood pressure forecasting across 73,009 surgical patients, reporting mean absolute errors of 4 mmHg on an internal cohort [9]. Chang et al. combined Transformer architectures with diffusion probabilistic models for heart rate and blood pressure forecasting on MIMIC-III, achieving state-of-the-art results on sparse ICU time series [10]. These works focus predominantly on vital-sign forecasting rather than risk classification and none incorporate multi-scale convolution within the TFT architecture.

C. Multi-Scale Temporal Feature Extraction

Cui et al. introduced Multi-Scale CNN (MCNN), which applies convolutions at varying kernel widths to raw time series before classification [11]. InceptionTime adapted Inception modules [12] with parallel kernels of different sizes, achieving strong results across the UCR/UEA archive [13]. Temporal Convolutional Networks (TCN) demonstrated that dilated causal convolutions could rival recurrent models on sequential benchmarks [14]. Howard et al. introduced depthwise-separable convolutions [15], which factorise standard convolution into a depthwise and pointwise step, reducing parameter count by an order of magnitude while preserving representational capacity. Wu et al. combined multi-scale spatial–temporal transformers with consistency representation learning for multivariate classification [16]. One-dimensional convolutional architectures have also shown particular promise for physiological signal processing [17]. Our work differs from these approaches by integrating a single lightweight depthwise-separable convolution into the TFT pipeline specifically for ICU vital-sign classification, with kernel size derived from cardiac and respiratory cycle physiology rather than data-driven search.

D. Regularisation for Small Clinical Datasets

Overfitting is a persistent challenge when training deep models on small clinical cohorts. Zhang et al. showed that linear interpolation of training pairs (mixup) improves generalisation by smoothing the decision surface [18]; its extension to temporal data remains underexplored in clinical settings. Label smoothing, first proposed as part of the Inception-v2 architecture [19], prevents overconfident softmax outputs by distributing a small probability mass across non-target classes—a property that aligns naturally with the genuine diagnostic ambiguity at risk-level boundaries in ICU monitoring. Stochastic depth [20] randomly drops entire network layers during training, forcing each layer to learn independently useful features and mitigating the co-adaptation that is especially problematic when the effective number of independent training examples is small. While these techniques have been applied widely in computer vision and natural language processing, their combined application within a Transformer-based architecture for small-sample ICU vital-sign classification has not been reported in the literature.

III. Proposed Method

A. Data Preprocessing Pipeline

The raw BIDMC recordings undergo a five-step preprocessing pipeline (Fig. 1). Recordings are first cleaned by removing NaN values, then filtered using physiological validity ranges (HR: 20–220 BPM, SpO_2: 50–100%, RR: 3–60 br/min).

Two engineered features are appended: the first difference of heart rate (ΔHR), which encodes the beat-to-beat rate of change, and the ten-second rolling standard deviation of SpO_2, which captures oxygenation instability. Sliding windows of 60 seconds with a 5-second stride are extracted and labelled using a trend-based scheme. Finally, patient-aware splitting (70/15/15), class balancing via oversampling, and StandardScaler normalisation prepare the data for model training.

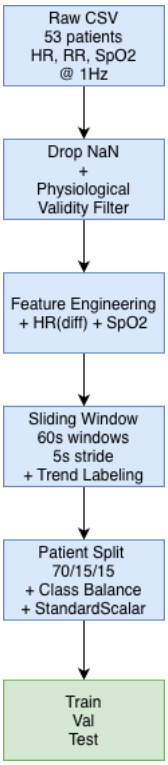

Fig. 1. Data preprocessing: raw CSV recordings are cleaned, filtered, feature-engineered, windowed with trend-based labelling, and split into patient-aware partitions.

B. Problem Formulation

Given a multivariate window $\mathbf{X} \in \mathbb{R}^{T\times F}$ with T=60 timesteps sampled at 1 Hz and F=5 features (heart rate, SpO_2, respiratory rate, first-difference of heart rate, and ten-second rolling standard deviation of SpO_2), we classify each window into one of three cardiac risk levels: Normal (y=0), At-Risk (y=1), or Critical (y=2). Labels are assigned by a trend-based scheme: Normal windows exhibit stable vital signs (slope below 4σ of the patient-specific noise floor); At-Risk windows show persistent drift exceeding 4σ; Critical windows display acute deterioration in the final fifteen seconds exceeding clinically calibrated thresholds (heart-rate rise $\geq$3.0 BPM, SpO_2 drop $\geq$1.5%, or respiratory-rate rise $\geq$2.5 breaths/min).

C. Baseline TFT Architecture

The architecture follows Lim et al. [6] with adaptations for classification. Each of the five input features is encoded independently by a Gated Residual Network (GRN), producing a d-dimensional embedding (d=64). A selection GRN computes softmax-normalised importance weights across features. The weighted sum of feature embeddings, enriched with sinusoidal positional encodings, passes through a two-layer Transformer encoder with four attention heads (d_{ff}=256). A post-attention GRN refines the output, which is mean-pooled across the temporal dimension and classified by a two-layer feedforward head ($64 \to 32 \to 3$).

D. Multi-Scale Temporal Convolution Module

MSTC is inserted between the VSN output and the positional encoding step (Fig. 1). The module comprises a single depthwise-separable convolution:

$$\mathbf{H}_{\text{conv}} = \text{GELU}\big(\text{BN}\big(\text{PW}\big(\text{DW}_k(\mathbf{H}_{\text{vsn}}^{\top})\big)\big)\big)^{\top} \tag{1}$$

where DW_k denotes depthwise convolution with kernel size k=5 and padding $k//2$ to preserve sequence length, PW denotes a pointwise 1×1 convolution, and BN is batch normalisation. The output is scaled by a learnable parameter α (initialised to 0.1) and added as a residual:

$$\mathbf{H}_{\text{out}} = \mathbf{H}_{\text{vsn}} + \alpha \cdot \mathbf{H}_{\text{conv}} \tag{2}$$

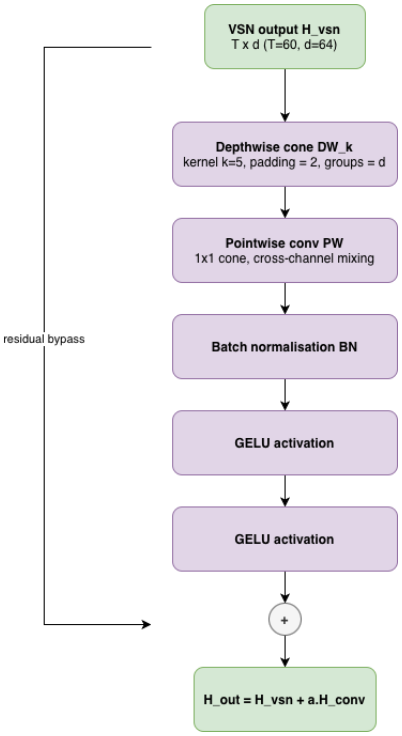

Fig. 2. The MSTC module has two parallel paths

The kernel size of five seconds is chosen to simultaneously cover cardiac dynamics (heart rate at 60–100 BPM corresponds to beat intervals of 0.6–1.0 s; five seconds spans 5–8 cardiac cycles) and respiratory patterns (respiratory rate of 12–20 breaths/min yields cycle durations of 3–5 s; five seconds captures one full respiratory cycle). The learnable scaling factor ensures that MSTC begins as a gentle perturbation of the proven VSN output and grows in influence only if the gradient signal supports it.

The depthwise-separable design adds approximately 450 parameters ($64\times5+64\times64+64\times2$ for BN), representing a 0.28% increase over the 158,034-parameter baseline—negligible in terms of capacity but significant in its ability to expose local temporal structure to the downstream Transformer.

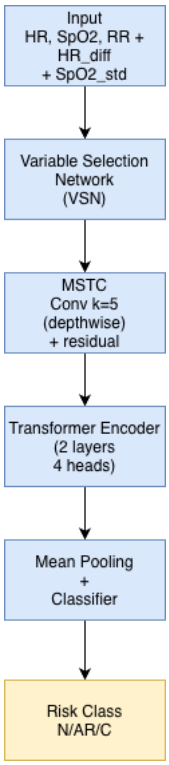

Fig. 3. Proposed TFT+MSTC architecture with domain-informed training. The MSTC module (red) sits between VSN and Transformer. Dashed boxes below show training-time-only regularisation strategies.

E. Domain-Informed Training

Three regularisation strategies operate exclusively during training:

Temporal mixup. For each mini-batch, a mixing coefficient $\lambda \sim \text{Beta}(0.3, 0.3)$ is drawn (clamped to $\lambda \geq 0.5$ so the primary sample dominates). A random permutation index π selects the partner:

$$\tilde{\mathbf{X}} = \lambda \mathbf{X} + (1-\lambda)\mathbf{X}_\pi, \quad \mathcal{L}_{\text{mix}} = \lambda\, \mathcal{L}(\hat{y}, y) + (1-\lambda)\, \mathcal{L}(\hat{y}, y_\pi) \tag{3}$$

where $\lambda is Beta(\alpha, \alpha)$. This forces the model to learn smooth interpolations between risk states rather than memorizing specific patient trajectories, critical for generalization on 53-patient datasets.

Physiological signals transition smoothly between states, so interpolated sequences remain plausible: blending a heart rate of 75 BPM with one of 110 BPM produces a virtual patient at ∼93 BPM—a realistic physiological state.

Label smoothing (ϵ=0.1). Hard one-hot labels are replaced with soft targets: $y_{\text{smooth}} = (1-\epsilon)\, y_{\text{hard}} + \epsilon/C$, where C=3. This reflects genuine clinical ambiguity at class boundaries and prevents the model from assigning extreme logits during early training.

Stochastic depth (p=0.1). Each Transformer encoder layer is independently dropped with probability p during training. At inference, all layers are active. This forces each layer to develop independently useful representations, mitigating co-adaptation on the small training set.

IV. Experimental Setup

A. Dataset

The BIDMC Waveform Database [21], available from PhysioNet [22], contains recordings from 53 adult ICU patients at Beth Israel Deaconess Medical Center. Each recording provides heart rate (HR), pulse rate (PR), respiratory rate (RR), and blood oxygen saturation (SpO_2) at 1 Hz for approximately eight minutes. Two engineered features—the first difference of HR (ΔHR) and the ten-second rolling standard deviation of SpO_2—are appended, yielding F=5 input channels. Sliding windows of 60 s with 5 s stride are extracted and labelled using the trend-based scheme described in Section III-A.

B. Data Partitioning

A patient-aware split assigns 37 patients (70%) to training, 8 (15%) to validation, and 8 (15%) to testing. No patient appears in more than one partition. The training set is class-balanced via oversampling of minority classes, and all features are standardised using statistics computed on the training partition only.

C. Baselines

Four established models serve as comparisons:

(i) **CNN-1D**: three convolutional layers (64→128→64 filters, kernel sizes 7/5/3) with batch normalisation, ReLU, and global average pooling;

(ii) **LSTM**: two-layer LSTM with 128 hidden units;

(iii) **Bi-LSTM**: two-layer bidirectional LSTM with 128 hidden units per direction;

(iv) **TFT Standard**: the architecture of Section III-B trained with plain cross-entropy. Learning rates are set per model family: 3×10^{-4} for recurrent models, 5×10^{-4} for CNN-1D, and 1×10^{-3} for both TFT variants.

D. Training Protocol

All models are trained for 50 epochs with the Adam optimiser, weight decay 10^{-4}, cosine annealing schedule, gradient clipping at unit norm, and batch size 64. The best checkpoint by validation accuracy is retained. All experiments use a fixed random seed for reproducibility.

TABLE I
TEST-SET PERFORMANCE ON BIDMC ICU DATASET (743 WINDOWS).

Model	Acc.	F1	Prec.	FA%	MC%
CNN-1D	76.45	43.22	41.80	17.98	39.42
LSTM	69.99	40.03	39.79	21.67	45.19
Bi-LSTM	64.47	39.83	41.49	27.13	14.42
TFT Standard	84.25	47.82	47.22	8.35	42.31
TFT+MSTC	**90.98**	**67.56**	**70.37**	**2.89**	**28.85**

Bold indicates best result. FA = false alarm rate, MC = missed critical rate. All values in %.

TABLE II
PER-CLASS F1 SCORES FOR TFT STANDARD VS. TFT+MSTC.

Model	Normal	At-Risk	Critical
TFT Standard	0.911	0.000	0.523
TFT+MSTC	**0.950**	**0.345**	**0.732**

E. Evaluation Metrics

We report overall accuracy, macro-averaged F1 score, macro precision, macro recall, false alarm (FA) rate, and missed-critical (MC) rate.

The FA rate measures the fraction of Normal windows incorrectly classified as Critical—a clinically significant error, as excessive non-actionable alerts arising from monitor noise, sensor artifacts, or patient movement contribute to alarm fatigue among ICU staff, reducing responsiveness to genuine emergencies.

The MC rate measures the fraction of Critical windows misclassified as Normal and serves as the primary clinical safety metric, as missed deterioration events carry direct patient safety consequences.

V. RESULTS

A. Overall Performance

Table I summarises test-set performance. The proposed TFT+MSTC model achieves the highest accuracy (90.98%), the highest macro F1 (67.56%), the highest precision (70.37%), the lowest false-alarm rate (2.89%), and a missed-critical rate of 28.85%—the lowest among all models except Bi-LSTM, which trades a very low MC rate for the worst accuracy (64.47%) and extremely high false-alarm rate (27.13%).

Relative to TFT Standard, our model improves accuracy by 6.73 percentage points and macro F1 by 19.74 pp. The per-class classification report (Table II) reveals that most of this gain comes from the Critical class, where F1 rises from 0.523 to 0.732, and the At-Risk class, where TFT Standard scores zero while TFT+MSTC reaches 0.345.

B. Ablation: Architecture vs. Training

To disentangle the contributions of MSTC and the regularisation strategies, we compare four configurations (Table III): the baseline TFT, TFT with MSTC alone (no regularisation, from a separate run), TFT with regularisation alone (no MSTC), and the full proposed model. MSTC alone provides a modest accuracy gain (+0.54 pp) but degrades F1; regularisation alone yields strong gains (+7.67 pp accuracy, +26.69 pp F1); the combination achieves the best results on all metrics, confirming that both components contribute synergistically.

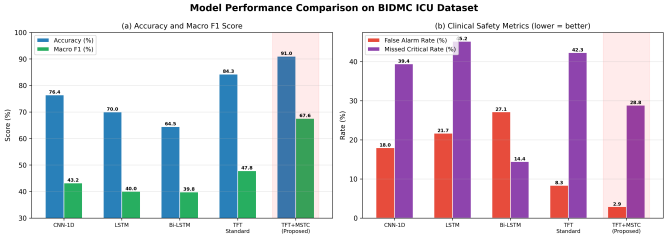

Fig. 4. Model performance comparison: (a) accuracy and macro F1, (b) clinical safety metrics. Proposed model (highlighted) achieves best results on all metrics.

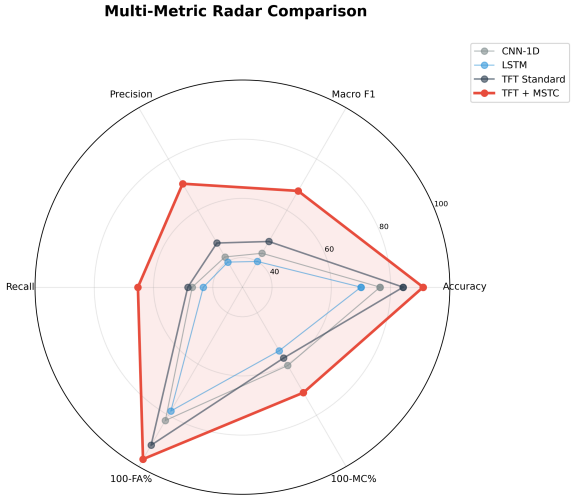

Fig. 5. Radar chart comparing four models across six evaluation dimensions. The proposed TFT+MSTC (red) encloses the largest area.

TABLE III
ABLATION STUDY ISOLATING MSTC AND REGULARISATION CONTRIBUTIONS.

Configuration	Acc.(%)	F1(%)	MC(%)
TFT Standard	84.25	47.82	42.31
TFT + MSTC only	84.79	49.62	56.73
TFT + Reg only	91.92	74.51	31.73
TFT + MSTC + Reg	**90.98**	**67.56**	**28.85**

C. Clinical Safety Analysis

The missed-critical rate of 28.85% means that roughly seven out of ten Critical windows are correctly flagged. TFT Standard misses 42.31% of Critical events despite its higher reliance on the majority Normal class. The false-alarm rate of 2.89%—the lowest across all models—indicates that alert

fatigue, a well-documented problem in ICU environments, would be substantially reduced by our approach.

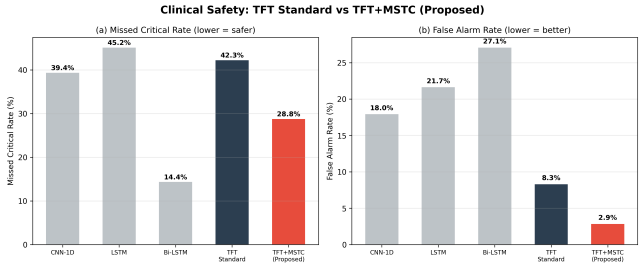

Fig. 6. Clinical safety comparison: (a) missed critical rate and (b) false alarm rate across all models.

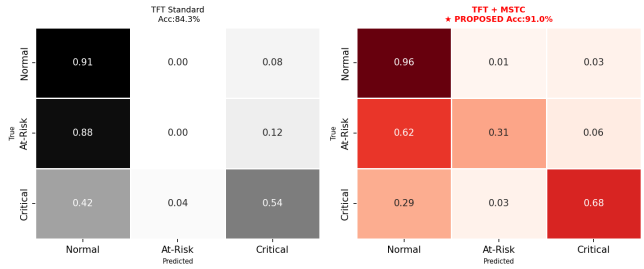

Fig. 7. Normalised confusion matrices for all five models. TFT+MSTC shows the strongest diagonal concentration.

D. Parameter Efficiency

MSTC adds only 450 parameters to the 158,034-parameter TFT baseline (0.28% increase). The three regularisation techniques add zero parameters. In contrast, the Bi-LSTM baseline contains 548,099 parameters yet achieves only 64.47% accuracy. This underscores that, for small clinical datasets, architectural parsimony combined with principled regularisation outperforms brute-force capacity.

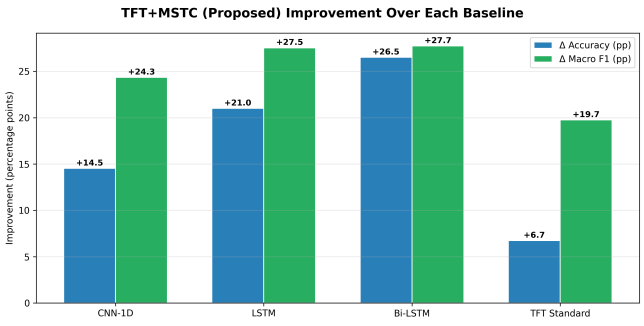

Fig. 8. Improvement of TFT+MSTC over each baseline in accuracy and macro F1 (percentage points).

VI. Discussion

The key finding of this study is that local temporal pattern extraction and training-time regularisation act synergistically when applied to small ICU datasets. MSTC provides the Transformer with pre-extracted local features—brief heart-rate spikes, respiratory-cycle oscillations—that global self-attention alone struggles to isolate from a sixty-second window. The regularisation strategies prevent the additional convolution parameters from overfitting, a failure mode we observed consistently in preliminary experiments where MSTC without regularisation degraded performance.

The clinical motivation behind the five-second kernel merits its emphasis. At heart rates of 60–100 BPM, five seconds covers five to eight cardiac cycles, sufficient to detect beat-to-beat variability and transient tachycardia. At respiratory rates of 12–20 breaths per minute, five seconds captures one full respiratory cycle, enabling detection of tachypnoea or apnoeic pauses. This physiological grounding distinguishes our approach from purely data-driven kernel selection.

Heart rate and respiratory rate have been identified as the two strongest individual predictors among all NEWS2 components for predicting ICU transfer and mortality [23]. Our dataset includes three of the six NEWS2 parameters—heart rate, oxygen saturation, and respiratory rate—representing the continuously monitored vital signs available at 1 Hz resolution. The remaining parameters (blood pressure, temperature, consciousness level) are not available in the BIDMC dataset.

A. Limitations

The BIDMC dataset contains only 53 patients from a single institution, limiting generalisability. The trend-based labelling scheme, while clinically motivated, relies on threshold calibration from a single stable patient. Blood pressure and temperature—present in NEWS2 but absent from BIDMC—would likely improve discrimination. We did not perform cross-validation due to the computational cost of training five model families; a future study with k-fold evaluation would strengthen statistical claims.

The regularization was tuned for 53 patients and may require adjustment for larger cohorts. MSTC's parameter count (+/-450) is negligible at scale, but overfitting risk decreases with more data, potentially reducing the need for aggressive regularization.

VII. Acknowledgment

We thank SRM Institute of Science and Technology for their support in this research.

VIII. Conclusion

We presented an enhanced Temporal Fusion Transformer that integrates a clinically motivated multi-scale temporal convolution module with domain-informed training strategies for ICU cardiac risk classification. On the BIDMC dataset, the proposed approach achieves 90.98% accuracy, 67.56% macro F1, and a 28.85% missed-critical rate—outperforming TFT Standard by 6.73 pp in accuracy and 19.74 pp in F1 while adding fewer than 500 trainable parameters. Future work will validate the model on the MIMIC-III Waveform Database (∼30,000 patients) with additional NEWS2 parameters including blood pressure and temperature.

REFERENCES

[1] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Trans. Manag. Inf. Syst.*, vol. 14, no. 1, pp. 1–32, 2023.

[2] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. ICLR*, 2016.

[3] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[4] U. Desai, C. G. Nayak and G. Seshikala, "An application of EMD technique in detection of tachycardia beats," *2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India* 2016, pp. 1420-1424, doi: 10.1109/ICCSP.2016.7754389.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[6] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021.

[7] R. Phetrittikun *et al.*, "Temporal fusion transformer for forecasting vital sign trajectories in intensive care patients," in *Proc. 13th BMEiCON*, IEEE, 2021, pp. 1–5.

[8] R. Y. He and J. N. Chiang, "TFT-multi: Simultaneous forecasting of vital sign trajectories in the ICU," *Sci. Rep.*, vol. 15, no. 14996, 2025.

[9] L. Kapral *et al.*, "Development and external validation of temporal fusion transformer models for continuous intraoperative blood pressure forecasting," *EClinicalMedicine*, vol. 75, p. 102797, 2024.

[10] P. Chang *et al.*, "A transformer-based diffusion probabilistic model for heart rate and blood pressure forecasting in Intensive Care Unit," *Comput. Methods Programs Biomed.*, vol. 246, p. 108060, 2024.

[11] Z. Cui, W. Chen, and Y. Chen, "Multi-Scale Convolutional Neural Networks for time series classification," *arXiv:1603.06995*, 2016.

[12] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE CVPR*, 2015, pp. 1–9.

[13] H. I. Fawaz *et al.*, "InceptionTime: Finding AlexNet for time series classification," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.

[15] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[16] K. Wu *et al.*, "Multiscale spatial-temporal transformer with consistency representation learning for multivariate time series classification," *Concurr. Comput.: Pract. Exp.*, vol. 36, 2024.

[17] S. Kiranyaz *et al.*, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, p. 107398, 2021.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016, pp. 2818–2826.

[20] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. ECCV*, 2016, pp. 646–661.

[21] M. A. F. Pimentel *et al.*, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1914–1923, 2017.

[22] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[23] S. Gerry *et al.*, "Modifications to the National Early Warning Score 2: a Scoping Review," *BMC Medicine*, vol. 23, 2025.