Hindawi
Journal of Sensors
Volume 2022, Article ID 2449956, 9 pages
https://doi.org/10.1155/2022/2449956

Research Article

ECG-ViT: A Transformer-Based ECG Classifier for Energy-Constraint Wearable Devices

Neha Shukla,[1,2] **Anand Pandey**,[1] **Anand Prakash Shukla**,[3] **and Sanjeev Chandra Neupane**[4]

[1] *CSE Department, SRM Institute of Science and Technology, Meerut Road, Modi Nagar, Delhi-NCR, India*
[2] *CS Department, KIET Group of institutions, Delhi-NCR, India*
[3] *Technical Education Department, Government of Uttar Pradesh, India*
[4] *Reconwithme, Nepal*

Correspondence should be addressed to Sanjeev Chandra Neupane; sanjeev@reconwithme.com

Received 1 March 2022; Revised 4 May 2022; Accepted 16 June 2022; Published 31 July 2022

Academic Editor: Pradeep Kumar Singh

The advancement in deep learning techniques has helped researchers acquire and process multimodal data signals from different healthcare domains. Now, the focus has shifted towards providing end-to-end solutions, i.e., processing these data and developing models that can be directly implemented on edge devices. To achieve this, the researchers try to solve two problems: (I) reduce the complex feature dependencies and (II) reduce the complexity of the deep learning model without compromising accuracy. In this paper, we focus on the later part of reducing the complexity of the model by using the knowledge distillation framework. We have introduced knowledge distillation on the Vision Transformer model to study the MIT-BIH Arrhythmia Database. A tenfold crossvalidation technique was used to validate the model, and we obtained a 99.7% F1 score and 99.3% accuracy. The model was further tested on the Xilinx Alveo U50 FPGA accelerator, and it is found fit for any low-powered wearable device implementation.

1. Introduction

Cardiovascular disease is an umbrella term that refers to cardiovascular disorders that are the leading cause of death worldwide. According to the World Health Organization (WHO), in 2017, Cardiovascular diseases (CVDs) were reported as the leading cause of death worldwide (WHO 2017). The report indicates that CVDs cause 31% of global deaths, out of which at least three-quarters of deaths occur in low- or medium-income countries [1]. One of the primary reasons behind this is the lack of primary healthcare support and the inaccessible on-demand health monitoring infrastructure. Electrocardiogram (ECG) is considered one of the essential attributes for continuous health monitoring required for identifying those at serious risk of future cardiovascular events or death [2–4].

The waveform of the ECG signal is illustrated in Figure 1. Every day, around 3 million ECGs are generated worldwide [5]. ECG readings give much information regarding the heartbeat's pace and rhythm. The ECG is evaluated clinically for a brief period using a graph of numerous consecutive cardiac cycles. The procedure starts with the discovery of an *R*-peak. It is often the most prominent portion of the ECG and hence the easiest to identify. The P-wave indicates the sinus rhythm, whereas a prolonged PR interval generally indicates a first-degree heart blockage [4, 6]. As a result, cardiologists consistently use ECG to assess the heart's condition and performance.

However, these signals are primarily collected by skin-contact ECG/BVP sensors, which may be uncomfortable and unpleasant for long-term monitoring [2, 7, 8]. The photoplethysmogram (PPG), an optical technique for monitoring changes in blood volume at the skin's surface, is regarded as a close substitute for ECG monitoring, which carries vital cardiovascular information [9]. For example, studies have shown a strong correlation between several

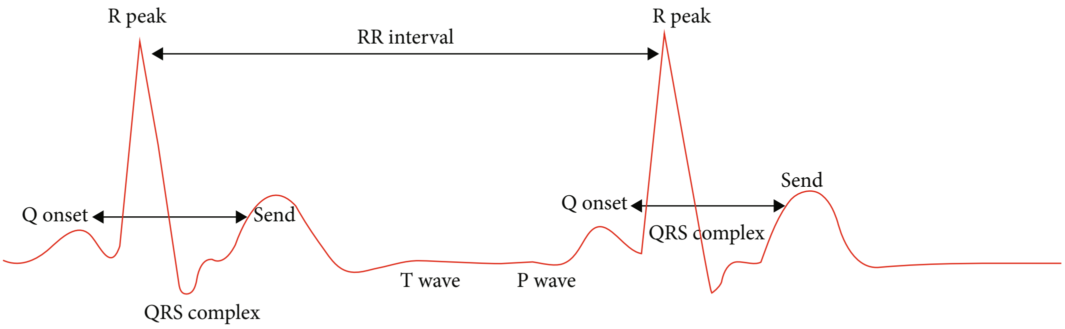

Figure 1: The illustrative waveform of the ECG signal.

features obtained from PPG (e.g., pulse rate variability) and similar metrics collected from ECG (e.g., heart rate variability), highlighting the reciprocal information between these two modalities. However, as smartwatches, smartphones, and other similar wearable and mobile devices have advanced, PPG has become the industry standard as a simple, wearable-friendly, and low-cost option for continuous heart rate (HR) monitoring for daily usage [10–12]. Nonetheless, PPG has inaccuracies in HR estimates and other limitations compared to standard ECG monitoring equipment, owing to skin tone, varied skin types, motion artifacts, and signal crossover.

However, many deep learning (DL) solutions are available to solve the ECG classification problem but most use manually crafted features. Some fully automated solutions require high computational resources like GPUs and TPUs [13–15]. So, they require high power consumption, i.e., they cannot be implemented on energy-constraint devices directly. These methods use a standard convolutional neural network (CNN) as their backbone network as they can perform very well when the input data have regular structure i.e., Euclidean. However, the ECG signals are non-Euclidean time series in nature; hence, processing them with conventional convolutional neural networks (CNNs) compromises accuracy. This motivates graph-based deep learning algorithms [16]. Graph neural network (GNN) is a general term used to denote these algorithms. Transformers are special categories of GNNs [17]. The development of Internet-of-things (IoT) devices requires bringing these complex deep learning architectures to energy and storage constraint devices.

Generally, FPGA is most suitable for implementing deep learning models as they achieve high resource utilization and lower power consumption than graphics processing unit (GPU) [18].

We have made the following contributions to this paper:

(i) A transformer neural network-based deep learning model (ECG-ViT) to solve the ECG classification problem

(ii) Cascade distillation approach to reduce the complexity of the ECG-ViT classifier

(iii) Testing and validating of the ECG-ViT model on FPGA

2. Background Study

The automated classification model can only be studied if a large ECG database with annotations is available. The MIT-BIH, ST-T, and AHA databases are used in the majority of contemporary ECG research [6, 19]. There is a single class for all of the ECG indications. Signal preprocessing is the foundational step in enhancing the quality of the ECG signal and the accuracy of the ECG analysis [20]. The subject of this investigation has been thoroughly researched. Several machine learning algorithms have been developed to assess the quality of an ECG signal. These methods mostly rely on ECG signal properties such as the RR interval and the form of the P- and T-waves [21].

2.1. ECG Classification.

2.1. ECG Classification. Applying deep learning models to ECG classification has gained growing attention [22, 23]. The state-of-the-art method for ECG heartbeat-level classification recently showed that superior results are reached by applying a ResNet model which classifies each heartbeat class separately [19, 21, 24]. In this work, we focus on developing a transformer-based method that is used for ECG classification. The comparison results with state-of-the-art methods have been shown in Section 4.

2.2. ECG Synthesis from PPG. To the best of our knowledge, only [25] has been published for the particular problem of PPG-to-ECG translation. This work did not use deep learning, instead used the discrete cosine transformation (DCT) technique to map each PPG cycle to its corresponding ECG cycle. First, onsets of the PPG signals were aligned to the *R*-peaks of the ECG signals, followed by a detrending operation to reduce noise. Next, each cycle of ECG and PPG was segmented, followed by temporal scaling using linear interpolation to maintain a fixed segment length. Finally, a linear regression model was trained to learn the relation between DCT coefficients of PPG and corresponding ECG segments. Despite several contributions, this study suffers from a few limitations. First, the model failed to produce reliable ECG in a subject-independent manner, which limits its application to only previously seen subject's data. Second, the relation between PPG and ECG segments is often not linear. Therefore, in several cases, this model failed to capture the nonlinear relationships between these two domains. Lastly, no experiments have been performed to indicate any performance enhancement gained from using the generated

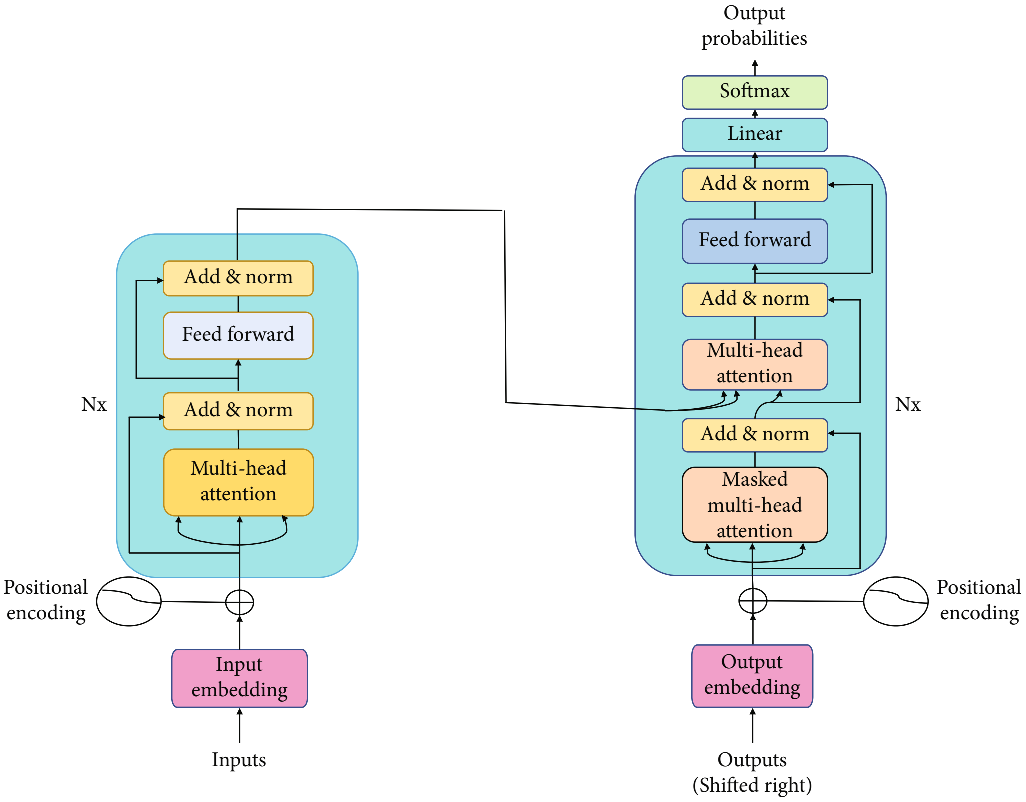

FIGURE 2: Structure of transformer as proposed by Vaswani et al. [32].

ECG instead of the available PPG (for example, a comparison of measured HR). Other works related to ECG and PPG are [26–31], but they do not show how to synthesis ECG from PPG.

2.3. Transformers in Image Classification. Transformers, deep neural networks introduced by Vaswani et al. [32], act as the reference models for the field of natural language processing. There are multiple transformer blocks with the same construction, as seen in Figure 2. An attention layer, feedforward network, skip connection, and normalization layer are present in each transformer block.

The *self-attention* mechanism of transformer is defined using equation (1). Q, K, and V are the query, key, and value vectors, respectively. d is the dimension of the model. It computes the score between input vectors by multiplying query vector to transpose of the key vector. Then, score is normalized for the stability of the gradient by dividing it with square root of dimension. In the original paper, there were eight multihead attentions. Softmax function is used to calculate the probabilities for classification, and the obtained score is multiplied with weight value matrix.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^{\top}}{\sqrt{d_k}}\right) \cdot V. \tag{1}$$

Multihead attention is a technique for enhancing the performance of the standard self-attention layer. Take note that as we go through a sentence, we often want to concentrate on multiple other words in addition to the reference word. A single-head self-attention layer constrains our capacity to concentrate on one or more particular positions without affecting our attention on other equally essential locations. This is accomplished by assigning distinct representation subspaces to attention layers. To be precise, distinct query, key, and value matrices are employed for each head, and these matrices might project the input vectors into a different representation subspace after training due to random initialization. Equation Equation (2) shows the multihead process.

$$\text{Multihead}\left(Q', K', V'\right) = \text{Concat}\,(\text{head}_1,\cdots,\text{head}_h)\text{W}^{\text{o}}, \tag{2}$$

where $\text{head}_i = \text{attention}\,(Q_i, K_i, V_i)$.

2.4. Knowledge Distillation (KD). Knowledge distillation (KD), commonly called student-teacher paradigm network, is a model compression technique used to reduce the complexity of neural networks. Rich supervision is critical when developing a machine learning or image recognition method, as it enables the model training in the present task to be accelerated by using the learning experience from relevant pretrained models. KD extracts several types of dark knowledge/privileged knowledge to aid the model's training process from the "data" perspective [33]. Depending on the

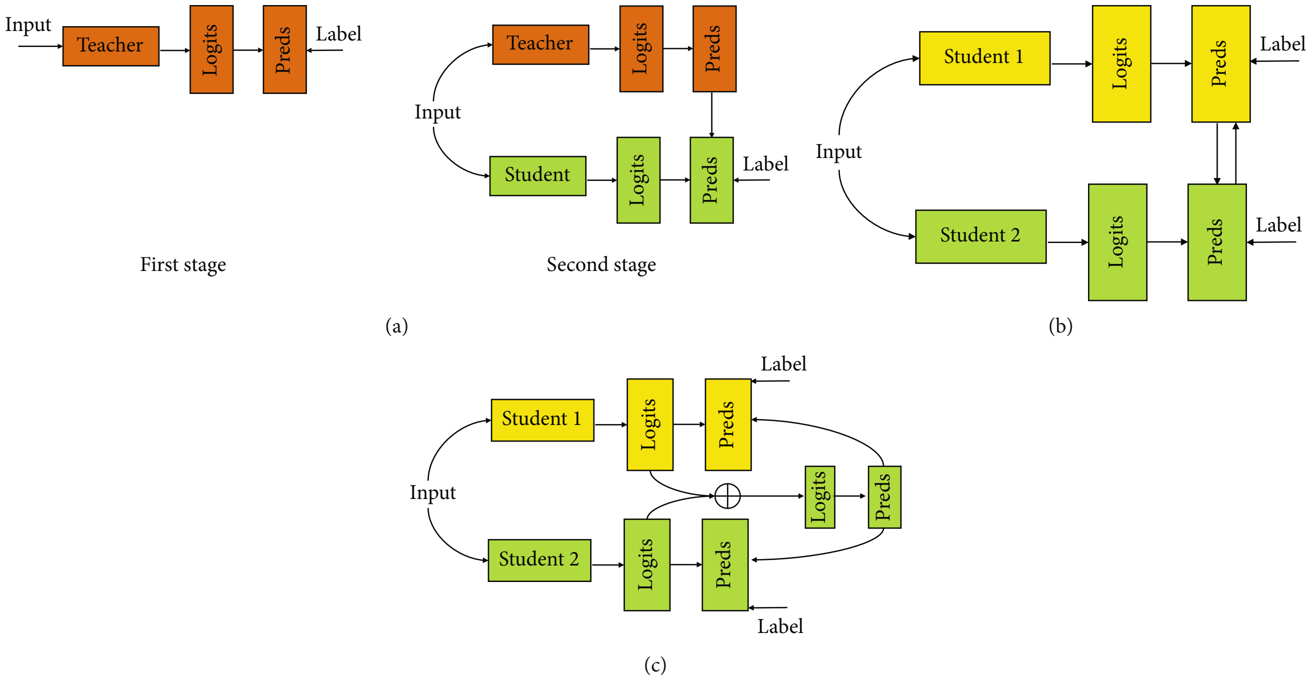

Figure 3: (a) Traditional knowledge distillation. (b) Two-stage optimization of distillation, which has to pretrain a large-scale teacher model. (c) Online distillation using either mutual learning or ensemble learning, which does not involve a teacher model.

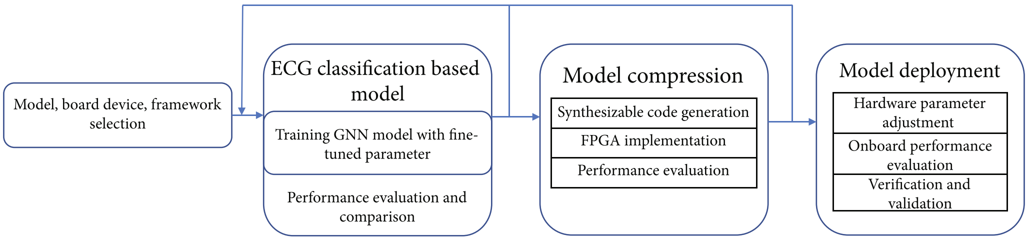
Model, board device, framework selection
ECG classification based model
Training GNN model with fine-tuned parameter
Performance evaluation and comparison
Model compression
Synthesizable code generation
FPGA implementation
Performance evaluation
Model deployment
Hardware parameter adjustment
Onboard performance evaluation
Verification and validation

Figure 4: Methodology of ECG classifier implementation on hardware.

teacher and student's training, the distillation technique is categorized as offline, online, and self-distillation Figure 3. In offline distillation, the teacher (complex) model is trained independently, and its knowledge is passed to the student (simpler) model, whereas in online distillation, both teacher and student models are trained simultaneously [34]. In this study, we have used self-distillation as it is more efficient in handling real-world situations where a large capacity teacher model is unavailable.

2.5. Field Programmable Gate Arrays (FPGA). Designers have traditionally turned to field-programmable gate arrays (FPGAs) to accelerate performance in hardware designs for compute-intensive applications such as computer vision, communications, industrial embedded systems, and increasingly the Internet of Things (IoT). Engineers who need to employ complex, compute-intensive algorithms often rely on FPGAs to accelerate execution without compromising tight power budgets [10, 11, 18]. FPGAs have emerged as a dominant platform for speeding artificial intelligence algorithms in edge-computing systems [14, 18, 35].

3. Methodology

Our work comprises mainly of three steps as demonstrated in Figure 4. We first train the ViT model with smaller patch size, as demonstrated by the accuracy which does not drop. Then, we use the knowledge distillation approach to reduce the complexity of the model. Further, the model is tested on Xilinx FPGA.

3.1. Transformer Model Architecture. The Vision Transformer (ViT) is a pure transformer that is used directly to image patch sequences for image categorization tasks. It adheres as closely as feasible to the transformer's original design. ViT's framework is shown in Figure 5. Following the ViT paradigm, a number of ViT versions have been developed to enhance performance on vision tasks. The primary techniques are to increase location, self-attention, and architectural design. Recently, academics have begun to focus on enhancing the modeling capabilities for local data [36].

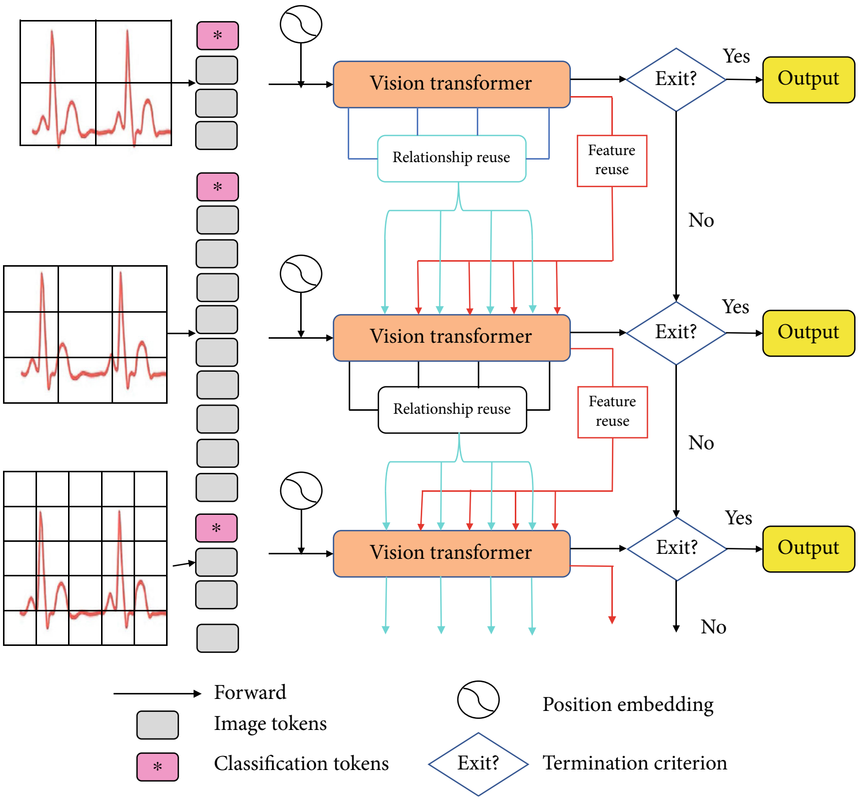

Figure 5: The ViT model for ECG arrythmia classification.

Self-attention layer, as a critical component of transformer, enables global interaction between visual patches. Numerous academics have been working on improving the computation of the self-attention layer. DeepViT suggests establishing crosshead communication in order to regenerate attention maps in order to improve variety at various levels. KVT introduces the k-NN attention to take use of the proximity of picture patches and to disregard noisy tokens by calculating attentions solely for the top-k comparable tokens [37]. Refiner investigates attention expansion in higher-dimensional space and uses convolution to enrich the attention maps' local patterns. We propose design similar to ViT without convolutional operations Figure 5.

3.1.1. Architectural Design. The ViT divides input pictures of size 224 into 16 by 16 non-overlapping patches of 14 by 14 pixels and embeds them using a convolutional stem into vectors of dimension $D_{\text{emb}} = 64N_h$. It then propagates the patches across 12 blocks that maintain the patches' dimension. Each block is comprised of an SA layer followed by a two-layer feed-forward network (FFN) with GeLU activation, both of which have residual connections. The ECG-ViT is essentially a ViT with the SA layers replaced by GPSA layers with a convolutional initialization in the first ten blocks.

Our ECG-Vit is based on the DeiT (Touvron et al., 2020) [38], an open-source hyperparameter-optimized version of the ViT. Due to its capacity to generate competitive results without the use of external data, the DeiT serves as a good baseline and is reasonably simple to train: the biggest model (DeiT-B) takes just a few days of training on eight GPUs. To simulate two, three, and four convolutional filters, we analyze three alternative ECG-ViT models with four, nine, and sixteen attention heads, respectively. Their attention heads are significantly more than those in Touvron et al., (2020) [38]. DeiT-Ti, ConViT-S, and ConViT-B utilize 4, 7, and 13 attention heads, respectively. To get models of comparable dimensions, we use two comparison techniques.

3.2. Knowledge Distillation. Traditionally, distillation works by transferring information from a clumsy instructor model to a nimble student model [39, 40]. As such, a large-scale model must be trained in advance, on the basis of which alternative knowledge definitions and transfer methodologies are recommended to improve the student model's performance [41, 42]. We augment the original embeddings with a new token, the distillation token (patches and class token). Our distillation token is similar to the class token in that it interacts with other embeddings through self-attention and is produced by the network after the final layer. The distillation component of the loss indicates its intended use. As with a traditional distillation, the distillation embedding enables our model to learn from the teacher's output while staying complimentary to the class embedding.

Interestingly, we notice that the learnt class and distillation tokens converge toward distinct vectors, with an average cosine similarity of 0.06 between these tokens. As the class and distillation embeddings are calculated at each layer, their similarity increases progressively across the network, until they reach the last layer, where their similarity is great ($\cos = 0.91$) but still less than one. This is to be anticipated,

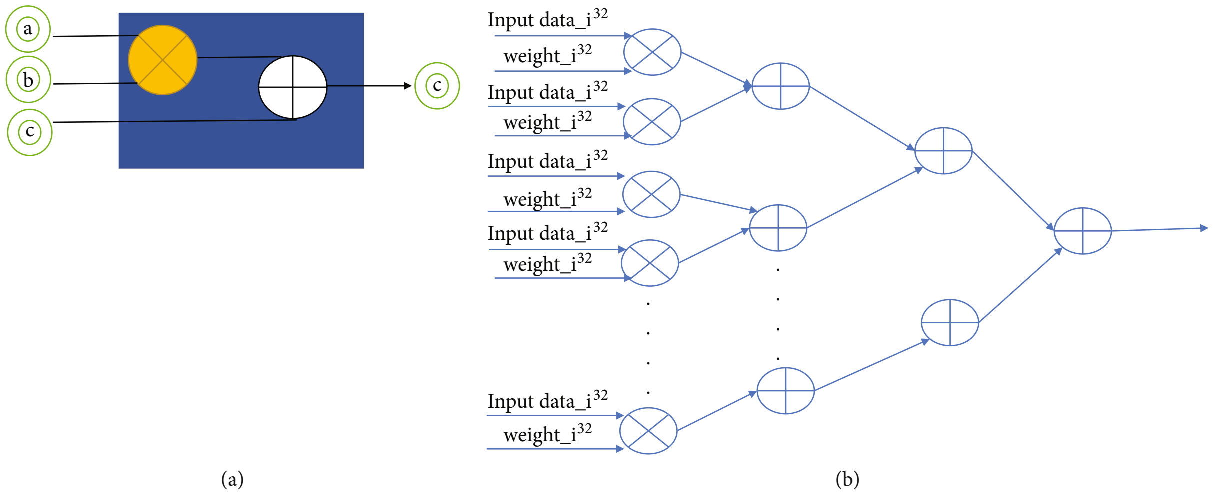

Figure 6: (a) Independent MAC unit. (b) MAC operations of neural network.

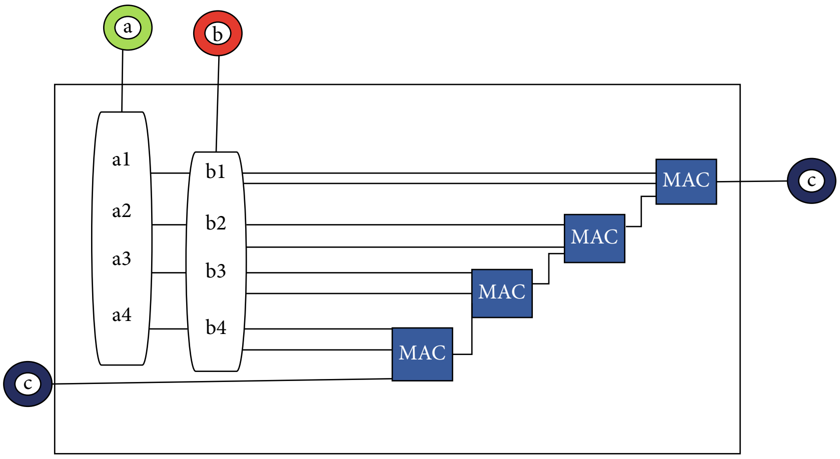

Figure 7: MAC_4 operations.

Table 1: ECG beat type present in MIT-BIH dataset.

ECG beat type	AAMI
Ventricular premature beat	VE
Supraventricular premature beat	SVE
Atrial fibrillation	AT
Unclassifiable beat	U

since they are attempting to create targets that are comparable but not identical.

At a greater resolution, we employ both the genuine label and teacher prediction during the fine-tuning step. We use a teacher with the same target resolution as the lower-resolution teacher, which is typically obtained using the Touvron et al. [43] method. We have also tried using solely true labels; however, this decreases the teacher's advantage and results in worse performance.

At test time, the transformer's class or distillation embeddings are coupled with linear classifiers and capable of inferring the picture label. Nonetheless, our referent technique is a late merger of these two distinct heads, to which we add the softmax output from the two classifiers.

Our distillation strategy results in a vision transformer that is comparable to the top ConvNets in terms of accuracy-throughput trade-off. Surprisingly, the distilled model beats its instructor in terms of the accuracy-throughput trade-off. Our best model on the MIT-BIH data-set has a top-1 accuracy of 99.7%.

3.3. Hardware Design. The core of our deep learning algorithm depends on general matrix multiplication step. It is a combination of multiplication and accumulation (MAC unit) of weights of the neural network as demonstrated in Figure 6.

MAC_4 is obtained by combining four MAC units as shown in Figure 7. By implementing 16 MAC_4 units on FPGA, we have obtained the ECG-ViT. There are a total of 64 operations performed by GEMM unit in 1 clock cycle which uses 64 multiplier and adder as shown in Figure 5.

We had to provide 4×4 matrices p and q, which equates to 32 scalars, to obtain 16 dot products of matrix r. Hence, we need to transfer only 2 scalars per dot product from memory on each update.

For efficient implementation, we have used 16-bit fixed-point representation. We have approximated the

Table 2: ECG-ViT outperforms other classifier [44] for two common classification tasks.

	VE			SVE		
	Sensitivity	Positive predicted value	F-score	Sensitivity	Positive predicted value	F-score
Wiens and Guttag [44]	99.0	87.4	99.1	63.5	53.7	58.2
ECG-ViT	99.2	89.6	99.4	68.9	57.6	60.2

Table 3: ECG-ViT outperforms Cong et al. classifier [45] for four common classification tasks.

	VE		SVE		AT		U	
	mP	mA	mP	mA	mP	mA	mP	mA
Cong et al. [45]	0.95	0.94	0.92	0.95	0.95	0.92	0.88	0.87
ECG-ViT	0.97	0.95	0.93	0.95	0.98	0.94	0.92	0.91

multiplication operations at the cost of accuracy to reduce the energy consumption, inference speed, and less area occupancy. We consumed 38% less area and 27% less energy to implement the general matrix multiplication. Since the multiplier circuit is more expensive than the adder circuit, approximations have been done for multiplication. While testing, we have analyzed that there is not much drop in accuracy.

4. Results and Discussion

Classifier performance was as follows: a thorough ablation study of our ECG-ViT model is performed on the MIT-BIH Arrhythmia Database (MITDB), a widely used benchmark. We preprocessed the data to obtain the sample at 128 Hz. Four classification tasks were proposed by the Association for the Advancement of Medical Instrumentation (AAMI) as shown in Table 1.

For these four classification tasks, we tested our proposed approach, and we report the results when tested on the records reported on in. Table 2 demonstrates the comparison of sensitivity, positive predicted value, and F-score of our ECG-ViT algorithm and Wiens and Guttag [44]. Our method clearly outperforms the classifier used by [4].

We compared our ECG-ViT with Cong et al. [45] on parameter of mean precision and mean accuracy as demonstrated in Table 3. All four classification tasks such as VE, SVE, AT, and U have been compared. Our classifier has clearly outperformed the previous classifier [4] by a significant margin.

5. Conclusion

In this paper, we provided a new way of implementing the ECG IoT monitoring system based on transformers. The model was compressed using knowledge distillation to reduce its complexity. The implemented algorithm was tested on Xilinx Alveo U50 FPGA and outperformed existing state-of-the-art methods. We have obtained accuracy of 99.7%. In the future work, we plan to reduce the area for hardware implementation i.e., to make it area aware so that it could be implemented on wearable devices to diagnose heartbeat.

Data Availability

The dataset can be found from the below mentioned link https://physionet.org/content/mitdb/1.0.0/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] U. Satija, B. Ramkumar, and M. S. Manikandan, “A new automated signal quality-aware ECG beat classification method for unsupervised ECG diagnosis environments,” *IEEE Sensors Journal*, vol. 19, no. 1, pp. 277–286, 2019.

[2] M. Risso, A. Burrello, D. J. Pagliari et al., “Robust and energy-efficient PPG-based heart-rate monitoring,” in *In 2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, Daegu, Korea, 2021.

[3] U. Neeraj, J. Satija, J. Mathew, and R. K. Behera, “A unified attentive cycle-generative adversarial framework for deriving electrocardiogram from seismocardiogram signal,” *IEEE Signal Processing Letters*, vol. 29, pp. 802–806, 2022.

[4] Z. Sun, C. Wang, Y. Zhao, and C. Yan, “Multi-Label ECG signal classification based on ensemble classifier,” *IEEE Access*, vol. 8, pp. 117986–117996, 2020.

[5] T. Golany, D. Freedman, and K. Radinsky, “SimGANs: simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification,” *International Conference on Machine Learning*, vol. 119, pp. 3597–3606, 2020.

[6] X. Tang and W. Tang, “An ECG delineation and arrhythmia classification system using slope variation measurement by ternary second-order delta modulators for wearable ECG sensors,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 1053–1065, 2021.

[7] X. Liu, H. Wang, Z. Li, and L. Qin, “Deep learning in ECG diagnosis: a review,” *Knowledge-Based Systems*, vol. 227, p. 107187, 2021.

[8] E. H. Houssein, I. E. Ibrahim, N. Neggaz, M. Hassaballah, and Y. M. Wazery, “An efficient ECG arrhythmia classification method based on manta ray foraging optimization,” *Expert Systems with Applications*, vol. 181, p. 115131, 2021.

[9] K. J. Chen, P.-C. Chien, Z.-J. Gao, and C.-H. Wu, “A fast ECG diagnosis by using non-uniform spectral analysis and the artificial neural network,” *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, 2021.

[10] J. Lu, D. Liu, Z. Liu et al., “Efficient hardware architecture of convolutional neural network for ECG classification in wearable healthcare device,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 7, pp. 2976–2985, 2021.

[11] H. B. Seidel, M. M. A. da Rosa, G. Paim, E. A. C. da Costa, S. J. M. Almeida, and S. Bampi, “Approximate pruned and truncated Haar discrete wavelet transform VLSI hardware for

energy-efficient ECG signal processing," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1814–1826, 2021.

[12] L. Sun, Y. Wang, Z. Qu, and N. N. Xiong, "BeatClass: a sustainable ECG classification system in IoT-based eHealth," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7178–7195, 2021.

[13] S. Ran, X. Yang, M. Liu et al., "Homecare-oriented ECG diagnosis with large-scale deep neural network for continuous monitoring on embedded devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[14] A. Canis, J. Choi, M. Aldham et al., "LegUp: high-level synthesis for FPGA-based processor/accelerator systems," in *In Proceedings of the 19th ACM/SIGDA international symposium on Field programmable gate arrays*, pp. 33–36, New York, 2011.

[15] A. K. Tiwari and N. Shukla, "Brain tumor segmentation using CNN," *Recent Trends Commun. Electron*, pp. 411–415, 2021.

[16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[17] Z. Qiao, Y. Fu, P. Wang et al., "RPT: toward transferable model on heterogeneous researcher data via pre-training," *IEEE Transactions on Big Data*, p. 1, 2022.

[18] M. Eldafrawy, A. Boutros, S. Yazdanshenas, and V. Betz, "FPGA logic block architectures for efficient deep learning inference," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 13, no. 3, 2020.

[19] C. Böck, P. Kovács, P. Laguna, J. Meier, and M. Huemer, "ECG beat representation and delineation by means of variable projection," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 2997–3008, 2021.

[20] U. Satija, B. Ramkumar, and M. S. Manikandan, "Automated ECG noise detection and classification system for unsupervised healthcare monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 722–732, 2018.

[21] M. A. Quiroz-Juárez, O. Jiménez-Ramírez, R. Vázquez-Medina, E. Ryzhii, M. Ryzhii, and J. L. Aragón, "Cardiac conduction model for generating 12 Lead ECG signals with realistic heart rate dynamics," *IEEE Transactions on Nanobioscience*, vol. 17, no. 4, pp. 525–532, 2018.

[22] J. Fayn, "A classification tree approach for cardiac ischemia detection using spatiotemporal information from three standard ECG leads," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 95–102, 2011.

[23] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019.

[24] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2016.

[25] W. Su, X. Zhu, Y. Cao et al., "Vl-bert: Pre-training of generic visual-linguistic representations," 2019, http://arxiv.org/abs/1908.08530.

[26] X. He, R. A. Goubran, and X. P. Liu, "Secondary peak detection of PPG signal for continuous Cuffless arterial blood pressure measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 6, pp. 1431–1439, 2014.

[27] K. Natarajan, R. C. Block, M. Yavarimanesh et al., "Photoplethysmography fast upstroke time intervals can be useful features for cuff-less measurement of blood pressure changes in humans," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 53–62, 2022.

[28] A. Chandrasekhar, M. Yavarimanesh, K. Natarajan, J.-O. Hahn, and R. Mukkamala, "PPG sensor contact pressure should be taken into account for cuff-less blood pressure measurement," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3134–3140, 2020.

[29] A. Hernando, M. D. Pelaez-Coca, M. T. Lozano et al., "Autonomic nervous system measurement in hyperbaric environments using ECG and PPG signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 132–142, 2019.

[30] Q. Zhu, X. Tian, C.-W. Wong, and M. Wu, "Learning your heart actions from pulse: ECG waveform reconstruction from PPG," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 16734–16748, 2021.

[31] J. Yu, S. Park, S.-H. Kwon, K.-H. Cho, and H. Lee, "AI-Based stroke disease prediction system using ECG and PPG bio-signals," *IEEE Access*, vol. 10, pp. 43623–43638, 2022.

[32] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[33] H.-J. Ye, S. Lu, and D.-C. Zhan, "Generalized knowledge distillation via relationship matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2022.

[34] S. Li, M. Lin, Y. Wang et al., "Distilling a powerful student model via online knowledge distillation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2022.

[35] T.-H. Tsai, Y.-C. Ho, and M.-H. Sheu, "Implementation of FPGA-based accelerator for deep neural networks," in *In 2019 IEEE 22nd International Symposium on Design and Diagnostics of Electronic Circuits \& Systems (DDECS)*, pp. 1–4, Cluj-Napoca, Romania, 2019.

[36] K. Han, Y. Wang, H. Chen et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2022.

[37] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: residual vision transformers for multi-modal medical image synthesis," *IEEE Trans. Med. Imaging*, p. 1, 2022.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers \& distillation through attention," *In International Conference on Machine Learning*, vol. 139, pp. 10347–10357, 2021.

[39] C. Wang, D. Chen, J.-P. Mei, Y. Zhang, Y. Feng, and C. Chen, "SemCKD: semantic calibration for cross-layer knowledge distillation," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2022.

[40] J. Song, Y. Chen, J. Ye, and M. Song, "Spot-adaptive knowledge distillation," *IEEE Transactions on Image Processing*, vol. 31, pp. 3359–3370, 2022.

[41] Z. Feng, J. Lai, and X. Xie, "Resolution-aware knowledge distillation for efficient inference," *IEEE Transactions on Image Processing*, vol. 30, pp. 6985–6996, 2021.

[42] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, p. 1, 2020.

[43] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," *Advances In Neural Information Processing Systems*, vol. 32, 2019.

[44] J. Wiens and J. V. Guttag, “Active learning applied to patient-adaptive heartbeat classification,” *Adv. Neural Inf. Process, Syst. 23 24th Annu. Conf. Neural Inf. Process. Syst*, pp. 1–9, 2010.

[45] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, and Z. Zhang, “High-Level synthesis for FPGAs: from prototyping to deployment,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 4, pp. 473–491, 2011.