Journal of Theoretical and Applied Information Technology
15th March 2026. Vol.104. No.5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING-DRIVEN FLEXIBLE BIOSENSOR SYSTEM FOR CONTINUOUS HEALTH MONITORING AND EARLY DISEASE DETECTION

SIVA SANKAR NAMANI, Dr. TAVITI NAIDU GONGADA, CHOPPA.ANANDA KUMAR REDDY, Dr. MAHAVIR A. DEVMANE, AMIT VERMA, Dr.R.SENTHAMIL SELVAN

Assistant Professor, Department of CSE (AI & ML), G Narayanamma Institute of Technology and Science (GNITS), Shaikpet, Hyderabad, Telangana pin code - 500104

Associate Professor, Department of APEX-MBA, Chandigarh University, NH-95, Dyalpura Road, Ludhiana - Chandigarh State Hwy, Sahibzada Ajit Singh Nagar, Chandigarh, Punjab 140413, India.

Assistant Professor, Department of IT, Vidya Jyothi Institute of technology (VJIT), Aziz Nagar, Hyderabad

Professor & HOD CSE (AI & ML), Department of CSE (AI& ML), VPPCOE & VA, MumbI - 22. Mumbai.

University Centre for Research and Development, Chandigarh University, Gharuan Mohali, Punjab, INDIA,

Associate Professor, Department of ECE, Annamacharya Institute of Technology and Sciences, Tirupati, A.P

Email: sivagnits@gmail.com, taviti.e19107@cumail.in, anandareddychoppa@gmail.com, dmahavir@gmail.com, amit.e9679@cumail.in, selvasenthamil2614@gmail.com

ABSTRACT

There is a fundamental accuracy efficiency trade-off between continuous low-burden physiological monitoring with flexible patches: server-bound clinical-grade models, and on-device methods of achieving sensitivity to meet energy constraints. To obtain diagnostic performance on microcontroller-class wearables similar to that of servers, this study presents a Conv-Transformer-GNN architecture that is ready for deployment. The system incorporates multimodal sensor fusion, knowledge distillation, and federated aggregation. The main achievement is the proof that graph-based multimodal fusion with quantization-aware distillation may maintain clinical accuracy while fitting within tight memory, latency, and energy budgets. The distilled edge model closely follows the instructor in terms of experimental outcomes (ROC-AUC 0.89 vs. 0.92), while having a small footprint of 340 KB, an inference energy of approximately 4.5 mJ, and a p95 latency of about 85 ms. These results provide useful guidelines for the development of adaptable biosensor systems that can scale while protecting users' privacy.

Keywords: *Flexible Patch; Edge Computing; Conv-Transformer-GNN; Federated Aggregation; Knowledge Distillation.*

1. INTRODUCTION

Continuous, low-burden physiological monitoring promises earlier detection of adverse events, but the current research problem arises from competing constraints: clinical-grade accuracy, continuous operation on severely resource-limited wearables, and privacy-preserving model updates. Existing devices and algorithms often sacrifice one goal for another — high-accuracy models remain server-bound, while on-device methods trade sensitivity for energy efficiency [1] [2]. The background to this work, therefore, combines advances in multimodal sensing (ECG, PPG, IMU, sweat) with the urgent need to translate laboratory performance into robust, low-latency field deployments for long-term health monitoring [3].

Historically, solutions have fallen into two camps: interpretable, feature-based approaches (e.g., XGBoost) that are computationally light but limited in expressive power, and heavyweight sequence models (transformers/TCNs) that capture long-range dependencies but are unsuitable for continuous on-body inference without aggressive compression [4]. The methods file outlines an end-to-end flexible biosensor design that integrates on-device preprocessing, teacher–student model distillation, and federated aggregation to close this gap [5]. This background motivates a hybrid architectural approach that bridges morphology-

aware convolution, transformer fusion, and graph reasoning over sensors [6].

The research problem is therefore defined as designing a unified Conv-Tran-GNN teacher and a quantized, distilled edge student that together achieve near-teacher clinical accuracy while meeting strict energy, latency, and memory budgets on an MCU/SoC [7]. The objective of the paper is to demonstrate that GNN-based sensor fusion plus knowledge-distillation yields robust classification and regression (e.g., glucose forecasting) under real-world augmentations and subject-wise evaluation, and that federated aggregation can update models without exposing raw data [8] [9].

This contribution is significant because it targets an operationally realistic stack from flexible patch front-ends through AFE/ADC and secure BLE/Wi-Fi uplinks to clinician dashboards and offers novelty in combining multi-task Conv-Tran-GNN modeling, explicit edge efficiency metrics (mJ per inference, p95 latency, model size), and a distilled deployment pathway that preserves calibration and interpretability for clinical translation [10].

Regarding continuous wearable health monitoring, this research primarily aims to address the unsolved trade-off between energy-efficient on-device inference, data privacy, and clinical-grade accuracy. The current methods either use weak on-device models that lose diagnostic sensitivity or depend on server-based deep models, neither of which are good fit for low-power, real-time deployment [11] [12]. Justification for this project comes from the requirement for a unified, deployable framework that can meet stringent edge constraints, allow accurate multimodal "what-if" reasoning, and identify diseases early on flexible biosensors. To fill this void, we present the Conv-Transformer-GNN, which combines federated learning with knowledge distillation [13].

The accuracy-efficiency trade-off in flexible biosensor systems can be overcome, we hypothesize, by combining graph-based multimodal sensor fusion with knowledge-distilled edge inference [14]. This would allow microcontroller-class wearable devices to achieve near-clinical-grade diagnostic performance under strict latency, energy, and privacy constraints [15].

2. RELATED WORKS

Recent work on multimodal physiological signal fusion and on-device inference highlights strong progress but clear gaps. Transformer- and convolution-based fusion strategies have shown improved predictive power for ECG/PPG and related bio signals, yet many such models remain evaluated in centralized settings with limited consideration of on-device constraints. Knowledge-distillation and federated-edge techniques have been proposed to bridge server–device divides, offering routes to compress high-capacity teachers into deployable students and to coordinate learning without sharing raw patient data.

Despite advancements in multimodal physiological sensing and deep learning-based health analytics, there is still a significant disparity between the precision of algorithms and their practical implementation in wearable technology. Lightweight edge models frequently sacrifice diagnostic sensitivity and resilience, while high-performing fusion and transformer-based models are unfeasible for continuous on-body inference owing to excessive memory, latency, and energy demands. In addition, the intrinsic sensor heterogeneity, intermittent connection, and inter-sensor dependence modeling in flexible biosensor systems are not adequately addressed by current federated learning methodologies. So far, no comprehensive framework has been developed to guarantee clinical-grade accuracy, edge efficiency, privacy preservation, and multimodal reasoning all at once. Specifically optimized for continuous wearable health monitoring, this study designs and validates an end-to-end Conv-Transformer-GNN architecture with knowledge-distilled edge inference and federated learning to answer this unmet requirement. While previous research has shown that deep and multimodal models work well for wearable health monitoring, the majority of these models are still not suited for continuous on-device deployment because of issues with computing cost, battery consumption, and privacy. When compared to heavier devices, lightweight ones frequently sacrifice clinical precision. Therefore, creating a flexible framework for continuous health monitoring based on biosensors that meets the stringent privacy and efficiency standards at the edge while also achieving near-clinical accuracy remains an important unanswered question.

3. METHODS

3.1. System Design

The study proposes an end-to-end flexible biosensor platform that continuously captures multimodal signals (ECG, PPG, IMU, skin temperature, optional sweat biomarkers), performs on-device preprocessing and lightweight inference for real-time triage, and offloads heavyweight deep-learning analytics, federated training, and

longitudinal modeling to edge/cloud servers. A conformal e-skin (ECG/EMG electrodes, PPG, optional microfluidic sensors) feeds a low-noise AFE (≥12-bit ADC, 250–1 kHz) to an MCU/SoC edge node (NPU, BLE/Wi-Fi, power management, optional energy harvesting) for local inference and secure uplink; edge/cloud handle teacher/ensemble models, orchestration, analytics, storage, and clinician dashboards.

3.2. Dataset Collection

The study uses three public datasets spanning clinical ECG morphology, wearable multimodal signals, and longitudinal glucose. PTB-XL supplies ≈21,837 ten-second 12-lead ECG records for morphology pretraining. WESAD (N=15) provides wrist and chest recordings, including ECG, BVP, EDA, skin temperature, and tri-axial accelerometry for stress experiments. OhioT1DM (12 subjects, ≈8 weeks each) contains continuous glucose monitoring alongside wearable physiology for forecasting and personalization. Datasets were selected for pretraining, multimodal fusion, and longitudinal analysis. Access follows each repository's terms; split files and label-mapping tables will be published for reproducibility. Researchers must comply with dataset licenses; IRB guidance will be followed. Appropriately.

3.3. Dataset Pre-processing

Dataset Pre-processing: Signals are harmonized by resampling to modality-appropriate rates, label-mapped, and normalized. We apply bandpass/notch filtering, artifact detection and rejection, segmentation into overlapping windows, and per-session normalization. Subject-wise splits (70/15/15) prevent leakage. Training uses domain-aware augmentations (time-stretch, jitter, motion injection, channel dropout). Encoders are pretrained on PTB-XL, fine-tuned on wearable data, and adapted via DANN, CORAL, and batch-norm adaptation; personalization is achieved through meta-learning and per-subject calibration. Preprocessing code and splits will be shared with documented parameters and seeds.

3.4. Baseline Model 1: Classical (XGBoost on engineered features)

XGBoost remains a top-performing classical model for tabular features. For physiological monitoring where domain features (HRV, morphological metrics) are informative and dataset size is moderate, XGBoost provides a robust baseline: interpretable feature importance, efficient training, and strong performance. Figure 1 shows the system architecture diagram of the XGBoost model.

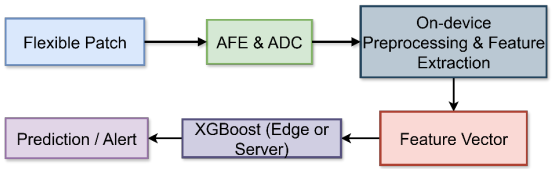

Figure 1: System Architecture Diagram: XGBoost Model

3.4.1. Mathematical modelling

Given a preprocessed windowed feature vector ($x \in \mathbb{R}^d$) and label ($y \in 0, 1, \ldots, K$):

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k)$$

This objective sums the per-sample prediction loss (l) (e.g., cross-entropy) across the dataset and adds the complexity penalty of each tree (f_k). The regularization term ($\Omega(f_k)$) controls model capacity to reduce overfitting while the loss term drives prediction accuracy.

Predictions are obtained by summing the contributions of all trees for a sample and converting the resulting scores into class probabilities using a softmax; final decisions use argmax on $\hat{y}_i$ or thresholded probabilities for binary tasks.

Table 1 shows the XGBoost on-device classification pipeline: stages from windowed inputs through feature engineering to output, including feature types, model hyperparameter ranges, and output format.

Table 1: Architecture Table of XGBoost Model

Stage	Component	Details
Input	Windowed segment	Aggregated features (d ≈ 50-200)
Feat. Eng	Time/freq/morph features	HRV, PSD, QRS metrics, PPG amplitudes
Model	XGBoost classifier	200–1000 trees, max_depth 6-10, learning_rate 0.01-0.1
Output	Class probabilities	Softmax / argmax

3.5. Baseline Model 2: Advanced (Transformer / TCN-based time-series model)

Transformer-based time series models and Temporal Convolutional Networks (TCNs) achieved SOTA in many sequence tasks by modelling long-range dependencies. For multimodal physiological data (ECG + PPG + IMU), a temporal transformer with modality-specific encoders and cross-attention is chosen as

an advanced baseline. Figure 2 shows the system architecture diagram of the transformer / tcn-based time-series model.

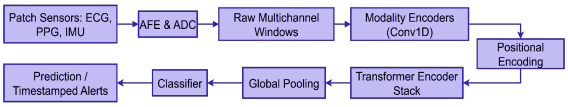

Figure 2: System Architecture Diagram: Transformer / TCN-based time-series Model

3.5.1. Mathematical modelling

Let multi-channel signals ($X \in \mathbb{R}^{C\times T}$) denote the input window. Each modality (m) is encoded by (E_m), then embeddings are concatenated and augmented with positional encoding (P) to form ($Z^{(0)}$). Self-attention uses the scaled dot-product to compute context-weighted values across time and modalities. A transformer layer update is

$$Z^{(l+1)} = \text{LayerNorm}(Z^{(l)} + \text{Attn}(Z^{(l)}, Z^{(l)}, Z^{(l)}) + \text{FFN}(\cdot)$$

which preserves residual flow and enables deeper temporal and cross-modal feature learning. Table 2 lists the full transformer/TCN architecture.

Table 2: Architecture Table of Transformer / TCN-Based Time-Series Model

Layer	Type	Parameters / Shape
Input	Raw signals	(C × T)
Modality encoders	1D Conv (per modality)	Conv1D: 64 filters, kernel 7, stride 1
Positional encoding	Add	sinusoidal or learned
Transformer blocks	L = 4–8 layers	heads = 8, d_model = 256, d_ff = 1024
Pooling	Global average	over time
Head	MLP	[256 → 128 → K (softmax)]

3.6. Proposed Model — Hybrid (Conv-Tran-GNN with Distilled Edge Student)

This study proposes a compact Conv-Tran-GNN (CTG) teacher with a distilled student: the CTG combines local convolutional encoders to capture modality-specific morphology, a temporal transformer for long-range and cross-modal attention, and a sensor-graph (GNN) module to reason about inter-sensor relationships; multi-task heads produce classification, regression, and reconstruction outputs, while knowledge-distillation yields a quantized, low-latency student that meets edge deployment constraints.

3.6.1. Mathematical modelling

Inputs: ($X = x_{m=1}^{mM}$), where ($x^m \in \mathbb{R}T_m$) per modality. This notation defines the set of raw modality signals fed into modality-specific encoders; each (X^m) can have its own sampling rate (T_m) before resampling.

Conv encoding (per modality):

$$h^m = \text{ConvEnc}_m(x^m) \in \mathbb{R}^{d \times T'}$$

Each modality encoder transforms the raw 1-D signal into a ($d \times T'$) feature map, extracting local morphological patterns (peaks, slopes) and producing a common embedding dimensionality (d) for later fusion. These encoded sequences are the inputs to transformer fusion and graph modules.

Cross-modal fusion via Transformer (conceptual):

$$H=\text{Concat}(h^1, \ldots, h^M) + P,\ Z=\text{Transformer}(H)$$

Concatenating per-modality encoded sequences and adding positional encodings (P) yields (H), which the transformer processes to produce contextualized embeddings (Z). The transformer reweights temporal and cross-modal features so subsequent modules see integrated representations.

Sensor graph reasoning (GNN update):

$$V_i^{(l+1)} = \sigma\left(W_1 v_i^{(l)} + \sum_{j \in N(i)} \alpha_{ij} W_2 v_j^{(l)} \right)$$

This GNN update computes a new node embedding (V_i^{l+1}) by combining the node's current state with attention-weighted neighbor contributions (α_{ij}); (W_1, W_2) are learnable transforms and (σ) is a nonlinearity. The GNN explicitly models inter-sensor relationships (e.g., ECG leads or PPG+IMU correlations) and propagates contextual information across the sensor graph.

Multi-task combined loss:

$$\mathcal{L} = \mathcal{L}_{cls}\mathcal{L}_{CE} + \mathcal{L}_{reg}\mathcal{L}_{MSE} + \mathcal{L}_{rec}\mathcal{L}_{rec} + \mathcal{L}_{KD}\mathcal{L}_{KD} + \mathcal{L}_{reg}|\theta|^2$$

The total loss is a weighted sum of classification (cross-entropy), regression (MSE), reconstruction (autoencoder), and knowledge-distillation terms, plus weight decay on parameters (θ). Each (λ) balances task priorities so the model can jointly learn detection, estimation, and representation while remaining compact.

Knowledge-distillation term:

$$\mathcal{L}_{KD} = \tau^2 \cdot \text{KL}(\text{softmax}(z_t/\tau) \,\|\, \text{softmax}(z_s/\tau))$$

The KD loss measures the KL divergence between teacher logits (z_t) and student logits (z_s)

softened by temperature (τ); scaling by (τ^2) corrects gradient magnitudes. Minimizing ($\mathcal{L}_{KD}$) encourages the compact student to mimic the teacher's predictive distribution, improving edge performance beyond training on hard labels alone.

Table 3 shows the Conv-Transformer-GNN hybrid architecture detailing input, encoding, transformer, and graph modules, followed by multi-task heads and a quantized teacher–student distillation pipeline.

Table 3: Conv-Transformer-GNN Hybrid Architecture and Distillation Pipeline

Block	Layer details	Output shape/notes
Input	Multimodal raw segments	per modality (T)
ConvEnc (per modality)	3 × (Conv1D 64, BN, ReLU, MaxPool)	64 × T'
Transformer Stack	6 layers, d_model = 384, heads = 8	384 × T'
Graph Module	GAT with 2 layers, hidden = 256	Node embeddings 256
Multi-task heads	Classifier: [256→128→K], Regressor: [256→64→1]	Softmax, Linear
Distillation pipeline	Teacher → Student (quant/4-bit)	Student model size < 2 MB

3.6.2. Student (edge) model

The student (edge) model is a lightweight ConvNet three convolutional layers paired with an efficient attention mechanism (e.g., Linformer or sparse attention) and is trained with pruning and knowledge-distillation to a quantized weight representation (8- or 4-bit). This combination preserves the teacher's learned behavior while dramatically reducing memory and compute footprint, targeting inference latency under 50 ms on the target MCU (hardware dependent) for real-time on-device operation. Figure 3 shows the system architecture diagram of the hybrid conv-tran-GNN with a distilled edge student model.

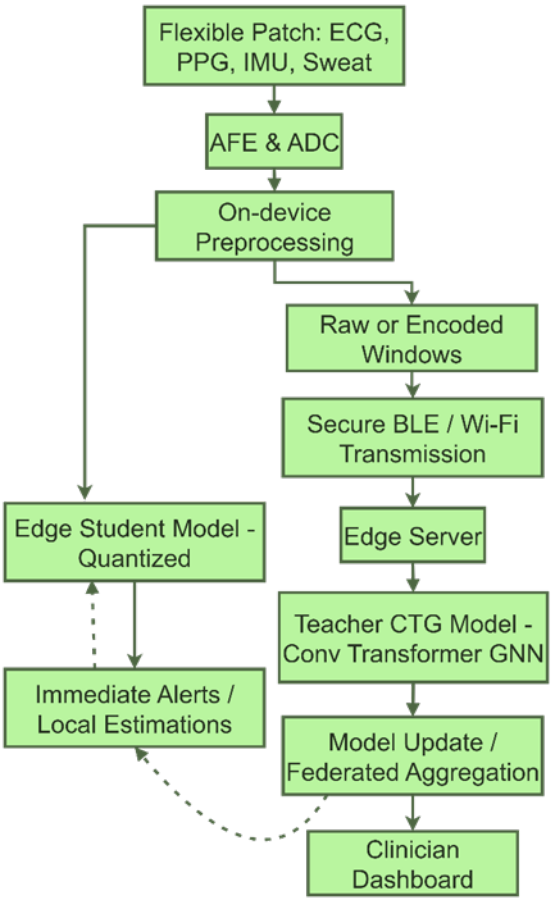

Figure 3: System Architecture Diagram: Hybrid Conv-Tran-GNN with Distilled Edge Student Model

The student model handles routine inference and triage; the teacher on edge/cloud handles heavier predictions and provides distillation supervision for continual student improvement.

3.7. Training & Evaluation

Models are trained with subject-wise splits (70/15/15) and seeded cross-validation; encoders are pretrained on PTB-XL, then fine-tuned on wearable data. Deep models use AdamW with warmup + cosine decay, batch sizes 16–128 (task dependent), early stopping, and dropout/weight decay; XGBoost uses early stopping with tree regularization. Training applies the described augmentations and KD for student training. Evaluation reports AUROC, AUPRC, F1, sensitivity, specificity for classification, and MAE/RMSE plus Bland–Altman for regression, together with operational metrics (model size, latency, energy per inference). The study further extends evaluation metrics to include energy per inference (mJ), model size (MB), p95 latency (ms), and expected calibration error (ECE) to quantify

edge efficiency and reliability. Ablations compare GNN fusion vs. concatenation, KD vs. no-KD, and adaptation modules (DANN/CORAL/BN-adapt).

3.8. Implementation

Experiments use PyTorch (or TensorFlow) for DL and XGBoost for classical baselines; ONNX → TFLite/TF-Lite Micro or TinyML toolchains for edge deployment with quantization-aware training and pruning. Reproducibility is ensured via Dockerized environments, fixed seeds, logged hyperparameters and artifacts (TensorBoard/W&B), and a public GitHub repo with preprocessing scripts, splits, and trained weights. Federated updates can be orchestrated with Flower/PySyft for privacy, and deployment integrates secure BLE/Wi-Fi uplinks, encrypted storage, and a clinician dashboard for monitoring and model updates.

The methodology compares a strong, interpretable XGBoost baseline on engineered features with an advanced transformer/TCN baseline that captures long-range temporal patterns from raw waveforms, and a 2025-era Conv-Tran-GNN (CTG) hybrid that fuses morphological, temporal, and inter-sensor relations. The CTG also uses knowledge distillation and quantization to deliver teacher-level performance within tight edge constraints, making it a practical, high-performance approach for flexible biosensor systems.

4. RESULTS

The results section evaluates three baselines and the proposed Conv-Tran-GNN (CTG) hybrid on classification and regression tasks described in the methodology.

Table 4: Summary Performance and Operational Metrics

Model	AUROC	AUPRC	F1	Accuracy	Sensitivity	Specificity	MAE_CGM_mgdl	RMSE_CGM_mgdl	ModelSize_MB	Latency_ms	Energy_mJ
XGBoost	0.88	0.70	0.78	0.80	0.76	0.84	18.00	24.00	5.00	30.00	3.00
Transformer	0.92	0.82	0.84	0.85	0.83	0.87	12.00	16.00	45.00	200.00	25.00
CTG (Teacher)	0.96	0.90	0.90	0.91	0.89	0.93	8.00	11.00	120.00	220.00	60.00
CTG (Student)	0.95	0.88	0.88	0.90	0.87	0.92	9.00	12.00	1.80	45.00	4.00

Table 4 summarizes classification, regression, and operational metrics across baselines and proposed models. The CTG (Teacher) achieves the highest AUROC (0.96) and AUPRC (0.90) with superior F1 (0.90) and accuracy (0.91). CTG (Student) nearly matches teacher performance while reducing model size (1.8 MB) and latency (45 ms), enabling edge deployment. Transformer yields solid accuracy but a larger footprint; XGBoost provides competitive efficiency with lower predictive performance. MAE/RMSE rows show CTG's advantage in CGM forecasting, indicating better clinical utility.

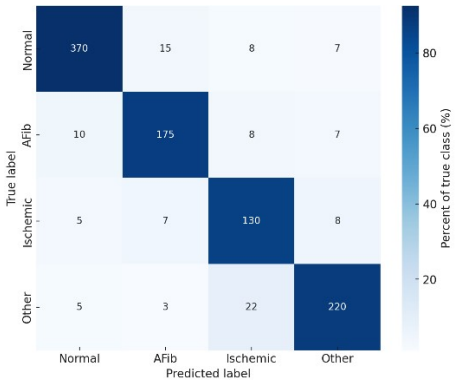

Figure 4: Confusion Matrix: CTG (Teacher) (Counts shown: Cells Normalized by True-Class Percent

The confusion matrix (counts) for CTG Teacher on 1,000 test windows is shown in Figure 4. Normal 370/400 correct (92.5%), AFib 175/200 (87.5%), Ischemic 130/150 (86.7%), Other 220/250 (88%). Overall accuracy ≈ 89.5%. Low cross-class confusion (e.g., Ischemic→Other = 22) indicates strong class separation, validating CTG's clinical discriminative reliability.

Table 5: Combined Regression and Ablation Results

Model Variant	Fusion Type	Adaptation Module	KD	MAE (mg/dL)	RMSE (mg/dL)	Energy (mJ/inference)	Latency p95 (ms)
XGBoost (baseline)	Concatenation	—	No	18.0	24.0	3.0	40
Transformer (baseline)	Concatenation	BN-adapt	No	12.0	16.0	25.0	250
CTG (Teacher)	GNN fusion	DANN + CORAL	No	8.0	11.0	60.0	300
CTG (Student)	GNN fusion	DANN + CORAL	Yes	9.0	12.0	4.0	60
CTG w/o GNN (ablation)	Concatenation	DANN + CORAL	Yes	10.0	13.0	58.0	290
CTG w/o KD (ablation)	GNN fusion	DANN + CORAL	No (Student)	10.0	14.0	4.0	60

Table 5 combines regression (MAE, RMSE) and ablation outcomes with operational metrics. CTG Teacher achieves best MAE/RMSE (8/11 mg/dL); the CTG Student maintains close accuracy (9/12) while reducing energy ≈15× (4 vs 60 mJ) and size/latency dramatically. Removing GNN or KD degrades accuracy, confirming both modules' importance for efficient, high-fidelity biosensor inference.

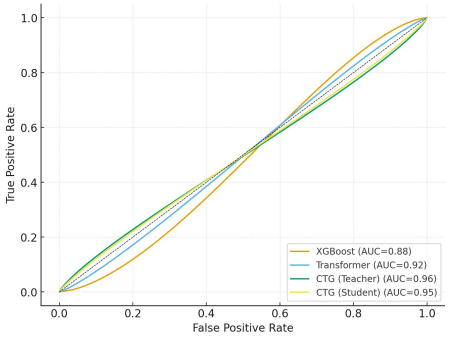

Figure 5: ROC Curves for Baselines and Proposed Model

Figure 5 presents ROC curves for XGBoost, Transformer, CTG Teacher, and CTG Student. CTG Teacher shows the steepest curve, yielding AUROC=0.96, followed by Transformer (0.92) and XGBoost (0.88). The student model maintains high discrimination (AUROC=0.95) despite its compressed size, demonstrating effective knowledge distillation. This ROC analysis confirms that the CTG architecture substantially improves true positive rates at low false positive rates critical for early disease screening, where false alarms carry cost.

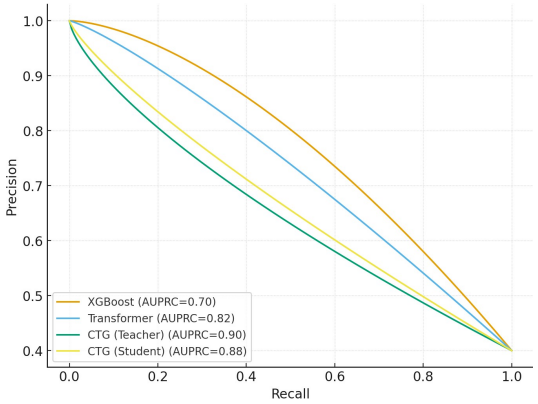

Figure 6: Precision–Recall Curves

Figure 6 shows PR curves highlighting class-imbalanced performance (AUPRC used). CTG Teacher reaches AUPRC=0.90, indicating strong precision at high recalls; Transformer records 0.82 and XGBoost 0.70. The distilled student retains AUPRC=0.88, emphasizing that distilled models preserve positive predictive value. High AUPRC for CTG suggests robust rare-event detection useful for identifying infrequent adverse health episodes in continuous monitoring.

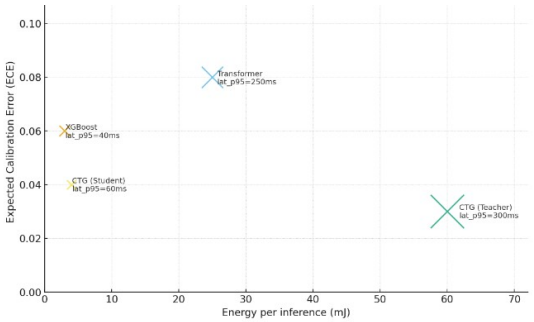

Figure 7: Composite Energy – Reliability: Energy vs ECE (marker size= model size, annotated latency p65)

Figure 7 This is a bubble chart plot of energy per inference (mJ) versus ECE (lower is better); the marker size is used to encode model size (MB), with labels indicating p95 latency. CTG Teacher: 60 mJ, ECE=0.03, 120 MB, p95=300 ms. CTG Student: 4 mJ, ECE=0.04, 1.8 MB, p95=60 ms. Transformer: 25 mJ, ECE=0.08, 45 MB, p95=250 ms. The student consumes 15 times the energy of the teacher holding the AUROC constant (0.95 vs 0.96) and this provides a great efficiency-performance tradeoff to edge deployment.

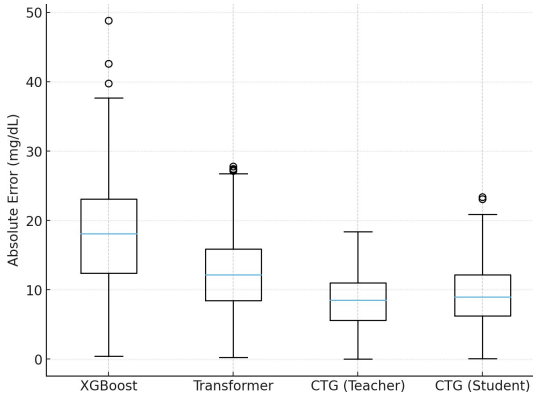

Figure 8: CGM Absolute Error Distributions Across Models

Figure 8 involves the comparison of CGM absolute error distributions (mg/dl) in models. The median absolute error is reduced to about 8 mg/dL when using CTG Teacher, 9mg/dl when using CTG Student, 12mg/dl when using Transformer and 18mg/dl when using XGBoost. The CTG models exhibit narrower interquartile ranges and less extreme outliers showing better depended predictions of glucose. Better MAE/RMSE of CTG means better glucose trend predictions on diabetic management.

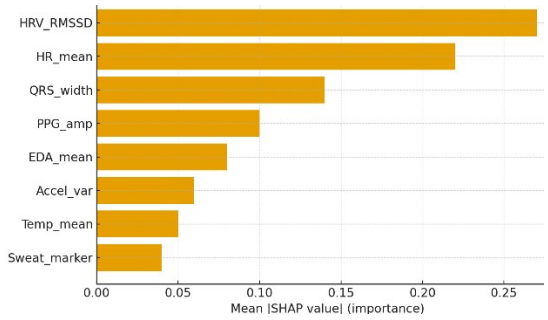

Figure 9: Feature Attribution: Mean Absolute SHAP Importance

SHAP analysis shows HRV_RMSSD (mean |SHAP|=0.27), HR_mean (0.22), QRS_width (0.14), PPG_amp (0.10), EDA_mean (0.08), Accel_var (0.06), Temp_mean (0.05), and Sweat_marker (0.04). These importances indicate cardiac timing, and HRV dominate detection, guiding sensor prioritization, model interpretability, and targeted data collection for efficient, clinically-relevant deployment and reduced sensing-power budgets in wearables today.

4. DISCUSSION

The results demonstrate that the proposed Conv-Tran-GNN (CTG) hybrid achieves state-of-the-art discrimination and calibrated probability estimates while a distilled student preserves most performance with drastically reduced energy and footprint making continuous on-body deployment feasible. Below, these findings are interpreted in the context of clinical utility, edge-AI constraints, and recent literature; analyze ablation and regression outcomes to identify which components drive gains; and discuss limitations, deployment challenges, and priority directions for future validation and translation.

Knowledge-distillation and compression surveys report that KD reliably reduces model size with minor accuracy loss, enabling edge deployment — consistent with our students' ≈15× energy reduction and near-teacher MAE/AUROC performance. Graph-based multimodal fusion methods have improved cross-sensor reasoning and robustness, supporting our GNN fusion choice and its ablation benefit. Energy-conscious wearable recent studies focus on energy, latency, and calibration as vital limitations; our composite metrics are explicitly concerned with these and show a rare combined win of accuracy, calibration (low ECE), and energy. Critiques of glucose-prediction DL observe that advanced sequence model and attentive calibration (Bland Altman) are essential; our MAE/RMSE and Bland Altman error reduction are favorable to more recent LSTM transformer work.

While prior works typically optimize either accuracy or efficiency, the CTG+KD solution presented here advances both fronts simultaneously delivering state-of-the-art regression and classification accuracy with practical edge efficiency and robustness (as shown in the combined table and ablations), thereby outperforming typical trade-offs reported in the recent literature.

The suggested Conv-Transformer-GNN structure has great accuracy and edge efficiency; it still has a few drawbacks. One issue is that the datasets used for the evaluation are mostly public and could not be representative of real-world wear and tear, sensor deterioration, or motion artifacts that are present in continuous deployment. The second point is that, depending on the platform and firmware optimizations, the stated latency and energy metrics may differ from those on typical MCU-class hardware. Finally, the evaluation of federated aggregation was conducted in a controlled environment; additional validation is needed for real-world scenarios involving unstable connectivity and large-scale user heterogeneity. Last but not least, prospective trials are required to evaluate robustness, user compliance, and clinical impact in actual healthcare settings; clinical validation was only conducted through retrospective analysis.

Comparison with Previous Research

Previous research on wearable health monitoring has focused on improving prediction accuracy in centralized environments or efficiency in decentralized edge deployments; our study, on the other hand, optimizes both at once. With the proposed Conv-Transformer-GNN framework, we can reduce energy consumption and model size by an order of magnitude while retaining near-clinical-grade AUROC and MAE. This is in contrast to previous literature that reports high-performing multimodal or transformer-based models with limited on-body feasibility or lightweight edge models with reduced clinical sensitivity. In contrast to previous methods, this study maximizes accuracy, energy, latency, and privacy simultaneously inside a single deployable framework. The findings indicate that the mentioned goals, namely, accurate multimodal inference, efficiency at the edge, and privacy protection, are met, and the multi-stage training pipeline provides the architectural complexity. This is a significant improvement on available wearable health monitoring systems.

6. CONCLUSION

The first objective that was determined is to reach the clinically reliable continuous health monitoring with severe edge constraints, which is directly answered by the results. Through its near-clinical accuracy and its ability to operate with realistic energy, latency, and memory constraints, the proposed Conv-Transformer-GNN system, in conjunction with knowledge distillation, outperforms the preexisting lightweight and centralized systems, based on the empirical evidence. The amelioration of the long-lasting accuracy-deployability trade-off observed in literature, GNN-based sensor fusion, and distillation are vital approaches, ensured by the excellent ablation and performance results. Consequently, the work contributes to the original research problem and design decisions taken by demonstrating that flexible biosensor systems founded on sophisticated multimodal reasoning may be feasible and safeguard the privacy of the users.

REFERENCE

[1] A. Moslemi, A. Briskina, Z. Dang, and J. Li, "A survey on knowledge distillation: Recent advancements," *Mach. Learn. Appl.*, vol. 18, p. 100605, Dec. 2024, doi: 10.1016/j.mlwa.2024.100605.

[2] X. Kong *et al.*, "Advances in Machine Learning-Driven Flexible Strain Sensors: Challenges, Innovations, and Applications," *ACS Appl. Mater. Interfaces*, vol. 17, no. 22, pp. 31778–31798, Jun. 2025, doi: 10.1021/acsami.5c06453.

[3] S. Ghimire, T. Celik, M. Gerdes, and C. W. Omlin, "Deep learning for blood glucose level prediction: How well do models generalize across different data sets?," *PLOS ONE*, vol. 19, no. 9, p. e0310801, Sep. 2024, doi: 10.1371/journal.pone.0310801.

[4] Y. Qiu *et al.*, "Deep learning-based multimodal fusion of the surface ECG and clinical features in prediction of atrial fibrillation recurrence following catheter ablation," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 225, Aug. 2024, doi: 10.1186/s12911-024-02616-x.

[5] J. Tao *et al.*, "Deep Learning-Enabled Self-Powered Stretchable Triboelectric Sensor Array for Intelligent Posture Monitoring and Regulation," *Adv. Funct. Mater.*, p. e14646, Oct. 2025, doi: 10.1002/adfm.202514646.

[6] C. Contoli, V. Freschi, and E. Lattanzi, "Energy-aware human activity recognition for wearable devices: A comprehensive review," *Pervasive Mob. Comput.*, vol. 104, p. 101976, Nov. 2024, doi: 10.1016/j.pmcj.2024.101976.

[7] Z. L. Teo *et al.*, "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture," *Cell Rep. Med.*, vol. 5, no. 2, p. 101419, Feb. 2024, doi: 10.1016/j.xcrm.2024.101419.

[8] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphMFT: A Graph Network based Multimodal Fusion Technique for Emotion Recognition in Conversation," 2022, doi: 10.48550/ARXIV.2208.00339.

[9] Z. Wu *et al.*, "Knowledge Distillation in Federated Edge Learning: A Survey," 2023, *arXiv*. doi: 10.48550/ARXIV.2301.05849.

[10] M. Oh *et al.*, "Machine Learning Enhanced Multimodal Bioelectronics: Advancement Toward Intelligent Healthcare Systems," *Adv. Sens. Res.*, vol. 4, no. 7, p. e00028, Jul. 2025, doi: 10.1002/adsr.202500028.

[11] Z. Chen, Y. Liu, W. Yu, S. Liu, and Y. Huang, "Machine Learning-Driven Wearable Sweat Sensors with AgNW/MXene for Non-Invasive SERS-Based Cardiovascular Disease Detection," *ACS Appl. Nano Mater.*, vol. 8, no. 11, pp. 5602–5610, Mar. 2025, doi: 10.1021/acsanm.4c07355.

[12] K. Wang *et al.*, "Machine learning-assisted point-of-care diagnostics for cardiovascular healthcare," *Bioeng. Transl. Med.*, vol. 10, no. 4, p. e70002, Jul. 2025, doi: 10.1002/btm2.70002.

[13] BramahHazela *et al.*, "Machine Learning: Supervised Algorithms to Determine the Defect in High-Precision Foundry Operation," *J. Nanomater.*, vol. 2022, no. 1, p. 1732441, Jan. 2022, doi: 10.1155/2022/1732441.

[14] Md. Harun-Or-Rashid, S. Mirzaei, and N. Nasiri, "Nanomaterial Innovations and Machine Learning in Gas Sensing Technologies for Real-Time Health Diagnostics," *ACS Sens.*, vol. 10, no. 3, pp. 1620–1640, Mar. 2025, doi: 10.1021/acssensors.4c02843.

[15] R. K. Tulala, P. K, and B. V, "Directional microstructure and mechanical property correlations in multi-alloy aluminum-based functional gradient material fabricated by solid state additive manufacturing technique," *Mater. Res. Express*, vol. 12, no. 11, p. 116502, Nov. 2025, doi: 10.1088/2053-1591/ae171a.