ECG-EmotionNet: Nested Mixture of Expert (NMoE) Adaptation of ECG-Foundation Model for Driver Emotion Recognition

Nastaran Mansourian[1], Arash Mohammadi[1,2], M. Omair Ahmad[1], M.N.S. Swamy[1]

***Abstract*— Driver emotion recognition plays a crucial role in driver monitoring systems, enhancing human-autonomy interactions and the trustworthiness of Autonomous Driving (AD). Various physiological and behavioural modalities have been explored for this purpose, with Electrocardiogram (ECG) emerging as a standout choice for real-time emotion monitoring, particularly in dynamic and unpredictable driving conditions. Existing methods, however, often rely on multi-channel ECG signals recorded under static conditions, limiting their applicability in real-world dynamic driving scenarios. To address this limitation, the paper introduces ECG-EmotionNet, a novel architecture designed specifically for emotion recognition in dynamic driving environments. ECG-EmotionNet is constructed by adapting a recently introduced ECG Foundation Model (FM) and uniquely employs single-channel ECG signals, ensuring both robust generalizability and computational efficiency. Unlike conventional adaptation methods such as full fine-tuning, linear probing, or low-rank adaptation, we propose an intuitively pleasing alternative, referred to as the nested Mixture of Experts (MoE) adaptation. More precisely, each transformer layer of the underlying FM is treated as a separate expert, with embeddings extracted from these experts fused using trainable weights within a gating mechanism. This approach enhances the representation of both global and local ECG features, leading to a 6% improvement in accuracy and a 7% increase in the F1 score, all while maintaining computational efficiency. The effectiveness of the proposed ECG-EmotionNet architecture is evaluated using a recently introduced and challenging driver emotion monitoring dataset. The proposed architecture outperforms its counterparts, achieving an average classification accuracy of 82.45% and an F1 score of 77.11% across five emotional states: anger, fear, neutral, sadness, and surprise.**

***Index Terms*— Autonomous Driving, ECG Signals, Emotion Recognition, Foundation Models, Mixture of Experts,**

I. Introduction

Recently, presence of partially/semi-Autonomous Vehicles (AVs) on the roads has increased considerably. Equipped with unparalleled capabilities in perceiving their surroundings, AVs aim to provide safer and more efficient Autonomous Driving (AD). Despite recent advancements in AD, however, building trust in the AI-driven decision-making processes of AVs remains a significant barrier to their widespread adoption. Recognizing driver emotion is one crucial factor to improve trust in AD, as emotional states such as stress, anger, or fatigue can impair decision-making and reaction times, significantly increasing the risk of accidents [1]. By monitoring driver emotions in real-time, Advanced Driver-Assistance Systems (ADAS) can detect hazardous states and provide interventions such as calming alerts or break suggestions [2]. Emotion-aware systems also improve Human-Autonomy Teaming (HAT) interactions in fully/semi AVs, ensuring smoother transitions and personalized driving experiences [3]. While several studies [4]–[6] have explored this domain, the application of recent advancements in Foundational Modelling (FM) [7] for emotion recognition in AD remains in its infancy, particularly in terms of accuracy, efficiency, and generalizability. This paper addresses this gap by proposing a novel approach that achieves comparable accuracy with significantly reduced complexity and improved generalizability to unseen data through the use of transfer learning.

Literature Review: Generally speaking, emotion recognition methods in dynamic driving scenario can be classified based on the input signals into physiological and non-physiological categories. Non-physiological signals, such as facial expressions, often face challenges due to individual differences, lighting conditions, and camera angles, leading to potential inaccuracies [8]. In contrast, physiological signals are involuntarily produced by the nervous and endocrine systems [9], making them less susceptible to external factors and providing a more accurate reflection of emotional states [10]. Various physiological signals, including Eelectroencephalograms (EEGs) [11], Electrodermal activity (EDA) [12], Electrocardiograms (ECGs) [13], and Electromyography (EMG) [14], are commonly utilized to assess human psychological states across diverse contexts, such as driving. Among these, ECGs are particularly advantageous for emotion recognition in dynamic environments due to their robustness against motion artifacts, continuous non-invasive monitoring capabilities, and the computational efficiency afforded by single-channel ECG usage [13].

Recently, there has been a surge of interest in using Machine Learning (ML) and Deep Learning (DL) techniques for ECG-based emotion recognition [15], [16]. Traditional ML approaches rely on manually extracting ECG features, such as Interbeat Interval (IBI), Heart Rate Variability (HRV), and Power Spectral Densities (PSD), which requires domain expertise and is time-intensive [17], [18]. In contrast, DL enables end-to-end emotion recognition from raw ECG signals, removing the need for manual feature engineering [16]. DL approaches, such as Temporal Convolutional Neural Networks (TCNNs) [19], have demonstrated high accuracy in classifying arousal and valence levels from ECG signals. However, several challenges persist. On the one hand, to the

*This work is partially supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.

[2]Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada.

best of our knowledge, most existing ECG-based emotion recognition methods [19]–[21] are designed for static environments, where no secondary tasks are involved. This makes such methods impractical for real-world dynamic driving scenarios. On the other hand is the significant limitation of reliance on large, labelled datasets for training. Acquiring large datasets is, typically, infeasible and time-consuming especially in the domain of AVs. Additionally, models trained in a fully supervised fashion may develop representations that are overly specific to the training data, resulting in limited generalizability to new, unseen data. To address these issues, self-supervised learning frameworks have been proposed [22] in other domains, enabling models to learn robust ECG representations without the need for extensive labelled data. However, the self-supervised learning process often involves multiple sequential training steps, making it both time-intensive and computationally demanding.

Contributions: Foundation Models (FM) have revolutionized Natural Language Processing (NLP) [23], computer vision [24], and speech recognition [25], demonstrating the effectiveness of pretraining on massive datasets. Models such as GPT [23] and CLIP [24] enable fine-tuning for diverse tasks with high accuracy and efficiency. Despite their success in other domains, however, foundation modeling for physiological signal analysis, particularly ECG-based emotion recognition in driving monitoring, remains largely unexplored [26]. To bridge this gap, we leverage a recently introduced ECG-FM [27], a transformer-based FM pre-trained on 2.5 million ECG samples, for driver emotion recognition, and introduce the ECG-EmotionNet architecture.

To preserve the pre-trained model's generalization capability while enhancing computational efficiency, ECG-FM is adapted via an intuitively pleasing approach, referred to as the nested Mixture of Experts (NMoE) adaptation. More specifically, we keep the FM's parameters frozen, but instead of using only the final transformer's embedding as input to a linear layer, we treat each transformer layer of the underlying ECG-FM as an independent expert. In summary, the paper makes the following two main contributions:

- Introduction of the ECG-EmotionNet, a novel adapted ECG foundation model for driver emotion recognition using single-channel ECG signals. To the best of our knowledge, this is the first study to leverage a FM for ECG-based emotion recognition in the AD context.
- Introduction of the NMoE adaptation mechanism. In the NMoE, the feature vector is sequentially processed by experts rather than being fed to all the experts simultaneously. NMoE adaptation improves parameter efficiency, robustness to noise, and hierarchical representation learning, resulting in richer contextual embeddings while maintaining computational efficiency. Additionally, by capturing sequential dependencies, the NMoE offers stronger generalization to unseen data, making it particularly effective for dynamic tasks such as ECG-based driver emotion recognition.

The performance of the proposed ECG-EmotionNet architecture is evaluated via a recently introduced benchmark and challenging dataset for driver emotion monitoring. ECG-EmotionNet outperforms existing methods, attaining an average classification accuracy of 82.45% and an F1 score of 77.11% across five emotional states: anger, fear, neutral, sadness, and surprise.

The remainder of the paper is organized as follows: Section II provides the required material and methods. The proposed ECG-EmotionNet framework is introduced in Section III. Experimental results and comparisons are presented in Section IV. Finally, Section V concludes the paper.

II. Materials and Methods

This section provides an overview of the backbone ECG-FM, the dataset used for training and evaluation, and preprocessing and data augmentation steps.

A. ECG-FM Architecture

The ECG-FM model [27] is a self-supervised, transformer-based foundation model designed for ECG signal analysis. It includes a feature extractor with four convolutional blocks that converts raw ECG signals into latent representations, incorporating relative positional embeddings for temporal awareness. In addition, it includes a transformer encoder, inspired by BERT-Large [28], which processes the extracted representations through self-attention mechanisms within a high-dimensional embedding space.

Pretraining incorporates multiple objectives, including the masking objective from wav2vec 2.0 [25], the Contrastive Multi-segment Coding (CMSC) objective from Contrastive Learning of Cardiac Signals (CLOCS) [29], and Random Lead Masking (RLM) [30], ensuring robust feature learning. Trained on 2.5 million ECG samples, ECG-FM effectively captures both local and global patterns, making it well-suited for downstream tasks such as emotion recognition in real-world settings, including driving scenarios explored in this study. Please refer to [27] for further details on the ECG-FM model's architecture.

B. Dataset

This study utilizes the manD 1.0 dataset [31], a recently released multimodal benchmark dataset for driver monitoring in autonomous driving. The manD 1.0 dataset includes synchronized physiological signals (ECG, EEG, EDA), vehicle dynamics, driver activities, and environmental factors from 50 participants (balanced by gender, and aged between 21–65) in a controlled setting. Participants drove through five scenarios designed to elicit neutral, anger, fear, sadness, and surprise, simulating real-world driving conditions.

For this study, we focused on ECG data to classify emotional states. A visual inspection was conducted to ensure signal quality, leading to the exclusion of specific low-quality signals, i.e., the 4th ECG from anger, the 17th and 23rd from fear, the 7th and 16th from neutral, the 29th from sadness, and the 14th, 22nd, and 30th from surprise.

C. Pre-processing and Data Augmentation

A structured pre-processing pipeline was applied to ensure high-quality ECG data for emotion recognition. First, a

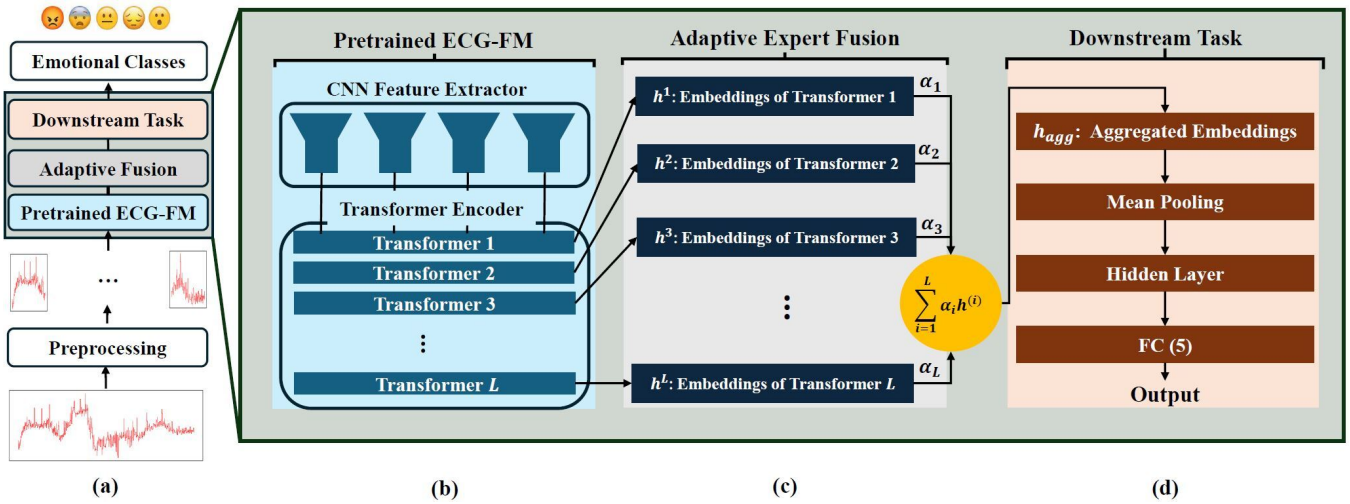

Fig. 1. A graphical representation of the proposed ECG-EmotionNet methodology for emotion recognition in dynamic driving scenarios. (a) The method begins with preprocessing raw ECG signals, extracting their representations using the pretrained ECG-FM model, followed by adaptive fusion layer and downstream emotion classification. (b) The pretrained ECG-FM model, comprising a CNN-based feature extractor and a transformer encoder with 12 layers, processes the signals to generate multi-layer embeddings from each transformer layer. (c) Adaptive Expert Fusion Layer aggregates embeddings from all transformer layers using trainable weights (α_l). Hooks are registered in each transformer encoder layer to dynamically capture intermediate outputs during the forward pass, enabling the integration of global and local features into a unified representation (h_{agg}). (d) Aggregated embeddings undergo pooling, are processed through a hidden layer, and are classified into five emotional categories via a fully connected layer.

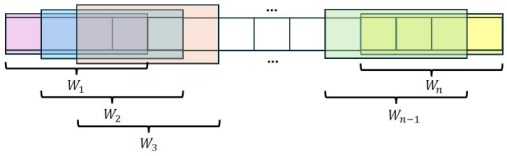

Fig. 2. Illustration of the overlapping window technique for data augmentation, where each window W_i captures a subset of the ECG signal with overlapping regions to generate augmented samples.

second-order Butterworth high-pass filter with a 0.8 Hz cut-off was used to remove baseline wander and low-frequency noise while preserving essential signal components. Subsequently, the signals were standardized using z-score normalization to minimize variability across recordings. Next, 10-second windows sampled at 256 Hz were extracted to preserve both temporal dynamics and spatial relationships crucial for emotion recognition.

Data augmentation is used to enhance the training dataset. For augmentation, one can rely on overlapping (sliding window) or Generative Adversarial Networks (GAN). Given the size of the available dataset, we applied the overlapping-window technique, generating additional samples by leveraging repeated patterns within each trial [32]. This approach expanded the training set by creating multiple, slightly shifted representations of the same data. Consequently, this method augments the dataset by progressively incorporating new temporal segments while partially discarding previous ones. As illustrated in Fig. 2, the overlapping window technique segments the ECG signal of length L into smaller overlapping windows W_i of fixed size N. The stride between consecutive windows determines the overlap percentage, ensuring that each segment captures both unique and shared patterns from the signal through positional changes in the previous signal. Such an approach increases the diversity of the training dataset, providing $n = \left\lfloor \frac{L-N}{\text{stride}} \right\rfloor + 1$ augmented samples while preserving temporal and spatial features critical for emotion recognition.

III. The ECG-EmotionNet

In this section, we present details of the proposed ECG-EmotionNet architecture to adopt the ECG-FM for the task of driver emotion recognition. Reference [27], which introduced the ECG-FM, introduced the following two approaches for its adaptation to downstream tasks: (i) *Full Fine-Tuning,* where the ECG-FM is initialized with the pretrained weights and then all model weights are updated using the dataset associated with the downstream task, and; (ii) *Linear Probing,* where the ECG-FM pretrained weights are frozen, instead the extracted embeddings from the last transformer layer of the pretrained model are provided as inputs to a single linear layer to generate predictions.

We propose to use an alternative approach, we refer to as the Nested Mixture of Experts (NMoE) adaptation inspired by [33]. Intuitively speaking, the idea is to preserve the pretrained model's generalization capability while optimizing its computational efficiency. For this purpose, we retain the FM's parameters frozen (similar to the aforementioned linear probing mechanism), however, instead of feeding only the last transformer's embedding to a linear layer, we treat each of the underlying transformer layers of the ECG-FM as a separate expert. Extracted embedding from these experts are then fused (mixed) using trainable weights. In other words, by introduction of such a multi-model fusion architecture, hidden layer embeddings from all transformers' outputs are leveraged to capture richer contextual representations.

Intuitively speaking, NMoE offers several advantages over its traditional counterparts. On the one hand is its enhanced feature refinement through hierarchical representation learning. More specifically, each expert in the nested structure receives the transformed output from the previous expert. This allows for a progressive abstraction of features, where

early experts focus on low-level patterns, while later experts extract higher-level representations, leading to a deeper contextual understanding of the data. Furthermore, NMoE reduces redundancy and improves parameter efficiency. In conventional MoE setups, experts may redundantly learn overlapping features since they all process the same raw input. By contrast, the nested structure forces specialization among experts, ensuring that each expert contributes uniquely to feature extraction. This diversification in learning improves the expressiveness of representations while maintaining computational efficiency.

We constructed a Nested MoE settings, i.e., the extracted feature vector x is provided as input to the first expert. The output embedding of the first expert $h_i(x)$ is used both as an output and as input to the second expert. This continues in a sequential fashion, where the output of each expert is both an output embedding and the input to the next expert. More specifically,

- Let $\boldsymbol{x}$ß be the input feature vector to the i^{th} expert (transformer layer), for $(1 \leq i \leq L)$.
- Let $h^{(i)}(\boldsymbol{x}$ß$) \in \mathbb{R}^d$ denote the hidden embeddings extracted from the i^{th} transformer layer. Here, d represents the embedding dimension, and L denotes the total number of Transformer layers.
- Let $G(\boldsymbol{x}) = (\alpha_1(\boldsymbol{x}_1), \alpha_2(\boldsymbol{x}_2), \ldots, \alpha_L(\boldsymbol{x}_L)$ be the gating function, where each αß$(\boldsymbol{x}$ß$)$ represents the weight assigned to expert i based on its input.

The final aggregated embedding $h_{agg}(\boldsymbol{x})$ is computed as

$$h_{agg}(\boldsymbol{x}) = \sum_{i=1}^{L} \alpha_i(\boldsymbol{x}_i) h^{(i)}(\boldsymbol{x}\text{ß}) \tag{1}$$

where $\alpha_i(\boldsymbol{x}_i)$ represents the trainable gating function output for expert i to determine its contribution to the final representation, ensuring that $\sum_{i=1}^{L} \alpha_i(\boldsymbol{x}_i) = 1$. The gating function for the i^{th} expert is computed as a softmax layer over a learned function W_h given by

$$\alpha_i(\boldsymbol{x}_i) = \frac{\exp(W_g^{(l)} \boldsymbol{x}\text{ß})}{\sum_{j=1}^{L} \exp(W_g^{(j)} \boldsymbol{x}\text{ß})} \tag{2}$$

where $W_g \in \mathbb{R}^{L \times d}$ is the trainable gating weight matrix.

The transformer-based contextualized encoder extracts global temporal patterns across the entire input sequence due to its global receptive field, whereas the local encoder, constrained by a limited receptive field, focuses on detailed localized features. By integrating embeddings from multiple transformer layers, our method balances local and global feature extraction, enhancing emotion recognition from ECG signals while minimizing overfitting and improving efficiency. By freezing the pretrained model parameters and employing a trainable weighted averaging strategy, we optimize feature selection for emotion recognition while preserving the pretrained model's knowledge. This domain-adaptive approach effectively refines both global and local features without altering the encoder's expressive capacity. Furthermore, this method significantly reduces computational complexity, leading to faster convergence and more efficient training.

To summarize, Fig. 1 illustrates the proposed ECG-EmotionNet architecture, which consists of a pretrained transformer backbone followed by a trainable weighted averaging mechanism that aggregates hidden outputs from all transformer layers. The resulting weighted embeddings undergo average pooling, reducing the temporal dimension to generate a fixed-size feature vector. This vector is then processed through a dense layer (128 units, ReLU activation), batch normalization, and dropout (0.3 probability) to improve generalization and mitigate overfitting. Finally, a fully connected layer maps the refined features to five emotion classes.

IV. Result and Discussion

In this section, we evaluate performance of the proposed ECG-EmotionNet architecture through a comprehensive set of experiments. For evaluation purposes, the dataset described in Section II-A was divided into 80% for training and 20% for testing, with 5-fold cross-validation. Models were trained for 10 epochs using a batch size of 32, the CrossEntropy objective function, and the Adam optimizer with a learning rate of 10^{-3}.

A. Full vs. Partial vs. NMoE-based Fine-Tuning

Table I summarizes the results of the proposed algorithm for emotion recognition using single-channel ECG under different fine-tuning strategies and overlapping percentages. For the augmented dataset (75% overlapping), both CNN fine-tuning and the NMoE-based model achieved superior performance while requiring significantly fewer trainable parameters compared to full or encoder fine-tuning. However, as the overlapping percentage decreases and data availability becomes more constrained, the NMoE-based model consistently outperforms CNN fine-tuning and other strategies. Notably, the performance gap is significantly larger in low-data scenarios than in augmented settings, further highlighting the NMoE model's effectiveness in dynamic environments.

B. Robustness and Efficiency Evaluation

To evaluate robustness, which is critical factor in driving scenarios, Gaussian noise with varying SNR levels was added to the data. All four fine-tuning strategies were tested under these conditions. As shown in Fig. 3, the NMoE-based approach consistently outperforms other strategies in terms of accuracy and F1 score, even in noisy environments.

To evaluate model efficiency, we compared the number of trainable parameters. Full fine-tuning, encoder fine-tuning, and CNN fine-tuning involve approximately 312 million, 302 million, and 1.6 million parameters, respectively. In contrast, ECG-EmotionNet architecture trains fewer than $200,000$ parameters per epoch, significantly improving computational efficiency without compromising performance. In summary, the advantages of NMoE-based model over fine-tuning it are: *(i) Superior Performance:* It consistently outperforms other fine-tuning approaches, especially in limited-data scenarios; ***(ii) Reduced Parameters:*** It requires significantly fewer trainable parameters compared to other strategies, and; ***(ii)***

TABLE I

RESULTS FOR DRIVER EMOTION CLASSIFICATION USING THE ALL TRANSFORMER HIDDEN STATES OF THE PRETRAINED ECG-FM. THIS TABLE COMPARES DIFFERENT FINE-TUNING STRATEGIES AND THE OVERLAPPING DATA AUGMENTATION TECHNIQUE, INCLUDING THE NUMBER OF PARAMETERS INVOLVED IN EACH STRATEGY.

Tuning Strategy	**Overlap**	**Validation**		**Test**		**Number of Parameters**
		Accuracy (%)	**F1 (%)**	**Accuracy (%)**	**F1 (%)**	
Full Fine-tuning	0%	73.03 ± 2.71	68.98 ± 2.91	73.80 ± 1.84	69.02 ± 1.86	≈ 312 million
	25%	76.60 ± 2.51	72.04 ± 2.90	74.26 ± 0.73	68.09 ± 0.83	
	50%	77.52 ± 1.33	72.71 ± 1.62	76.88 ± 0.75	71.67 ± 0.87	
	75%	78.24 ± 1.33	74.00 ± 1.87	79.45 ± 0.42	76.15 ± 0.53	
CNN Fine-tuning	0%	78.13 ± 0.93	72.98 ± 1.16	76.15 ± 1.22	70.96 ± 1.33	≈ 1.6 million
	25%	79.08 ± 1.79	73.72 ± 2.39	76.80 ± 1.12	70.86 ± 1.16	
	50%	80.08 ± 1.82	75.63 ± 2.78	79.03 ± 0.82	73.03 ± 0.92	
	75%	82.13 ± 1.59	77.62 ± 2.75	82.16 ± 0.73	77.72 ± 0.80	
Encoder Fine-tuning	0%	73.28 ± 1.38	68.94 ± 1.91	72.87 ± 0.80	68.06 ± 0.87	≈ 302 million
	25%	76.36 ± 1.90	71.74 ± 2.69	73.62 ± 0.50	66.98 ± 0.72	
	50%	77.35 ± 2.27	72.73 ± 2.63	76.15 ± 1.2	71.12 ± 1.16	
	75%	77.15 ± 1.06	72.47 ± 1.55	78.78 ± 0.72	74.40 ± 0.54	
The NMoE-based ECG-EmotionNet	0%	79.38 ± 2.93	75.12 ± 2.92	78.64 ± 0.80	75.06 ± 0.77	**< 200,000**
	25%	79.84 ± 1.80	75.70 ± 3.11	78.17 ± 0.31	70.85 ± 0.59	
	50%	81.26 ± 1.53	76.70 ± 2.61	80.98 ± 0.38	75.02 ± 0.31	
	75%	**80.88 ± 1.32**	**77.11 ± 1.93**	**82.45 ± 0.45**	**77.20 ± 0.44**	

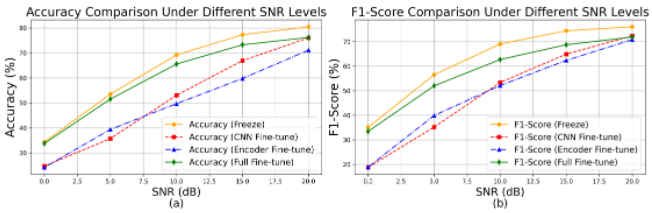

Fig. 3. Comparison of Accuracy: (a) and F1-Score. (b) under different additional noise levels for four strategies: NMoE fine-tuning, CNN fine-tuning, Encoder fine-tuning, and full fine-tuning.

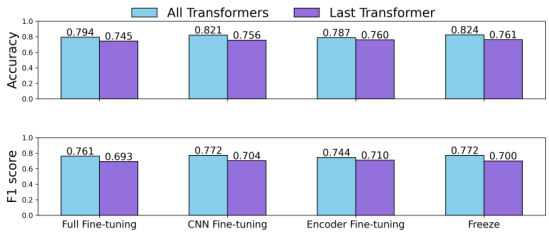

Fig. 4. Comparison of Accuracy and F1 Score for Models Using All Transformers Embeddings vs. Only the Last Transformer Embeddings.

Robustness: It performs better in the presence of noise, making it suitable for real-world driving scenarios.

C. NMoE vs. Final Transformer Layer

We evaluated model performance using the final transformer layer outputs against the NMoE that uses embeddings from all transformer layers. As shown in Fig. 4, leveraging all transformer embeddings consistently improved the accuracy across all fine-tuning strategies. Analyzing the learned weights of $\alpha_i(\boldsymbol{x}ß)$ (Fig. 5) revealed that middle layers (e.g., layers $6 - 8$) contribute most significantly to ECG-based emotion recognition, while the local encoder (layer 0) and deeper layers (layers $10 - 12$) are less impactful. This underscores the middle layers' role in balancing local signal features and global contextual information.

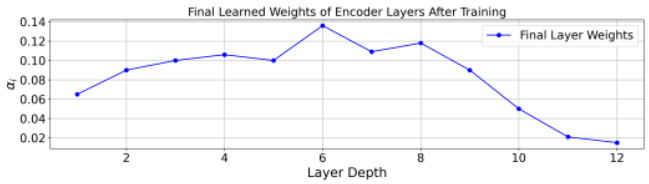

Fig. 5. Final learned weights of α_i of the encoder layers after training, illustrating the importance of each layer in the model.

TABLE II

COMPARISON OF THE DRIVER EMOTION RECOGNITION APPROACHES.

Methodology	**Accuracy**	**F1 score**
TCNN [19]	41.81%	29.15%
Self-Supervised Learning [22]	58.21%	33.64%
ECG-EmotionNet (ours)	**82.45%**	**77.20%**

D. Comparison with Existing Methods

To compare the proposed ECG-EmotionNet architecture with state-of-the-art, we have implemented the TCNN [19] and the self-supervised learning framework of Reference [22] on our dataset for driver emotion recognition. These are proposed for ECG-based emotion recognition in static environments. Although these methods perform well in static environments, as shown in Table II, our proposed architecture significantly outperforms them in scenarios involving secondary tasks such as driving.

Moreover, the proposed model's use of single-channel ECG for five-class emotion classification offers significant computational efficiency. This is while comparable performance is achieved compared to existing methods proposed for driver emotion recognition relying on alternative modalities, such as multi-channel EEG, or facial expressions. For example, Chen *et al.* achieved 75.26% accuracy in a three-class task using 32-channel EEG signals, with a maximum F1 score of 76% [4]. Similarly, Gursesli *et al.* utilized image-based facial expressions from multiple datasets, including FER-2013, RAF-DB, and AffectNet, achieving an average

accuracy of 67% across seven emotion classes [5]. Additionally, Xiang *et al.* [6] explored multi-modal data fusion for driver emotion recognition Their findings showed that Blood Volume Pulse (BVP) signals alone achieved 77.01% accuracy and an F1-score of 76.43%, surpassing facial video, which achieved 72.56% accuracy and a 71.76% F1-score.

V. Conclusion and Future Works

In this paper, we introduced ECG-EmotionNet, a novel deep learning framework tailored for driver emotion recognition using single-channel ECG signals. Unlike conventional approaches that rely on multi-channel ECG data recorded in static environments, our proposed model adapts a pretrained-fondation model to dynamic driving conditions, ensuring improved robustness, generalizability and computational efficiency. Through the NMoE adaptation, we effectively utilize all transformer layers as independent experts, enhancing both global and local feature representations. The experimental results demonstrated that ECG-EmotionNet achieves an average classification accuracy of 82.45% and an F1 score of 77.11%, outperforming state-of-the-art approaches while maintaining a significantly lower computational cost. These findings suggest that ECG-EmotionNet can serve as a practical solution for ADAS and AD applications, contributing to enhanced driver monitoring and human-autonomy interaction. Despite its promising results, ECG-EmotionNet presents several opportunities for future improvements. Since variations in ECG signals across individuals and driving conditions may impact recognition performance, real-time evaluations can provide deeper insights into its adaptability. Moreover, while ECG-EmotionNet effectively captures physiological signals, integrating multimodal emotion recognition, combining ECG with EDA, EEG, and/or facial expressions, can lead to an improved emotion detection model.

References

[1] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–30, 2020.

[2] V. Tavakkoli, K. Mohsenzadegan, and K. Kyamakya, "Leveraging Context-Aware Emotion and Fatigue Recognition Through Large Language Models for Enhanced Advanced Driver Assistance Systems (ADAS)," in *Recent Advances in Machine Learning Techniques and Sensor Applications for Human Emotion, Activity Recognition and Support*. Springer, 2024, pp. 49–85.

[3] W. Li, G. Li, R. Tan, C. Wang, Z. Sun, Y. Li, G. Guo, D. Cao, and K. Li, "Review and Perspectives on Human Emotion for Connected Automated Vehicles," *Automotive Innovation*, vol. 7, no. 1, 2024.

[4] J. Chen, X. Lin, W. Ma, Y. Wang, and W. Tang, "EEG-based emotion recognition for road accidents in a simulated driving environment," *Biomedical Signal Processing and Control*, vol. 87, p. 105411, 2024.

[5] M. C. Gursesli, *et al.*, "Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets," *IEEE Access*, 2024.

[6] G. Xiang, *et al.*, "A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals," *Eng. Applications of Artificial Intelligence*, vol. 130, p. 107772, 2024.

[7] H. Gao, Z. Wang, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," *arXiv preprint arXiv:2402.01105*, 2024.

[8] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Information*, vol. 15, no. 3, p. 135, 2024.

[9] L. Shu, *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[10] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer methods and programs in biomedicine*, vol. 140, pp. 93–110, 2017.

[11] J. Chen, X. Lin, W. Ma, Y. Wang, and W. Tang, "EEG-based emotion recognition for road accidents in a simulated driving environment," *Biomedical Signal Processing and Control*, vol. 87, p. 105411, 2024.

[12] Y. R. Veeranki, *et al.*, "Non-Linear Signal Processing Methods for Automatic Emotion Recognition using Electrodermal Activity," *IEEE Sensors Journal*, 2024.

[13] A. Abdou and S. Krishnan, "Horizons in single-lead ECG analysis from devices to data," *Frontiers in Signal Processing*, vol. 2, p. 866047, 2022.

[14] V. K. Barigala, *et al.*, "Evaluating the effectiveness of machine learning in identifying the optimal facial electromyography location for emotion detection," *Biomedical Signal Processing and Control*, vol. 100, p. 107012, 2025.

[15] M. Baghizadeh, K. Maghooli, F. Farokhi, and N. J. Dabanloo, "A new emotion detection algorithm using extracted features of the different time-series generated from ST intervals poincaré map," *Biomedical Signal Processing and Control*, vol. 59, p. 101902, 2020.

[16] S. Nita, S. Bitam, M. Heidet, and A. Mellouk, "A new data augmentation convolutional neural network for human emotion recognition based on ECG signals," *Biomedical Signal Processing and Control*, vol. 75, p. 103580, 2022.

[17] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[18] S. Qiu, *et al.*, "A novel two-level interactive action recognition model based on inertial data fusion," *Information Sciences*, vol. 633, 2023.

[19] T. C. Sweeney-Fanelli and M. H. Imtiaz, "Ecg-based automated emotion recognition using temporal convolution neural networks," *IEEE Sensors Journal*, 2024.

[20] J. Chen, *et al.*, "Graph enhanced low-resource ECG representation learning for emotion recognition based on wearable internet of things," *IEEE Internet of Things Journal*, 2024.

[21] T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, and N. Jiang, "A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition," *Computers in Biology and Medicine*, vol. 159, p. 106938, 2023.

[22] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541–1554, 2020.

[23] T. Brown, *et al.*, "Language models are few-shot learners," *Advances in neural inf. processing systems*, vol. 33, pp. 1877–1901, 2020.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] F. Del Pup and M. Atzori, "Applications of self-supervised learning to biomedical signals: A survey," *IEEE Access*, 2023.

[27] K. McKeen, L. Oliva, S. Masood, A. Toma, B. Rubin, and B. Wang, "Ecg-fm: An open electrocardiogram foundation model," *arXiv preprint arXiv:2408.05178*, 2024.

[28] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5606–5615.

[30] J. Oh, *et al.*, "Lead-agnostic self-supervised learning for local and global representations of electrocardiogram," in *Conference on Health, Inference, and Learning*. PMLR, 2022, pp. 338–353.

[31] K. Dargahi Nobari and T. Bertram, "A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation," *Scientific data*, vol. 11, no. 1, p. 327, 2024.

[32] I. Majidov and T. Whangbo, "Efficient classification of motor imagery electroencephalography signals using deep learning methods," *Sensors*, vol. 19, no. 7, p. 1736, 2019.

[33] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv:2104.03502*, 2021.