Arrhythmia Detection Using Transformer Models and ECG Images

Devika Gumaste
Department of Electrical Engineering
Columbia University
New York, USA
dg3370@columbia.edu

Jessica Marshall
Department of Electrical Engineering
Columbia University
New York, USA
jm5679@columbia.edu

***Abstract*—This study explores the use of Vision Transformers (ViTs) for the detection of cardiac arrthythmia in 12-lead Electrocardiogram (ECG) images, aiming to facilitate early detection and treatment of arrhythmias from non-clinical settings. We investigate several transformer architectures, including BEiT, Swin Tiny, Google ViT, and DEiT, leveraging their robust feature extraction capabilities for medical image analysis. Our approach incorporates a knowledge distillation framework where a larger, pre-trained transformer (teacher) model imparts learned representations to a more compact convolutional neural network (CNN, student) model, optimizing for both performance and computational efficiency. The dataset comprises 16,884 12-lead ECG images categorized into arrhythmia and non-arrhythmia (i.e., abnormal and normal heartbeats), processed and evaluated using a 5-fold cross-validation scheme. The primary metrics for model performance evaluation include precision, recall, accuracy, F1-score, and a confusion matrix. Results from this study are expected to provide insights into the scalability of transformer-based models in healthcare applications, particularly for in-home monitoring and early diagnosis.**

I. Introduction

Cardiovascular disease, commonly known as heart disease, significantly impacts global health, leading to a variety of complications such as coronary artery disease, heart failure, and myocardial infarction. Ventricular arrhythmia accounts for 75–80% of sudden cardiac death cases [8]. The ability to detect heart disease accurately and early is crucial in clinical practice, as it enables timely and targeted interventions, which can significantly alter patient outcomes. ECG is a vital, non-invasive tool widely used in the diagnosis of cardiac arrhythmia and other heart conditions, offering a quick and convenient method for monitoring heart activity [6]. ECG recordings can be broadly categorized into two types: Multi-lead and single-lead ECGs. Multi-lead ECGs involve multiple electrodes placed on the patient's chest and limbs, providing a comprehensive view of the heart's electrical activity from various angles. This is the standard clinical setup for a detailed heart examination.

In recent years, the integration of deep learning techniques has significantly transformed the field of cardiac diagnostics. Recent advancements in medical technology have enhanced our ability to detect heart disease earlier and more accurately. These advancements include the integration of sophisticated computational approaches such as machine learning algorithms, data mining techniques, and predictive modeling frameworks. These technologies leverage extensive clinical and physiological data, enhancing diagnostic accuracy and improving risk stratification.

In this context, deep learning, and particularly Transformer models, have emerged as powerful tools in various domains, including medical imaging. The self-attention mechanism in transformers allows for the efficient processing of sequences by capturing global dependencies within the data [7]. This capability makes them particularly suited for complex tasks such as interpreting ECG images, where traditional approaches may falter.

This work focuses on employing ViT models for the detection of heart disease using ECG images, aiming to harness their robust feature extraction capabilities to enhance diagnostic precision. Specifically, we explore the application of Vision Transformer architectures such as Google-ViT, Microsoft-BeiT, and Swin-Tiny to classify ECG images into normal and abnormal heartbeats. The introduction of advanced models into the medical field holds promise not only for improved diagnostic accuracy but also for the development of accessible tools that could be deployed in non-clinical settings, potentially democratizing advanced health monitoring technologies and facilitating early and widespread disease detection.

II. Summary of the Original Paper

A. Methodology of the Original Paper

We took inspiration from Qin et al. [5] who utilized the dataset of 1937 individual patient records from ECG devices across various healthcare facilities in Pakistan, covering conditions like COVID-19, abnormal heartbeats, myocardial infarction (MI), and normal heart activity. The dataset includes both detailed patient data and 12-lead ECG images, making it a rich resource for research on various cardiovascular conditions.

The preprocessing involves selecting regions of interest (ROI) from the ECG images, specifically the waveform sections. These images are then binarized to eliminate unnecessary color information and background noise, enhancing the focus on waveform data. This process also involves cropping the ECG into 12 individual lead images for detailed analysis.

The study employs three advanced vision transformer models—Google-Vit, Swin, and BEiT—to analyze the processed ECG images. These models utilize self-attention mech-

anisms to process image patches sequentially, capturing intricate relationships within the data.

B. Key Results of the Original Paper

The results section of the study details the performance evaluation of vision transformer models used for detecting diseases from clinical ECG waveform images. A total of 817 clinical ECG images, each split into 12 leads resulting in 9,804 total images, were analyzed using Google-Vit, Swin, and BEiT models. These models were trained and tested using a 5-fold cross-validation technique to mitigate potential biases and ensure a robust evaluation of the models' generalization capabilities.

Key training parameters set for the models included varying epochs, batch sizes, learning rates, warmup ratios, and the AdamW optimizer. These parameters were systematically fine-tuned through grid search techniques to optimize model performance. The training process was carefully monitored to avoid overfitting, as indicated by stable loss curves across training iterations.

Performance metrics such as accuracy, precision, recall, and F1-score were employed to provide a comprehensive assessment of each model. The results demonstrated that all three vision transformer models performed effectively in ECG disease detection tasks. Among them, the BEiT model exhibited superior performance across all metrics, marking it as a particularly promising tool for disease detection in ECG images.

III. Methodology

A. Objectives and Technical Challenges

The primary objective of our project is to develop an effective heart disease detection system using vision transformers applied to ECG images. This approach aims to leverage the capabilities of transformer models to capture intricate patterns within ECG data that are indicative of various cardiac abnormalities.

Effective preprocessing to transform the multi-lead ECGs into suitable formats for transformer models poses a challenge given the original image format. There is also difficulty in fine-tuning deep learning models that were primarily designed for image data to work effectively with ECG signals, which are essentially time-series data represented in image format. Ensuring the models are not only accurate but also interpretable, is crucial in medical applications for trust and reliability.

B. Problem Statement and Design Description

The ECG images are first converted to grayscale and undergo thresholding to enhance feature visibility. This is followed by a cropping step where the image is segmented into individual ECG leads.

```
Preprocess(image):
  1. Convert to grayscale
  2. Apply thresholding
  3. Crop to separate individual leads
  Return processed_image
```

We use a combination of different transformer architectures (Google Vision Transformer, Swin-Tiny, BEiT, and DEiT) to analyze these processed images. Each model treats the segmented leads as separate input instances, predicting potential abnormalities. The system consists of three main blocks:

1) *Input Module*: Handles the loading and preprocessing of 12-lead ECG data.
2) *Processing Module*: Comprises the transformer models that process the preprocessed sinlge-lead ECG images.
3) *Output Module*: Aggregates the predictions from different models and determines the final diagnosis.

```
Algorithmic Steps:
For each ECG image in the dataset:
  processed_image = Preprocess(ECG image)
  predictions = []
  For each model in [ViT, Swin-Tiny, BEiT, DEiT]:
    prediction = model(processed_image)
    predictions.append(prediction)
  final_prediction = aggregate_predictions(predict
  Store or display final_prediction
```

This project represents a partial reproduction of the methodologies described in the referenced study, focusing on the adaptability and integration of multiple transformer models. The expected outcome is a robust system capable of high accuracy in predicting cardiac arrhythmia from single-lead ECG images.

IV. Implementation

This section provides a detailed discussion about the implementation of the project, including data handling, the deep learning network configuration of the chosen models, and software design.

A. Data

The project utilizes a specifically curated dataset consisting of ECG images, which have been preprocessed to fit the input requirements of transformer models. The data is sourced from public health databases and has been anonymized to ensure privacy. It includes thousands of 12-lead ECG images labeled with various heart conditions. Images are normalized and resized to uniform dimensions to ensure consistency in model input.

B. Deep Learning Network

This subsection details the architecture and training algorithms of the deep learning network used in this project.

Flowchart: A flowchart of the training process is included as Figure 1, illustrating the step-by-step data flow from input to output through various transformation stages.

C. Software Design

This subsection describes the software design, including a flowchart.

GitHub Repository: All code and documentation for this project are maintained in a GitHub

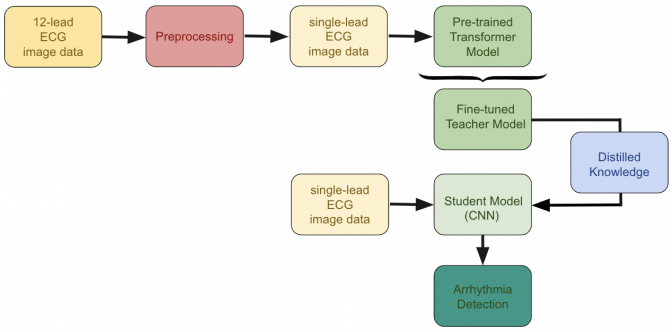

Fig. 1. Flowchart of the Training Process

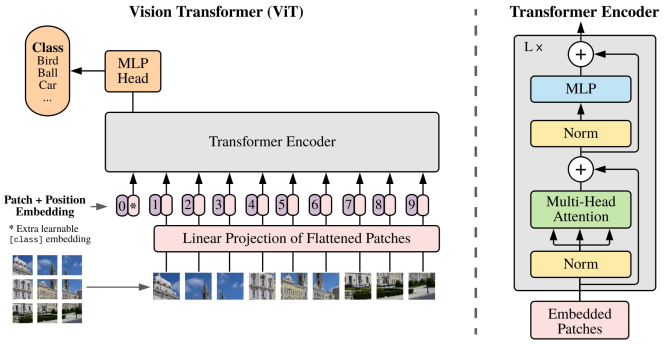

Fig. 2. Vision Transformer Architecture [1]

repository available at: https://github.com/ecbme6040/e6691-2024spring-project-DGJM-dg3370-jm5679.git

D. Proposed Models

1) Vision Transformer: Introduced by Dosovitskiy et al. [1] in their paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" in 2021, the ViT represents a significant advancement in the realm of image classification. Departing from the conventional convolutional neural networks (CNNs), ViT adopts a transformer-based architecture to process images.

In essence, ViT operates by segmenting an input image into fixed-size patches, similar to breaking down a puzzle into smaller, manageable pieces. Each patch undergoes linear embedding, transforming the pixel values into high-dimensional vectors. These vectors, along with positional embeddings to encode spatial information, are then fed into a Transformer Encoder to capture long-range dependencies and extract meaningful features from the image.

One distinctive aspect of ViT is its use of a special token, named the "classification token," which is learned alongside the image patches. This token acts as a global representation of the entire image and facilitates the classification task by aggregating information from all patches. To train ViT effectively, large-scale datasets such as ImageNet are utilized. Through extensive training, ViT learns to discern intricate patterns and features within images, achieving remarkable performance.

For practical implementation, we can readily access pre-trained ViT models through resources like the Hugging Face transformers library. By fine-tuning these pretrained models on specific datasets or tasks, we can leverage ViT's capabilities for a wide range of applications, from image classification to object detection and beyond.

In conclusion, ViT's ability to capture global dependencies, combined with the flexibility of transfer learning, paves the way for exciting advancements and applications in the field of visual recognition.

2) BEiT: BEiT, short for "Bidirectional Encoder Representation from Image Transformers" [2] marks a significant stride in the domain of computer vision and natural language processing convergence. Developed as an extension of the influential BERT (Bidirectional Encoder Representations from Transformers) model, BEiT bridges the gap between textual and visual understanding. Unlike its predecessors that predominantly focus on text or image data independently, BEiT amalgamates these modalities, offering a unified framework for comprehensive understanding of multimodal data.

BEiT exhibits remarkable versatility, finding applications across a spectrum of tasks ranging from image captioning to visual question answering.

The pre-training process of BEiT involves exposing the model to vast quantities of multimodal data, enabling it to learn rich representations that encapsulate the nuanced relationships between images and texts. This pre-training phase equips the model with a foundational understanding of the underlying semantics and associations within multimodal inputs. Leveraging self-supervised learning techniques akin to those employed in BERT, BEiT learns to predict missing elements within multimodal inputs, encouraging the development of robust representations that capture the intricate interplay between textual and visual information.

Similar to BERT, BEiT uses a masked image modeling task to pretrain vision transformers. Each image has two views - *image patches* and *visual tokens*. The raw images are tokenized into visual tokens. Some image patches are randomly masked and fed into the transformer. The pre-training task is to recover the original visual tokens based on the corrupted image patches.

The pre-trained BEiT model can then be fine-tuned on specific downstream tasks, further enhancing its performance and adaptability across various applications in both computer vision and natural language processing domains.

3) DEiT: The DeiT (Data-efficient Image Transformer) [3] architecture was developed as a response to the computational demands of traditional transformer-based models.

One of the hallmark features of DeiT is its utilization of distillation techniques during training. By distilling knowledge from a larger, pre-trained model, DeiT can learn to mimic the behavior of the larger model while requiring significantly fewer parameters. This distillation process allows DeiT to achieve remarkable performance on various image recognition

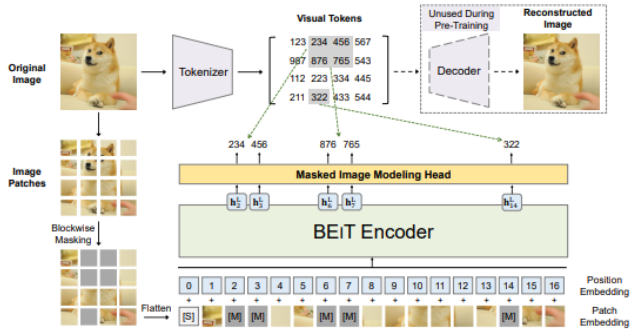

Fig. 3. Overview of BEiT Pretraining [2]: BEiT employs an "image tokenizer" trained via autoencoding-style reconstruction to tokenize images into discrete visual tokens. During pre-training, each image undergoes masking of randomly selected patches (shown in grey), replaced with a special mask embedding (M), before being fed into a vision Transformer to predict the original image's visual tokens based on encoding vectors of the corrupted image.

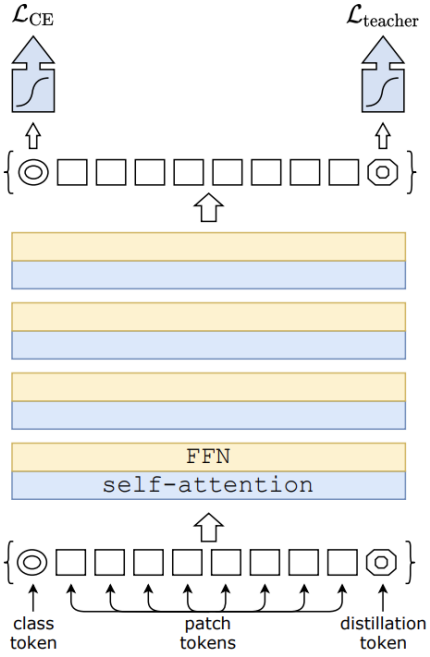

Fig. 4. Data-Efficient Image Transformer

tasks while being trained on substantially smaller datasets compared to its predecessors.

DeiT introduces innovations in training strategies to further enhance its efficiency and effectiveness. Techniques such as knowledge distillation, self-distillation, and data augmentation play pivotal roles in shaping DeiT's learning process, enabling it to generalize well across various domains and achieve competitive performance with fewer computational resources.

4) Swin Tiny: The Swin Transformer architecture was introduced in 2021 [4] by Liu et al. Swin Transformer stands for "Shifted Windows Transformer," emphasizing its unique approach to processing images through a hierarchical window-based mechanism. This novel architecture addresses the limitations of capturing global context in large images while maintaining computational efficiency.

At the core of the Swin Transformer is the concept of hierarchical processing. Unlike ViT, which treats the entire image as a sequence of tokens, Swin Transformer divides the image into non-overlapping patches or "windows." These windows are processed hierarchically through multiple stages or "layers," where each layer refines the features extracted from the previous stage. By incorporating a hierarchical approach, Swin Transformer can efficiently capture both local and global context, leading to improved performance in tasks such as image classification and object detection.

Swin Transformer introduces a novel mechanism called "swin blocks" that efficiently process information within each window. Swin blocks consist of two key components: a shifted window self-attention mechanism and a cross-window token aggregation step. This design enables the model to capture both local and global context within each window, effectively leveraging the hierarchical structure of the image. By shifting the windows across the image, Swin Transformer effectively aggregates information from neighboring patches, enhancing its ability to capture fine-grained details and long-range dependencies. For fine-tuning, the model microsoft/swin-tiny-patch4-window7-224 is readily available in the HuggingFace library.

E. Knowledge Distillation

V. Methods: Knowledge Distillation

In this project, we utilize a knowledge distillation technique inspired by recent advances in enhancing single-lead ECG interpretation models [5]. We apply Multi-View Knowledge Transferring (MVKT) to transfer critical diagnostic features from a BEiT teacher model to a MobileNetV3 student model.

A. Teacher-Student Model Configuration

The teacher model is a pre-trained BEiT loaded with weights fine-tuned on a comprehensive multi-lead ECG dataset. The student model is a lightweight MobileNetV3, optimized for less computational intensity and quicker inference times, making it suitable for deployment in mobile or remote health monitoring devices. The code was implemented by importing the pre-trained student and teacher models, then loading the teacher model with the fine-tuned weights.

B. Mixed Knowledge Distillation (MKD)

The distillation process involves a mixed loss function that combines cross-entropy for accuracy and Kullback-Leibler divergence for feature consistency, enhancing the student's ability to replicate the teacher's diagnostic performance. This approach controls the balance between fitting the true labels and mimicking the teacher.

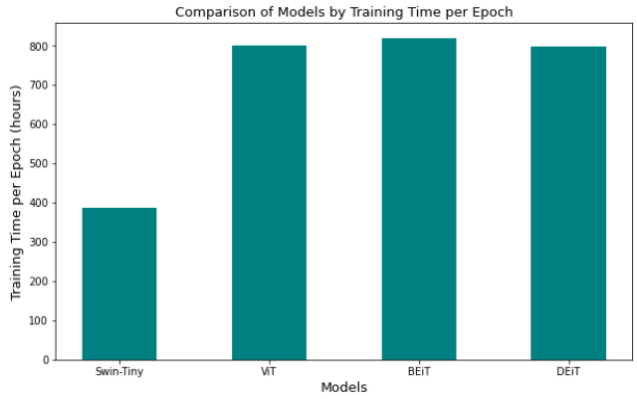

Fig. 5. Comparison of training time per epoch

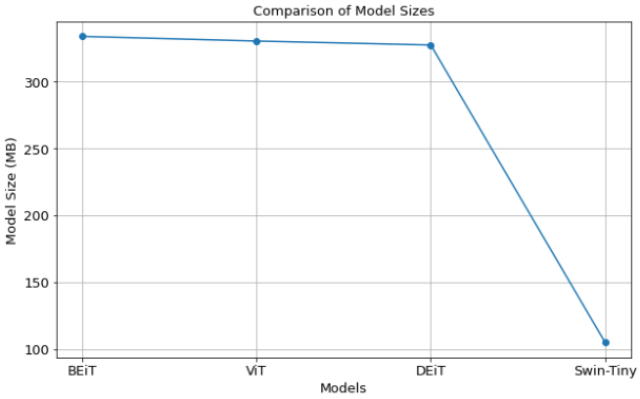

Fig. 6. Comparison of model size in MB

Mixed Knowledge Distillation Loss:

$$\text{MKD Loss} = \alpha \times \text{CE Loss} + (1-\alpha) \times \text{KL Divergence} \times T^2 \quad (1)$$

where α is a hyperparameter balancing the trade-off between fitting the true labels (hard targets) and mimicking the teacher's output (soft targets), and T is the temperature parameter moderating the softening of probabilities.

VI. Results

The evaluation of four transformer models—Google-ViT, Swin, BEiT, and DEiT—on the binary classification task of normal versus abnormal heartbeat detection using ECG images yields compelling insights. Across precision, recall, F1-score, and accuracy metrics, all models exhibit impressive performance. Swin notably stands out, boasting the highest scores across most metrics: precision of 99.64%, recall of 95.96%, F1-score of 97.77%, and accuracy of 98.02%. BEiT closely follows, achieving a precision of 99.63%, recall of 95.61%, F1-score of 97.58%, and accuracy of 97.86%. DEiT also performs admirably, with a precision of 99.28%, recall of 96.14%, F1-score of 97.68%, and accuracy of 97.94%.

Comparing these experimental outcomes with the results from the original papers reveals improvements in performance metrics. For instance, Google-ViT, which originally reported a precision, recall, F1-score, and accuracy of 94.3%, now achieves notably higher scores across the board, with an average improvement of approximately 4%. Swin, BEiT, and DEiT similarly demonstrate enhancements from their original implementations.

Examining additional factors such as training time and loss curve dynamics sheds light on the behavior of the models. Despite its shorter training duration, DEiT exhibits competitive performance, achieving high classification metrics with a training time of approximately 11973.28 seconds. Swin's ability to deliver superior results while maintaining a relatively smaller model size, coupled with a training time of approximately 13481.89 seconds, underscores its computational efficiency and suitability for resource-constrained environments.

Analyzing the nuances of the training and validation loss curves highlights potential challenges faced by each model. While Google-ViT and BEiT display stable learning dynamics, Swin's validation loss curve exhibits greater variability, suggesting the need for further investigation into its generalization capabilities. Similarly, DEiT's slight increase in validation loss hints at potential overfitting concerns, despite its impressive classification performance.

Overall, Swin emerges as the frontrunner in this comparative analysis, boasting the highest precision, recall, F1-score, and accuracy metrics. However, each model offers unique strengths and considerations, ranging from computational efficiency to generalization capabilities, thereby enriching the landscape of deep learning solutions for abnormal heartbeat detection in ECG data.

TABLE I
Experiment Parameters

Parameters	Google-Vit	Swin	BEiT	DEiT
Epochs	30	35	25	15
Batch	64	80	64	64
Learning Rate	9×10^{-6}	4×10^{-5}	6×10^{-5}	4×10^{-5}
Warmup	0.1	0.1	0.08	0.08
Optimizer	AdamW	AdamW	AdamW	AdamW

TABLE II
Original Paper Parameters

Parameters	Google-Vit	Swin	BEiT
Epochs	30	35	25
Batch	64	80	64
Learning Rate	9×10^{-6}	4×10^{-5}	6×10^{-5}
Warmup	0.1	0.1	0.08
Optimizer	AdamW	AdamW	AdamW

TABLE III
Experiment Results

Models	Precision	Recall	F1-score	Accuracy
Google-Vit	0.9780	0.9368	0.9570	0.9619
Swin	0.9964	0.9596	0.9777	0.9802
BEiT	0.9963	0.9561	0.9758	0.9786
DEiT	0.9928	0.9614	0.9768	0.9794

TABLE IV
ORIGINAL PAPER RESULTS

Models	Precision	Recall	F1-score	Accuracy
Google-Vit	0.943	0.943	0.942	0.943
Swin	0.955	0.955	0.954	0.955
BEiT	0.959	0.959	0.959	0.959

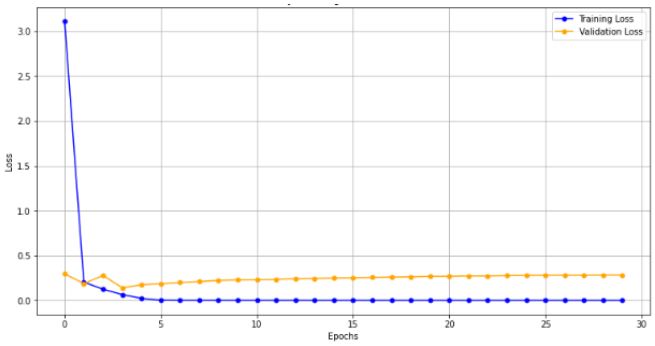

Fig. 7. ViT: Training and Validation Loss

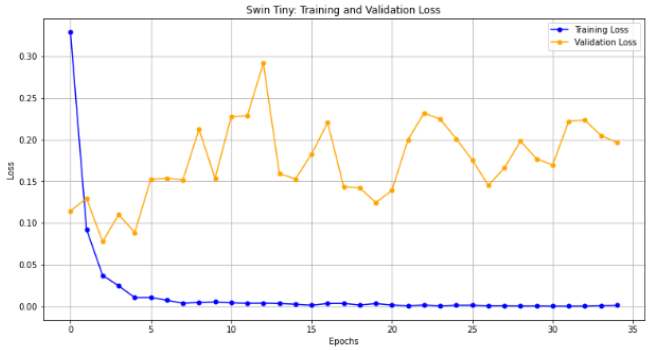

Fig. 8. Swin-Tiny Training and Validation Loss

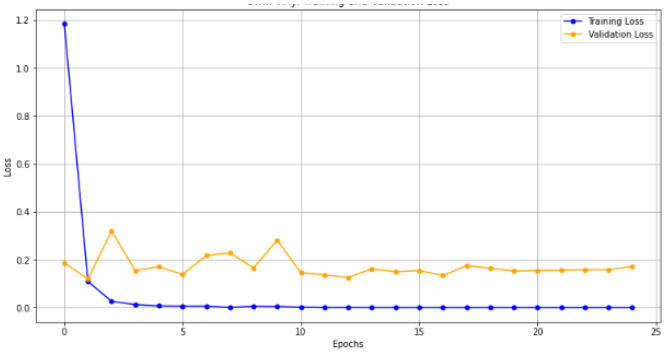

Fig. 9. BEiT: Training and Validation Loss

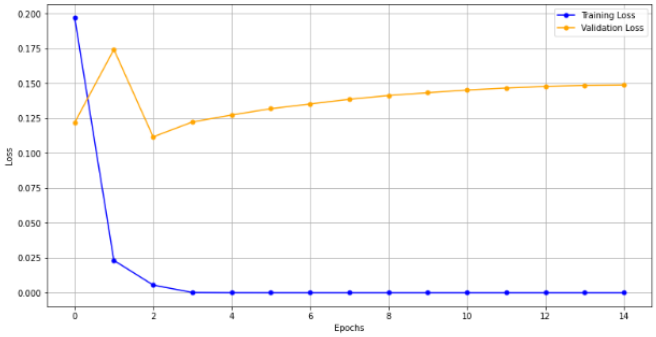

Fig. 10. DEiT: Training and validation loss

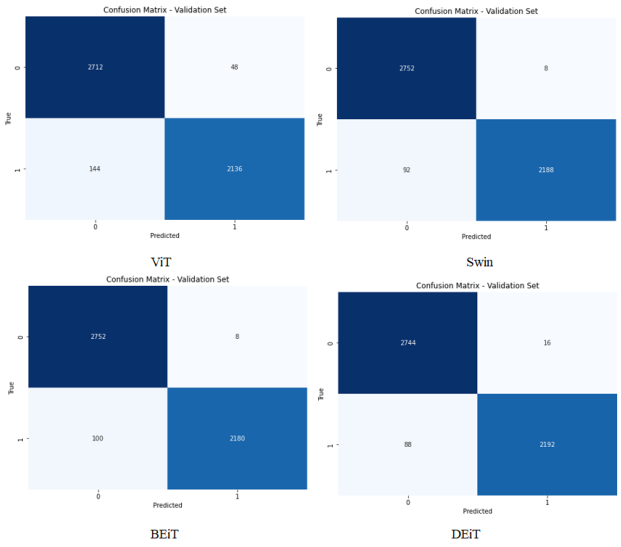

Fig. 11. Confusion Matrices of all four models

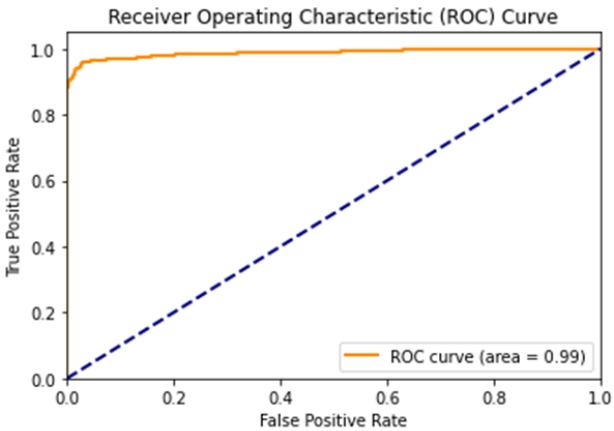

Fig. 12. ROC curve for ViT

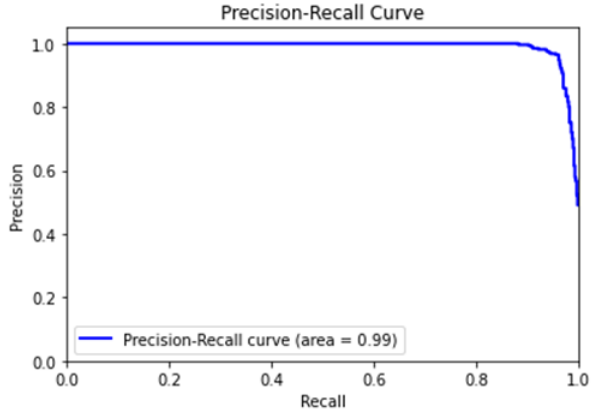

Fig. 13. Precision-Recall Curve for ViT

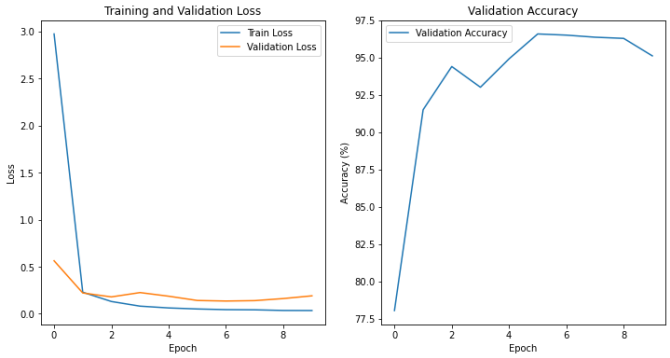

Fig. 14. Knowledge Distilled Model Loss

VII. Future Work

In future work, integrating real-time ECG readings as input into our model presents an intriguing avenue for real-time arrhythmia detection. The utilization of such data could offer a more comprehensive understanding of cardiac health, enabling earlier detection of issues and facilitating timely intervention.

Another future work interest involves comparing the accuracy of predictions across the 12 different leads. These leads provide distinct perspectives on cardiac activity, each offering unique insights into the heart's functioning. By analyzing the performance of our model across these leads, we can identify which ones yield better predictive outcomes. This comparative analysis will enable us to optimize our model by focusing on the leads that contribute most effectively to accurate predictions.

Exploring other student models for knowledge distillation presents an avenue for refining and improving our predictive capabilities. By leveraging insights from a diverse range of student models, we can gain a deeper understanding of effective distillation techniques and refine our approach to model optimization.

VIII. Conclusion

Based on the results presented, the project demonstrates the robust performance of ViTs and knowledge distillation in ECG classification. The comparison between the original paper's results and those achieved in this project illustrates a clear enhancement in precision, recall, F1-score, and accuracy for all models used. Furthermore, the knowledge distillation approach effectively conveyed expertise from a well-trained teacher model to a student model, achieving high accuracies in both teacher and student models. The confusion matrix and validation accuracy graphs underline the effectiveness of the model training and validation processes.

In conclusion, this project not only validates the efficacy of using vision transformers for complex medical image classification tasks but also showcases the potential of knowledge distillation techniques in optimizing model performance. The improvements in diagnostic capabilities presented in this study could significantly enhance automated ECG analysis tools, potentially leading to better patient outcomes through faster and more accurate diagnosis. This research sets a benchmark for future studies in medical imaging and diagnostic automation.

Acknowledgment

A big thank you to Professor Kostic for his guidance throughout the course, as well to Chengbo and Sanjeev for all the helpful assistance they both provided. We're also grateful for our classmates, whose collaboration enriched our learning experience.

Appendix

A. *Individual Student Contributions in Fractions*

	jm5679	dg3370
Last Name	Marshall	Gumaste
Fraction of total contribution	1/2	1/2
What I did 1	Literature Review	Literature Review
What I did 2	Implemented Knowledge Distillation and Data Processing	Fine-tuned pre-trained models on ECG data
What I did 3	Report and Presentation	Report and Presentation

REFERENCES

[1] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2021).

[2] Songhao Piao Hangbo Bao Li Dong and Furu Wei. "BEIT: BERT Pre-Training of Image Transformers". In: *arXiv:2106.08254v2* (2022).

[3] et. al Hugo Touvron Matthieu Cord. "Training data-efficient image transformers distillation through attention". In: *arXiv:2012.12877v2* (2021).

[4] Ze Liu. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *arXiv:2103.14030* (2021).

[5] Yuzhen Qin et al. "MVKT-ECG: Efficient Single-lead ECG Classification on Multi-Label Arrhythmia by Multi-View Knowledge Transferring". In: *arxiv:2310.12630* (2023).

[6] Majid Sepahvand et al. "A Novel Multi-Lead ECG Personal Recognition Based on Signals Functional and Structural Dependencies Using Time-Frequency Representation and Evolutionary Morphological CNN". In: *Biomedical Signal Processing and Control* (2021).

[7] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.

[8] Marta Załeska-Kociecka et al. "Epicardial Fat and Ventricular Arrhythmias". In: *Journal Name* Volume Number.Issue Number (Publication Year), Page Range. DOI: DOI.