Transformer Meets Gated Residual Networks To Enhance PICU's PPG Artifact Detection Informed by Mutual Information Neural Estimation

Thanh-Dung Le, *Senior Member, IEEE,* Clara Macabiau, Kevin Albert,
Symeon Chatzinotas, *Fellow, IEEE,* Philippe Jouvet, and Rita Noumeir, *Member, IEEE*

***Abstract*—This study delves into the effectiveness of various learning methods in improving Transformer models, focusing mainly on the Gated Residual Network (GRN) Transformer in the context of pediatric intensive care units (PICU) with limited data availability. Our findings indicate that Transformers trained via supervised learning are less effective than MLP, CNN, and LSTM networks in such environments. Yet, leveraging unsupervised and self-supervised learning on unannotated data, with subsequent fine-tuning on annotated data, notably enhances Transformer performance, although not to the level of the GRN-Transformer. Central to our research is analyzing different activation functions for the Gated Linear Unit (GLU), a crucial element of the GRN structure. We also employ Mutual Information Neural Estimation (MINE) to evaluate the GRN's contribution. Additionally, the study examines the effects of integrating GRN within the Transformer's Attention mechanism versus using it as a separate intermediary layer. Our results highlight that GLU with sigmoid activation stands out, achieving 0.98 accuracy, 0.91 precision, 0.96 recall, and 0.94 F1 score. The MINE analysis supports the hypothesis that GRN enhances the mutual information (MI) between the hidden representations and the output. Moreover, using GRN as an intermediate filter layer proves more beneficial than incorporating it within the Attention mechanism. In summary, this research clarifies how GRN boosters GRN-Transformer's performance surpasses other techniques. These findings offer a promising avenue for adopting sophisticated models like Transformers in data-constrained environments, such as PPG artifact detection in PICU settings.**

***Index Terms*—clinical PPG signals, Transformers, Gated Residual Networks, imbalanced classes, and mutual information.**

I. Introduction

Recently, the PICU at CHU Sainte-Justine (CHUSJ) has achieved significant progress by establishing a high-resolution research database (HRDB) [1], [2]. This state-of-the-art database seamlessly integrates biomedical signals from various monitoring devices into the electronic patient record, enhancing data continuity throughout a patient's PICU stay [3]. The implementation of HRDB has notably improved the Clinical Decision Support System (CDSS) at CHUSJ, elevating patient safety and supporting decision-making with substantial evidence [4]. A critical aim of the CDSS at CHUSJ is the prompt and precise diagnosis of acute respiratory distress syndrome (ARDS). Monitoring Oxygen saturation (SpO2) values, crucial for ARDS diagnosis, is key in predicting and managing ARDS [5], [6]. These values are also vital in determining respiratory support strategies [7]–[9]. Additionally, the ability to predict SpO2 from Photoplethysmography (PPG) waveforms and non-invasive blood pressure estimation is increasingly acknowledged as crucial for enhancing CDSS functionalities [10], [11]. Therefore, accurately identifying and discarding erroneous waveforms and SpO2 values from CDSS inputs is of utmost importance. Maintaining the accuracy of these inputs is crucial for the operation of the CDSS, thereby directly influencing patient outcomes and the efficiency of care.

In the current landscape of clinical data application, recent researches focus on improving PPG artifact detection through advanced machine learning (ML) techniques to enhance diagnostic accuracy in settings like the PICU. While previous studies, including those by Macabiau et al. [12], have explored ML approaches in PPG artifact detection, they highlight ongoing challenges such as limited data availability and significant class imbalances, which restrict the performance of fully supervised ML models. Traditional Transformer models, while promising due to their powerful attention mechanisms, have proven less effective in such constrained scenarios. Compared to these models, semi-supervised methods like label propagation and traditional algorithms such as K-Nearest Neighbors (KNN) have provided reasonable accuracy but still fall short in environments with high variance and limited annotations.

Another work introduces a novel architecture to address these challenges by integrating GRN into the Transformer model, creating the GRN-Transformer hybrid [13]. This model seeks to leverage GRN's structural advantages in handling limited data and class imbalances, aiming to improve artifact detection accuracy and reliability over traditional approaches. However, the performance of supervised approaches remains tied to the availability of annotated data, a well-known limitation in clinical environments [14].

In response to this constraint, an alternative approach explored self-supervised learning (SSL) as an alternative to

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC), in part by the Institut de Valorisation des données de l'Université de Montréal (IVADO), in part by the Fonds de la recherche en sante du Quebec (FRQS).

Thanh-Dung Le is with the Biomedical Information Processing Lab, Ecole de Technologie Superieure, University of Quebec, Canada, and also is with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg (Email: thanh-dung.le@uni.lu).

Symeon Chatzinotas is with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg.

Kevin Albert, and Philippe Jouvet are with the CHU Sainte-Justine Research Center, University of Montreal, Montréal, Québec, Canada.

Clara Macabiau and Rita Noumeir is with the Biomedical Information Processing Lab, École de Technologie Supérieure, University of Québec, Montréal, Québec, Canada.

supervised training to enhance the Transformer's performance under data-limited conditions [14]. Although SSL helped improve the model's robustness, it did not match the performance gains observed with the GRN-Transformer regarding artifact detection accuracy. Therefore, as shown in Table I, a comparative analysis of supervised, unsupervised, and self-supervised methods reveals the GRN-Transformer as the most proficient model, particularly in its accuracy and recall performance, underscoring the value of the GRN-Transformer's integration.

The motivation behind this study extends further into understanding the role of the GLU within the GRN structure. Studies have shown GLU's effectiveness in various NLP tasks, often attributed to its ability to optimize perplexity scores and enhance language understanding capabilities [15]. In the context of the GRN-Transformer, GLU serves as a core component, yet its specific impact on artifact detection within our framework remains largely unexplored. First, we seek to clarify this role by examining different GLU activation functions to determine the most effective configurations for handling limited clinical data. Then, to quantify the GRN's contribution to the model's performance more rigorously, we employ MINE [16]. MINE enables us to quantify and analyze the flow of information between the GRN and the Transformer components, revealing how effectively the GRN captures and transfers information across layers. By facilitating reliable MI estimation, MINE aids in identifying dependencies that contribute to performance improvements. These insights are valuable for understanding the GRN's role and optimizing the overall network architecture for tasks that require handling PPG artifact detection.

Despite the growing availability of high-resolution PICU waveforms, reliable PPG-artifact detection remains elusive because (i) labelled data are scarce and highly imbalanced, and (ii) standard Transformer blocks overfit or under-utilise such limited information. Our objective is therefore two-fold: (1) to determine whether a Transformer augmented with a GRN and carefully chosen gated activations can overcome these data constraints, and (2) to quantify, through MI analysis, the extent to which the GRN improves representation quality and downstream diagnostic performance. Our explicit contributions include:

- Systematic activation analysis: testing 11 gated or smooth activations inside the GRN to identify functions best suited to noisy, low-volume clinical data.
- GRN-Transformer architecture: embedding a GRN layer - acting as an intermediate representation filter that down-weights noisy features and amplifies the most predictive patterns, into both standard Attention and lightweight Gated Attention Unit (GAU) stacks [17], yielding a family of models that trade off accuracy and compute.
- Bayesian justification of Sigmoid gating: We show that the Sigmoid activation in the GRN layer arises naturally as a Bayes-optimal mapping from evidence to posterior probability, enabling minimal error, effective noise suppression, and preservation of task-relevant signals.
- Information-theoretic explainability: by applying MINE and t-SNE, we show that the GRN increases feature–label MI by x2.6 and produces cleaner class clusters, directly linking architectural changes to performance gains.

TABLE I: A summary of Transformer's performance for PPG artifact detection at CHUSJ.

Transformer-based Models	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)
Supervised [12]	0.95	0.85	0.86	0.85
Unsupervised (AE) [14]	0.97	0.89	0.93	0.91
Self-supervised [14]	0.97	**0.93**	0.92	**0.93**
GRN-Transformer [13]	**0.98**	0.90	**0.97**	**0.93**

Bold denotes the best values.

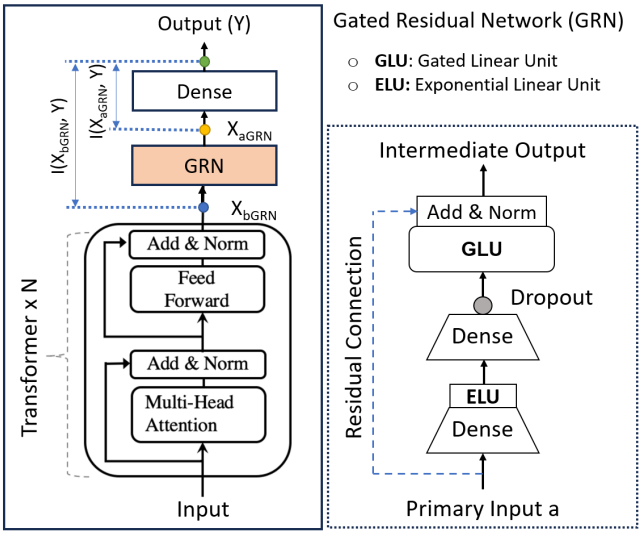

Fig. 1: An end-to-end process diagram workflow demonstration. *Left*: an N-layer Transformer encoder processes the input sequence, after which a GRN is inserted as an intermediate representation layer that filters and re-weights the encoder features before the final dense output. *Right*: the GRN itself couples an ELU-activated dense projection with a GLU gate inside a Residual-Add-&-Norm block; this design adaptively amplifies informative patterns while suppressing noise, providing a cleaner feature space for the downstream classifier.

- Benchmarking on PICU data: using CHUSJ's high-resolution database, we demonstrate that the proposed GRN-Transformer outperforms supervised, semi-supervised, and self-supervised baselines while maintaining bedside-friendly compute requirements.

The remainder of the manuscript is organised as follows. Section II introduces the GRN-Transformer architecture, and Section III-A details the CHUSJ dataset and the artifact-annotation protocol. Section III-B discusses the activation-selection methodology, the training procedure, and the MINE-based explainability pipeline. Section IV presents quantitative results, MI analyses, and t-SNE visualisations. Section V discusses clinical implications and outlines deployment considerations. Section VI concludes and sketches future directions for extending GRN-enhanced Transformers to other vital-sign modalities and low-resource healthcare settings.

II. Gated Residual Network (GRN)

Training Transformer models with small datasets pose significant challenges. These models typically exhibit a generalization gap and tend toward sharp minima in such contexts [18]. Furthermore, their efficacy diminishes when dealing with imbalanced and small PPG signals [12].

To mitigate these issues, various strategies have been proposed. One approach involves altering the attention mechanism and employing data augmentation methods [19]. Alternatively, integrating Convolutional Neural Networks (CNNs) with the Transformer's attention mechanism has been explored [20]. However, these solutions are not without drawbacks:

1) **Computational Complexity [21]**: Transformers are inherently resource-intensive, with the self-attention mechanism's computational demands scaling quadratically with input length. Adding CNNs can further increase these demands, particularly for lengthy data sequences, potentially making it unfeasible in certain scenarios.
2) **Sequential Processing in CNNs [22]**: CNNs process data sequentially, focusing on small, localized regions. This approach hinders their ability to capture long-range dependencies effectively.

In addition to these methods, study [13] introduces the GRN as a key component of a Transformer-based classifier. Termed the GRN-Transformer, this integration handles small datasets and ambiguous input-target relationships. The GRN effectively handles uncertain input-target relationships, enabling nonlinear processing when necessary. A crucial feature of our GRN is the use of GLU [23], which dynamically emphasizes or suppresses information based on task requirements. Such gating techniques have been utilized in various models, including Gated Transformer Networks [24] and Temporal Fusion Transformers [25]. Its benefits are not limited to time-series data [24], [25] but extend to a wide range of data types [23], [26]. Incorporating GRN into the Transformer architecture marks a significant innovation of our work, greatly enhancing model performance and generalization across various domains.

According to [25], the GRN processes an input a, as illustrated in Fig. 1, with the following output:

$$\text{GRN}\omega(a) = \text{LayerNorm}(a + \text{GLU}\omega(\theta_1)), \tag{1}$$

$$\theta_1 = W_{1,\omega}\ \theta_2 + b_{1,\omega}, \tag{2}$$

$$\theta_2 = \text{ELU}(W_{2,\omega}\ a + b_{2,\omega}) \tag{3}$$

Here, θ_1 and θ_2 represent intermediate layers, LayerNorm denotes standard layer normalization, and ω indicates shared weights. The Exponential Linear Unit (ELU) activation function, defined as follows for $0 < \alpha$, is also employed:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0 \end{cases} \tag{4}$$

Additionally, the GRN uses a GLU in its gating layers for architectural flexibility. Given input η, the GLU operates as follows:

$$\text{GLU}_{\omega(\eta)} = \sigma(W_{3,\omega}\eta + b_{3,\omega}) \odot (W_{4,\omega}\eta + b_{4,\omega}), \tag{5}$$

where $W_{(.)}$ and $b_{(.)}$ represent the weights and biases, respectively, $\odot$ signifies the element-wise Hadamard product, and $\sigma(.)$ denotes the sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{6}$$

The GLU enables the GRN to regulate the extent of its contribution to the input a. It can effectively bypass the layer by setting the GLU outputs near zero, thus suppressing nonlinear contributions. During training, dropout is applied before the gating layer and layer normalization, specifically to θ_1 in Eq. (2). It has proven that this GRN enhances model robustness and helps prevent overfitting [26].

As established in prior research, the GLU has demonstrated a critical role in achieving lower perplexities in de-noising tasks. When integrated into Transformer architectures, it has consistently enhanced performance across multiple downstream language understanding applications. Nevertheless, the mechanisms by GLU contribute to these improvements remain insufficiently understood. As noted in [15], "*We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence.*" This study seeks to address this gap by investigating two central aspects. **Firstly**, we hypothesize that the activation function is pivotal. In neural networks, the selection of activation functions directly impacts the learning process's efficiency and effectiveness [27]. Our work will experimentally evaluate the influence of the GLU activation function on Transformer architecture performance and examine how it affects associated learning dynamics. **Secondly**, we propose leveraging MINE to quantify and assess information flow between the GRN and the Transformer layers. By applying MINE, we aim to determine how effectively the GRN captures and conveys information across layers, shedding light on the dependency structures that underpin the model's performance gains. We anticipate that these insights will elucidate the role of GRN in the network and inform optimization strategies for applications requiring sophisticated dependency, such as PPG artifact detection.

III. Materials and Methods

A. Clinical PPG Data at CHUSJ

The CHUSJ-PICU has established a HRDB that links biomedical signals from patient monitors to electronic patient records, enabling comprehensive physiological data capture through invasive and non-invasive monitoring techniques. This extensive dataset includes measurements from pulse oximetry, providing PPG signals and blood pressure readings obtained via various methodologies. This study, approved by the ethics board of CHUSJ, University of Montreal (approval number eNIMP:2023-4556), focuses on pediatric patients aged 0 to 18 admitted between September 2018 and September 2023. It incorporates electrocardiogram (ECG), PPG, and arterial blood pressure (ABP) waveforms, with data exclusions applied for recordings beyond the fourth day of hospitalization and patients undergoing Extracorporeal Membrane Oxygenation (ECMO) or experiencing multiple admissions. From this selection criteria, data from 1,573 patients were retained, comprising continuous 96-hour ECG, PPG, and blood pressure recordings. These were recorded at 5-second intervals, with a sampling rate of 128 Hz for PPG and 512 Hz for both blood pressure and ECG, and a focused 30-second window of PPG signals was explicitly extracted for analysis.

Data preprocessing in this study consists of four main stages: filtering, segmentation, resampling and normalization,

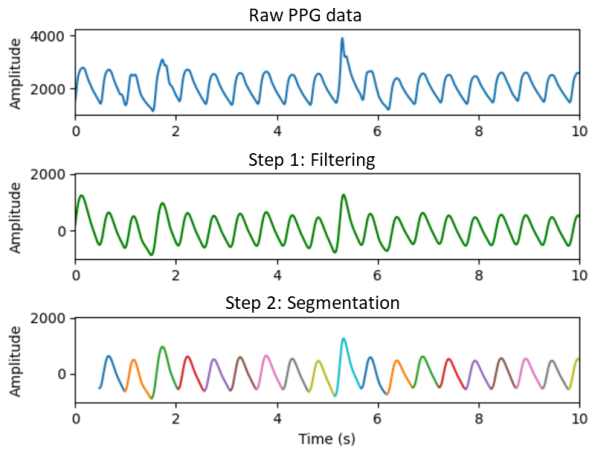

Fig. 2: An example for the first two steps of the preprocessing from a 10-second raw PPG signal (top), corresponding filtered signal (middle), and segmented signal (bottom) [12].

and feature extraction. Initially, a bandpass filter is applied to each signal using a Butterworth filter with cut-off frequencies set between 0.5 Hz and 5 Hz, implemented with a forward-backward approach to preserve signal integrity while reducing noise. The segmentation step follows, whereby the PPG signal is divided into individual pulse segments by identifying local minima, facilitating in-depth artifact analysis within each pulse. Each pulse segment is then uniformly oversampled to 256 samples, representing a 1-second heart cycle, with linear interpolation employed to maintain consistent sampling across pulses. Normalization is subsequently applied to achieve uniform feature scaling. Finally, temporal features are extracted at 4 millisecond intervals within each segment, resulting in 256 samples per pulse. This process effectively captures the temporal dynamics of the PPG signal, preparing it for subsequent analysis and classification using ML methods from traditional to advanced architectures.

The study's data annotation process begins with a healthcare professional manually annotating PPG signal pulses to establish a reliable ground truth, which is crucial for evaluating classification algorithms and classifier performance. To augment this, an automated algorithm, acting as a pseudo-expert, reannotates 10% of the data initially reviewed by the human expert, using statistical techniques to ensure the pulses fall within expected parameters [12]. This dual-annotation approach enhances the accuracy and reliability of the motion artifact annotations. Furthermore, to optimize the ML algorithms for automatic artifact classification. This approach, involving 1,571 signals with over 81,000 pulses and 256 features per pulse, helps identify the most efficient subset size for annotation. The statistical characteristics of this data, including key metrics such as signal distribution and feature variability, are summarized in Table II. For a comprehensive overview of the data preprocessing methods employed in this study, we direct the reader to the following studies [12]–[14].

TABLE II: Statistical summary of the dataset.

Statistic	Overall	Non-artifact	Artifact
Count	8190	6753	1437
Mean	13.53	14.98	6.70
Standard Deviation	329.36	285.95	439.81
Minimum	-1784.64	-1590.12	-1686.26
25th Percentile	-185.99	-165.09	-254.19
50th Percentile (Median)	-5.05	-1.61	-12.73
75th Percentile	203.48	180.45	279.35
Maximum	2016.81	1644.23	1981.25
Skewness	0.23	0.39	-0.00
Kurtosis	3.04	3.28	1.70

B. Gated Linear Unit with Different Activation Functions

As confirmed, GLU is vital in producing better perplexities for the de-noising objective and better results on downstream language-understanding tasks from Transformer. However, it is still unclear how GLU helps the Transformer in such an experiment [15]. One to experimentally explain this is the activation function. In neural networks, the choice of activation functions plays a pivotal role in determining the effectiveness and efficiency of learning algorithms [27]. Table III summarizes various activation functions, each with its unique equation and characteristics, essential for different neural network applications. The classic Sigmoid function, known for its smooth gradient, is represented alongside its variant, the Hard Sigmoid, which offers a computationally simpler alternative. The Soft-Sign Sigmoid provides a balanced approach, while the Snake Periodic Function introduces a periodic component to the activation. The Linearly Scaled Hyperbolic Tangent (LiSHT) enhances the traditional tanh function with linear scaling. Widely used in deep learning, the Rectified Linear Unit (ReLU) and its variations like the Exponential Linear Unit (ELU), Gaussian Error Linear Unit (GELU), and Scaled Exponential Linear Unit (SELU) offer different approaches to handling negative input values. The Sigmoid-Weighted Linear Units (Swish) and Self Regularized Non-Monotonic Neural (Mish) functions further extend the repertoire of activation functions, providing flexibility and adaptability in neural network design and performance optimization [28], [29].

Table IV compares various GLU models, each paired with a distinct activation function to create specialized GLU form functions. The BilinearGLU model employs a linear activation function, resulting in a straightforward GLU form. The standard GLU model uses the sigmoid function, while the hardGLU adapts the hard sigmoid (hard_σ) for its gating mechanism. The SoftsignGLU integrates the SoftSign activation, and the SnakeGLU incorporates the periodicity of the Snake function. LiGLU utilizes the Linearly Scaled Hyperbolic Tangent (LiSHT), adding a non-linear, scaled twist. The ReGLU, EGLU, GEGLU, and SeGLU models apply the ReLU, ELU, GELU, and SELU functions, each adding unique characteristics to the gating process. SwiGLU and MiGLU explore the dynamics of Sigmoid-Weighted Linear Units (Swishβ) and the Self Regularized Non-Monotonic Neural (Mish) functions, respectively. Each model's GLU form function follows a similar pattern, combining the chosen activation function with linear transformations to modulate the

TABLE III: List of activation functions [28], [29]

Name	Equation
Sigmoid	$\frac{1}{1+e^{-x}}$
Hard Sigmoid (hard_σ)	$\max\left(0, \min\left(1, \frac{(x+1)}{2}\right)\right)$
Soft-Sign Sigmoid (SoftSign)	$x/(1+\lvert x\rvert)$
Snake Periodic Function (Snake)	$x + \sin^2(ax)/a$
Linearly Scaled Hyperbolic Tangent (LiSHT)	$x.\tanh(x)$
Rectified Linear Unit (ReLU)	$\max(0, x)$
Exponential Linear Unit (ELU)	$\max(0, x) + \min(0, \alpha(e^x - 1))$
Gaussian Error Linear Unit (GELU)	$x\mathcal{P}(X \leq x), X \sim \mathcal{N}(0, 1)$
Scaled Exponential Linear Unit (SELU)	$\gamma(\max(0, x) + \min(0, \alpha(e^x - 1)))$
Sigmoid-Weighted Linear Units (Swishβ)	$\frac{x}{1+e^{-\beta x}}$
Self Regularized Non-Monotonic Neural (Mish)	$x\tanh(\log(1+e^x))$

TABLE IV: GLU functions with different activations

Models	**Activation functions**	**GLU form functions**
BilinearGLU	Linear	$(xW+b) \odot (xV+c)$
GLU	Sigmoid	$\sigma(xW+b) \odot (xV+c)$
hardGLU	hard_σ	hard_$\sigma(xW+b) \odot (xV+c)$
SoftsignGLU	SoftSign	Softsign$(xW+b) \odot (xV+c)$
SnakeGLU	Snake	Snake$(xW+b) \odot (xV+c)$
LiGLU	LiSHT	LiSHT$(xW+b) \odot (xV+c)$
ReGLU	ReLU	$\max(0, xW+b) \odot (xV+c)$
EGLU	ELU	ELU$(xW+b) \odot (xV+c)$
GEGLU	GELU	GELU$(xW+b) \odot (xV+c)$
SeGLU	SELU	SELU$(xW+b) \odot (xV+c)$
SwiGLU	Swishβ	Swish$\beta(xW+b) \odot (xV+c)$
MiGLU	Mish	Mish$(xW+b) \odot (xV+c)$

TABLE V: Gated Non-Linear Unit (GnLU) functions with different activations

Models	**Activation functions**	**GnLU functions**
GnLU	Sigmoid	$\sigma(xW+b) \odot \sigma(xV+c)$
LiGnLU	LiSHT	LiSHT$(xW+b) \odot$ LiSHT$(xV+c)$
MiGnLU	Mish	Mish$(xW+b) \odot$ Mish$(xV+c)$
SeGnLU	SELU	SELU$(xW+b) \odot$ SELU$(xV+c)$
SwiGnLU	Swishβ	Swish$\beta(xW+b) \odot$ Swish$\beta(xV+c)$

input signal effectively.

To further investigate non-linear transformations, Table V presents various Gated Non-Linear Unit (GnLU) models. Like the GLU forms in Table IV, each GnLU variant combines two identical activation functions to apply a non-linear transformation, effectively modulating the input signal. The standard GnLU model utilizes the sigmoid function, while LiGnLU, MiGnLU, SeGnLU, and SwiGnLU models apply LiSHT, Mish, SELU, and Swishβ functions, respectively. This approach aims to enhance model expressiveness by employing these activation functions in a symmetric gating mechanism.

C. Mutual Information Neural Estimation

Following our initial experiments with various activation functions, we employ MINE [16] as the next step to further investigate the GRN's role within the Transformer model. MINE provides a robust method for estimating MI between high-dimensional continuous variables using neural networks trained via gradient descent. Its scalability concerning both dimensionality and sample size, as well as its compatibility with back-propagation, makes MINE particularly well-suited for complex neural network applications. Several studies have applied MINE in neural networks to enhance interpretability and performance. For example, [30] utilized MINE to maximize MI for representation learning, improving feature robustness in neural network embeddings. Similarly, [31] applied MINE to analyze neural network layers, demonstrating its utility in evaluating MI for better network interpretability.

The motivation for using MINE lies in its capacity to quantify relationships within complex architectures; specifically, in explaining how the GRN functions within a neural network, MINE provides insights into how information flows and is retained or lost across layers. By measuring MI, MINE elucidates the GRN's effectiveness in capturing and transferring information, offering a deeper understanding of its contribution to the Transformer's overall performance. This analysis reveals the GRN's role in network performance, aiding our understanding of its integration into the Transformer.

MINE estimates MI between high-dimensional continuous random variables using a neural network trained with gradient ascent. The process begins with initializing the neural network's parameters, denoted as θ, which govern the function $T_\theta(x, z)$ that MINE uses to approximate MI. The estimation procedure iteratively computes a lower bound on the MI by leveraging samples from the joint distribution $\mathbb{P}_{XZ}$ and the marginal distribution $\mathbb{P}_Z$. In each iteration, a minibatch of b paired samples $(x^{(i)}, z^{(i)})$ is drawn from $\mathbb{P}_{XZ}$, representing the joint distribution of variables X and Z. Additionally, b independent samples $\tilde{z}^{(i)}$ are drawn from the marginal distribution $\mathbb{P}_Z$.

To estimate the MI, MINE computes the empirical lower

bound $V(\theta)$ as follows:

$$V(\theta) = \frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, z^{(i)}) - \log\left(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(x^{(i)}, \tilde{z}^{(i)})}\right) \quad (7)$$

This equation represents a lower bound on the MI between X and Z, where $T_\theta(x, z)$ is a neural network parameterized by θ that estimates dependencies between the variables. The first term, $\frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, z^{(i)})$, estimates the expectation over the joint distribution $\mathbb{P}_{XZ}$. In contrast, the second term approximates the expectation over the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$. The gradient $G(\theta)$ of the lower bound $V(\theta)$ is then calculated with respect to θ to obtain a bias-corrected estimate:

$$G(\theta) = \nabla_\theta V(\theta) \quad (8)$$

This gradient is used to update the network parameters through gradient ascent:

$$\theta \leftarrow \theta + \alpha G(\theta) \quad (9)$$

where α represents the learning rate. This iterative process continues until convergence, yielding a flexible estimator for MI; the pseudo-code is summarized in Algorithm 1.

To evaluate the impact of the GRN on MI within the Transformer architecture, we implement MINE following the critical steps summarized in Algorithm 2. This process measures the MI before and after the GRN is applied, offering insights into how effectively the GRN contributes to information retention and representation quality. By comparing MI values, we can assess the GRN's role in enhancing the sequential processing capabilities of the Transformer.

To compute MI using MINE, we model the dependencies between encoded representations and target labels in the GRN-Transformer framework. Let **before_grn** and **after_grn** represent encoded representations before and after the Gated Residual Network (GRN) is applied, with **y** denoting the target labels. The joint distribution $\mathbb{P}_{XY}$ represents pairs (X, Y) where X is the encoded data and Y is the target. To break dependencies, we also define the marginal distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$, which pairs each encoded representation X with a shuffled target label Y.

MINE estimates MI by leveraging the Donsker-Varadhan representation of the KL-divergence, providing a lower bound on the mutual information $I(X;Y)$ between X and Y:

$$I(X;Y) \geq \mathbb{E}_{\mathbb{P}_{XY}}[T_\theta(X,Y)] - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}\left[e^{T_\theta(X,Y)}\right] \quad (10)$$

where $T_\theta(X, Y)$ is a neural network parameterized by θ that learns to distinguish between samples from the joint distribution $\mathbb{P}_{XY}$ and the marginal distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$. For each minibatch, MINE estimates the joint and marginal predictions. The joint prediction, **joint_pred**, is computed by applying T_θ to pairs (X, Y) sampled from $\mathbb{P}_{XY}$, while the marginal prediction, **marginal_pred**, is obtained by pairing X with shuffled Y to simulate the marginal distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$.

The empirical mutual information $\hat{I}(X;Y)$ for each minibatch is computed as:

Algorithm 1 Mutual Information Neural Estimation (MINE)

1: $\theta \leftarrow$ Initialize network parameters
2: **repeat**
3: Sample a minibatch of b pairs $(x^{(i)}, z^{(i)})$ from the joint distribution $\mathbb{P}_{XZ}$
4: Independently sample b instances $\tilde{z}^{(i)}$ from the marginal distribution $\mathbb{P}_Z$
5: Compute the empirical lower bound $V(\theta)$:

$$V(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, z^{(i)}) - \log\left(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(x^{(i)}, \tilde{z}^{(i)})}\right)$$

6: Calculate bias-corrected gradient estimates:

$$G(\theta) \leftarrow \nabla_\theta V(\theta)$$

7: Update the network parameters using gradient ascent:

$$\theta \leftarrow \theta + \alpha G(\theta)$$

8: **until** convergence is reached

$$\hat{I}(X;Y) = \frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, y^{(i)}) - \log\left(\frac{1}{b}\sum_{i=1}^{b} \exp(T_\theta(x^{(i)}, \tilde{y}^{(i)}))\right) \quad (11)$$

where b is the minibatch size, $x^{(i)}$ and $y^{(i)}$ are samples from the joint distribution, and $\tilde{y}^{(i)}$ is a shuffled sample used to approximate the marginal distribution. The first term, $\frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, y^{(i)})$, estimates the expectation over the joint distribution $\mathbb{P}_{XY}$, while the second term, $\log\left(\frac{1}{b}\sum_{i=1}^{b} \exp(T_\theta(x^{(i)}, \tilde{y}^{(i)}))\right)$, approximates the expectation over the marginal distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$.

To assess the GRN's impact on MI, we compute $\hat{I}_{\text{before}} = \hat{I}(X;Y)$ using **before_grn** and **y**, and $\hat{I}_{\text{after}} = \hat{I}(X;Y)$ using **after_grn** and **y**. By comparing $\hat{I}_{\text{before}}$ and $\hat{I}_{\text{after}}$, we can quantify the GRN's influence on information retention and its contribution to representation quality in the Transformer.

D. Integrating GRN to Attention Mechanism

Inspired by modifications to the attention mechanism proposed in [17], namely GAU, we investigate integrating the GRN directly into the attention mechanism rather than placing it solely in intermediate layers. This adaptation aims to enhance the model's ability to capture complex temporal dependencies, potentially improving performance on tasks that rely on time-dependent patterns. Figure 3 illustrates the progression of the attention mechanism in our Transformer model, evolving from the original multi-head attention to a GRN-integrated variant. In the standard setup, multi-head attention computes queries, keys, and values through linear transformations, applying softmax to their dot products for attention weighting, followed by concatenation and a dense layer to generate outputs. Our modified approach introduces a GRN between the softmax and concatenation steps, enabling the model to capture sequential dependencies within the data better. The final Transformer model with multi-head GRN-Attention retains the core structure but replaces the conventional attention with the GRN-enhanced variant.

TABLE VI: Comparison of state-of-the-art and the proposed framework, detailing network architecture (Attention, GAU, and State Space Model (SSM)), activation functions, explainability methods, application domain, and key implementation insights.

Study	Architecture	Activation	Explainability	Applications	Remarks
STFT [32]	Attention	GELU	n/a	Large-scale traffic forecasting	Using GLU as an intermediate representation layer.
LMHaze [33]	SSM	Swishβ	n/a	Imaging dehazing	Swishβ is used in each SSM block sequentially.
Agda [34]	Attention	Swishβ	n/a	Proof formalization	SwiGLU + residual connections + prenormalization with RMSNorm.
Read-ME [35]	MoE-Attention	Swishβ	n/a	Resource-constrained LLM inference	Inserting a MoE router into the Transformer.
GLUSE [36]	Attention	Sigmoid	n/a	Onboard satellite EO image classification	GLU as a residual block plus SE for adaptive attention.
Adapter [37]	Attention	Sigmoid	n/a	Clinical notes classification	Low-Rank Adaptation layers run alongside Transformer layers.
This study	**Attention, GAU**	**11 functions**	**MINE,-t-SNE**	**PICU PPG artifact detection**	**Experiments on activations/attention with explainability.**

Algorithm 2 MINE for a GRN-Transformer

Require: Encoded representations **before_grn** and **after_grn**, Target labels **y**

procedure MINE
  function INITIALIZEMINEMODEL(input_shape)
    Define a neural network with input shape *input_shape* and output of size 1, representing $T_\theta(X, Y)$
    return Initialized neural network
  end function
  function COMPUTEMI(data, y, model)
    shuffled_y $\leftarrow$ **shuffle(y)** ▷ Shuffle y to create samples from the marginal distribution
    joint_pred $\leftarrow$ **model**$([data, y])$ ▷ Compute predictions for the joint distribution, $T_\theta(X, Y)$
    marginal_pred $\leftarrow$ **model**$([data,$ **shuffled_y**$])$ ▷ Compute predictions for the marginal distribution, $T_\theta(X, \tilde{Y})$
    Compute empirical mutual information estimate MI as:

$$MI \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, y^{(i)}) - \log\left(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(x^{(i)}, \tilde{y}^{(i)})}\right)$$

    where b is the minibatch size, $(x^{(i)}, y^{(i)})$ are joint samples, and $(x^{(i)}, \tilde{y}^{(i)})$ are marginal samples.
    return MI
  end function
  $model\_before\_grn \leftarrow$ INITIALIZEMINEMODEL$(shape_of($**before_grn**$[1]))$
  $model\_after\_grn \leftarrow$ INITIALIZEMINEMODEL$(shape_of($**after_grn**$[1]))$
  for $epoch = 1$ to num_epochs **do**
    Train $model\_before\_grn$ on **before_grn** and **y** ▷ Optimize T_θ to learn dependencies for the representation **before_grn**
    Train $model\_after\_grn$ on **after_grn** and **y** ▷ Optimize T_θ to learn dependencies for the representation **after_grn**
  end for
  $MI\_before \leftarrow$ COMPUTEMI(**before_grn**, **y**, $model\_before\_grn$)
  $MI\_after \leftarrow$ COMPUTEMI(**after_grn**, **y**, $model\_after\_grn$)
  Print("MI before GRN: ", MI_before)
  Print("MI after GRN: ", MI_after)
end procedure

Table VI presents a comprehensive comparison of our proposed framework against state-of-the-art studies employing the GRN with Transformer-based approaches, highlighting differences in network structure, activation choices, explainability tools, and target applications. Our work is the first to address physiological-signal PPG artifact detection in the PICU, a notoriously low-data, high-noise setting. It performs a systematic search over 11 activation functions - Linear, Sigmoid, hard_σ, SoftSign, Snake, LiSHT, ReLU, ELU, GELU, Swishβ, and Mish—and proves their impact across two distinct network structures, standard Attention and the lightweight GAU. Explainability is built in from the outset by coupling MINE to quantify feature relevance with t-SNE plots that reveal clear class separation. In short, this study does not merely port an existing technique to a new dataset; it introduces architectural, activation-level, and interpretability innovations that unlock Transformer-style models for a previously out-of-reach healthcare use case.

IV. Experimental Results

All experiments were conducted on the PICU e-Medical infrastructure, the Miircic Server at CHUSJ. A GPU Quadro RTX 6000 with 24 Gb of memory provided computational capacity for these experiments. The experiments were conducted using the *scikit-learn* library [38] and *Keras* [39]. Data was split into 70% for training and 30% for testing. Given the complexity of training Transformer models, we focused on four essential hyperparameters—model size, learning rate, batch size, and maximum sequence length—which have been shown to impact the training dynamics of Transformers critically [40]. Additionally, dropout with a probability of 0.25 [41], the GlorotNormal initializer [42], and batch normalization [43],

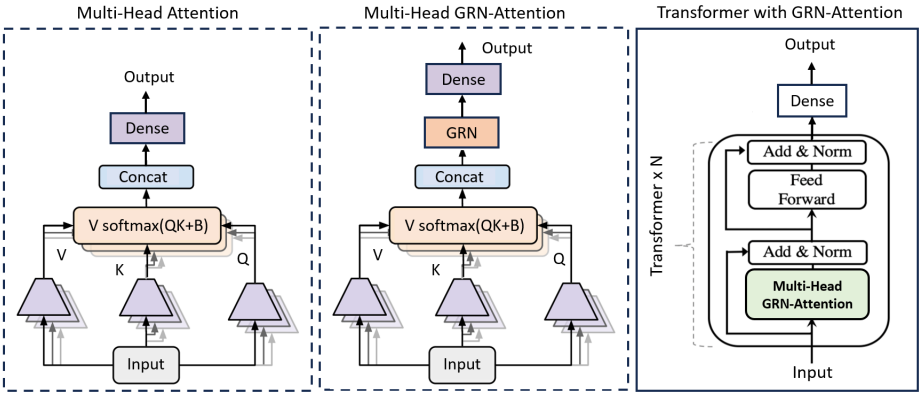

Fig. 3: An evolution of the attention mechanism in a Transformer from standard multi-head attention (left) to a GRN-enhanced variant (middle), and the Transformer with GRN-Attention (right).

TABLE VII: Hyperparameters of classifiers

Hyperparameters	Transformer
Hidden layers	4
Number of neurons	128
Number of multi-heads attention	4
Batch size	96
Dropout	0.25
Learning rate	6e-04
Optimizer	Adam

[44] were incorporated to enhance model stability. To address class imbalance, we employed the oversampling technique ADASYN [45]. These hyperparameters were fine-tuned to optimize performance while minimizing the risk of overfitting.

To effectively assess the performance of our method, metrics including accuracy, precision, recall (or sensitivity), and F1 score. These metrics are defined as follows:

$$\text{Accuracy (acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{Precision (pre)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall/Sensitivity (rec)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{F1-Score (f1)} = \frac{2^{\star}\text{Precision}^{\star}\text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

where TN and TP stand for true negative and true positive, respectively, and they are the number of negative and positive patients that are classified correctly. Whereas FP and FN represent false positive and false negative, respectively, and they represent the number of positive and negative patients that were wrongly predicted.

Table VIII comprehensively compares various Gated Linear Unit (GLU) model variants across four key performance metrics. Among these, the GLU and MiGLU models emerge as the top performers. Both models achieve the highest accuracy, scoring 0.98, indicating their superior overall prediction correctness. The MiGLU model leads in precision, boasting a score of 0.91, which suggests it is particularly effective in making accurate positive predictions with the least false positives. On the other hand, the GLU model excels in recall with a top score of 0.97, highlighting its capability to identify the majority of true positive cases correctly. Furthermore, both

TABLE VIII: GLU Models' Performance

Models	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)
BilinearGLU	0.97	0.90	0.94	0.92
GLU	**0.98**	0.90	**0.97**	**0.93**
hardGLU	0.97	0.89	0.93	0.91
SoftSignGLU	0.97	0.91	0.93	0.92
SnakeGLU	0.97	0.91	0.92	0.92
LiGLU	0.97	0.91	0.94	0.92
ReGLU	0.97	0.89	0.94	0.92
EGLU	0.97	0.93	0.92	0.92
GEGLU	0.97	0.88	0.94	0.91
SeGLU	0.97	0.90	0.93	0.92
SwiGLU	0.97	0.91	0.93	0.92
MiGLU	**0.98**	**0.91**	0.95	**0.93**

Bold denotes the best values.

these models share the highest F1 score of 0.93, illustrating their optimal balance between precision and recall. The GLU and MiGLU models demonstrate the best overall performance among the variants, making them potentially more effective for tasks requiring high accuracy, precise predictions, and reliable identification of true positives.

Additionally, the proposed GRN enhances the information-theoretic relevance of Transformer features, yielding a 2.6x increase in mutual information with class labels (Fig. 6). This quantitative gain is reinforced by t-SNE visualizations (Fig. 7), where post-GRN embeddings exhibit markedly tighter intra-class cohesion and more apparent inter-class separation, evidencing improved feature discriminability. Such representational sharpening directly translates to more robust artifact detection in PPG signals, mitigating the risk of false alarms or missed detections in clinical monitoring pipelines. Consequently, the GRN not only improves statistical performance metrics but also strengthens the reliability of downstream vital-sign estimation, a factor with direct implications for patient safety and clinical decision-making.

Table IX compares various GnLU model variants, evaluated on key performance metrics. Among these, the standard GnLU model distinctly outperforms its counterparts. It achieves the highest accuracy at 0.98, indicating superior overall prediction correctness. It leads in precision with a score of 0.91, highlighting its effectiveness in making accurate positive predictions with minimal false positives. Additionally,

TABLE IX: GnLU Models' Performance

Models	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)
GnLU	**0.98**	**0.91**	**0.96**	**0.94**
LiGnLU	0.97	0.91	0.92	0.91
MiGnLU	0.97	0.90	0.94	0.92
SeGnLU	0.97	0.91	0.93	0.92
SwiGnLU	0.97	0.90	0.92	0.91

Bold denotes the best values.

the GnLU model excels in recall with the highest score of 0.96, demonstrating its ability to identify the vast majority of true positive cases correctly. Furthermore, it achieves the top F1 score of 0.94, indicating an optimal balance between precision and recall. This comprehensive performance makes the GnLU model highly effective and reliable in diverse scenarios, particularly when recall and accurate identification of true positives are crucial.

The effectiveness of the sigmoid activation in the GnLU gating mechanism can be derived directly from Bayesian decision theory. Let V denote the feature evidence generated by the network, which encodes the likelihood that a given input belongs to the positive class ($Y = 1$) rather than the negative class ($Y = 0$). The gating function $g(V)$ aims to modulate the contribution of the feature signal U according to the posterior probability $p(Y = 1 \mid V)$. From Bayes' theorem, this posterior is expressed as:

$$p(Y = 1 \mid V) = \frac{p(V \mid Y = 1)p(Y = 1)}{p(V)}. \tag{16}$$

In many practical settings, including neural network representations, the log-odds of the posterior are approximately affine in V:

$$\log \frac{p(Y = 1 \mid V)}{p(Y = 0 \mid V)} \approx aV + b, \tag{17}$$

where a and b are learnable parameters. Solving for the posterior yields:

$$p(Y = 1 \mid V) = \sigma(aV + b) = \frac{1}{1 + e^{-(aV+b)}}, \tag{18}$$

demonstrating that the sigmoid function arises naturally as the Bayes-optimal mapping from evidence to posterior probability. This property ensures that the gating operation minimizes the expected error under standard log-likelihood-based loss functions and effectively suppresses noise while preserving task-relevant features. Furthermore, introducing a temperature parameter τ in the gating function $\sigma(V/\tau)$ allows explicit control of the gate's sharpness, enabling empirical validation of the Bayes-consistent operating point.

To empirically assess this theoretical prediction, we designed two complementary experiments. In the first approach, the temperature parameter τ was treated as a learnable variable, jointly optimized with the model parameters, allowing the network to discover its own posterior calibration scale. If the theoretical framework is valid, the learned τ should converge near 1.0, corresponding to the Bayes-consistent regime where the gating function best approximates $p(Y = 1 \mid V)$. In the second experiment, we performed a controlled ablation by fixing τ to predefined values ($\tau \in \{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$) and retraining the model under identical conditions. This systematic sweep was designed to quantify the impact of gating sharpness on predictive performance across standard evaluation metrics and to test the hypothesis that the optimum occurs near $\tau = 1$.

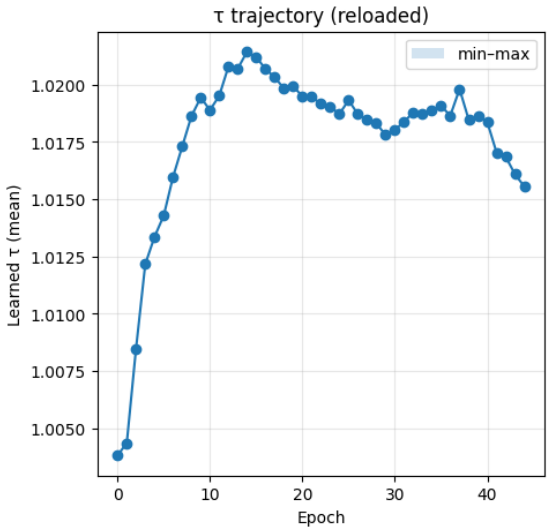

Fig. 4: Evolution of the learnable temperature parameter τ during training. The model consistently converges near $\tau = 1.0175 - 1.02$, aligning with the Bayes-consistent regime predicted by theory.

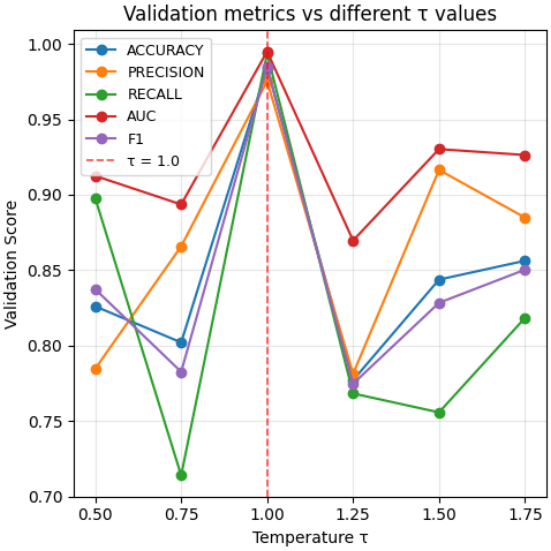

Fig. 5: Model performance across fixed temperature values $\tau \in \{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$.

The results provide strong empirical support for the theoretical analysis. When τ was learned end-to-end, it consistently converged to approximately 1.018 (Fig. 4), closely matching the Bayes-optimal prediction. Fixed-τ experiments further confirmed this behavior: all evaluation metrics peaked at $\tau = 1.0$ (Fig. 5), with performance degrading for both lower and higher values due to overly sharp gating ($\tau < 1$) or insufficient noise suppression ($\tau > 1$). This alignment between theory and observation demonstrates that the sigmoid

gate with $\tau \approx 1$ is not merely empirically effective but also theoretically grounded as a Bayes-optimal mechanism. Moreover, this finding is consistent with recent advances in neural network theory, which show that models converge toward Bayes-optimal inference under suitable conditions [46]. Thus, the GnLU layer can be interpreted as a probabilistic filter approximating $p(Y = 1 \mid V)$, providing a principled approach to balancing noise attenuation and information retention.

The provided Fig. 8 depicts the performance of a GRN-Transformer with a GnLU across various training epochs, showing metrics for both training and validation data. The loss graph shows a typical sharp decline in loss for both training and validation at the start of training, which stabilizes as epochs increase, indicating that the model is learning and generalizing well without signs of overfitting. The AUC graph reveals high and stable values for both sets, suggesting excellent class separation capability. Precision and recall start low but rise quickly to plateau at high values, demonstrating that the model accurately identifies true positives and covers a high proportion of actual positive samples. The proximity of training and validation curves across all metrics indicates that the model generalizes well to new data. There's some fluctuation in validation loss, but it's within normal bounds, suggesting some variability in the validation set or learning process. These results point to a well-performing model with strong classification abilities, pending further evaluation in the context of specific project goals and benchmarks.

Fig. 9 shows the comparison of Transformer-based models trained under different paradigms: supervised [12], unsupervised (AE) [14], self-supervised [14], with GRN [13] . Experimental results confirm that both GRN variants, GLU [13] and GnLN (this study), form the outer-most polygon, showing the highest recall (0.97 – 0.98) while maintaining near-perfect accuracy (0.98) and strong precision/F1 (0.94). Standard supervised, unsupervised, and self-supervised baselines trail behind, especially on recall and F1. Consequently, GRN-Transformer models combine top-tier accuracy with the best recall, delivering the most balanced F1. Their high recall means far fewer false negatives, an essential property for clinical decision support, where missing a positive case can be costly. This superiority is further reinforced by the confusion matrix in Fig. 10, where GRN-Transformers variants display the lowest false-negative and false-positive cases.

The Fig. 11 displays the performance of a Transformer model with GRN-Attention over 140 epochs, showing key metrics for training and validation datasets. The loss graph indicates an initial decrease in training and validation losses, with subsequent fluctuations in validation loss hinting at potential overfitting. The AUC scores begin high and plateau, but a widening gap between training and validation suggests the model may be learning training-specific patterns that don't generalize well. Precision and recall metrics are initially volatile but stabilize, though they exhibit considerable noise, especially in the validation set. This noise implies inconsistency in the model's predictive accuracy and sensitivity to positive samples. While the model shows initial solid learning, the observed metrics suggest it may be overfitting to the training data. Regularization, hyperparameter tuning, or early stopping may be necessary to improve the model's generalization to new data and ensure stable performance across all metrics.

V. Clinical and Translational Impact

Embedding the proposed GRN-Transformer artifact-detection module into the CHUSJ's CDSS stack will supply bedside teams with cleaner SpO_2 and PPG information, reducing false alarms, minimising manual waveform checks, and accelerating recognition of incipient hypoxaemia or ARDS. When fused with algorithms already validated at CHUSJ - including occlusion-segmentation of monitor traces [47], heart-rate/temperature coupling analysis [48], absence-of-heart-failure extraction from clinical notes [5], [18], [49]–[51], hypoxaemia prediction [6], and automated chest x-ray [52], [53], the system can deliver an integrated, continuously updated cardiorespiratory score that helps clinicians adjust ventilatory settings or order confirmatory tests earlier and with greater confidence.

Our proposed GRN-Transformer model contains only 111,561 parameters with an inference memory footprint of 0.43 MB and an average runtime of 1.012 s per sample on an NVIDIA Quadro RTX 6000 GPU. These results demonstrate that the gated residual mechanism adds minimal computational overhead, offering a lightweight and resource-efficient solution suitable for real-time clinical deployment. Deploying the composite CDSS in a clinical setting entails several operational challenges. First, the system must process 128 Hz waveform streams in real time so that alerts are not delayed. Second, it must interface reliably with diverse bedside monitors and the existing HRDB. Third, robust model governance, encompassing periodic retraining, performance audits, and version control, is essential to prevent drift as hardware or patient populations evolve. Fourth, all data flows must comply with institutional and provincial regulations on cybersecurity and patient privacy. Finally, the user interface should minimise alarm fatigue and integrate smoothly with established PICU workflows as proven [48], [54]. This evaluation will guide further system refinements and improvements to quantify latency, false-alarm reduction, and clinician acceptance before the system becomes a routine part of PICU care.

VI. Conclusion

This study explores the Transformer model's performance in various learning environments, focusing on GRN-Transformer. Our research analyzes different activation functions for the GLU, a crucial component of the GRN structure. The study employs MINE to verify the effectiveness of GRN. Additionally, it investigates the positional impact of GRN within the Transformer architecture, particularly examining its role as an intermediate layer within the Attention mechanism instead of an external intermediary layer. Results show that the GnLU with Sigmoid gating consistently yields the best accuracy, precision, recall, and AUC; MINE confirms the GRN's insertion increases feature-label MI; and positioning the GRN as an intermediate representation filter layer - rather

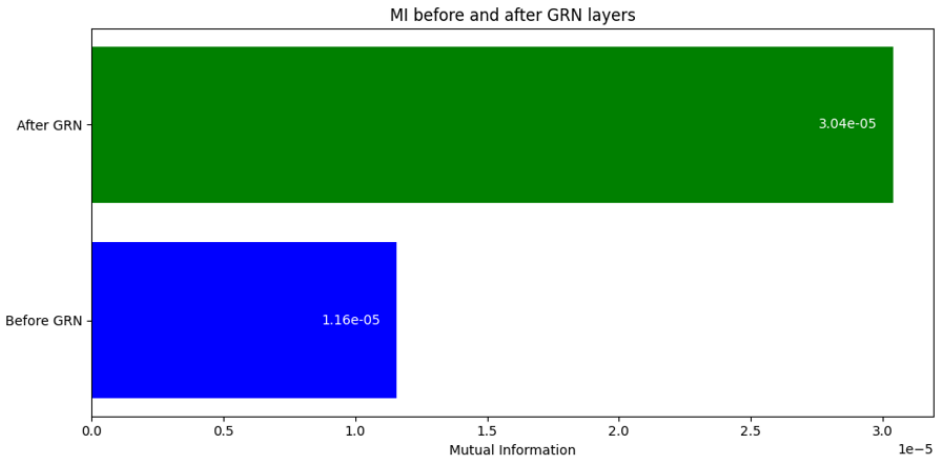

Fig. 6: MI Estimation between latent features and class labels before and after the GRN filter block (see Fig. 1 for its position in the pipeline). The GRN raises MI, showing that its sigmoid-gated transformation suppresses noise and concentrates task-relevant content, consistent with the subsequent gains in AUC, precision, and recall.

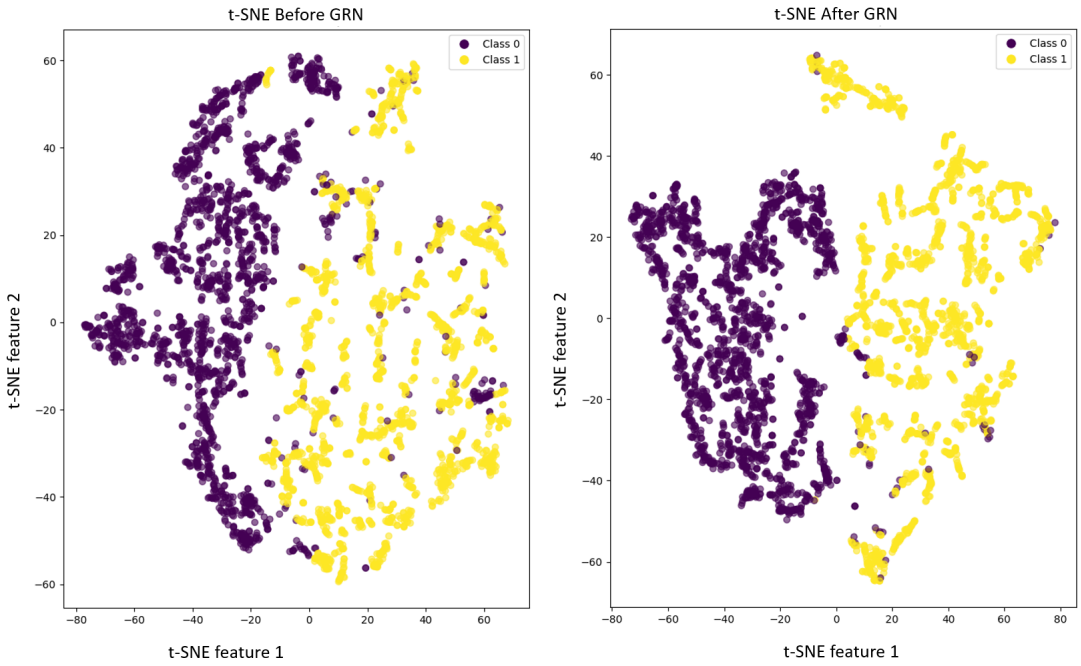

Fig. 7: t-SNE visualisation of latent features before (left) and after (right) the Gated Residual Network (GRN) layer. After GRN insertion, the purple (Class 0) and yellow (Class 1) clusters tighten and move farther apart, visibly fewer points straying across the class boundary. This cleaner separation confirms that the GRN transformation filters out noise and concentrates discriminative patterns, mirroring the MI gain in Fig. 6.

than embedding it inside the Attention block - more effectively suppresses noise and amplifies clinically relevant patterns.

This work explicates how a GRN acts as an intermediate-representation filter suppressing noise, amplifying salient features, and measurably increasing feature-label MI to boost Transformer performance when labelled data are scarce. By hybridizing this architectural unit with a systematic activation analysis and an MI-based interpretability pipeline, we deliver methodological advances that extend beyond the PICU use-case: the GRN block is model-agnostic, computationally lightweight, and can be dropped into any encoder–decoder stack or temporal-sequence model where high noise and low annotation density prevail (e.g. wearable sensing, industrial IoT, or low-resource speech). These findings, therefore, contribute to core neural-network design and analysis while simultaneously offering a practical pathway for deploying models in challenging clinical and other real-world environments.

Our study has two key limitations related to clinical and technical generalization. First, the model was evaluated using data from a single institution (CHUSJ), which may not fully capture the variability in patient populations, monitoring devices, and workflows seen across different clinical settings. Multi-site validation will therefore be essential to confirm robustness and external applicability. Second, the GRN-

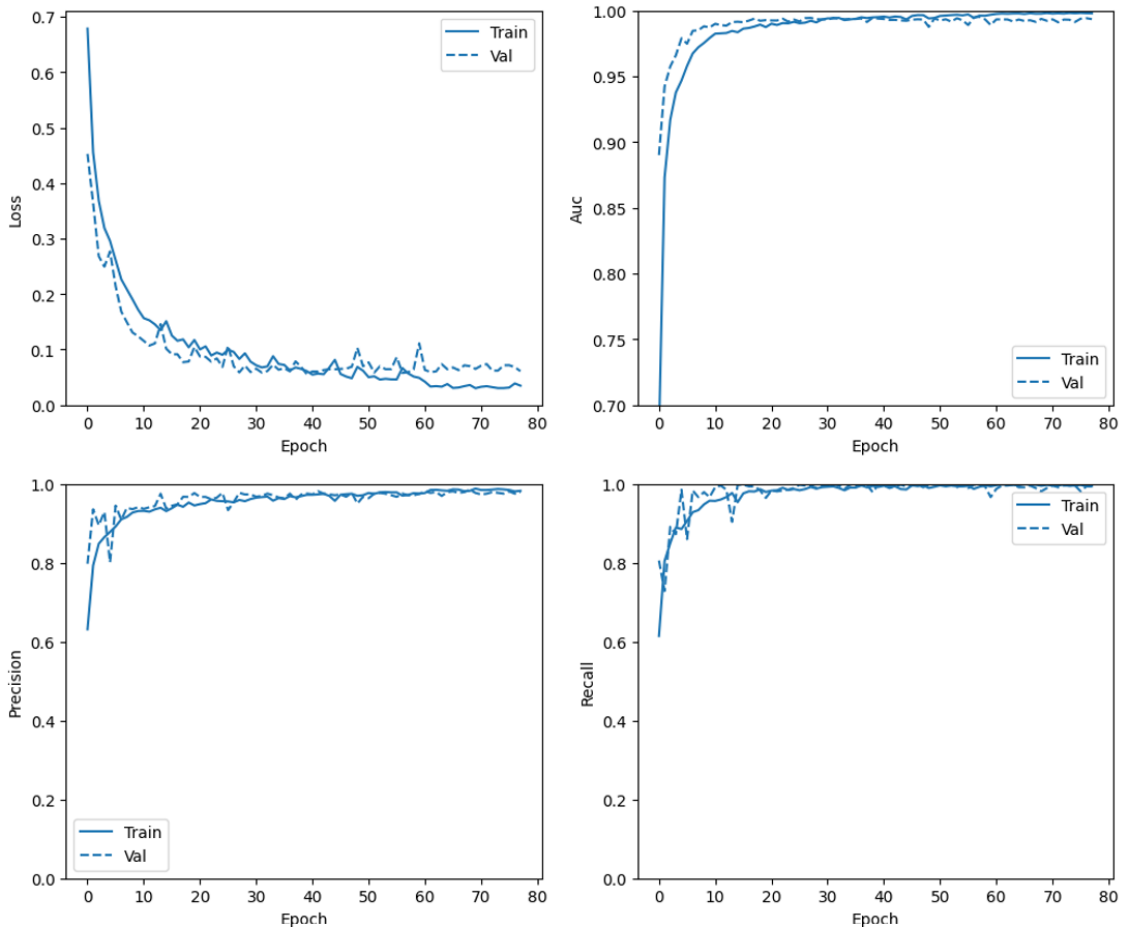

Fig. 8: Learning curve during training and validation for the GRN-Transformer that uses a non-linear GLU unit $\sigma(xW+b)\odot\sigma(xV+c)$. The model converges rapidly: validation AUC surpasses 0.98 within ten epochs, while precision and recall stabilise above 0.94 with no divergence between training and validation curves, indicating strong generalisation and performance.

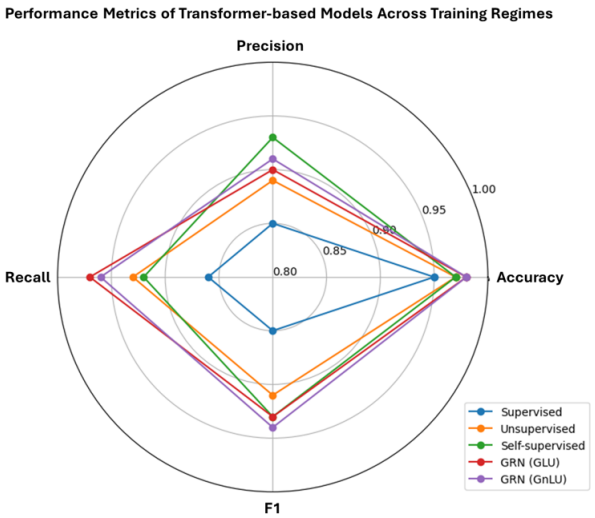

Fig. 9: Performance results from Transformer-based models with different learning paradigms, supervised [12], unsupervised [14], self-supervised [14], and GRN variants i) with linear GLU unit [13] ($(xW+b)\odot(xV+c)$), and ii) with non-linear GLU unit ($\sigma(xW+b)\odot\sigma(xV+c)$).

Attention integration showed reduced performance compared to the GnLU-based gating, likely due to over-suppression of informative features and instability in attention weights when interacting with high-dimensional contextual signals. Future work will involve a more in-depth analysis of these failure cases and the development of hybrid gating-attention mechanisms to improve both interpretability and generalization.

ACKNOWLEDGMENT

The Research Center at CHU Sainte-Justine Hospital, University of Montreal, provided the clinical PPG data. The authors thank Clara Macabiau and Dr. Kevin Albert for their support in data preprocessing and annotating for this research. Data and reproducible codes are available upon reasonable request to Prof. Philippe Jouvet, M.D., PhD. (Email: philippe.jouvet.med@ssss.gouv.qc.ca).

REFERENCES

[1] D. Brossier, *et al.*, "Creating a high-frequency electronic database in the picu: the perpetual patient," *Pediatr. Crit. Care Med.*, vol. 19, no. 4, pp. e189–e198, 2018.

[2] N. Roumeliotis, *et al.*, "Reorganizing care with the implementation of electronic medical records: a time-motion study in the picu," *Pediatr. Crit. Care Med.*, vol. 19, no. 4, pp. e172–e179, 2018.

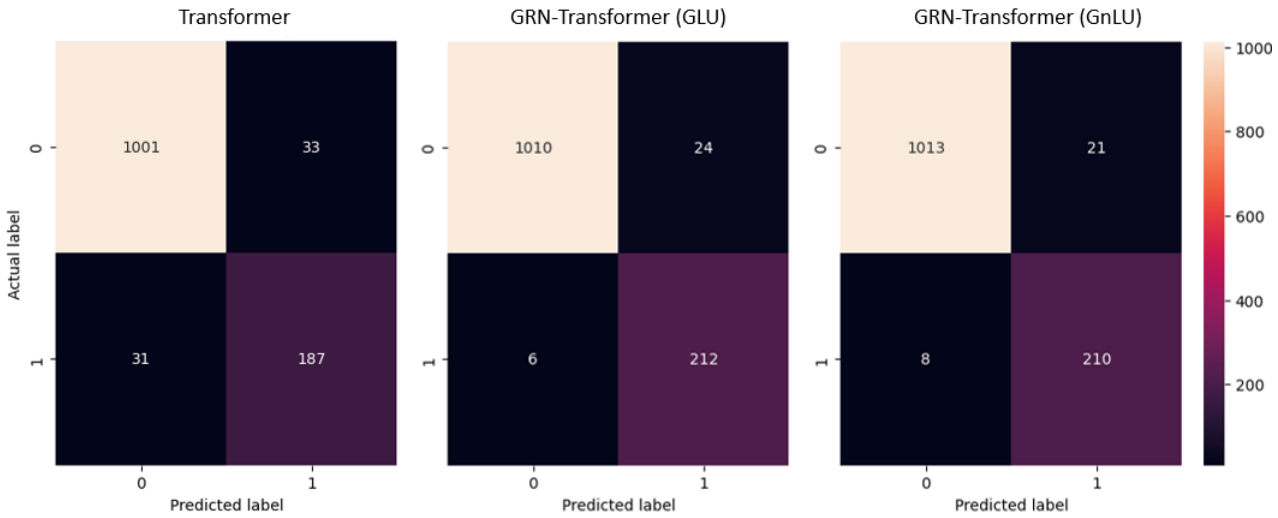

Fig. 10: Confusion matrix of the classification task with: *Left*: original Transformer, *Middle*: GRN-Transformer with linear GLU unit ($(xW + b) \odot (xV + c)$), and *Right*: GRN-Transformer with non-linear GLU unit ($\sigma(xW + b) \odot \sigma(xV + c)$).

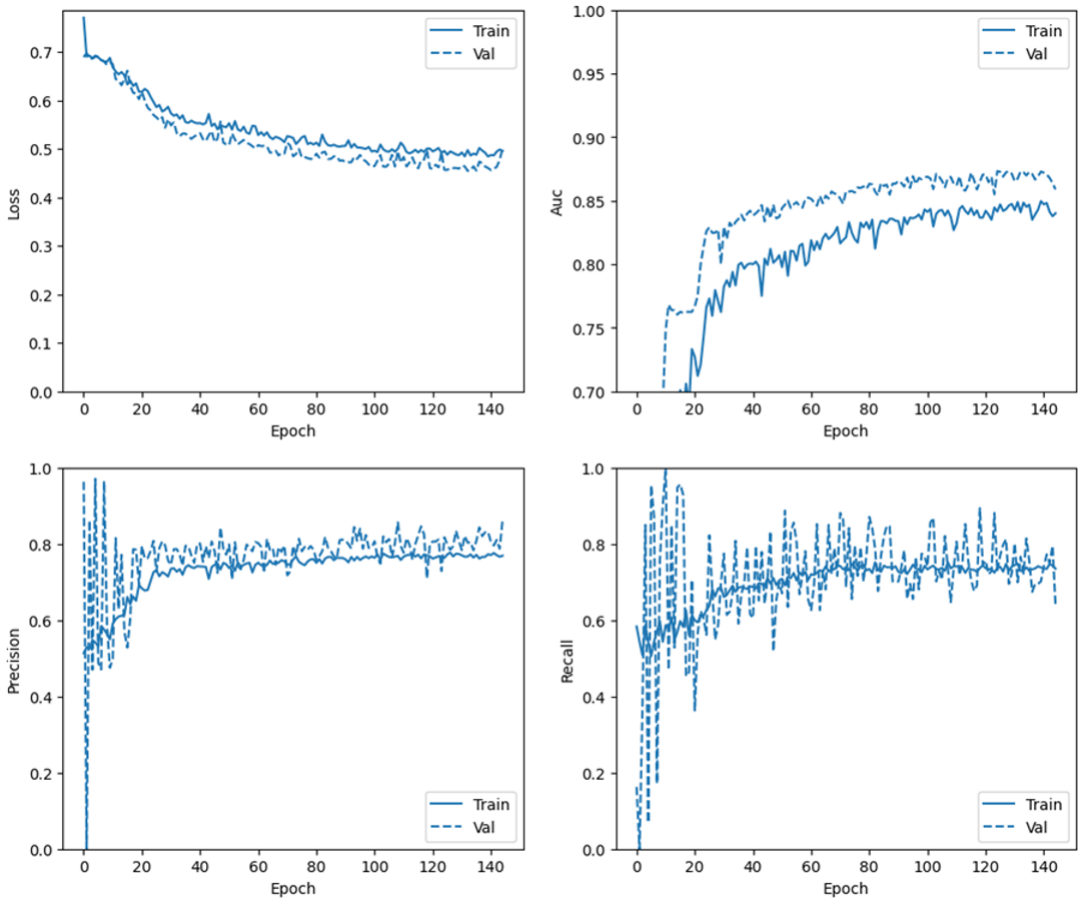

Fig. 11: Transformer's learning curve performance with GRN-Attention structure. The higher loss, lower and fluctuating AUC, recall, and precision compared with the conventional Attention (Fig. 8) indicate that the hybrid GRN-Attention cannot exploit the scarce PICU data as effectively as the conventional Attention design.

[3] A. Mathieu and et. al., "Validation process of a high-resolution database in a pediatric intensive care unit—describing the perpetual patient's validation," *J. Eval. Clin. Pract.*, vol. 27, no. 2, pp. 316–324, 2021.

[4] A. C. Dziorny and et. al., "Clinical decision support in the picu: Implications for design and evaluation," *Pediatr. Crit. Care Med.*, vol. 23, no. 8, pp. e392–e396, 2022.

[5] T.-D. Le and et. al., "Detecting of a patient's condition from clinical narratives using natural language representation," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 142–149, 2022.

[6] M. Sauthier, *et al.*, "Estimated pao2: A continuous and noninvasive method to estimate pao2 and oxygenation index," *Critical care explorations*, vol. 3, no. 10, p. e0546, 2021.

[7] G. Emeriaud, *et al.*, "Executive summary of the second international guidelines for the diagnosis and management of pediatric acute respiratory distress syndrome (palicc-2)," *Pediatr. Crit. Care Med.*, vol. 24, no. 2, p. 143, 2023.

[8] P. Jouvet and et. al., "A pilot prospective study on closed loop controlled ventilation and oxygenation in ventilated children during the weaning phase," *Critical Care*, vol. 16, no. 3, pp. 1–9, 2012.

[9] M. Wysocki, P. Jouvet, and S. Jaber, "Closed loop mechanical ventilation," *J. Clin. Monit. Comput.*, vol. 28, pp. 49–56, 2014.

[10] B. L. Hill and et. al., "Imputation of the continuous arterial line blood pressure waveform from non-invasive measurements using deep learning," *Scientific reports*, vol. 11, no. 1, p. 15755, 2021.

[11] F. Fan and et. al., "Estimating spo 2 via time-efficient high-resolution harmonics analysis and maximum likelihood tracking," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1075–1086, 2017.

[12] C. Macabiau, *et al.*, "Label propagation techniques for artifact detection in imbalanced classes using photoplethysmogram signals," *IEEE Access*, vol. 12, pp. 81 221–81 235, 2024.

[13] T.-D. Le, *et al.*, "GRN-Transformer: Enhancing motion artifact detection in PICU photoplethysmogram signals," *arXiv preprint arXiv:2308.03722*, 2023.

[14] T. D. Le, *et al.*, "A novel transformer-based self-supervised learning method to enhance photoplethysmogram signal artifact detection," *IEEE Access*, 2024, DOI: 10.1109/ACCESS.2024.3488595.

[15] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[16] M. I. Belghazi, *et al.*, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.

[17] W. Hua, *et al.*, "Transformer quality in linear time," in *International conference on machine learning*. PMLR, 2022, pp. 9099–9117.

[18] T.-D. Le, P. Jouvet, and R. Noumeir., "Improving transformer performance for french clinical notes classification using mixture of experts on a limited dataset," *arXiv preprint arXiv:2303.12892*, 2023.

[19] S. H. Lee and et. al., "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[20] R. Shao and X.-J. Bi, "Transformers meet small datasets," *IEEE Access*, vol. 10, pp. 118 454–118 464, 2022.

[21] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 156–171, 2020.

[22] T. Sattler and et. al., "Understanding the limitations of cnn-based absolute camera pose regression," in *IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3302–3312.

[23] Y. N. Dauphin and et. al., "Language modeling with gated convolutional networks," in *International Conference on ML*, 2017, pp. 933–941.

[24] M. Liu and et. al., "Gated transformer networks for multivariate time series classification," *arXiv preprint arXiv:2103.14438*, 2021.

[25] B. Lim, *et al.*, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[26] P. Savarese and D. Figueiredo, "Residual gates: A simple mechanism for improved network optimization," in *Int. Conf. Learn. Represent.*, 2017.

[27] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in *6th International Conference on Learning Representations, ICLR 2018,*. OpenReview.net, 2018.

[28] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, "A review of activation function for artificial neural network," in *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 2020, pp. 281–286.

[29] J. Lederer, "Activation functions in artificial neural networks: A systematic overview," *arXiv preprint arXiv:2101.09957*, 2021.

[30] R. D. Hjelm, *et al.*, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations, ICLR*, 2019.

[31] B. Poole, *et al.*, "On variational bounds of mutual information," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019,*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 5171–5180.

[32] Z. Wang, *et al.*, "Spatiotemporal fusion transformer for large-scale traffic forecasting," *Information Fusion*, vol. 107, p. 102293, 2024.

[33] R. Zhang, *et al.*, "Lmhaze: Intensity-aware image dehazing with a large-scale multi-intensity real haze dataset," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, 2024, pp. 1–1.

[34] K. Kogkalidis, O. Melkonian, and J.-P. Bernardy, "Learning structure-aware representations of dependent types," *Advances in Neural Information Processing Systems*, vol. 37, pp. 65 095–65 118, 2024.

[35] R. Cai, *et al.*, "Read-ME: Refactorizing llms as router-decoupled mixture of experts with system co-design," *Advances in Neural Information Processing Systems*, vol. 37, pp. 116 126–116 148, 2024.

[36] T.-D. Le, *et al.*, "GLUSE: Enhanced channel-wise adaptive gated linear units se for onboard satellite earth observation image classification," *arXiv preprint arXiv:2504.12484*, 2025.

[37] T.-D. Le, T. T. Nguyen, and V. N. Ha, "The impact of LoRA adapters for llms on clinical nlp classification under data limitations," *arXiv preprint arXiv:2407.19299*, 2024.

[38] F. Pedregosa and et. al, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[39] F. Chollet and et. al., "keras," 2015.

[40] M. Popel and et. al., "Training tips for the transformer model," *arXiv preprint arXiv:1804.00247*, 2018.

[41] N. Srivastava and et. al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[42] X. Glorot and et. al., "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[43] S. Ioffe and et. al., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[44] N. Bjorck and et. al., "Understanding batch normalization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[45] H. He and et. al., "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE international joint conference on neural networks*, 2008, pp. 1322–1328.

[46] A. Maillard, *et al.*, "Bayes-optimal learning of an extensive-width neural network from quadratically many samples," *Advances in Neural Information Processing Systems*, vol. 37, pp. 82 085–82 132, 2024.

[47] M. F. Munoz, *et al.*, "Hybrid deep learning-based enhanced occlusion segmentation in PICU patient monitoring," *IEEE open Journal of Engineering in Medicine and Biology*, 2024.

[48] E. Lu, *et al.*, "Heart rate and body temperature relationship in children admitted to picu-a machine learning approach," *IEEE Transactions on Biomedical Engineering*, 2025.

[49] T.-D. Le, *et al.*, "Adaptation of autoencoder for sparsity reduction from clinical notes representation learning," *IEEE Journal of Translational Engineering in Health and Medicine*, 2023.

[50] B. A. Lompo and T.-D. Le, "Numerical attributes learning for cardiac failure diagnostic from clinical narratives–a lesa-camembert-bio approach," *arXiv preprint arXiv:2404.10171*, 2024.

[51] ——, "Multi-objective representation for numbers in clinical narratives using camembert-bio," *arXiv preprint arXiv:2405.18448*, 2024.

[52] N. Zaglam, *et al.*, "Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs," *Computers in biology and medicine*, vol. 52, pp. 41–48, 2014.

[53] M. Yahyatabar, P. Jouvet, and F. Cheriet, "Dense-unet: a light model for lung fields segmentation in chest x-ray images," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1242–1245.

[54] N. Yakob, *et al.*, "Data representation structure to support clinical decision-making in the pediatric intensive care unit: Interview study and preliminary decision support interface design," *JMIR Formative Research*, vol. 8, p. e49497, 2024.

Thanh-Dung Le (Senior Member, IEEE) received a B.Eng. degree in mechatronics engineering from Can Tho University, Vietnam, an M.Eng. degree in electrical engineering from Jeju National University, S. Korea, and a Ph.D. in electrical engineering (major in Applied Artificial Intelligence) from École de Technologie Supérieure (ETS), Canada. He was a postdoctoral fellow at the Biomedical Information Processing Laboratory, ETS. He is a Research Associate at The Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg. His research interests include applied machine learning approaches for critical decision-making systems, enhancing the model's interpretability with semantic feature extraction and mutual information techniques. He received the merit doctoral scholarship from Le Fonds de Recherche du Quebec Nature et Technologies. He also received the NSERC-PERSWADE fellowship, in Canada, and a graduate scholarship from the Korean National Research Foundation, S. Korea.

Clara Macabiau is a double degree student in Canada. After three years at the École nationale supérieure d'électrotechnique, d'électronique, d'informatique, d'hydraulique et des télécommunications (ENSEEIHT) engineering school in Toulouse, she is completing her master's degree in electrical engineering at École de Technologie Supérieure (ETS), Canada. Her master's project focuses on the detection of artifacts in photoplethysmography signals. She interests in signal processing, machine learning, and electronics.

Kevin Albert is a physiotherapist who graduated from EUSES School of Health and Sport (2018 - Girona, Spain). He developed clinical expertise in the field of function rehabilitation after neurotraumatic injury (France) and in cardio-respiratory rehabilitation (Swiss). He is currently enrolled in the Master's Biomedical Engineering program at the University of Montreal. He has joined the Clinical Decision Support System (CDSS) laboratory under the supervision of Prof. P. Jouvet, M.D. Ph.D. in the Pediatric Intensive Care Unit at Sainte-Justine Hospital (Montréal, Canada) since May 2023. His primary research interest is the application of new technologies of support care system tools with artificial intelligence, especially in ventilatory support. His research program is supported by the Sainte-Justine Hospital and the Quebec Respiratory Health Research Network (QRHN).

Symeon Chatzinotas (Fellow, IEEE) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, U.K., in 2006 and 2009, respectively. He is currently a Full Professor/Chief Scientist I and the Head of the Research Group SIGCOM, Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg. In parallel, he is an Adjunct Professor with the Department of Electronic Systems, Norwegian University of Science and Technology and a Collaborating Scholar with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos." In the past, he has lectured as a Visiting Professor with the University of Parma, Italy and contributed in numerous research and development projects for the Institute of Telematics and Informatics, Center of Research and Technology Hellas and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has authored more than 700 technical papers in refereed international journals, conferences, and scientific books and has received numerous awards and recognitions, including the IEEE Fellowship and an IEEE Distinguished Contributions Award. He is currently on the editorial board of the IEEE Transactions on Communications, IEEE Open Journal of Vehicular Technology, and the International Journal of Satellite Communications and Networking.

Philippe Jouvet received the M.D. degree from Paris V University, Paris, France, in 1989, the M.D. specialty in pediatrics and the M.D. subspecialty in intensive care from Paris V University, in 1989 and 1990, respectively, and the Ph.D. degree in pathophysiology of human nutrition and metabolism from Paris VII University, Paris, in 2001. He joined the Pediatric Intensive Care Unit of Sainte Justine Hospital—University of Montreal, Montreal, QC, Canada, in 2004. He is currently the Deputy Director of the Research Center and the Scientific Director of the Health Technology Assessment Unit, Sainte Justine Hospital–University of Montreal. He has a salary award for research from the Quebec Public Research Agency (FRQS). He currently conducts a research program on computerized decision support systems for health providers. His research program is supported by several grants from the Sainte-Justine Hospital, Quebec Ministry of Health, the FRQS, the Canadian Institutes of Health Research (CIHR), and the Natural Sciences and Engineering Research Council (NSERC). He has published more than 160 articles in peer-reviewed journals. Dr. Jouvet gave more than 120 lectures in national and international congresses.

Rita Noumeir (Member, IEEE) received master's and Ph.D. degrees in biomedical engineering from École Polytechnique of Montreal. She is a Full Professor at the Department of Electrical Engineering, École de Technologie Superieure, University of Quebec, Canada. She is a Research Chair in artificial intelligence (AI) in healthcare/ Digital health and life sciences from the Fonds de Recherche du Québec-Santé (FRQS). Concurrently, she is the Co-Director of the research cluster on AI Applied to Acute Child Care, FRQS. Her main research interest is applying AI methods to create decision support systems. She has extensively worked in healthcare information technology and image processing. She has also provided consulting services in large-scale software architecture, healthcare interoperability, workflow analysis, and technology assessment for several international software and medical companies, including Canada Health Infoway.