Cross-Modal Distillation by Additive Importance Measure In HITL Autonomous Driving

Saira Bano, Pietro Cassarà, Claudio Gennaro, Alberto Gotta

***Abstract*—With the advent of Advanced Driver Assistance Systems (ADAS) and intelligent transport system applications, recognizing driver emotions has become essential for a decision support system (DSS) with humans in the loop (HITL). Multimodal approaches using visual cues, speech, physiological signals, and driving patterns improve emotion recognition but are challenging in resource-constrained environments where only a subset of modalities is available. This work addresses these challenges by combining multi-modal benefits with single-modality inference for emotion recognition using unlabeled external road condition data. Unlike traditional methods that average teachers' contribution, the proposed cross-modal distillation (CMD) weights teachers thanks to the Shapley additive global explanation (SAGE) aid, which improves the student model's accuracy and provides an interpretation of it. Experimental evaluations of the PPB-Emo dataset show that XA-CMD improves emotion recognition accuracy with other baselines and provides deeper insights into decision-making.**

***Index Terms*—Emotion Recognition, Cross-Modal Distillation, Feature Explanation, Hitl, Autonomous Driving**

I. Introduction

With advances in artificial intelligence for intelligent transport systems and intelligent vehicle applications, improving driver comfort and experience [1] is gaining momentum. Recognizing driver emotions is crucial for adapting automated systems based on attention levels, as emotions significantly influence driving behavior and road safety [2]. Negative emotions can lead to rash decisions that increase the risk of accidents [3]. Real-time recognition systems that analyze factors such as the driver's psychophysiological state, facial expressions, and the external environment are essential for safer roads. These systems combine contact-based methods (e.g. wearable devices) and non-contact techniques (e.g. facial expressions, speech, and driving behavior) and often use multimodal ensemble techniques to increase accuracy [4]. Multimodal approaches offer a more robust solution. However, their high computational costs pose a challenge for practical use in resource-constrained environments such as vehicles [5].

To overcome these challenges, Knowledge Distillation (KD) techniques offer a promising alternative by transferring knowledge from computationally intensive ensembles to lightweight student models. Cross-modal distillation (CMD) extends this concept by using teacher models trained on different modalities (e.g., visual, physiological, or behavioral data) to train a student model that operates on a single modality. CMD techniques have been successfully applied to tasks such as human activity recognition [6] and emotion recognition [7]. Furthermore, ensemble distillation methods aggregate predictions of multiple teachers, improving the accuracy and robustness of the student model [8], demonstrating their potential in distributed and resource-constrained environments.

In this work, we present a method for driver emotion recognition that, through vehicular mobile EDGE technologies, can be used to transfer knowledge from a distributed multimodal ensemble information into a student model located in a vehicle. We focus on how the surrounding environment, such as accidents, aggressive driving of others, or scenic views, influence drivers' emotional states. The main concept is to infer the driver status conditioned by external road factors instead of relying only on internal vehicle data. In this regard, the proposed method uses an ensemble of teacher models, each specialized in a different modality (e.g., facial expressions, physiological signals, and behavioral data of the car), to distill knowledge into a single student model that enables emotion recognition from unlabeled external road videos.

Effective knowledge transfer in ensemble-based models requires appropriate weighting of teacher models, as treating all teachers equally can overlook individual strengths and allow weak teachers to degrade performance. Sophisticated weighting schemes, such as constant weights [9] or adaptive methods exploring gradient diversity [10], improve accuracy but often treat all data samples equally, ignoring sample-specific relevance. Additionally, current ensemble-based distillation methods lack interpretability, functioning as black-box models with limited information about which teacher features contribute to decision-making.

To this aim, SAGE (Shapley Additive Global Explanation) is an eXplainable-AI (XAI) method for interpreting machine learning models by quantifying feature importance. SAGE is an algorithm based on the Shapley value from game theory that has recently been used to interpret black-box machine learning models. It is designed to measure the importance of each feature by evaluating its contribution to predictive performance while accounting for complex feature interactions.

In this work, we introduce a XAI-Aided CMD (XA-CMD) approach that incorporates SAGE into the distillation process to overcome the above challenges. Exploiting the capability of SAGE to calculate the global importance of features and assigning predictive performance to individual input variables, we want to weight teacher models based on their overall contributions in the distillation process. In contrast to previous approaches, our method emphasizes the contributions of individual features of each teacher model, providing deeper insights into the decision-making of the ensemble. By using the PPB-Emo dataset, we show that XA-CMD improves both the accuracy and explainability of emotion recognition tasks and provides valuable insights into the role of individual teacher models.

The remainder of the paper is structured as follows. The

S. Bano, P. Cassarà, C. Gennaro, and A. Gotta, are with the Institute of Information Science and Technologies (ISTI), National Research Council, Pisa, Italy - e-mails: name.surname@isti.cnr.it.

proposed method is presented in Section II, while the experimental results are discussed in Section III. Finally, Section IV concludes our paper.

II. The Proposed Method

Distilling knowledge from multi-modal teacher models can improve student performance, but weak teachers can negatively impact the process. To optimize multi-modal ensemble distillation, we proposed the XA-CMD to weight the teacher models based on their feature contributions, which we compute using the SAGE algorithm [11] to measure feature importance. Then, we use this information to ensure that the most valuable teacher models are prioritized during the distillation, maximizing the accuracy of the ensemble and improving the student's performance.

XA-CMD approach is presented in Algorithm 1.This begins by training each teacher model independently on its respective modality. Each teacher m uses convolutional layers for extracting the features $\{x_{g_m}\} \mid |\{x_{g_m}\}| = F_m$, which are then passed through a classifier with parameters $\{\theta_m\}$ to predict the correct class. The extracted features are analyzed using the SAGE algorithm to quantify the importance of each feature. These importance scores dynamically determine the weight assigned to each teacher model in the ensemble and reflect their contribution to the predictions. Once the teachers' training phase is complete, each teacher can generate its own prediction on a validation set. These predictions are aggregated using weights calculated according to the contribution of each teacher model's features to the true class label. The resulting unified output serves as a distillation target for training the student model.

In the following, we discuss the key concepts behind the generation of the weights calculated as shown in Algorithm 1.

1) SAGE Value Computation (Feature Importance): The important step in XA-CMD is to evaluate the relevance of each feature for a given teacher m with model prediction $f(\theta_m, x_{g_m})$ in determining a specific class y_k (i.e., emotion) using the SAGE algorithm. As shown by the authors in [11], SAGE values, when the loss function is cross entropy or mean squared error, represent the feature importance in terms of the mutual information between the class labels y_k and the feature x_{g_m} given that the subset of features $S \subseteq F_m$ is already selected, as shown in Equation 1:

$$\phi_{g_m}(k) = \sum_{S \subseteq F_m \setminus x_{g_m}} |F_m|^{-1} I(y_k; x_{g_m} | S). \tag{1}$$

These values represent the contribution of each feature to the teacher model's prediction and reflect the impact of each feature in determining the class label. Furthermore, the SAGE algorithm verifies the additive predictive power, i.e., the contribution of each feature in reducing the risk over the mean prediction of the class label can be evaluated considering one feature by time. So, Equation 1 provides a measure of the reduction in uncertainty for describing the class labels through the model $f(\theta_m, x_{g_m})$ adopting x_{g_m} given the subset of features S .

2) Class-Specific Weighting Strategy: Once the feature importance values have been calculated using SAGE, the next step is calculating a scalar weight for each teacher model and class. The scalar weight, $W_{m,k}$, for each teacher m and each class k is calculated by summing the feature importance values across all features in the embedding, as the following equation shows $W_{m,k} = \sum_{\{x_{g_m}\}} \phi_{g_m}(k)$. According to the Equation (1) and the properties of the conditional mutual information, the weight can be written as: $W_{m,k} = \frac{1}{F_m} \sum_{\{x_{g_m}\}} \sum_{S \subseteq F_m \setminus x_{g_m}} \mathbb{E}_S[I(y_k; x_{g_m})]$.

According to the previous equation, $W_{m,k}$ represents the average contribution of the teacher model m in terms of mutual information to the classification of the class k. The weight assigned to each teacher model is proportional to the relative importance of its features, ensuring that models with features that contribute significantly to the predicted class have a more significant influence on the ensemble's final output. This approach allows the ensemble to prioritize more effective teachers while maintaining a balanced and interpretable contribution framework.

After aggregation, we normalize the aggregated weights using the softmax function:

$$\hat{W}_{m,k} = \frac{e^{\mathrm{W}_{m,k}}}{\sum_k e^{\mathrm{W}_{m,k}}}, \forall\, m, k$$

3) Distillation with SAGE-Weighted Predictions: We utilize the normalized weights to combine teachers' models and create a weighted ensemble target. Under the supervision of this weighted model, the student is trained on potentially unlabeled data. Indeed, as explained by the authors in [7], the weighted model is then used in inference mode to generate predictions on the local data of the student, which are used as soft labels for a backpropagation procedure used to update the weights of the student's neural network during the distillation procedure. Assuming that the vehicles have a resource-constrained device that performs computation and communication tasks, a splitting method of the teacher model can be implemented.

Algorithm 1: Class-Specific Weight Computation using SAGE

Input: Labels $\{y_k\}$, Features sets $\{x_{g_1}\} \cdots \{x_{g_m}\}$, Teacher Classifiers $f(\theta_1, x_{g_1}) \cdots (f(\theta_m, x_{g_m})$
Output: Class-specific weights $\hat{W}\ \forall\ k, m$
for m **do**
  for k **do**
    Compute Global Feature Importances:

$$\phi_{g_m}(k) \leftarrow \text{SAGE}(f(\theta_m, x_{g_m}), y_k)$$

    Compute Scalar Mean Across Feature Importances:

$$W_{m,k} = \sum_{\{x_{g_m}\}} \phi_{g_m}(k)$$

  end for
end for
Normalize Weights across models:

$$\hat{W}_{m,k} = \frac{e^{W_{m,k}}}{\sum_m e^{W_{m,k}}}, \quad \forall m, k.$$

Return: $\hat{W}$

III. Results

A. Dataset

We conducted multi-label emotion classification experiments using the PPB-Emo dataset [12], the first publicly available multi-modal dataset for emotion classification while driving. It includes physiological signals, facial video data, and the driving behavior of 40 participants in 240 driving scenarios. The participants performed driving tasks after watching emotion-eliciting videos simulating real road situations. The dataset contains seven emotions: anger, happiness, surprise, disgust, neutral, fear, and sadness. As far as we know, this is the only publicly available dataset focusing on emotions in driving tasks and containing stimulating road videos.

B. Data Preprocessing

Before training the teacher models for their respective modalities, we preprocessed the data as follows. For the facial teacher, we used the videos of the PPB emo dataset sampled at 30 fps and processed them with OpenFace [13] to extract temporal facial features. The behavior teacher model was trained using 11 parameters from the dataset that capture driving dynamics, such as acceleration, pedal force, and speeds. To increase robustness, we also derived additional features, such as the magnitude of acceleration, rate of speed change, braking effect, cumulative speed, the interaction between acceleration and speed, and sudden stops. Finally, we extracted trend curves from 32 EEG channels for the physiological teacher to represent different emotional states. These features were then used to train the corresponding teacher models for each modality. After feature extraction, common pre-processing steps include standardizing features to a common scale, segmenting the data into 2-second non-overlapping windows, and splitting the dataset into 80% for training and 20% for validation.

C. Ensemble Training

The ensemble consists of three specialized teacher models, i.e., $m = 1, 2, 3$, each designed for a specific modality: EEG signals, facial features, and driving behavior data. The features $\{x_{g_m}\}$ for training the teacher model $f(\theta_m, x_{g_m})$ are obtained by a feature extractor. This is a one-dimensional CNN architecture optimized to capture modality-specific features, while θ_m helps to map these features to the corresponding emotion classes. To train the ensemble, the dataset $X \in \mathbb{R}^{n \times c \times t}$ and labels $Y \in \{1, 2, \ldots, k\}^n$ were used. The CNN architecture of each feature extractor was optimized using Pytorch Optuna with a learning rate of 0.0001, the Adam optimizer, a batch size of 64, and early stopping to prevent overfitting. Each teacher classifier was trained to predict the correct class y_k, and the best-performing models were selected based on cross-entropy validation loss. For example, the physiological model achieved a validation accuracy of 91.95% with a loss of 0.1897. The results for all teacher models are summarized in Table II. These specialized teachers were then distilled into a unified student model to leverage their combined predictive capabilities.

D. Student Training

We used three approaches to train the student model on external road videos for emotion recognition: supervised learning as a baseline, an ensemble-based CMD technique, and our proposed XA-CMD method with SAGE-based feature weighting. For the student models video frames were extracted at 30 fps and resized to 224x224 pixels. To account for the limited size of the dataset, data augmentation techniques such as rotation, flipping, scaling, and brightness adjustments were applied. The architecture of the student model is shown in Table I. Training was performed with a learning rate of 0.0001 using the Adam optimizer. Early stopping with a patience of 30 epochs was used to prevent overfitting.

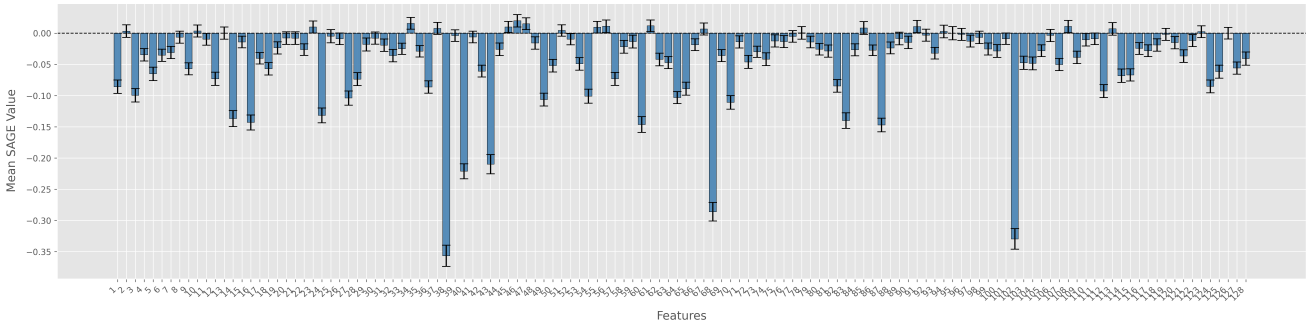
Table I: The Student Model Architecture.

Layer	No. of Neurons (filters)	Kernel Size	Data Size
conv1	32	3x3	224x224
pool1	-	2x2	112x112
conv2	64	3x3	112x112
pool2	-	2x2	56x56
conv3	128	3x3	56x56
pool3	-	2x2	28x28
FC1	512	-	512

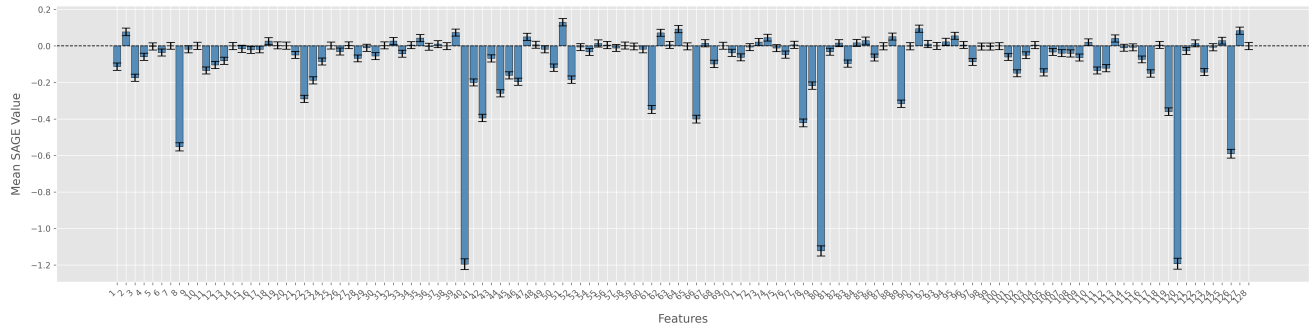
1) Supervised Learning: We trained the student model using the stimulus video labels in a supervised manner and set this as the baseline for comparison with our proposed XA-CMD method. Despite using data augmentation, the baseline model achieved an accuracy of 59.61% and a loss of 2.15. These results suggest that overfitting still occurred, likely due to the limited diversity of training examples and the complexity of the model relative to the available data.

2) Ensemble-based CMD: In this approach, each teacher model generates its own predictions, which are then aggregated into a single set of teacher output logits. All teacher models are uniformly weighted into the ensemble without favoring one model over the others. The student model is trained to match this aggregated prediction and learn from the combined knowledge of all teachers. This resulted in an accuracy of 55.14% and a slight reduction in loss to 1.45, helping to reduce overfitting through the guidance provided by the teacher models.

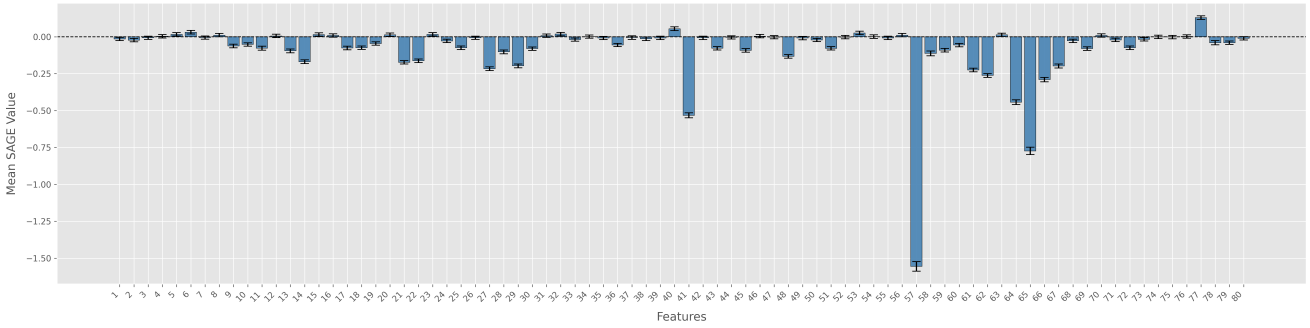
3) XAI-Aided-CMD: In XA-CMD, the student model learns from the teacher ensemble by distilling knowledge based on feature importance values computed by the SAGE algorithm. For each class, SAGE identifies the most influential features of the teacher models, which are then used to dynamically weight their contributions during the distillation process. The weights of the teacher models are determined according to the algorithm 1. These weights were used to adjust the predictions of the teacher models for each batch during the training of the student model. The distillation loss was calculated using the KL divergence between the student model's predicted log-likelihoods and the teacher models' weighted ensemble output. The results presented in Table II confirm the effectiveness of the XA-CMD method and show that it almost reaches the baseline accuracy with reduction in loss as shown in Figure 2. Despite the strong performance of the teacher models in their respective modalities, the transfer of knowledge to the student model remains a challenge due to the limited number of stimulus videos available for training. It is worth noting that all experiments were conducted with both overlapping and non-overlapping datasets. In all cases, the non-overlapping



(a) Physiological Teacher showing the Mean SAGE values for class label 0 (anger)



(b) Behavioral Teacher showing the Mean SAGE values for class label 0 (anger)



(c) Facial Teacher showing the Mean SAGE values for class label 0 (anger)

Figure 1: Feature Importance using SAGE algorithm for each modality specific teacher

dataset performed better than the overlapping one, as shown in Table II.

Figures 1a, 1b and 1c show the SAGE values for the three teacher models for a class label corresponding to the emotion "anger". The SAGE algorithm evaluates features' importance by calculating their contributions' mean and standard deviation. This provides valuable insights into how the features influence the performance of the model. Features with positive SAGE values contribute to improved prediction accuracy, while features with negative values can reduce accuracy and may distort predictions. Additionally, Figure 3 quantifies the contributions of the teacher models for each class label as determined by the SAGE algorithm. For the class label "anger", the behavioral model shows a more significant contribution as most of its features have positive Mean SAGE values, as shown in Figure 1b. This indicates that these features generally improve the prediction accuracy. In contrast, the EEG and facial models have primarily negative Mean SAGE values, suggesting that only a small subset of their features positively affects predictions, while the rest may degrade performance. This dynamic adjustment of teacher model weights based on SAGE-derived feature contributions ensures that the ensemble method favors the most effective predictors. By fine-tuning the weights of the teacher models during training, the XA-CMD method not only optimizes the transfer of knowledge but also improves the explainability of the models. It provides a clear understanding of the nature and extent of the feature contributions of the individual teacher models and thus aligns model performance with the most effective predictors.

Table II: Validation Metrics for Different Teacher models and the student model

Model	**50% Overlap**		**No Overlap**	
	Val. Loss	**Val. Accuracy**	**Val. Loss**	**Val.Accuracy**
EEG	0.1992	0.9484	0.1897	0.9195
Behavioural	0.2344	0.9294	1.6347	0.6162
Facial	0.0944	0.9795	0.4503	0.9188
Baseline	6.1845	0.2669	2.1515	0.5961
CMD	3.6058	0.2641	1.4580	0.5514
XA-CMD	4.3588	0.2719	1.6210	0.5931

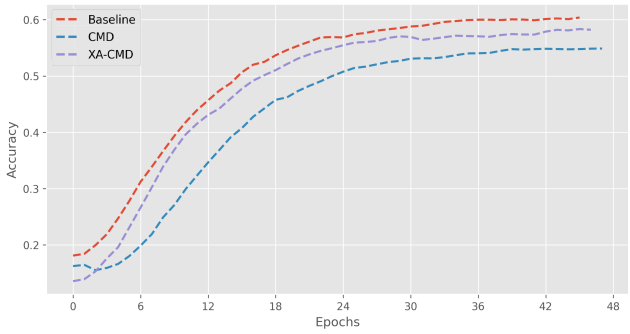


Figure 2: Validation accuracy of student with labels (baseline), ensemble-based CMD and XA-CMD method

IV. Conclusions

This paper proposes a novel approach to dynamically distill multi-modal knowledge from an ensemble of teacher models while providing feature-level explanations, especially for a driver emotion recognition task. The effectiveness of the proposed method is evaluated on the PPB-Emo dataset and compared with baseline models as well as with traditional ensemble-based CMD approaches. The experimental results show that the proposed approach using unlabeled student data performs comparably to the baseline models when labeling information is available. Furthermore, it provides valuable insights into the decision-making processes of the individual teacher models and thus improves the explainability of the system.

In the future, we would like to extend this work by transferring knowledge from HITL approaches, such as facial expressions and physiological signals, to vehicle-centered models that focus on driving behavior. The goal is to identify important thresholds for vehicular analytics, such as speed and acceleration, to infer the driver's emotional state. By integrating these perspectives, we seek to improve the explainability of the system and gain deeper insights into how human emotions influence driving behavior.

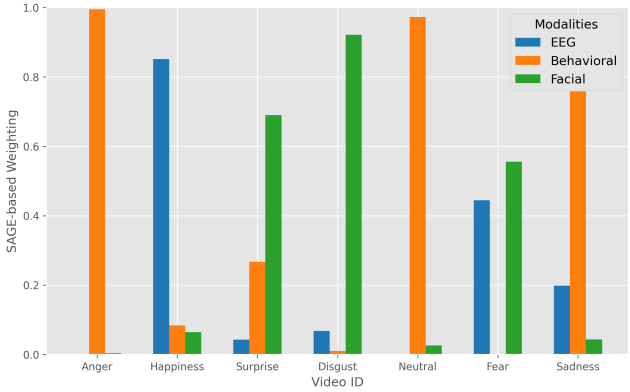


Figure 3: Contribution of each teacher model to different emotion classes

Acknowledgments

Work supported by the European Union under the Italian National Recovery and Resilience Plan (PNRR) PE00000001 - program "RESTART", and Mission 4 Issue 2 Investment 1.4 "Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S" CN00000023 – "Sustainable Mobility Center (CNMS)".

References

[1] W. Li, G. Zeng, J. Zhang, Y. Xu, Y. Xing, R. Zhou, G. Guo, Y. Shen, D. Cao, and F.-Y. Wang, "Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit," *IEEE Transactions on Computational Social Systems*, 2021.

[2] V. De Caro, S. Bano, A. Machumilane, A. Gotta, P. Cassarà, A. Carta, R. Semola, C. Sardianos, C. Chronis, I. Varlamis *et al.*, "Ai-as-a-service toolkit for human-centered intelligence in autonomous driving," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 91–93.

[3] L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain, and W. Gao, "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2970–2981, 2023.

[4] X. Xiong, A. Arnab, A. Nagrani, and C. Schmid, "M&m mix: A multimodal multiview transformer ensemble," *arXiv preprint arXiv:2206.09852*, 2022.

[5] Z. Dong, C. Hu, S. Zhou, L. Zhu, J. Wang, Y. Chen, X. Lv, and X. Ji, "Decnet: A non-contacting dual-modality emotion classification network for driver health monitoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[6] J. Ni, R. Sarbajna, Y. Liu, A. H. Ngu, and Y. Yan, "Cross-modal knowledge distillation for vision-to-sensor action recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4448–4452.

[7] S. Bano, N. Tonellotto, P. Cassarà, and A. Gotta, "Fedcmd: A federated cross-modal knowledge distillation for drivers' emotion recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–27, 2024.

[8] G. Radevski, D. Grujicic, M. Blaschko, M.-F. Moens, and T. Tuytelaars, "Multimodal distillation for egocentric action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5213–5224.

[9] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition." in *Interspeech*, 2016.

[10] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *advances in neural information processing systems*, 2020.

[11] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 212–17 223, 2020.

[12] W. Li, R. Tan, Y. Xing, G. Li, and S. Li, "A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks," *Scientific Data*, vol. 9, no. 1, p. 481, 2022.

[13] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*.