IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 16 August 2024, accepted 28 August 2024, date of publication 2 September 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3452781

RESEARCH ARTICLE

Modality- and Subject-Aware Emotion Recognition Using Knowledge Distillation

MEHMET ALI SARIKAYA AND GÖKHAN INCE
Department of Computer Engineering, Istanbul Technical University, 34469 Istanbul, Türkiye

Corresponding author: Gökhan Ince (gokhan.ince@itu.edu.tr)

ABSTRACT Multimodal emotion recognition has the potential to impact various fields, including human-computer interaction, virtual reality, and emotional intelligence systems. This study introduces a comprehensive framework that enhances the accuracy and computational efficiency of emotion recognition by leveraging knowledge distillation and transfer learning, incorporating both unimodal and multimodal models. The framework also combines subject-specific and subject-independent models, achieving a balance between localization and generalization. Subject-independent models include EEG-based, non-EEG-based (i.e., electromyography, electrooculography, electrodermal activity, galvanic skin response, skin temperature, respiration, blood volume pulse, heart rate, and eye movements), and multimodal models trained on all training subjects, capturing a broader context. Subject-specific models, including EEG-based, non-EEG-based, and multimodal models, are trained on individual subjects to provide localized knowledge. The proposed framework then distills knowledge from these teacher models into a student model, utilizing six different distillation losses to combine both subject-independent and subject-specific insights. This approach makes the model subject-aware by using local patterns and modality-aware by incorporating unimodal data, enhancing the robustness and generalizability of emotion recognition systems to varied real-world scenarios. The framework was tested on two well-known datasets, SEED-V and DEAP, as well as an immersive three-Dimensional (3D) Virtual Reality (VR) dataset, GraffitiVR, which captures emotional and behavioral responses from individuals experiencing urban graffiti in a VR environment. This broader application provides insights into the effectiveness of emotion recognition models in both 2D and 3D settings, facilitating a wider range of assessment. Empirical results demonstrate that the proposed knowledge distillation-based model significantly elevates performance across all datasets when compared to traditional models. Specifically, the model demonstrated improvements ranging from 6.56% to 24.59% over unimodal models and from 1.56% to 4.11% over multimodal approaches across the SEED-V, DEAP, and GraffitiVR datasets. These results underscore the robustness and effectiveness of the proposed approach, suggesting that it significantly enhances emotion recognition processes across various environmental settings.

INDEX TERMS Brain-computer interface, emotion recognition, knowledge distillation, transfer learning, virtual reality, multimodal models, EEG-based models, cross-modal distillation, subject-specific models, subject-independent models.

I. INTRODUCTION

In recent years, the study of emotion recognition through physiological signals has gained significant attention due to its applicability in psychology, healthcare, and human-computer interaction. A variety of physiological signals are employed to detect emotions in these research areas. These include ElectroEncephaloGraphy (EEG), ElectroMyoGraphy (EMG), ElectroOculoGraphy (EOG), ElectroDermal Activity (EDA), Galvanic Skin Response (GSR), SKin Temperature (SKT), RESPiration (RESP), Blood Volume Pulse (BVP), Heart Rate (HR), and eye movements [1], [2], [3].

Recently, the popularity of Virtual Reality (VR) headsets has increased, leading to the development of emotion-driven applications in areas such as gaming, education, meditation/therapy, and physical activity [4]. To tailor these

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

applications to individual users, it is crucial to accurately assess their emotional states and interactions within the virtual environment. Traditional facial recognition systems are ineffective in VR settings due to the use of headsets. Instead, Brain-Computer Interfaces (BCIs) are utilized to detect emotions through physiological signals like EEG, EMG, EOG, EDA, GSR, SKT, RESP, BVP, HR, and eye movements, with numerous studies exploring these methods [1], [5]. Creating reliable emotion recognition models involves extensive data collection from many subjects and experiments, which can be time-consuming and expensive. Consequently, it is important to use previously gathered data and pre-trained models to support new research in VR settings.

Transfer learning techniques facilitate the use of pre-existing models and datasets. These methods were initially applied to image recognition, where they successfully enhanced recognition accuracy [6]. However, physiological signals exhibit significant variability [7], and factors like an individual's current mood and stress levels can affect the reusability and generalization of models [8]. Consequently, a novel transfer learning approach is needed to adapt models with limited data while maintaining their learning capacity despite individual differences.

Subject-specific emotion data may not capture the general emotional patterns of an individual within a limited recording period, often leading to subject-specific models overfitting to this limited data [9]. This overfitting significantly prevents their generalization capabilities to new, unseen emotion recognition scenarios. On the other hand, subject-independent models, while useful in broader contexts, fail to capture individual-specific patterns essential for personalized emotion recognition. This dichotomy highlights the challenges in using subject-specific data alone to model complex emotion recognition mechanisms effectively. Additionally, the extensive calibration time required to collect subject-specific data for all individuals makes the development of subject-independent models a more practical approach for zero-calibration and plug-and-play applications [10].

Taking into account the nuances between EEG modalities and non-EEG based sensory data is also crucial when designing an emotion recognition system. EEG signals are covert and vary significantly across individuals, making them highly subject-dependent. Conversely, non-EEG based sensors, such as EMG, EOG, EDA, GSR, SKT, RESP, BVP, HR, and eye movements tend to exhibit more consistent characteristics across different subjects, which can be universally applied in emotion recognition. This distinction allows for the strategic use of subject-independent non-EEG-based models and subject-specific EEG-based models as teachers in our proposed approach. Moreover, by integrating various modality-subject combinations, our method aims to develop a model that is both modality-aware and subject-aware, enhancing its generalizability and effectiveness.

To tackle these issues and improve the training schemes for both subject-specific and subject-independent models, we propose a novel end-to-end Knowledge Distillation (KD) based training scheme. This approach enhances the performance of individual subjects by leveraging emotion-recognition mechanisms derived from subject-specific and subject-independent models, as well as from modality-based models.

This article introduces a novel approach leveraging the power of KD, a technique widely recognized for its efficacy in combining multimodal data and transferring knowledge constructively from one network to another during training [11]. The versatility of KD makes it especially beneficial in complex scenarios that require the integration of diverse data types, thereby paving the way for advancements in the field of emotion recognition.

This research makes several significant contributions to the field of emotion recognition. Firstly, it introduces the concept of KD-based Transfer Learning, a pioneering approach to enhancing multimodal emotion recognition models by targeting both modality and subject awareness first time in the literature. This methodology offers a novel application, enabling a comparative analysis across various emotion recognition settings. Secondly, the study establishes a comprehensive benchmarking framework to evaluate emotion recognition performance in both 2D and 3D environments. This framework demonstrates the effectiveness of the KD-based multimodal emotion recognition architecture, resulting in improved accuracy and overall system performance across different contexts.

In addressing these aspects, the research sets a new standard for emotion recognition, offering insights into the potential of transfer learning and knowledge distillation to enhance generalization ability and model accuracy across both traditional 2D and immersive 3D settings.

The remainder of this paper is structured as follows. In Section II, we review the related works in the field of emotion recognition, emphasizing recent advances in transfer learning and knowledge distillation. Section III describes the proposed framework for multimodal emotion recognition using KD-based transfer learning. Section IV details the datasets and the extracted features utilized in our experiments. Section V presents a comprehensive analysis of the relationships between sensory data and emotion across multiple datasets. In Section VI, we outline the experimental setup and methodologies used to test our framework. Section VII presents the results obtained from the experiments. Section VIII explores the implications of our findings and compares them with existing studies. Finally, Section IX summarizes our contributions and suggests directions for future research.

II. RELATED WORKS

Advancements in sensor technology, signal processing, and Artificial Intelligence (AI), especially within the domain of BCI, have significantly propelled emotion recognition research forward. This section critically reviews the literature on the impact of transfer learning and knowledge distillation on BCIs and emotion recognition, with a focus on their role in improving generalization ability and overall performance.

A. TRANSFER LEARNING IN BCI

Transfer learning has been increasingly applied to enhance the performance of BCIs in various contexts, including emotion recognition. These applications often seek to leverage data from multiple subjects or sessions to overcome the challenge of limited labeled data in a single-subject context [12]. One notable study by Li et al. introduced a multisource transfer learning method, which significantly reduced the need for labeled data by selecting appropriate sources and integrating source models, resulting in a 12.72% improvement in emotion recognition accuracy on the SEED dataset [13]. Xue et al. proposed a transfer framework based on feature analysis using Transfer Component Analysis (TCA). This method projected the source and target domains into a Kernel Hilbert space to reduce the distance between them, achieving a mean accuracy of 58.49% in emotion recognition, an improvement over previous studies [14]. Similarly, Zheng et al. developed a subject transfer framework using TCA and Kernel Principal Component Analysis, reaching an accuracy of 79.83% for emotion recognition. Their approach aimed to address the structural and functional variability between subjects, enhancing the robustness of affective brain-computer interfaces [15].

Yin et al. presented a transfer recursive feature elimination method for selecting robust EEG features across different subjects. Their approach significantly improved emotion classification performance, demonstrating statistical gains in accuracy on the DEAP database [16]. Ruan et al. proposed a transfer discriminative dictionary pair learning approach to handle individual differences in EEG data. This method enhanced cross-subject emotion classification accuracy by projecting data into a domain-invariant subspace and constructing a discriminative classifier in the target domain [17]. Lin and Jung introduced a conditional transfer learning framework to improve EEG-based emotion classification accuracy. Their framework selectively leveraged data from other subjects with similar EEG signatures, resulting in a 15% improvement in valence and a 12% improvement in arousal classification [18].

Li et al. developed the Fast Online Instance Transfer (FOIT) method for efficient instance transfer in EEG emotion recognition. FOIT significantly improved classification accuracy in cross-subject and cross-session scenarios, offering a fast and practical solution for enhancing the generalization of affective brain-computer interfaces [19]. Zhong and Jianhua proposed a subject-generic feature selection method using transfer recursive feature elimination to identify robust EEG features for emotion classification across different subjects. Their method demonstrated significant improvements in classification accuracy on the DEAP database, outperforming several recent works [20].

Li et al. developed a multi-source transfer learning approach based on style transfer mapping to reduce EEG differences between the target domain and each source domain. This method significantly improves emotion recognition accuracy by facilitating the fast deployment of models with minimal labeled data requirements [13]. Chai et al. introduced an unsupervised domain adaptation strategy called Multi-Subject Subspace Alignment (MSSA). By utilizing differential entropy features, MSSA performs subspace alignment for non-stationary EEG signals in multi-subject emotion recognition scenarios, demonstrating its effectiveness on the SEED dataset [21]. Xiao et al. proposed a group division approach that compares the emotion signals of different individuals and calculates the similarity of EEG signals. This method creates a personalized machine emotion expression system for robots, enhancing the dynamic emotional interaction between humans and machines [22].

Lan et al. introduced the Maximum Independence Domain Adaptation (MIDA) method to reduce intra- and inter-subject variance in multiple emotion classification tasks. MIDA significantly improves accuracy, with gains of up to 13.40% compared to baseline accuracy, highlighting its effectiveness in various datasets and cross-dataset applications [23]. Santana et al. employed multi-objective genetic programming and novel fitness functions that assess transferability to train a cross-subject classifier for predicting stimulants based on brain activity. Their approach focuses on evolving transferable classifiers that improve classification accuracy over classical classifiers incorporating domain adaptation methods [24].

B. DEEP LEARNING-BASED TRANSFER LEARNING

A Deep Neural Network (DNN) is a machine learning method that consists of an input layer, an output layer, and several hidden layers. It is frequently applied to EEG-based emotion recognition tasks [25], [26]. DNNs are also used in various transfer learning tasks, including network adaptation, feature transfer, and parameter transfer. Convolution-based transfer learning, such as Convolutional Neural Networks (CNNs), reuses pre-trained components in the source domain, while adversarial-based transfer learning employs adversarial techniques, such as Generative Adversarial Networks (GANs). Olamat et al. utilized deep learning methods including convolutional neural networks, such as AlexNet, DenseNet-201, ResNet-101, and ResNet50, to achieve remarkable classification accuracies of up to 100% on the SEED dataset. Their study highlights the effectiveness of transfer learning methods in emotion recognition tasks [27]. Özdenizci et al. introduced an adversarial inference approach to reduce inter-subject variability in a publicly available motor imagery EEG dataset [28].

The ASTDF-Net proposes a spatial-temporal dual-stream fusion network, specifically addressing the complex spatiotemporal dynamics inherent in EEG signals. This method incorporates a collaborative embedding module, stacked spatial and temporal attention streams, and a hybrid feature fusion module, demonstrating superior performance on public datasets such as DEAP and MAHNOB-HCI [29]. Similarly, SGLNet, a Spiking Neural Network combined with adaptive graph convolution and Long Short-Term Memory (LSTM), effectively captures intricate spatiotemporal patterns in EEG signals for robust emotion recognition. This approach utilizes a learnable spike encoder, adaptive graph convolution, and spike-based LSTM units, evaluated on datasets like DEAP and PhysioNet, showcasing significant improvements over existing EEG classification algorithms [30]. Furthermore, the Multi-modal Mood Reader introduces a pre-trained model that excels in cross-subject emotion recognition. By integrating masked brain signal modeling with an interlinked spatial-temporal attention mechanism, this approach achieves notable performance enhancements across various datasets. This model also provides valuable insights into emotion-related brain areas through attention visualization, illustrating the power of multimodal cross-scale fusion [31].

Lu et al. introduced a Domain adaptation with Few-shot Fine-tuning Network (DFF-Net) designed to improve cross-subject emotion recognition accuracy. By combining domain adaptation and fine-tuning techniques, their approach achieved 93.37% accuracy on the SEED dataset and 82.32% on the SEED-IV dataset, demonstrating significant improvements in handling inter-individual differences in EEG signals [32]. Zhang et al. proposed a subject-independent approach to evaluate sleep quality by applying Kullback–Leibler (KL) divergence to deep autoencoders to calculate the difference between source and target domains during feature extraction. Their experimental results indicated that the deep transfer learning model achieved superior classification accuracy compared to an Support Vector Machine (SVM) based baseline model [33].

Sidharth et al. employed Resnet50 along with a novel feature combination approach for EEG-based emotion detection. Their method achieved a subject-dependent accuracy of 93.1% and a subject-independent accuracy of 71.6%, utilizing mean phase coherence and magnitude squared coherence in combination with differential entropy features. This study underscores the potential of advanced feature combinations in enhancing emotion classification performance [34]. Li et al. developed a domain adaptation method for cross-subject emotion recognition models, utilizing adversarial training to adapt the marginal distributions in the early layers of the DNN model and employing association reinforcement to adapt the conditional distributions in the final layers [35]. Aldayel et al. investigated the use of deep transfer learning to improve the classification accuracy of EEG-based preference recognition. By transferring knowledge from emotion recognition tasks, their approach achieved a high accuracy of 93%, showcasing the versatility and effectiveness of transfer learning in various EEG classification tasks [36].

C. KNOWLEDGE DISTILLATION IN EMOTION RECOGNITION

Knowledge distillation has emerged as a pivotal strategy in the enhancement of EEG-based classification systems, particularly within the realm of emotion recognition. By transferring knowledge from complex, high-capacity models to simpler, more efficient ones, this technique has been instrumental in addressing the inherent challenges of EEG data, such as high dimensionality and subject-specific variability. For instance, Wang et al. developed a lightweight domain adversarial neural network utilizing knowledge distillation to effectively manage domain shifts in EEG data, thereby significantly boosting the accuracy of cross-subject emotion recognition [37]. This model leverages a deep, transformer-based architecture as the teacher model, guiding a lighter Bidirectional Long Short-Term Memory (Bi-LSTM) based student model to mimic robust feature representations, which in turn enhances domain-invariant feature learning for emotion classification.

Similarly, the application of knowledge distillation extends beyond EEG to other physiological signals, as demonstrated by Joshi et al. [38]. They proposed a cross-modal framework that employs EEG-trained models to enhance electrocardiogram based sleep staging. This approach not only showcases the versatility of knowledge distillation but also highlights its potential in leveraging diverse physiological data for improved diagnostic accuracy. Moreover, the framework achieved significant improvements in sleep staging performance, affirming the viability of knowledge distillation in clinical applications.

Further advancing the field, Wang et al. optimized a residual network for EEG-based emotion recognition through knowledge distillation [39]. This approach significantly reduced the model's complexity while maintaining high accuracy, demonstrating the potential for deploying such models in resource-constrained environments like embedded systems. The effectiveness of this methodology was validated on standard datasets, showcasing substantial improvements in performance over traditional models.

Liu et al. introduced EmotionKD, a cross-modal knowledge distillation framework that enhances the performance of models using GSR signals by distilling knowledge from EEG signals [40]. This hybrid approach not only addresses the challenges of multimodal emotion recognition but also reduces the reliance on EEG data, making the technology more accessible and feasible for real-world applications. The framework incorporates an adaptive feedback mechanism that dynamically adjusts based on performance, further optimizing the distillation process.

Zhang and Etemad introduced a pioneering approach to streamline EEG models for real-time applications on smart devices [41]. They developed a knowledge distillation pipeline where a heavily pre-trained teacher model distills crucial EEG representations into a more compact student model through a capsule-based architecture. This method effectively handles the complexity of large-scale EEG datasets and demonstrates its efficacy by achieving state-of-the-art results, enhancing the model's performance while maintaining a lower computational footprint.

Building on the cross-modal potential of knowledge distillation, another study by Zhang et al. explored the integration of visual and EEG modalities to improve emotion recognition [42]. They leveraged a cascade of CNN and Temporal Convolutional Network (TCN) architecture for the teacher model, trained on visual data, to enhance a TCN-based student model trained on EEG data. This visual-to-EEG distillation not only improved the accuracy of emotion prediction but also offered insights into the brain's synchronized activity across various areas, highlighting the contribution of fast beta and gamma waves in emotional processing. The success of this approach, evidenced by significant statistical validation, illustrates the advantages of integrating diverse modalities through knowledge distillation.

Further addressing the challenges inherent in brain-machine interfaces, Wu et al. tackled the issue of epileptic seizure prediction by proposing a novel training scheme that utilizes knowledge distillation to merge the benefits of patient-specific and patient-independent models [43]. Their approach distilled informative features from a large corpus of multi-subject data into a patient-specific model, significantly improving the accuracy and generalizability of seizure prediction. This method not only demonstrated superior performance over traditional models but also bridged the gap between personalized and universal modeling approaches in medical applications.

The integration of transfer learning and knowledge distillation within BCIs for emotion recognition offers a significant research avenue. These methodologies address critical challenges like data variability and model complexity, contributing to the development of more robust and adaptable emotion recognition systems. However, there remains a distinct gap in the application of these techniques specifically within 3D VR environments, where modality-aware and subject-aware data integration is essential.

Review of the existing literature reveals a focus on enhancing emotion recognition models primarily outside the context of immersive environments. Key studies demonstrate the application of knowledge distillation to improve model efficiency and accuracy but often restrict their scope to traditional, non-immersive settings or single-modal data. These methodologies, while advancing the field in significant ways, do not fully address the complexities of integrating multiple types of modalities and subject-specific data, which is crucial for comprehensive emotion recognition in VR environments.

To bridge these gaps, our work expands the application of knowledge distillation by implementing a modality-aware and subject-aware framework that effectively integrates both EEG and non-EEG sensory data across different environments. This approach not only caters to the conventional 2D datasets but is also adept at handling immersive 3D VR environments. By employing both subject-independent and subject-specific strategies within our knowledge distillation framework, we enhance the adaptability and accuracy of emotion recognition systems, ensuring they are robust across various sensory inputs and user experiences.

III. KNOWLEDGE DISTILLATION FOR MULTIMODAL EMOTION RECOGNITION MODELS

KD has emerged as a potent technique for enhancing the performance of machine learning models by transferring knowledge from complex, often cumbersome models (teachers) to simpler, more efficient ones (students) [11]. In the context of emotion recognition, leveraging KD can help address the challenges of multimodal data integration and the variability inherent in subject-specific responses. This section introduces our novel KD framework designed to optimize multimodal emotion recognition models by harnessing both subject-specific and subject-independent data sources.

A. SYSTEM ARCHITECTURE

As depicted in Figure 1, our proposed framework presents a structured approach to knowledge distillation for training multimodal emotion recognition models. The architecture is designed to optimize the integration of EEG and non-EEG data across different subjects and modalities. Below, we define and discuss each component in the architecture:

1) SUBJECT-INDEPENDENT MODEL TRAINING

The subject-independent model training component is designed to generalize across individuals, providing a robust baseline for comparison and further refinement through knowledge distillation:

- **EEG-based Model:** This model is trained on aggregated data from multiple subjects, focusing on capturing common features that are not overly subject-specific.
- **Non-EEG-based Model:** Similarly, this model processes aggregated non-EEG data (such as EMG, EOG, EDA, GSR, SKT, RESP, BVP, HR, and eye movements) to identify general emotional indicators.
- **Multimodal Model:** Combining both EEG and non-EEG data from various subjects, this model aims to maximize the extraction of generalizable features.

2) SUBJECT-SPECIFIC MODEL TRAINING

These models are tailored to individual subjects, enhancing the personalized handling of EEG and non-EEG data:

- **EEG-based Models:** Each subject-specific EEG model is adapted to the unique brainwave patterns of individuals, aiming to capture detailed emotional states from EEG signals.
- **Non-EEG-based Models:** These models process non-EEG data that are indicative of emotional states, providing a complementary perspective to the EEG data.

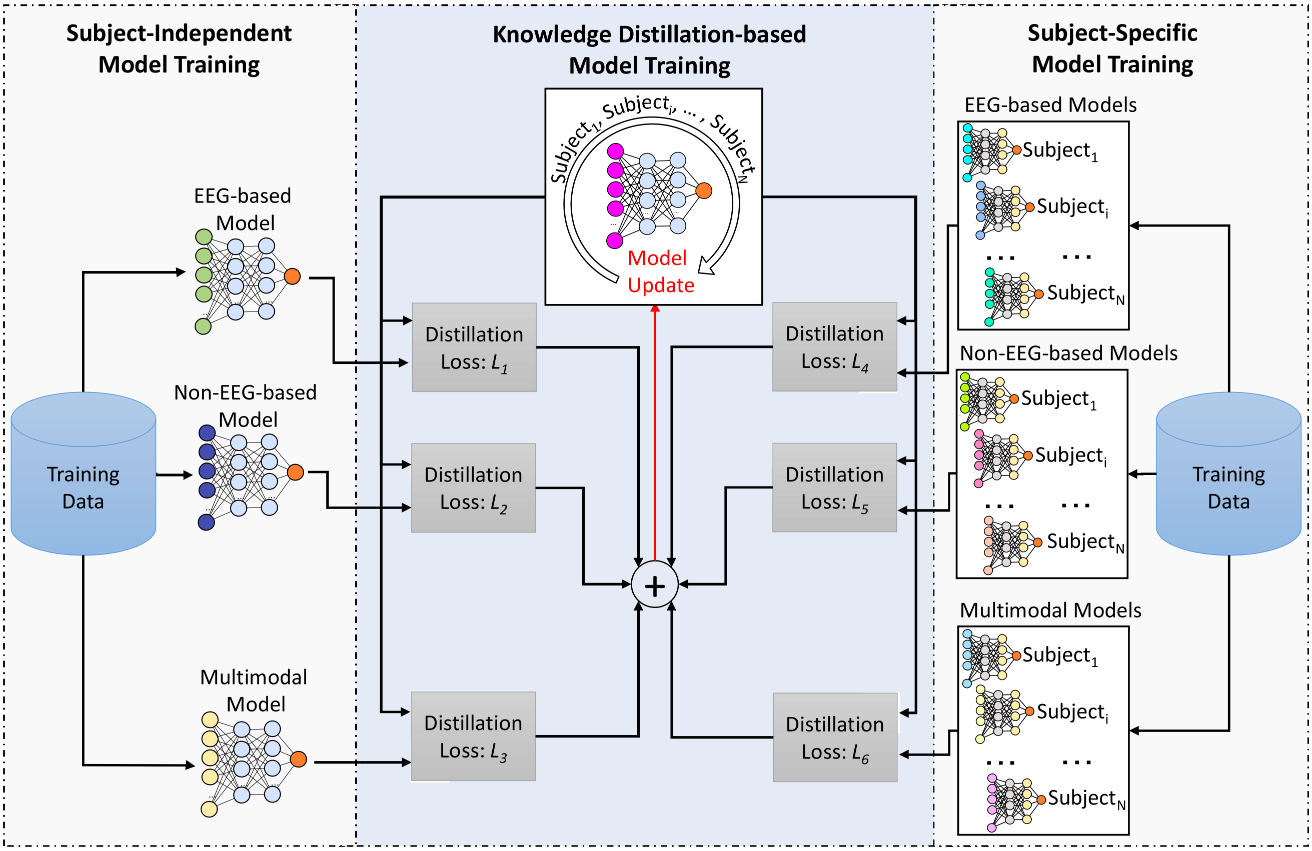

FIGURE 1. Architecture of knowledge distillation-based transfer learning in emotion recognition models.

- **Multimodal Models:** Integrating both EEG and non-EEG data, these models offer a holistic view of emotional states, leveraging the strengths of each modality to improve accuracy and robustness.

3) KNOWLEDGE DISTILLATION-BASED MODEL TRAINING

Our framework's knowledge distillation component revolves around a single deep neural network model that is iteratively updated through a systematic training process. This model does not begin as pre-trained; on the contrary, it is configured to adapt and enhance its parameters continuously through exposure to a series of distillation losses. These losses are derived from both subject-specific and subject-independent models across each step of the training cycle.

As the model processes data from each subject, it undergoes a sequence of updates where distillation losses play a critical role. These losses are meticulously calculated to ensure that the model refines its learning, tuning into the nuanced emotional states captured in the EEG and non-EEG data. Distillation losses such as L_1, L_2, L_3, for generalization, and L_4, L_5, L_6, for personalization, facilitate this adaptive learning process. How each loss function contributes to the model's learning and updating process is explained below:

Losses L_1, L_2, L_3: These losses are used during the initial stages of the model updates. They are calculated by comparing the predictions of the student model against the predictions derived from the subject-independent teacher models for the corresponding subject's data. These losses help in aligning the student model more closely with generalized patterns observed across different subjects.

Losses L_4, L_5, L_6: Utilized in subsequent updates, these losses compare the student model's predictions with those from the subject-specific teacher models. This comparison is crucial for ensuring that the student model adapts to the unique emotional patterns of each subject, thus personalizing the learning process. The adjustments made via these losses help in refining the integration of multimodal data specific to each subject, ensuring that the model not only learns from broad patterns but also captures essential characteristics unique to individual modalities and subjects.

This structured approach to employing distillation losses ensures that the student model dynamically integrates and adapts to both subject-independent generalizations and subject-specific nuances, resulting in a robust, adaptable final model optimized for accurate emotion recognition.

This comprehensive framework aims to leverage the robustness of multimodal data and the personalization potential of subject-specific models, driven by the efficiency of knowledge distillation. By combining these approaches, our model addresses several gaps in current literature, particularly in the integration of diverse data types and the customization of emotion recognition systems to individual variability in physiological responses.

B. KNOWLEDGE DISTILLATION-BASED MODEL GENERATION

The training process begins with collecting data from various training subjects. This data includes both EEG signals and non-EEG sensory information, such as EMG, EOG, EDA, GSR, SKT, RESP, BVP, HR, and eye movements.

In the following sections, we explain the algorithm for training the proposed multimodal emotion recognition model using knowledge distillation and define the knowledge distillation loss function used to transfer knowledge from the teacher models to the student model during training.

1) ALGORITHM OF KNOWLEDGE DISTILLATION

Algorithm 1 outlines the proposed KD framework for multimodal emotion recognition models. The algorithm consists of four main steps: 1) training subject-independent teacher models, 2) training subject-specific teacher models, 3) knowledge distillation, and 4) evaluation. The algorithm takes as input a set of training subjects $S = \{s_1, s_i, \ldots, s_N\}$, where s_i represents the i-th subject and N is the total number of subjects, and outputs the final multimodal emotion recognition model M_{final}.

In the first step, subject-independent teacher models are trained using EEG data, non-EEG sensory data, and a feature level fusion of both data types. The Subject-Independent EEG-based Model (M_{SIE}) is trained using EEG data from all training subjects, while the Subject-Independent Non-EEG-based Model (M_{SINE}) is trained using non-EEG sensory data from all training subjects. The Subject-Independent Multimodal Model (M_{SIM}) is trained using a fusion of EEG and non-EEG sensory data from all training subjects.

In the second step, subject-specific teacher models are trained for each subject in the training set. Subject-Specific EEG-based Models (M^i_{SSE}), Subject-Specific Non-EEG-based Models (M^i_{SSNE}), and Subject-Specific Multimodal Models (M^i_{SSM}) are trained using EEG data, non-EEG sensory data, and a fusion of both data types from subject s_i.

In the third step, knowledge distillation is performed to transfer knowledge from the subject-independent and subject-specific teacher models to the student model. For each subject s_i in the training set, distillation losses are computed between the student model ($M_{student}$) and the teacher models (subject-independent models M_{SIE}, M_{SINE}, M_{SIM}, and the subject-specific models M^i_{SSE}, M^i_{SSNE}, M^i_{SSM}). The distillation losses are combined to calculate a single distillation loss, which is used to update the student model ($M_{student}$). Once updated, $M_{student}$ is finalized as M_{final}. In this context, α functions as a hyperparameter that modulates the trade-off between the student loss and the distillation loss.

In the final step, the performance of the final multimodal model M_{final} is evaluated using test subjects. The model is tested on unseen data to assess its generalization and performance on new subjects.

Algorithm 1 Knowledge Distillation for Multimodal Emotion Recognition

```
Input: Training subjects $S = \{s_1, s_i, \ldots, s_N\}$
Output: Final model $M_{final}$
Step 1: Train Subject-independent Teacher Models:
Train subject-independent EEG-based model $M_{SIE}$ using EEG data from all training subjects.
Train subject-independent non-EEG-based model $M_{SINE}$ using non-EEG sensory data from all training subjects.
Train subject-independent Multimodal model $M_{SIM}$ using a fusion of EEG and non-EEG sensory data from all training subjects.
Step 2: Train Subject-specific Teacher Models:
for each subject $s_i \in S$ do
    Train subject-specific EEG-based model $M^i_{SSE}$ using EEG data from subject $s_i$.
    Train subject-specific non-EEG-based model $M^i_{SSNE}$ using non-EEG sensory data from subject $s_i$.
    Train subject-specific Multimodal model $M^i_{SSM}$ using a fusion of EEG and non-EEG sensory data from subject $s_i$.
end for
Step 3: Knowledge Distillation:
for each subject $s_i \in S$ do
    Compute Distillation Losses:
    $\mathcal{L}_1 = \mathcal{L}_{KD}(M_{SIE}, M_{student})$
    $\mathcal{L}_2 = \mathcal{L}_{KD}(M_{SINE}, M_{student})$
    $\mathcal{L}_3 = \mathcal{L}_{KD}(M_{SIM}, M_{student})$
    $\mathcal{L}_4 = \mathcal{L}_{KD}(M^i_{SSE}, M_{student})$
    $\mathcal{L}_5 = \mathcal{L}_{KD}(M^i_{SSNE}, M_{student})$
    $\mathcal{L}_6 = \mathcal{L}_{KD}(M^i_{SSM}, M_{student})$
    Calculate combined distillation loss:
    $\mathcal{L}_{\text{distillation}} = \frac{\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3+\mathcal{L}_4+\mathcal{L}_5+\mathcal{L}_6}{6}$
    Compute the total loss:
    $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{student}} + (1 - \alpha) \cdot \mathcal{L}_{\text{distillation}}$
    Update the proposed model $M_{student}$ using the computed loss.
end for
Assign the updated model to the final model:
$M_{\text{final}} = M_{\text{student}}$
Step 4: Evaluation:
Evaluate $M_{final}$ using test subjects.
```

2) KNOWLEDGE DISTILLATION LOSS

The knowledge distillation loss function defines the loss function used to transfer knowledge from the teacher models to the student model during training. The Kullback-Leibler divergence, denoted as KL, is a measure from information theory that quantifies how one probability distribution diverges from a second, expected probability distribution [44]. In the context of knowledge distillation, KL is used to calculate the loss between the softened predictions of student model (proposed multimodal model) and the teacher models (subject-independent and subject-specific models). The distillation losses for each pair of teacher-student predictions are computed using the scaled softmax function with a temperature parameter T, and are then averaged to compute the final distillation loss [11]. The complete loss function, including the trade-off between student loss and distillation loss controlled by the hyperparameter α, is defined as follows:

$$\mathcal{L}_k = \text{KL}\left(\text{softmax}(P^k_{\text{teacher}}/T),\right.$$
$$\left.\text{softmax}(P_{\text{student}}/T)\right) \cdot T^2 \tag{1}$$

$$\mathcal{L}_{\text{distillation}} = \frac{1}{6}\sum_{k=1}^{6}\mathcal{L}_k \tag{2}$$

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{student}} + (1-\alpha)\cdot \mathcal{L}_{\text{distillation}}, \tag{3}$$

where:

- P^k_{teacher} and P_{student} are the softmax predictions of the teacher and student models, respectively, for each set of models.
- KL is the Kullback-Leibler divergence function applied to the softened predictions of the teacher and student models to compute the distillation loss.
- T is the temperature parameter that scales the softmax function, controlling the smoothness of the output probability distribution and influencing the magnitude of gradients during training.
- α is a hyperparameter that balances the contribution of the student's loss and the distillation loss in the overall training loss.

The distillation loss is used to update the final multimodal model by minimizing the difference between the teacher and student model predictions. By transferring knowledge from the teacher models to the student model, the final multimodal model can capture both subject-independent and subject-specific features, leading to improved performance and generalization.

IV. DATASETS AND EXTRACTED FEATURES

A. SEED-V

The SEED-V dataset, gathered by Liu et al. [45], includes eye movements and EEG signals from 16 participants. This dataset includes a wide range of eye-tracking features, such as pupil diameter, fixation, saccades, and blinks. The eye-tracking features extracted include the mean and standard deviation of pupil diameter, and Differential Entropy (DE) in four frequency bands (0–0.2Hz, 0.2–0.4Hz, 0.4–0.6Hz, 0.6–1.0Hz). Additionally, the dataset records the mean and standard deviation of dispersion, fixation duration, blink duration, and saccade duration and amplitude. Event statistics such as blink frequency, fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, and saccade latency average are also included.

EEG signal processing applied to the SEED-V dataset involves downsampling the raw data to 200 Hz, applying a bandpass filter within the 1-75 Hz range, and extracting Differential Entropy (DE) features across five frequency bands. The dataset comprises 310 dimensions from 62 EEG channels, 33 eye movement dimensions, and 5 emotional class labels: Happy, Sad, Disgust, Neutral, and Fear. For the SEED-V dataset, the proposed framework utilizes eye movement data as the non-EEG sensory data.

B. DEAP

The DEAP dataset consists of physiological and EEG recordings from 32 participants exposed to 60-second video stimuli intended to provoke emotional responses. These recordings are categorized into dimensions such as arousal, valence, like/dislike, dominance, and familiarity, as described in [5]. The EEG data, collected from 32 channels, cover the theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-45 Hz) frequency bands. The DEAP dataset EEG recordings are limited to a frequency range of 3 to 47 Hz, which inherently excludes the delta rhythm, typically below 3 Hz, from consideration in the analysis. The upper limit of the gamma frequency band is set at 45 Hz to mitigate the risk of signal contamination from electrical noise and other higher frequency interferences. Additionally, the dataset includes various non-EEG sensory data such as horizontal and vertical EOG, zygomaticus major EMG, trapezius EMG, GSR, RESP amplitude, BVP via plethysmograph, and SKT.

For our analysis, we selected the first 16 subjects from the dataset, aligning with the SEED-V dataset. The emotional responses were classified within the Valence-Arousal space and divided into four quadrants: High Valence, High Arousal (HVHA), High Valence, Low Arousal (HVLA), Low Valence, Low Arousal (LVLA), and Low Valence, High Arousal (LVHA). These ratings, ranging from 1 to 9, were split using the median value as the threshold to categorize the responses into these quadrants.

Regarding EEG signal processing, we extracted DE features from the theta, alpha, beta, and gamma bands across all 32 EEG channels, using 3-second time windows with 2-second overlaps, focusing on the power spectrum of EEG signals sampled at 128 Hz. This sampling rate was chosen to adequately capture the EEG frequency components relevant to emotional processing, while also considering the practical aspects of data handling and analysis.

For processing the non-EEG sensory data, we extracted eight distinctive features: mean, standard deviation, minimum, maximum, first differences, second differences, power spectrum, and average gradient. This process was similar to the EEG windowing approach, using 3-second time windows with 2-second overlaps. As a result, the feature extraction yielded 128 features for EEG and 64 features for non-EEG sensory data, providing a comprehensive overview of the participants' physiological and emotional states.

C. GRAFFITIVR

The GraffitiVR dataset is a carefully curated collection aimed at analyzing the relationship between human emotional responses and their behavioral manifestations in response to urban graffiti, using VR to create immersive experiences for participants [46]. This study involved 39 participants, focusing on emotional reactions, specifically fear and pleasure, elicited by facial expressions depicted in graffiti.

For EEG data collection, the study utilized the Looxid Link package for Vive, an integrated accessory for the HTC Vive VR headset. This wireless, dry device features six channels and captures brainwave frequencies including delta (1–3 Hz), theta (3–8 Hz), alpha (8–12 Hz), beta (12–38 Hz), and gamma (38–45 Hz) from the brain's prefrontal area through channels AF3, AF4, AF7, AF8, Fp1, and Fp2 at a sampling rate of 500 Hz [47]. To quantify the electrical activity across these bands, the average band powers were computed using Welch's periodogram [48]. This method involved aggregating the areas under parabolas fitted to the power spectral density estimates for each frequency band, providing an accurate measure of the power within each EEG signal frequency band.

In the GraffitiVR dataset, we selected the initial 16 subjects to maintain consistency with the SEED-V and DEAP datasets. Emotional responses were evaluated within the Valence-Arousal framework and categorized into four quadrants: HVHA, HVLA, LVLA, and LVHA, aligning with the labels in the DEAP dataset. Ratings, ranging from 1 to 7, were divided by the median score to systematically categorize the responses into these quadrants.

Behavioral data within the dataset were meticulously recorded through video capturing and analyzed using the robust Lucas/Kanade optic flow algorithm [49]. This technique was instrumental in extracting detailed patterns of head movements. The analysis provided a comprehensive set of 15 features derived from the changes in yaw, pitch, and roll directions of head movements, including the minimum, maximum, mean, median, and standard deviation for each direction. These behavioral metrics offer a granular view of how participants' physical responses align with their emotional experiences. For the GraffitiVR dataset, the proposed framework uses head movements as the non-EEG sensory data.

In total, the EEG portion of the dataset utilizes 35 features, which include 30 features from the five frequency bands across six channels and an additional set of 5 features: attention, relaxation, asymmetry, left brain activity, and right brain activity. Combined with the 15 features from head movement analysis, the dataset offers 50 features in total. This comprehensive set enables the exploration of the dynamic interplay between emotional states and corresponding physical responses within VR environments.

V. ANALYSIS OF CORRELATIONS BETWEEN SENSORY DATA AND EMOTIONAL STATES

This section presents a comprehensive analysis of three key datasets: SEED-V, DEAP, and GraffitiVR. The analysis separately explores the correlations between EEG data and emotional states, as well as between non-EEG sensory data and emotional states. By examining these datasets, we aim to highlight the relationships and patterns that emerge across different modalities in emotion recognition. Correlations are extracted and statistically analyzed; those found to be insignificant (p-value not lower than 0.05) are discarded. The resulting correlation coefficients are visualized, with red colors indicating positive correlations and blue colors indicating negative correlations. The analysis of both EEG and non-EEG data provides a nuanced view of how different physiological measures correlate with emotional states, indicating complex interactions between various brain regions and bodily responses to emotions.

A. CORRELATION ANALYSIS IN THE SEED-V DATASET

The SEED-V dataset includes EEG and non-EEG data related to five emotional states: disgust, fear, happy, neutral, and sad. This analysis explores the relationships between these emotional states and various physiological signals.

Figure 2 presents a heatmap of the EEG feature correlations for the SEED-V dataset. From the 310 EEG features analyzed, the ten features included are those with correlation coefficients that are not only the highest but also statistically significant (p-value lower than 0.05). These significant features are alpha rhythms at electrodes FC5, PO8, F6, F5, F3, and F1; delta rhythms at electrode CPZ; gamma rhythms at electrodes TP8 and CB2; and theta rhythms at electrode PO6. This selection showcases a wide array of brain activity patterns that are important for understanding emotional responses.

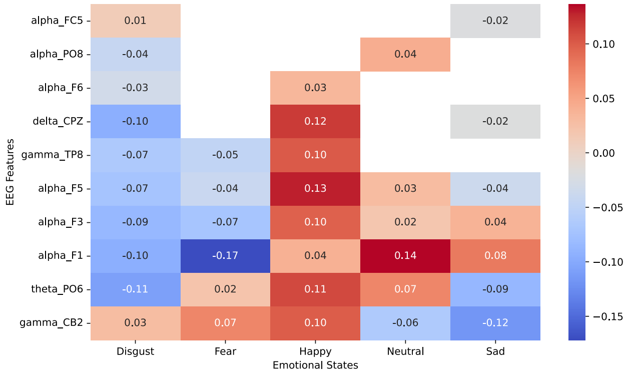

FIGURE 2. Heatmap of statistically significant eeg feature correlations on SEED-V.

The alpha band at electrode F5 shows a positive correlation with the happy state (0.13) and a negative correlation with the sad state (−0.07). Alpha rhythms are often linked to relaxed, wakeful states, which explain their positive association with the happy state [50]. Conversely, their reduction in sad emotional states suggests a shift in brain activity patterns during low mood or depressive states [51].

Figure 3 displays correlations of non-EEG features, including eye movement features like pupil size and saccadic movements. Among the 33 non-EEG sensory data features analyzed in the SEED-V dataset, the ten features with the highest correlation coefficients are pupil x-axis mean (pupil_x_mean), average saccade duration (sacc_dur_avg), saccade frequency (sacc_freq), maximum fixation dispersion (fix_disp_max), total fixation dispersion (fix_disp_total), standard deviation of pupil y-axis (pupil_y_sd), standard deviation of pupil x-axis (pupil_x_sd), pupil y-axis mean (pupil_y_mean), standard deviation of y-axis dispersion (disp_y_sd), and average saccade latency (sacc_lat_avg). Each of these features distinctly correlates with various emotional states, showcasing their pivotal role in emotion recognition.

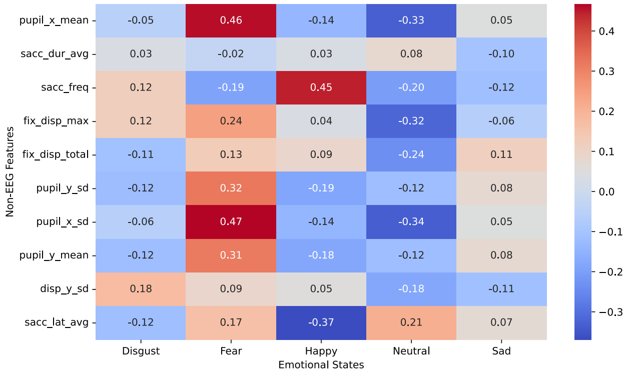

FIGURE 3. Heatmap of statistically significant non-EEG feature correlations on SEED-V.

For instance, the pupil x standard deviation shows a strong positive correlation with the fear state (0.47) and a strong negative correlation with the happy state (−0.34). Increased pupil dilation is generally associated with heightened emotional arousal, which is consistent with the response to fear [52]. The negative correlation with the happy state reflects a more relaxed and content state, leading to less pupil dilation [53].

B. CORRELATION ANALYSIS IN THE DEAP DATASET

The DEAP dataset includes EEG and non-EEG data related to arousal and valence. This section explores the correlation between these physiological signals and the emotional states: HVHA, HVLA, LVHA, LVLA.

Figure 4 presents the heatmap of statistically significant EEG feature correlations. Among the 128 EEG features analyzed in the DEAP dataset, the ten features with the highest correlation coefficients include theta rhythms at electrodes O1, Fp1 and Oz, beta rhythms at electrodes F4, O1, Oz, and Fp2, alpha rhythms at electrodes O1 and Oz, and gamma rhythms at electrode O2.

For instance, beta waves at electrode Oz show varying correlations across emotional states: a positive correlation with HVHA (0.04), and a negative correlation with LVLA (−0.05). Beta waves are associated with active, alert states and cognitive engagement, which aligns with their positive correlation in HVHA conditions [54]. The negative correlation in LVLA conditions indicates reduced cognitive activity or relaxation [55].

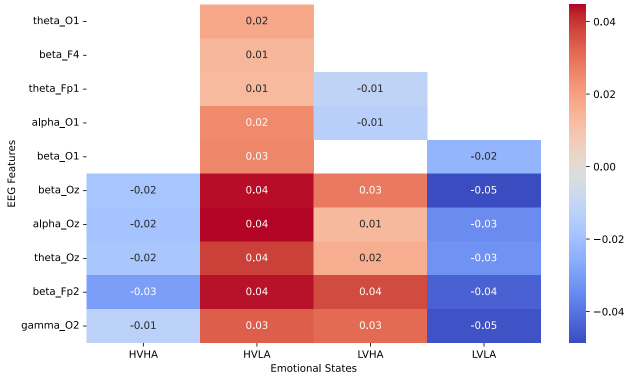

FIGURE 4. Heatmap of statistically significant EEG feature correlations on DEAP.

Figure 5 illustrates the heatmap for non-EEG feature correlations such as GSR and EMG. Among the 64 non-EEG sensory data features analyzed on the DEAP dataset, the ten features with the highest correlation coefficients are the mean horizontal electrooculography (hEOG_mean), maximum galvanic skin response (GSR_max), first and second differences in GSR (GSR_diff1, GSR_diff2), power spectrum of GSR (GSR_psd), mean, minimum, and maximum trapezius electromyography (tEMG_mean, tEMG_min, tEMG_max), and the second difference and power spectrum of temperature (Temperature_diff2, Temperature_psd).

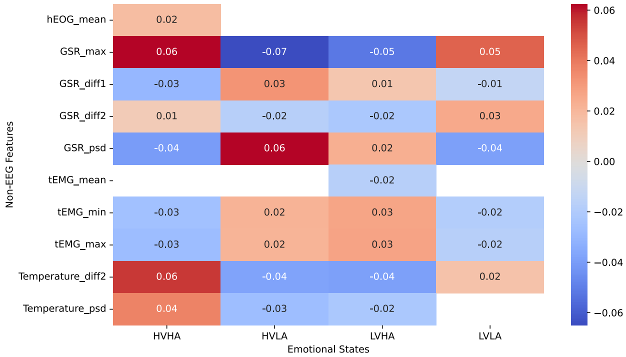

FIGURE 5. Heatmap of statistically significant non-EEG feature correlations on DEAP.

GSR maximum exhibits a strong positive correlation with HVHA (0.06) and a negative correlation with HVLA (−0.07). Increased GSR typically reflects heightened physiological arousal, which is consistent with its positive correlation in high arousal states [56]. The negative correlation in HVLA suggests complex interactions between physiological arousal and emotional valence.

C. CORRELATION ANALYSIS IN THE GRAFFITIVR DATASET

The GraffitiVR dataset offers a unique 3D immersive environment for emotion recognition. This section explores the EEG and non-EEG responses to emotional stimuli within this immersive setting, highlighting distinctive findings that emerge from the data.

Figure 6 shows a heatmap of statistically significant EEG feature correlations. Among the 35 EEG features examined in the GraffitiVR dataset, the ten features with the highest correlation coefficients are delta rhythms at AF3, measures of relaxation and attention, gamma and beta rhythms at AF8, alpha and beta rhythms at Fp1, gamma rhythms at Fp1, beta rhythms at AF4, and right brain activity.

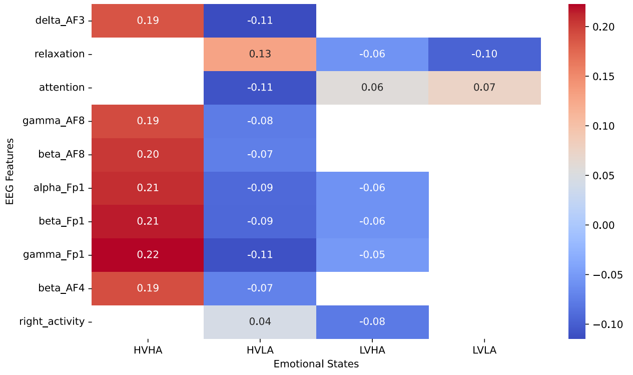

FIGURE 6. Heatmap of statistically significant EEG feature correlations on GraffitiVR.

There is a strong positive correlation of gamma wave activity at electrode Fp1 with HVHA states (0.22), contrasting with its negative correlation with HVLA states (−0.11). Gamma waves are linked to high-level information processing and cognitive functioning, which aligns with intense engagement in high valence conditions [57]. The negative correlation in LVHA states might reflect cognitive stress or disengagement in challenging emotional contexts [58]. In immersive settings, all frontal brain waves exhibit positive correlations with HVHA conditions, indicating that increases in either positive valence or arousal are associated with enhanced frontal brain activity. This pattern underscores the significant role of the frontal regions in modulating emotional responses within immersive environments [59].

Figure 7 illustrates the heatmap for non-EEG feature correlations, such as head movement metrics. Among the 15 non-EEG sensory data features analyzed in the GraffitiVR dataset, the statistically significant features are Z_min, X_mean, X_median, Y_median, and Y_sd. These features represent measurements of head movements along the Z, X, and Y axes respectively, reflecting the dynamic interaction between physical positioning and emotional responses within immersive VR settings.

X_mean, X_median and Y_median show slight positive correlations with HVLA. Such metrics may indicate physical responsiveness to emotional stimuli in the virtual environment, with greater head movements possibly reflecting positive engagement or emotional reactions [60]. Conversely, Y_sd show slight positive correlations with LVHA, which may suggest increased physical instability or agitation in more stressful or negatively valenced situations [61].

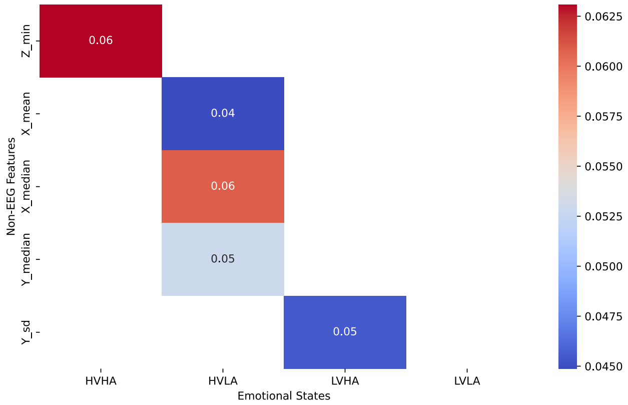

FIGURE 7. Heatmap of statistically significant non-EEG feature correlations on GraffitiVR.

VI. EXPERIMENTS

A. HARDWARE AND SOFTWARE

Development was conducted on a laptop featuring an Intel Core i7-12650H CPU, 16GB RAM, and an NVIDIA GeForce RTX 3060 GPU. The software tools utilized for development included Python 3.10, PyTorch 2.1, Keras 2.12.0, NumPy 1.23, Matplotlib 3.7, Scikit-learn 1.2.1, and Visual Studio Code. The experimental code was executed on Google Colab with a GPU runtime, specifically using a Tesla T4 with 15 GB of VRAM and 12 GB of system RAM.

B. ARCHITECTURE AND PARAMETER DETAILS

The experimental framework is designed to evaluate the efficacy of the proposed KD-based emotion recognition architecture, as illustrated in Figure 1. This setup adopts a standardized approach to optimizing hyperparameters across different datasets, streamlining the development of robust emotion recognition models.

For the datasets SEED-V, DEAP, and GraffitiVR, the neural network architectures were specifically tailored to align with the complexity of the input features and the requirements of the emotion recognition tasks. Across all models—including subject-specific EEG-based, subject-specific non-EEG-based, subject-specific multimodal, subject-independent EEG-based, subject-independent non-EEG-based, subject-independent multimodal, and knowledge distillation-based—the architecture employed the same network structure to ensure consistency in model evaluation. The architecture configurations for the datasets are as follows:

- Layers of $(128, 64)$ for both SEED-V and DEAP, reflecting their similar complexity and feature set sizes.
- A reduced configuration of $(64, 32)$ for GraffitiVR, adapted to its unique demands and smaller feature set.

These dimensions were carefully chosen to balance model complexity with training efficiency and generalization capability across diverse emotional datasets.

The neural network architecture for each dataset is structured as follows, blending performance with computational efficiency:

- A dense layer with hidden size corresponding to the first dimension, using ReLU activation, tailored to the feature size of the dataset.
- A dropout layer set at 0.5 to mitigate overfitting by randomly omitting a portion of the feature detectors on each iteration.
- Another dense layer with the second dimension, again using ReLU activation.
- A second dropout layer at 0.5 to further enhance the model's generalization.
- A final dense layer with a single output unit, employing a sigmoid activation function to predict the probability of an emotional state.

1) SUBJECT-INDEPENDENT EXPERIMENT SETTINGS

The Subject-One-Leave-Out (SOLO) approach is utilized, where for each subject's testing, the model is trained on data from the other 15 subjects. A network architecture similar to that of the subject-specific and knowledge-distillation-based models is used. There is one subject-independent EEG-based, one non-EEG-based, and one multimodal model evaluated for each subject.

2) SUBJECT-SPECIFIC EXPERIMENT SETTINGS

For the subject-specific settings, each subject's data is used to train a model individually. The knowledge distillation approach involves using a loss derived from the predictions of this model. To ensure fair comparisons, data for each subject are split 50/50 for training and testing. This setup is used exclusively to compare the subject-specific and knowledge distillation results related to the respective subject. There is one subject-specific EEG-based, one subject-specific non-EEG-based, and one subject-specific multimodal model evaluated for each subject.

3) KNOWLEDGE DISTILLATION-BASED EXPERIMENT SETTINGS

The knowledge distillation-based approach is designed to improve the performance of the student model through guidance from both subject-specific and subject-independent teacher models. Key hyperparameters were selected based on heuristic methods to optimize the training process and ensure model stability while preventing overfitting. The hyperparameters include:

- $\alpha = 0.4$ which allocates the weight to student loss and $(1 - \alpha)$ to distillation loss.
- Temperature of 4 for softening probability distributions, facilitating a smoother knowledge transfer.
- The Adam optimizer, renowned for its effectiveness in handling sparse gradients and adaptive learning rate adjustments.

This configuration not only supports the rigorous comparison of different uni-modal and multi-modal emotion recognition models but also underscores the robustness of the KD framework in accurately recognizing emotional states across varied environments.

C. EVALUATION CRITERIA

Accuracy remains a fundamental metric for evaluating the performance of classification models in various fields, including emotion recognition. It is defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined. The mathematical expression for accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where the terms are defined as follows:

- True Positives (TP) represents the number of instances the model correctly identified as positive.
- True Negatives (TN) denotes the number of instances the model correctly identified as negative.
- False Positives (FP) stands for the number of instances the model incorrectly identified as positive.
- False Negatives (FN) refers to the number of instances the model incorrectly identified as negative.

Accuracy is especially pertinent in the evaluation of our knowledge distillation-based multimodal emotion recognition models. It provides a straightforward and intuitive measure of the model's overall effectiveness in correctly predicting emotional states across various datasets. In our study, accuracy is utilized to assess the performance enhancements brought about by the KD approach on the SEED-V, DEAP, and GraffitiVR datasets. These datasets have been chosen for their diversity in emotional content and representation, which tests the robustness and adaptability of our models. Employing accuracy helps to ensure that the improvements in model performance are not only statistically significant but also relevant in practical scenarios where balanced class distribution is common.

VII. RESULTS

A. EMOTION RECOGNITION RESULTS USING DIFFERENT MODELS

To understand the key elements in emotion recognition, this study benchmarks the performance of seven distinct models across three datasets: SEED-V, DEAP, and GraffitiVR. These models include: a subject-independent EEG-based model, a subject-independent non-EEG-based model, a subject-independent multimodal model, and a knowledge distillation-based model. Additionally, subject-specific models are incorporated to demonstrate the target-only results of the transfer learning methodology. To facilitate fair comparisons, subject-specific data are split into two equal parts, with one half

TABLE 1. Emotion recognition results of different models with respect to the datasets.

Model	SEED-V	DEAP	GraffitiVR
Subject-specific EEG-based	0.63 ± 0.07	0.61 ± 0.04	0.61 ± 0.04
Subject-independent EEG-based	0.61 ± 0.05	0.61 ± 0.04	0.61 ± 0.05
Subject-specific Non-EEG-based	0.65 ± 0.06	0.61 ± 0.05	0.63 ± 0.06
Subject-independent Non-EEG-based	0.66 ± 0.06	0.61 ± 0.04	0.61 ± 0.08
Subject-specific Multimodal	0.67 ± 0.07	0.62 ± 0.04	0.66 ± 0.05
Subject-independent Multimodal	0.72 ± 0.06	0.64 ± 0.04	0.73 ± 0.12
Knowledge Distillation-based	**0.74 ± 0.08**	**0.65 ± 0.04**	**0.76 ± 0.11**

used for training and the other for testing. The performance outcomes, summarized in Table 1, reveal the comparative efficacy of each model.

For the SEED-V dataset, the subject-independent EEG-based and non-EEG-based models record accuracies of 0.61 and 0.66, respectively. The knowledge distillation-based model outperforms the others with a mean accuracy of 0.74, representing an improvement of 21.31% over the subject-independent EEG-based model and about 12.12% over the subject-independent non-EEG-based model. The subject-specific EEG-based model illustrates a slight improvement over its subject-independent counterpart. This increment indicates the potential of tailoring models to individual emotional profiles, which can yield better performance, though at the expense of generalizability. The subject-independent multimodal model, which integrates EEG data with non-EEG sensory data, shows improved accuracy at 0.72, suggesting the benefits of a multimodal approach in enhancing model robustness. When the proposed model is compared to the subject-independent multimodal model, the improvement is more modest at approximately 2.78%. These enhancements suggest that the knowledge distillation approach leverages both unimodal and multimodal models, demonstrating a superior ability to generalize emotion recognition across subjects by utilizing a modality-aware and subject-aware approach.

The DEAP dataset shows less variability in model performances with the knowledge distillation-based model marginally leading at 0.65. All models exhibit relatively similar accuracy levels around the 0.61 mark, with the multimodal approach showing a slight advantage at 0.64. The knowledge distillation-based model shows a 6.56% improvement over the unimodal models and 1.56% over the multimodal model. This minimal increase indicates that while the proposed model does provide an advantage, the unique characteristics of the DEAP dataset may limit the effectiveness of the enhancements brought by knowledge distillation.

The study was extended to include our proprietary GraffitiVR dataset, which analyzes emotional and behavioral responses to urban graffiti in a 3D VR setting. The knowledge distillation-based model again ranks highest with an accuracy of 0.76, which supports its effectiveness in diverse and potentially more immersive environments. This is an improvement of 24.59% over the subject-independent EEG-based and non-EEG-based models, both of which scored 0.61. The subject-independent multimodal model also shows robust performance with an accuracy of 0.73. Compared to the subject-independent multimodal model, the improvement stands at 4.11%. These results underscore the effectiveness of the knowledge distillation-based model in handling the complex and immersive environments typical of virtual reality, suggesting that the integration of multiple modalities or advanced knowledge distillation techniques might be particularly beneficial in virtual reality settings where emotional cues can be highly varied.

B. EMOTION RECOGNITION RESULTS USING SEED-V DATASET

Table 2 depicts the subject-based emotion recognition results, displaying the performance of various classifiers across subjects from the SEED-V dataset. Each cell in the table reflects the accuracy of a specific classifier for a given subject. Each row corresponds to a different subject ID, and the final row labeled 'mean' indicates the average accuracy for each classifier across all subjects.

The knowledge distillation-based model consistently outperforms the other models, as evidenced by its superior performance across multiple subjects. For instance, the models achieved remarkable accuracies of 0.81, 0.84, 0.83, and 0.88 on subjects 8, 9, 13 and 15, respectively, which are significantly higher compared to their counterparts in other models. These results are indicative of the robustness and efficacy of the knowledge distillation approach, particularly in subjects where other models struggle to maintain high accuracy.

This comprehensive analysis not only underscores the advantages of the knowledge distillation-based model but also demonstrates its capability to generalize across different individuals in emotion recognition tasks. The consistent outperformance across subjects highlights the model's adaptability and its potential in real-world applications where subject variability can pose significant challenges.

C. EMOTION RECOGNITION RESULTS USING DEAP DATASET

Table 3 illustrates subject-based emotion recognition results on DEAP dataset. The DEAP dataset exhibits less variability in performance among the models, suggesting that the emotional states reflected in this dataset may not distinctly favor one model over another due to similar mapping across all modalities.

TABLE 2. Comparison of emotion recognition models across subjects on SEED-V.

Subject ID	Subject-specific			Subject-independent			Knowledge Distillation-based
	EEG-based	Non-EEG-based	Multimodal	EEG-based	Non-EEG-based	Multimodal	
1	0.70	0.70	0.76	0.59	0.72	0.77	**0.78**
2	0.59	0.64	0.63	0.58	0.58	0.62	**0.69**
3	0.60	0.61	0.67	0.56	0.68	**0.81**	0.79
4	0.59	0.63	0.66	0.59	0.60	0.60	**0.66**
5	0.57	0.59	0.66	0.58	0.65	0.68	**0.70**
6	0.58	0.54	0.63	0.60	0.62	**0.72**	0.71
7	0.64	0.56	0.63	0.63	0.70	0.71	**0.71**
8	0.68	0.65	0.68	0.72	0.73	0.75	**0.81**
9	0.68	0.72	0.75	0.57	0.58	0.80	**0.84**
10	0.47	0.62	0.48	0.52	0.54	**0.61**	0.56
11	0.62	0.63	0.68	0.59	0.63	0.70	**0.70**
12	0.69	0.73	0.68	0.58	0.68	0.70	**0.76**
13	0.70	0.70	0.81	0.66	0.74	0.75	**0.83**
14	0.64	0.73	0.69	0.62	0.66	**0.74**	0.71
15	0.54	0.70	0.67	0.66	0.74	0.78	**0.88**
16	0.73	0.71	0.73	0.64	0.70	0.72	**0.75**
Mean	0.63	0.65	0.67	0.61	0.66	0.72	**0.74**

TABLE 3. Comparison of emotion recognition models across subjects on DEAP.

Subject ID	Subject-specific			Subject-independent			Knowledge Distillation-based
	EEG-based	Non-EEG-based	Multimodal	EEG-based	Non-EEG-based	Multimodal	
1	0.62	0.65	0.65	0.62	0.62	0.62	**0.66**
2	0.62	0.61	0.64	0.63	0.66	0.67	**0.68**
3	**0.65**	0.57	0.62	0.61	0.58	0.61	0.61
4	0.53	0.57	0.56	0.68	0.61	0.69	**0.71**
5	0.56	0.58	0.58	0.59	0.59	0.61	**0.64**
6	0.64	0.69	0.68	0.65	0.69	0.69	**0.71**
7	0.55	**0.67**	0.56	0.63	0.65	0.66	0.66
8	0.63	0.52	0.67	0.59	0.60	0.67	**0.67**
9	0.52	0.57	0.58	0.59	0.59	0.60	**0.64**
10	0.63	0.64	0.67	0.59	0.59	0.62	**0.70**
11	0.62	0.64	0.66	0.65	0.63	0.66	**0.66**
12	0.65	0.54	0.59	0.64	0.65	0.66	**0.67**
13	0.65	0.65	0.63	0.62	0.60	0.65	**0.65**
14	0.58	**0.63**	0.59	0.54	0.53	0.56	0.56
15	0.63	0.60	**0.65**	0.55	0.60	0.63	0.64
16	0.63	0.59	**0.64**	0.57	0.55	0.59	0.60
Mean	0.61	0.61	0.62	0.61	0.61	0.64	**0.65**

Despite the close performance values, the knowledge distillation-based model slightly leads in accuracy across several subjects. For instance, while the emotion recognition accuracy for Subject 4 reached 0.71, this is an improvement over the 0.68, 0.61, and 0.69 recorded by the subject-independent EEG-based, non-EEG-based, and multimodal models respectively. Similar improvements are observed for Subject 6, where the proposed model achieves a notable accuracy of 0.71, significantly surpassing the accuracies achieved by the other models.

Furthermore, subject-specific models display varying degrees of success; for example, the non-EEG-based model of Subject 7 reached the highest accuracy among subject-specific classifiers, demonstrating that personalized models can occasionally match or exceed the performance of more generalized approaches, particularly in handling unique or nuanced emotional data.

A particularly exceptional case is Subject 10, where the accuracy with the knowledge distillation-based model is 0.70, surpassing the 0.59 accuracy achieved with both the subject-independent EEG-based and non-EEG-based models and significantly outperforming the 0.62 accuracy with the multimodal model. This instance highlights the potential of knowledge distillation in certain contexts to effectively enhance emotion recognition tasks within the DEAP dataset.

D. EMOTION RECOGNITION RESULTS USING GRAFFITIVR DATASET

Table 4 illustrates subject-based emotion recognition results on GraffitiVR dataset. The knowledge distillation-based model generally exhibits superior performance across the dataset, demonstrating notable enhancements in accuracy, especially in immersive VR environments. For example, Subject 5 shows a dramatic improvement, where the proposed model scores 0.96, outperforming the 0.57 of the subject-independent EEG-based model, 0.76 of the subject-independent non-EEG-based model, and 0.95 of the subject-independent multimodal model. Similar significant improvements are observed in Subject 8, where the proposed

TABLE 4. Comparison of emotion recognition models across subjects on GraffitiVR.

Subject ID	Subject-specific			Subject-independent			Knowledge Distillation-based
	EEG-based	Non-EEG-based	Multimodal	EEG-based	Non-EEG-based	Multimodal	
1	0.61	0.68	0.68	0.57	0.52	**0.70**	0.65
2	0.58	0.58	0.62	0.56	0.56	0.67	**0.71**
3	0.65	0.63	0.63	0.65	0.57	0.65	**0.67**
4	0.58	0.57	0.57	0.54	0.57	0.58	**0.65**
5	0.65	0.62	0.65	0.57	0.76	0.95	**0.96**
6	0.70	0.70	0.70	0.70	0.68	0.72	**0.73**
7	0.57	0.56	0.60	**0.70**	0.68	0.68	0.68
8	0.68	0.79	0.79	0.65	0.80	0.94	**0.96**
9	0.57	0.57	0.58	0.63	0.50	**0.75**	0.67
10	0.63	0.61	0.69	0.57	0.55	0.58	**0.72**
11	0.67	0.67	0.67	0.65	0.57	0.93	**0.96**
12	0.53	0.53	0.63	0.56	0.62	0.67	**0.67**
13	0.60	0.63	0.70	0.63	0.62	**0.80**	0.78
14	0.60	0.63	0.63	0.57	0.63	0.63	**0.85**
15	0.61	0.67	0.67	0.59	0.54	0.74	**0.74**
16	0.57	0.66	0.70	0.58	0.58	0.67	**0.77**
Mean	0.61	0.63	0.66	0.61	0.61	0.73	**0.76**

model also achieves an accuracy of 0.96, compared to the respective accuracies of the other models.

Although subject-specific models offer tailored approaches to emotion recognition, they generally exhibit lower accuracies compared to subject-independent and knowledge distillation-based models in the GraffitiVR dataset. This result arises from the complex sensory data inherent in VR environments, which may contain intricate patterns that require extensive data for effective model training—data that is often insufficient in subject-specific settings.

Subjects 11 and 14 also highlight the effectiveness of the proposed model, with accuracies of 0.96 and 0.85 respectively, showcasing its capability to leverage the complex sensory data inherent in VR settings more effectively than other models.

These results underscore the knowledge distillation-based model's ability to surpass traditional and multimodal approaches in VR contexts, where immersive experiences may significantly influence emotional responses. The data demonstrates the potential of this model to facilitate high-fidelity emotion recognition in challenging VR environments, making it particularly suited for applications where understanding nuanced emotional dynamics is critical.

E. COMPARATIVE ANALYSIS WITH RELATED WORKS

This comparative analysis assesses the effectiveness of various transfer learning approaches in EEG-based emotion recognition. Improvements are quantified and ordered from lowest to highest in Table 5, which includes selected research studies that detail the extent of their improvements using transfer learning.

Ma et al. improved emotion recognition accuracy by 2.8% through the Cross-subject Source Domain Selection (CSDS) method [62]. Ren et al. demonstrated a 2-3% improvement in cross-subject emotion recognition scenarios through their multisource instance transfer learning framework [63]. Li et al. presented a Transferable Attention Neural Network (TANN) that optimizes emotional recognition by prioritizing EEG samples with high transferability; the mean improvement is about 2.5% and 3.4% [64]. Luo et al. employed a Manifold-based Domain Adaptation with Dynamic Distribution (MDDD) method to effectively align source and target domains, showing an average improvement of 3.54% over traditional non-deep learning methods [65]. The integration of transfer learning with dictionary learning in Gu et al.'s Multi-source Domain Transfer Discriminative Dictionary Learning (MDTDDL) enhances the adaptability of EEG features across different emotional states and subjects, improving accuracy by 3.94% to 4.02% [66]. Wang et al. enhanced emotion classification accuracy by 4.51% through automated feature extraction using Electrode-Frequency Distribution Maps (EFDMs) and a Residual Block-based Deep CNN [67]. Xue et al. utilized TCA to improve feature alignment between different subjects, thereby demonstrating a 6.43% improvement in the accuracy of emotion recognition [14]. Wang et al. incorporated an inter-subject contrastive loss and a pairwise similarity mechanism in their transfer learning model, achieving substantial generalization improvements with an 8% increase on the target domain [68]. Tang et al. demonstrated the effectiveness of their meta-transfer learning model in improving average accuracy by 6.23% on the SEED dataset and 10.43% on the SEED-IV dataset, underscoring its potential in cross-subject scenarios [69]. The multisource transfer learning approach introduced by Li et al. improves emotion recognition accuracy by 12.72% [13].

In comparison to existing methodologies, which predominantly utilize only EEG data, our model not only achieves a significantly broader range of accuracy enhancements—from 6.56% to 24.59% over unimodal models and 1.56% to 4.11% over multimodal approaches—but also demonstrates these improvements across multiple datasets, including SEED-V, DEAP, and GraffitiVR. By employing pre-trained models and integrating multimodal data, our approach enhances generalizability across different datasets and subjects. Furthermore, this multi-modal strategy, particularly its application to the GraffitiVR dataset, underscores its potential in immersive 3D

TABLE 5. Comparative analysis of EEG-based emotion recognition improvement rates using transfer learning.

Source	Improvement	Methods Used	Dataset
Ma et al. [62]	2.8%	CSDS	SEED
Ren et al. [63]	2% - 3%	Instance transfer learning	SEED, SEED-IV
Li et al. [64]	2.5% - 3.4%	TANN with attention	SEED-IV
Luo et al. [65]	3.54%	MDDD	SEED, SEED-IV
Gu et al. [66]	3.94% - 4.02%	MDTDDL	SEED, DEAP
Wang et al. [67]	4.51%	Deep CNNs with EFDMs	SEED
Xue et al. [14]	6.43%	TCA	SEED
Wang et al. [68]	8%	Contrastive Loss with Pairwise Similarity	SEED, SEED-IV
Tang et al. [69]	6.23% - 10.43%	Meta-transfer learning	SEED, SEED-IV
Li et al. [13]	12.72%	Multisource transfer	SEED
Our Study	**6.56% - 24.59%**	**KD-based**	**SEED-V, DEAP, GraffitiVR**

environments, leveraging the synergistic effects of various physiological signals to enhance robustness and accuracy in diverse real-world scenarios. This comprehensive application promises significant advancements in the field of EEG-based emotion recognition, adapting effectively to the complexities of real-world applications.

VIII. DISCUSSION

The proposed knowledge distillation framework integrates the strengths of subject-independent and subject-specific models to enhance multimodal emotion recognition. This integration allows the final model to effectively capture a comprehensive range of features that are both generic and unique to individual subjects. By employing a combination of EEG and non-EEG sensory data, the framework accommodates diverse data modalities, enhancing its capability to predict emotion accurately. Optimization of the model involves a dual approach using both student loss and distillation loss, which facilitates the effective transfer of knowledge from the teacher models to the student model. This approach ensures that the final multimodal model is not only robust but also demonstrates high performance on unseen test data, thereby confirming its generalization capabilities.

Empirical evidence from the study underscores the efficacy of the proposed KD framework in the field of emotion recognition. The framework exhibited commendable performance on the SEED-V, DEAP, and GraffitiVR datasets, surpassing benchmark results set by existing models. These results are particularly significant, demonstrating that the use of knowledge distillation can indeed enhance the accuracy of multimodal emotion recognition systems in line with existing studies such as [40], and [42]. The improvement in accuracy highlights the framework's ability to leverage distilled knowledge effectively, optimizing the recognition process across varied datasets.

The success of the proposed KD framework in achieving superior performance across multiple datasets illustrates its potential as a scalable and versatile tool in emotion recognition. By harmonizing knowledge from both subject-independent and subject-specific models, the framework not only boosts performance but also enriches the model's adaptability to new and varying contexts. This adaptability is crucial for applications that require high fidelity in emotion recognition, such as in virtual reality and healthcare, where personalized and accurate emotion assessment is key. The methodology's emphasis on optimizing both student and distillation losses further refines the model's efficiency, making it a valuable contribution to the advancements in emotion recognition technology.

IX. CONCLUSION

In this study, we introduced a novel knowledge distillation framework tailored for multimodal emotion recognition. This innovative approach harnesses both subject-independent and subject-specific data to comprehensively capture emotional features, thereby enhancing model performance and generalization capabilities. By adeptly transferring knowledge from varied models into a cohesive multimodal model, our framework not only excels at accurately predicting emotion labels but also significantly outperforms existing emotion recognition models. The efficacy of this framework has been robustly demonstrated on three datasets: SEED-V, DEAP, and particularly GraffitiVR, which emphasizes its substantial advantages in complex VR environments.

In conclusion, the knowledge distillation framework proposed in this article marks a significant advancement in the field of emotion recognition by enabling enhanced model performance through the strategic integration of subject-specific and subject-independent data. The promising results obtained from rigorous testing on benchmark datasets, especially in VR settings with the GraffitiVR dataset, underscore the potential of knowledge distillation as a powerful tool for boosting the performance and generalization abilities of emotion recognition models. This research sets a new standard for future studies and paves the way for the practical application of multimodal emotion recognition technologies in diverse and immersive environments.

Moving forward, our future research will incorporate cutting-edge techniques such as attention mechanisms and reinforcement learning to refine the performance of multimodal emotion recognition models further. We also plan to explore the application of transfer learning and meta-learning strategies to boost generalization across various settings.

An extensive evaluation of our framework on larger and more diverse datasets, especially in VR contexts, will be essential to verify its scalability and robustness in real-world applications.

REFERENCES

[1] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. Da Silva, ''A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals,'' *IEEE Access*, vol. 7, pp. 140990–141020, 2019.

[2] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, ''A multimodal database for affect recognition and implicit tagging,'' *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[3] R. A. Calvo and S. D'Mello, ''Affect detection: An interdisciplinary review of models, methods, and their applications,'' *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[4] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, ''Emotion recognition in immersive virtual reality: From statistics to affective computing,'' *Sensors*, vol. 20, no. 18, p. 5163, Sep. 2020.

[5] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, ''DEAP: A database for emotion analysis; Using physiological signals,'' *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[6] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, ''Transfer learning: A friendly introduction,'' *J. Big Data*, vol. 9, no. 1, p. 102, Oct. 2022.

[7] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, ''A review on transfer learning approaches in brain–computer interface,'' in *Signal Processing and Machine Learning for Brain-Machine Interfaces*. London, U.K.: Institution of Engineering and Technology, 2018, pp. 81–98.

[8] H. A. Abbass, J. Tang, R. Amin, M. Ellejmi, and S. Kirby, ''Augmented cognition using real-time EEG-based adaptive strategies for air traffic control,'' in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 58. Los Angeles, CA, USA: SAGE, 2014, pp. 230–234.

[9] S. Saha and M. Baumert, ''Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: A review,'' *Frontiers Comput. Neurosci.*, vol. 13, p. 87, Jan. 2020.

[10] L. M. Zhao, X. Yan, and B. L. Lu, ''Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition,'' in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 863–870.

[11] G. Hinton, O. Vinyals, and J. Dean, ''Distilling the knowledge in a neural network,'' 2015, *arXiv:1503.02531*.

[12] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song, ''Can emotion be transferred—A review on transfer learning for EEG-based emotion recognition,'' *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 833–846, Sep. 2022.

[13] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, ''Multisource transfer learning for cross-subject EEG emotion recognition,'' *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Jul. 2020.

[14] B. Xue, Z. Lv, and J. Xue, ''Feature transfer learning in EEG-based emotion recognition,'' in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2020, pp. 3608–3611.

[15] W.-L. Zheng, Y.-Q. Zhang, J.-Y. Zhu, and B.-L. Lu, ''Transfer components between subjects for EEG-based emotion recognition,'' in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 917–922.

[16] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, ''Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination,'' *Frontiers Neurorobotics*, vol. 11, p. 19, Apr. 2017.

[17] Y. Ruan, M. Du, and T. Ni, ''Transfer discriminative dictionary pair learning approach for across-subject EEG emotion classification,'' *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 899983.

[18] Y.-P. Lin and T.-P. Jung, ''Improving EEG-based emotion classification using conditional transfer learning,'' *Frontiers Hum. Neurosci.*, vol. 11, p. 334, Jun. 2017.

[19] J. Li, H. Chen, and T. Cai, ''FOIT: Fast online instance transfer for improved EEG emotion recognition,'' in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2618–2625.

[20] Y. Zhong and Z. Jianhua, ''Subject-generic EEG feature selection for emotion classification via transfer recursive feature elimination,'' in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 11005–11010.

[21] X. Chai, Q. Wang, Y. Zhao, X. Liu, D. Liu, and O. Bai, ''Multi-subject subspace alignment for non-stationary EEG-based emotion recognition,'' *Technol. Health Care*, vol. 26, pp. 327–335, May 2018.

[22] G. Xiao, Y. Ma, C. Liu, and D. Jiang, ''A machine emotion transfer model for intelligent human-machine interaction based on group division,'' *Mech. Syst. Signal Process.*, vol. 142, Aug. 2020, Art. no. 106736.

[23] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, ''Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets,'' *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[24] R. Santana, L. Marti, and M. Zhang, ''GP-based methods for domain adaptation: Using brain decoding across subjects as a test-case,'' *Genetic Program. Evolvable Mach.*, vol. 20, no. 3, pp. 385–411, Sep. 2019.

[25] M. A. Sarikaya and G. Ince, ''Emotion recognition from EEG signals through one electrode device,'' in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.

[26] M. Yasemin, M. A. Sarikaya, and G. Ince, ''Emotional state estimation using sensor fusion of EEG and EDA,'' in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5609–5612.

[27] A. Olamat, P. Ozel, and S. Atasever, ''Deep learning methods for multi-channel EEG-based emotion recognition,'' *Int. J. Neural Syst.*, vol. 32, no. 5, May 2022, Art. no. 2250021.

[28] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, ''Learning invariant representations from EEG via adversarial inference,'' *IEEE Access*, vol. 8, pp. 27074–27085, 2020.

[29] P. Gong, Z. Jia, P. Wang, Y. Zhou, and D. Zhang, ''ASTDF-net: Attention-based spatial–temporal dual-stream fusion network for EEG-based emotion recognition,'' in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 883–892.

[30] P. Gong, P. Wang, Y. Zhou, and D. Zhang, ''A spiking neural network with adaptive graph convolution and LSTM for EEG-based brain-computer interfaces,'' *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1440–1450, 2023.

[31] Y. Dong, X. Chen, Y. Shen, M. Kwok-Po Ng, T. Qian, and S. Wang, ''Multi-modal mood reader: Pre-trained model empowers cross-subject emotion recognition,'' 2024, *arXiv:2405.19373*.

[32] W. Lu, H. Liu, H. Ma, T.-P. Tan, and L. Xia, ''Hybrid transfer learning strategy for cross-subject EEG emotion recognition,'' *Frontiers Hum. Neurosci.*, vol. 17, Nov. 2023, Art. no. 1280241.

[33] X.-Z. Zhang, W.-L. Zheng, and B.-L. Lu, ''EEG-based sleep quality evaluation with deep transfer learning,'' in *Proc. Int. Conf. Neural Inf. Process.* Guangzhou, China: Springer, 2017, pp. 543–552.

[34] S. Sidharth, A. A. Samuel, H. Ranjana, and J. T. Panachakel, ''Emotion detection from EEG using transfer learning,'' in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4.

[35] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, ''Domain adaptation for EEG emotion recognition based on latent representation similarity,'' *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 344–353, Jun. 2020.

[36] M. S. Aldayel, M. Ykhlef, and A. N. Al-Nafjan, ''Electroencephalogram-based preference prediction using deep transfer learning,'' *IEEE Access*, vol. 8, pp. 176818–176829, 2020.

[37] Z. Wang, Y. Wang, J. Zhang, Y. Tang, and Z. Pan, ''A lightweight domain adversarial neural network based on knowledge distillation for EEG-based cross-subject emotion recognition,'' 2023, *arXiv:2305.07446*.

[38] V. Joshi, S. Vijayarangan, S. P. Preejith, and M. Sivaprakasam, ''A deep knowledge distillation framework for EEG assisted enhancement of single-lead ECG based sleep staging,'' 2021, *arXiv:2112.07252*.

[39] P. Wang, C. Guo, S. Xie, X. Qiao, L. Mao, and X. Fu, ''EEG emotion recognition based on knowledge distillation optimized residual networks,'' in *Proc. IEEE 6th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Oct. 2022, pp. 574–581.

[40] Y. Liu, Z. Jia, and H. Wang, ''EmotionKD: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals,'' in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 6122–6131.

[41] G. Zhang and A. Etemad, ''Distilling EEG representations via capsules for affective computing,'' *Pattern Recognit. Lett.*, vol. 171, pp. 99–105, Jul. 2023.

[42] S. Zhang, C. Tang, and C. Guan, ''Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition,'' *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108833.

[43] D. Wu, J. Yang, and M. Sawan, ''Bridging the gap between patient-specific and patient-independent seizure prediction via knowledge distillation,'' *J. Neural Eng.*, vol. 19, no. 3, Jun. 2022, Art. no. 036035.

[44] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[45] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, “Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition,” *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022.

[46] T. Karakas, B. N. Dumlu, M. A. Sarikaya, D. Yildiz Ozkan, Y. Demir, and G. Ince, “The impact of urban graffiti with facial expressions on human behavioral and emotional experiences in a VR environment,” *Archnet-IJAR, Int. J. Architectural Res.*, vol. 18, no. 2, pp. 409–431, Jun. 2024.

[47] A. Jo and B. Y. Chae, “Introduction to real time user interaction in virtual reality powered by brain computer interface technology,” in *Proc. ACM SIGGRAPH Real-Time Live*, New York, NY, USA, Aug. 2020, p. 1.

[48] P. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Trans. Audio Electroacoustics*, vol. AE-15, no. 2, pp. 70–73, Jun. 1967.

[49] A. Bruhn, J. Weickert, and C. Schnörr, “Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods,” *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 1–21, Feb. 2005.

[50] W. O. A. S. Wan Ismail, M. Hanif, S. B. Mohamed, N. Hamzah, and Z. I. Rizman, “Human emotion detection via brain waves study by using electroencephalogram (EEG),” *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1005–1011, Dec. 2016.

[51] K. S. Park, H. Choi, K. J. Lee, J. Y. Lee, K. O. An, and E. J. Kim, “Emotion recognition based on the asymmetric left and right activation,” *Int. J. Med. Med. Sci.*, vol. 3, no. 6, pp. 201–209, 2011.

[52] D. T. Burley, N. S. Gray, and R. J. Snowden, “As far as the eye can see: Relationship between psychopathic traits and pupil response to affective stimuli,” *PLoS ONE*, vol. 12, no. 1, Jan. 2017, Art. no. e0167436.

[53] T. Partala and V. Surakka, “Pupil size variation as an indication of affective processing,” *Int. J. Hum.-Comput. Stud.*, vol. 59, nos. 1–2, pp. 185–198, Jul. 2003.

[54] A. K. Engel and P. Fries, “Beta-band oscillations—Signalling the status quo?” *Current Opinion Neurobiol.*, vol. 20, no. 2, pp. 156–165, 2010.

[55] T. A. Nguyen and Y. Zeng, “Analysis of design activities using EEG signals,” in *Proc. Int. Design Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, vol. 44137, 2010, pp. 277–286.

[56] T. Terkildsen and G. Makransky, “Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence,” *Int. J. Hum.-Comput. Stud.*, vol. 126, pp. 64–80, Jun. 2019.

[57] C. Başar-Eroglu, D. Strüber, M. Schürmann, M. Stadler, and E. Başar, “Gamma-band responses in the brain: A short review of psychophysiological correlates and functional significance,” *Int. J. Psychophysiol.*, vol. 24, nos. 1–2, pp. 101–112, Nov. 1996.

[58] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, “Stress assessment by prefrontal relative gamma,” *Frontiers Comput. Neurosci.*, vol. 10, p. 101, Sep. 2016.

[59] C. Krogmeier, B. S. Coventry, and C. Mousas, “Affective image sequence viewing in virtual reality theater environment: Frontal alpha asymmetry responses from mobile EEG,” *Frontiers Virtual Reality*, vol. 3, Jul. 2022, Art. no. 895487.

[60] T. Xue, A. E. Ali, G. Ding, and P. Cesar, “Investigating the relationship between momentary emotion self-reports and head and eye movements in HMD-based 360° VR video watching,” in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, vol. 338, May 2021, pp. 1–8.

[61] A. S. Won, B. Perone, M. Friend, and J. N. Bailenson, “Identifying anxiety through tracked head movements in a virtual classroom,” *Cyberpsychol., Behav., Social Netw.*, vol. 19, no. 6, pp. 380–387, Jun. 2016.

[62] Y. Ma, W. Zhao, M. Meng, Q. Zhang, Q. She, and J. Zhang, “Cross-subject emotion recognition based on domain similarity of EEG signal transfer learning,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 936–943, 2023.

[63] R. Ren, Y. Yang, and H. Ren, “EEG emotion recognition using multisource instance transfer learning framework,” in *Proc. Int. Conf. Image Process., Comput. Vis. Mach. Learn. (ICICML)*, Oct. 2022, pp. 192–196.

[64] Y. Li, B. Fu, F. Li, G. Shi, and W. Zheng, “A novel transferability attention neural network model for EEG emotion recognition,” *Neurocomputing*, vol. 447, pp. 92–101, Aug. 2021.

[65] T. Luo, J. Zhang, Y. Qiu, L. Zhang, Y. Hu, Z. Yu, and Z. Liang, “MDDD: Manifold-based domain adaptation with dynamic distribution for non-deep transfer learning in cross-subject and cross-session EEG-based emotion recognition,” 2024, *arXiv:2404.15615*.

[66] X. Gu, W. Cai, M. Gao, Y. Jiang, X. Ning, and P. Qian, “Multi-source domain transfer discriminative dictionary learning modeling for electroencephalogram-based emotion recognition,” *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 6, pp. 1604–1612, Dec. 2022.

[67] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, and S. Coleman, “Emotion recognition with convolutional neural network and EEG-based EFDMs,” *Neuropsychologia*, vol. 146, Sep. 2020, Art. no. 107506.

[68] Y. Wang, Q. Li, J. Jia, and R. Zhang, “A novel transfer learning model for cross-subject emotion recognition using EEGs,” in *Proc. 6th Int. Conf. Comput. Sci. Artif. Intell.*, Dec. 2022, pp. 217–223.

[69] H. Tang, G. Jiang, and Q. Wang, “Deep neural network for emotion recognition based on meta-transfer learning,” *IEEE Access*, vol. 10, pp. 78114–78122, 2022.

MEHMET ALI SARIKAYA received the B.S. and M.S. degrees in computer engineering from Istanbul Technical University, where he is currently pursuing the Ph.D. degree. His research interests include BCI-based emotion recognition and emotion recognition in VR environments.

GÖKHAN INCE received the B.S. degree in electrical engineering from Istanbul Technical University, Türkiye, in 2004, the M.S. degree in information engineering from Darmstadt University of Technology, Germany, in 2007, and the Ph.D. degree from the Department of Mechanical and Environmental Informatics, Tokyo Institute of Technology, Japan, in 2011. From 2006 to 2008, he was a Researcher at the Honda Research Institute Europe, Offenbach, Germany, and from 2008 to 2012, he was at the Honda Research Institute Japan Company Ltd., Saitama, Japan. Since 2012, he has been an Associate Professor with the Department of Computer Engineering, Istanbul Technical University. His current research interests include human–computer interaction, robotics, artificial intelligence, and signal processing.