ActDiffNet: Multisensor Affective State Recognition by Actively Synthesizing Minority Patterns via Conditional Diffusion

Jatin Chhabria*

Advised by Prof. Sreyasee Das Bhattacharjee
University at Buffalo, State University of New York

**This work is for fulfillment of credits for the master's project. The research paper has been approved for publishing with authors Vamsi Kumar Naidu Pallapothula, Sreyasee Das Bhattacharjee at IEEE/ACM CHASE 2025.*

Abstract

A key challenge in personalized ubiquitous healthcare is developing efficient wearable platforms that accurately classify biosignals while remaining adaptive to the evolving data patterns particularly highlighting individual's personal and other exterior context dynamics. However, several challenges plague machine learning applications involving biomedical signals, including limited data, imbalanced classes, difficulty accessing reliable annotated data, and noisy measurements. To this end, we propose an active learning model *ActDiffNet* for affective state recognition from multisensor signals that first leverage only a small annotated data collection to build an initial classifier, and later iteratively upgrade via a shortlisted set of synthesized 'hard' signals conditionally diffused by those unique signal patterns, on which the model has not been sufficiently trained yet. The proposed *ActDiffNet* converges faster, achieving comparable classification performance with 1-2 orders of magnitude fewer labeled samples than fully supervised approaches to attain a state-of-the-art accuracy of 78%. An effective *Context Conditioned Synthetic Signal Generation* module that employs multiple sensor-specific copies of the conditioned U-Net to facilitate synthesizing signals that closely mimic the sensor and class-specific patterns of shortlisted 'hard' signals within its generated outputs. Extensive evaluation using two public datasets WESAD and CASE reports outperformance (e.g., $1.5-3\%$ improved accuracy) of the proposed *ActDiffNet* against state-of-the-art supervised or self-supervised models while delivering a consistently robust generalization all across.

Keywords: Multisensor Signal, Wearable Device, Emotion Recognition, Data Augmentation, Diffusion

1 Introduction

In recent years, the research community has shown considerable interest in biosignal-based approaches to human affective state recognition [4, 5, 6, 18, 23, 31]. The interest extends much beyond Computer Science and emotion AI [16, 19, 27]. This recent surge of interest in physiological data collection has been driven by the advent of compact, user-friendly wearable devices, which are capable of measuring electrocardiogram (ECG), photoplethysmography (PPG), electrodermal activity (EDA), and skin temperature and have therefore facilitated a streamlined and accelerated process of real-time data collection. Studies show, that physiological signals, which originate from the autonomic nervous system (ANS), are often involuntary during certain emotional states, and thereby can provide more reliable information for emotion recognition than observable physical expressions [3, 33].

Despite these, the task of continual monitoring of multiple physiological signals often is presented with several pragmatic challenges. For example, often the acquired signals are noisy or may as well be missing due to users' personal lifestyle choices or environmental interferences. This may impact the system's inference performance. While most existing works do not consider such realistic challenges into algorithm building, ensuring satisfactory performance in

several such practical problem-settings continues to face manifold challenges: *First*, Emotional perception is subjective and may vary from person to person. For example, in WESAD dataset, subjects S14 and S17 seem to express their emotions relatively uniquely compared to other subjects who participated in the study. However, having a sufficiently large training collection that can deliver comprehensive representations of all such unique expressions for each emotion category may not always be guaranteed; *Secondly*, due to limited annotated samples, ensuring model generalization is another persistent concern; *Finally*, imbalances in datasets (e.g., The CASE dataset shows a severely skewed sample distribution in its three categories that has 4.44% negative samples; 81.2% neutral samples; and 14.36% positive samples) may pose as yet another serious challenge, especially for complex tasks like recognizing and tracking human emotions; a task that is inherently evolving as well as continually influenced by an individual's personal and other exterior context dynamics.

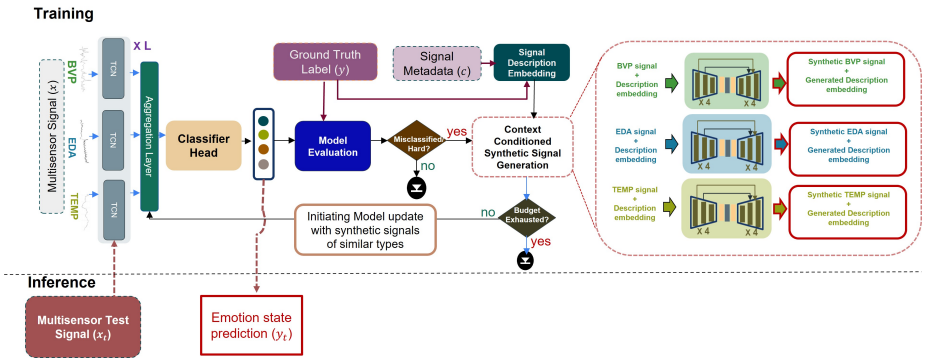

Figure 1: An overview of the proposed *ActDiffNet* model

To this end, we propose an active learning model *ActDiffNet* for affective state recognition from multiple biosensors that first leverages only a small annotated data collection to build an initial classifier and later iteratively upgrade via a collection of synthesized 'hard' signals conditionally diffused by those unique signal patterns, on which the model has not been sufficiently trained yet. The proposed multisensor signal augmentation approach is inspired by the recently proposed Stable Diffusion models[21], which report promising performance in the multi-channel image synthesis task. In this work, we adopt this approach in the process of multi-sensor biosignal generation wherein the class and instance-specific details described using a short text regulate the sensor-specific diffusion modules to generate signals that closely mimic the patterns of shortlisted 'hard' signals (i.e., signals which were not correctly classified by the existing version of classifier). To summarize, the primary contributions of the proposed *ActDiffNet* include:

1. *A multisensor affective state recognition framework* that is capable of quickly adapting to rare domain-specific context details using only a shortlisted set of signals, appears more effective in handling the challenges in an unbalanced training setting.

2. *An effective iterative multimodal training* approach that is particularly designed to capture instance-level 'hardness' both from mode-specific and multimodal perspectives within the learned model, demonstrates the power of continual model improvement within an evolving problem environment.

3. *Extensive evaluation* reporting outperformance of the proposed *ActDiffNet* against state-of-the-art supervised or self-supervised models in identifying various human mental health states like stress and arousal.

2 Related Works

Early emotion recognition research using physiological signals focused on task-specific feature engineering for effective decision-making. However, these methods have limited generalizability across different signals, and the task is further complicated by the high dimensionality of data from multiple wearable devices [10]. Consequently, recent research has explored deep learning models to predict emotional states from various physiological signals, including EEG, EMG, EDA, and ECG [2, 12, 13, 14, 20, 22, 26]. To capture broader temporal contexts, some studies have also proposed hybrid models for biosignal-based emotion analysis [9, 15].

Important to note that unlike traditional emotion recognition models that rely on potentially controllable facial, audio, and textual responses [1, 11], physiological signals are involuntary and less susceptible to manipulation for social or personal acceptability [37] and thereby have attracted significant attention from researchers in the last decade. However, due to the lack of a well-balanced data collection available to train a sophisticated and robust prediction system, relying exclusively on biosignals has always been challenging.

Recent studies have explored semi-supervised and self-supervised learning for emotion recognition [7, 28, 34]. These approaches leverage limited labeled data to define the classification task, while self-supervised contrastive learning facilitates pre-training. However, data from wearable devices is frequently noisy and intermittently missing, and existing models often fail to adequately address these practical challenges. Moreover, designing an optimized SSL framework in a multisensor data environment is still a challenge. Although a limited set of existing research[7, 34] propose SSL-based methods for multisensor signals, data imbalance continues to pose a challenge, which makes these methods less applicable in real-life problem scenarios.

To this end, we propose an active learning model *ActDiffNet* for multisensor biosignal-based affective state recognition that first, leverage a small annotated data collection to build an initial classifier, and later upgrade using multiple synthesized samples mimicking the sensor-specific data patterns observed in a shortlisted set of 'hard' signals. An effective context-conditioned diffusion mechanism that parallelly empowers *ActDiffNet* to learn multiple sensor-specific fine-grained data patterns while also enabling the model to preserve the high-level category details within the generated class-specific multi-sensor biosignals.

As proved with experimental validations, the backbone U-Net architecture is sufficiently generic and can seamlessly handle a variety of sensor-specific signal components without having to rely upon a large annotated and well-balanced multiclass training collection.

3 Proposed Method

Figure 1 gives an overview of the proposed multisensor fusion network *ActDiffNet* and this section introduces its four modules: *Signal Embedding*; *Multisensor Classifier*;*Active Learning-based Model Training*; and *Context Conditioned Synthetic Signal Generation* .

3.1 Problem Definition

Given the dataset $\mathcal{D}$ of multisensor signals, where each sample $x_i \in \mathcal{D}$ is represented by M different sensor-specific signal components and c_i is the metadata specific to x_i, the problem objective is to evaluate the emotional state y_i of the subject from whom the multisensor biosignal x_i is collected. In our work, signals generated from three sensors are utilized to represent the emotional state of an individual. These include: electrodermal activity (EDA), blood volume pressure (BVP), and skin temperature (TEMP). In particular, each x_i as a multisensor signal is presented as $x_i = \{s_i^1, s_i^2, \ldots, s_i^M\}$ and $s_i^m \in R^{N\times 1}$ represents a 1D time-domain signal from one of the M different sensors (in our work, $M = 3$), where N is the signal length. To ensure notational simplicity, unless sensor specification is required, we omit the superfix m in s_i^m and denote it as s_i instead.

3.2 Signal Embedding

Each normalized s_i is encoded using a Temporal Convolution Network (TCN) [17] to obtain its encoded representation $\mathbf{e}_i \in R^{N\times d_e}$, where $d_e >> 1$ is the embedding dimension determined by the chosen filter length in the last TCN layer of the network.

Adapted from classical convolutional neural networks (CNNs), temporal convolutional networks (TCNs) are specifically designed for sequence modeling. Their ability to capture long-range temporal dependencies and key signal dynamics make them superior to traditional recurrent neural networks (RNNs) for various time-series analysis tasks and thus well-suited as an encoder network for our purposes. Each input signal s_i is processed through a 1×1 convolutional layer followed by two dilated, causal convolutional layers, each with batch normalization, a non-linear activation, and dropout. Residual skip connections are employed at each layer to maintain consistent input/output dimensionality. Stacking multiple layers of residual blocks of TCN helps in building multi-level temporal contexts, as the dilation in each subsequent l^{th} block grows exponentially larger by the factor $d_l = 2^l - 1$ and the layer's convolution is defined as:

$$F_i = \sum_{j=0}^{k} f[j] s_i[t - d_l * j] \tag{1}$$

where f is a convolution filter of size k applied at time t. To preserve the sequence length, appropriate zero padding is applied at the beginning of the sequence. The output of the last TCN layer is fed into a pre-projection layer of dimension 128 to derive $\mathbf{e}_i$.

3.3 Multisensor Classifier

Given $x_i = \{\mathbf{e}_i^1, \mathbf{e}_i^2, \ldots, \mathbf{e}_i^M\}$, a multi-sensor signal represented by its sensor-specific TCN generated encoders, the compact multisensor representation of x_i is defined as $\mathbf{t}_i = \mathbf{e}_i^1 \oplus \mathbf{e}_i^2 \oplus \ldots \mathbf{e}_i^M$. The collection $\{\mathbf{t}_i, y_i\}$ is used to train the classifier head (θ), which is comprised of a three-layer perceptron with *GeLU* activation (followed by dropout) and we apply the *Cross Entropy classification loss* on the model's prediction output $P(\mathbf{t}_i|\theta)$ (where θ represents the classifier parameter) with ground truth label y_i as $\mathcal{L}_{CE} = CrossEntropyLoss\Big(P(\mathbf{t}_i|\theta), y_i\Big)$.

3.4 Active Learning-based Model Training

This paper adopts a classifier-agnostic active learning approach to address the challenge of unbalanced data collection, where only a limited number of 'hard' samples representing certain minority classes or other rare patterns are shortlisted for further explorations. In other words, it establishes an automated mechanism for selecting a subset of samples, the patterns of which the model requires revisiting during its later training phase. A simple yet effective uncertainty sampling mechanism is employed to identify ambiguous data points, wherein the 'hard' samples are selected using one or more of the following criteria: (1) samples that are misclassified by the present classifier; (2) selecting samples near the classifier's decision boundary; and (3) choosing samples where the difference in confidence scores for possible labels is low.

Intuitively, a low maximum confidence score for a sample suggests the model struggles to understand that type of data. Conversely, a small difference in class confidence scores indicates the model lacks the discriminative power to identify those data patterns effectively. Therefore, our interactive sample selection strategy emphasizes updating the model using these 'hard' data patterns. To ensure robustness, we use a high threshold to identify the most relevant ambiguous samples, which are tagged as 'hard'. Each of these shortlisted samples is then passed as input to the following *Context Conditioned Synthetic Signal Generation* module (described next) to create synthetic signals of similar types, which are later used to update the existing version of *ActDiffNet* end-to-end.

More precisely, the uncertainty scores for an unannotated sample $x_i \in \mathcal{D}$ is computed as follows:

$$U_1(x_i|\theta) = max(P(y_i^p = y_i|x), P(y_i^p \neq y_i|x))$$

$$U_2(x_i|\theta) = |P(y_i^p = y_i|x) - P(y_i^p \neq y_i|x))|$$

where y_i^p represents the predicted class label for x_i by the underlying classifier (θ). Across all experiments conducted, we have identified a sample (x_i) as 'hard' if $U_1(x_i|\theta) < \beta$ or $U_2(x_i|\theta) < \eta$ and we chose $\beta = 0.7$ and $\eta = 0.5$. At every iteration of the active learning, we utilize a smaller subset of top-K 'hard' samples selected using their total uncertainty score $U_1 + U_2$ to upgrade the existing classifier, and a total of 3 active learning iterations are performed in all experiments reported in this paper.

3.5 Context Conditioned Synthetic Signal Generation

The proposed context-conditioned synthetic signal generation module extends the conventional BioDiffusion model [21] by conditioning the model's unconditional framework with the BERT embedding ($\mathbf{C}_i$) of a comprehensive metadata description detailing the unique context (c_i) of the input signal (x_i). For example, a sample in a WESAD dataset is described using text context as *"This is an EDA signal classified as 0. The subject is a 27-year-old male with a height of 175 cm and a weight of 80 kg. The subject is right-handed. The signal has a mean of 0.77, a median of 0.79, and a standard deviation of 0.22. It ranges from a minimum of 0.39 to a maximum of 1.18. On the day of the study, the subject had no coffee, not within the last hour, was not in any sports activity, and is a non-smoker. The subject did not smoke within the last hour and reported feeling healthy on the day of the study."* The inclusion of this detailed description which also includes its class information, not only regulates the diffusion process but also allows for more targeted signal synthesis. Multiple sensor-specific copies of the U-Net module are designed, wherein each is specialized to synthesize particular sensor and class-specific patterns within its generated output signal. Given the signal representation provided by the datasets we used for our experiments, three sensor-specific signal-synthesizing U-nets are used for the multi-sensor biosignal generation process.

In particular, during *forward phase*, within a U-Net architecture (detailed next), each residual block is augmented with both the text embedding vector $\mathbf{C}_i$ and the ongoing diffusion timestep. In the *backward phase*, the diffusion model ingests noise drawn from a normal distribution, augmented with two additional inputs: a text description of the metadata requirement. For example, a sample from the CASE dataset is described as *"Signal from a participant in 30-34 age range, who is Female, recorded using BVP signal. Emotional state: 2 arousal, 2 valence. The signal has a mean of -0.01, median of -0.07, and standard deviation of 0.60. It ranges from a minimum of -1.16 to a maximum of 1.80."* and an example signal the pattern of which the model is required to mimic while synthesizing. Following this combination of condition inputs, a convolutional layer refines the result, ensuring it matches the original signal's structure. The rest of the reverse diffusion process focuses on removing the remaining noise and reconstructing a clean signal that resembles the original.

U-Net Architecture: The U-Net model designed as the backbone of the

proposed context-conditioned synthetic signal generation module is a convolutional neural network with an encoder-decoder architecture, designed specifically for effective signal processing. It features a symmetric architecture, consisting of two main paths: a contraction path (encoder) and an expansion path (decoder).

The encoder uses convolutional and max pooling layers to extract contextual information from the input signal. This compression reduces the input's dimensionality, enabling the model to learn inherent patterns and features. The architecture is composed of several blocks (in our work, we have used 4 encoder blocks), each with a convolutional layer followed by a residual block. These residual blocks help learn an identity function and prevent performance degradation as the network deepens. An attention layer follows each residual block, guiding the model to focus on the most important features for reconstruction.

The decoder pathway expands the compressed feature representation, enabling precise signal localization and reconstruction. In a U-Net, upsampling layers within the decoder increase the resolution of the bottleneck output. Each upsampling step is followed by a convolutional operation that generates high-resolution features. A key feature of the U-Net architecture is the use of skip connections, which concatenate feature maps from the encoder path to the corresponding decoder layers. This integration of high-level and low-level features facilitates accurate localization. By combining the generalized features from the contraction path with the detailed features from the expansion path, the network can produce a more precise reconstruction.

The U-Net architecture is central to both training and inference in the proposed signal synthesis module. At each time step, the signal is combined (concatenated) with embeddings representing that time step n, and potentially other conditional information like low-quality signal data or class labels. These embeddings provide context, guiding the diffusion process to generate the desired type of signal. More precisely, the U-Net is trained to reverse the diffusion process by learning to generate a less noisy signal at time $n-1$ from a more noisy signal at time n. This denoising process is repeated iteratively, stepping back from the fully noisy state at time N to the original, clean signal at time $n=0$. This reverse process mirrors the forward diffusion, allowing the model to reconstruct the original signal and complete the U-Net's training within the diffusion model framework.

Table 1: Distribution of Samples

Dataset	Task	Category (no. of samples)
WESAD	Emotion-3	baseline (58692), stress (33221), amusement (18584)
CASE	Valence-3	negative (3958), neutral (72283), positive (12785)
CASE	Arousal-3	low (2228), medium (75738), high (11060)

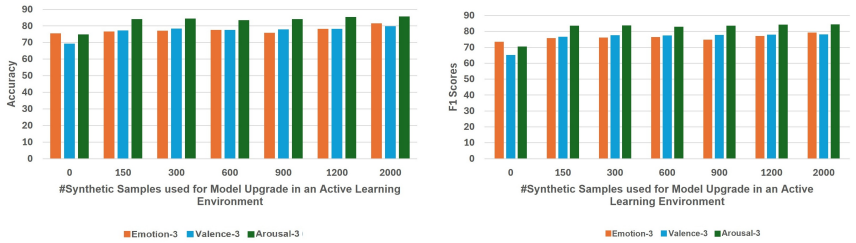

Figure 2: The Ablation analysis graphs show the performance improvement of the proposed *ActDiffNet* with the increased size of synthetic sample collections presented via active learning-based training.

4 Experiments

4.1 Datasets

We evaluate our model on the two largest publicly available biosignal-based affective datasets - CASE [30] and WESAD [29]. These datasets exhibit significant class imbalance across all tasks. The WESAD dataset is primarily composed of baseline samples, followed by stress samples, with amusement samples being the least frequent. The CASE dataset shows an even greater imbalance, with neutral samples heavily dominating both valence and arousal tasks, and negative valence and low arousal categories being significantly underrepresented. This substantial class imbalance poses a major challenge for building robust emotion recognition models by the existing methods and thereby prove themselves as the best testbeds for the proposed *ActDiffNet*.

4.2 Data Preprocessing

Following the preprocessing protocol followed by Wu et al. [35], CASE and WESAD dataset signals are collected through sensors with different sampling frequencies. To have a common sampling frequency, we have downsampled all signals in these datasets to 4Hz. Then they are segmented into 60s windows, with 99.5% overlap for WESAD and 99% overlap for CASE dataset. For segments with multiple labels, the majority label was chosen as the final label, consistent with previous work [8]. Z-score normalization, as described in [28], was applied to each subject's recording to reduce inter-subject variability in physiological responses. Table 1 shows the class label distribution for the WESAD and CASE datasets, highlighting the representation of each class. The observed significant class imbalance makes the multi-class classification task, the focus of this paper, particularly challenging.

The WESAD dataset categorizes physiological responses into three classes: amusement, stress, and baseline. The CASE dataset supports two distinct classification tasks: 1) classifying physiological signals into three valence levels (negative, neutral, positive), and 2) classifying the same signals into three arousal levels (low, medium, high). These datasets and tasks provide a comprehensive framework for investigating the complex relationships between physiologi-

cal signals and emotional states, advancing the fields of affective computing and physiological research.

Table 2: Performance Comparison of the proposed *ActDiffNet* network model with multiple state-of-the-art methods using the accuracy and F1-score metrics.

Dataset	Task	Method	Accuracy	F1
WESAD	Emotion-3	WESAD-Wrist [29]	75.21	64.12
		SimpDCNN [24]	78.3	74.59
		RF [29]	76.17	66.33
		LDA [29]	68.85	58.18
		SigRep [8]	78.13	77.35
		SSL [35]	78.7	75.98
		S&T[25]	69.84	73.86
		ActDiffNet (ours)	**81.66**	**79.30**
CASE	Valence-3	SSL [35]	78.99	76.66
		SimpDCNN [24]	59.2	51.95
		SigRep [8]	64.83	60.25
		MULT [32]	63.14	62.5
		CorrNet [36]	65.14	53.00
		S&T[25]	70.28	59.87
		ActDiffNet (ours)	**79.85**	**78.15**
CASE	Arousal-3	SSL[35]	85.38	82.63
		SimpDCNN [24]	56.8	53.85
		SigRep [8]	65.07	61.08
		MULT [32]	62.15	58.48
		CorrNet [36]	58.22	55.00
		S&T[25]	68.36	58.22
		ActDiffNet (ours)	**85.64**	**84.37**

4.3 Results

We measured model performance using both F1-score and Accuracy metrics, following established evaluation protocols. As outlined in Section 4.1, our analysis accounts for varying emotion/affective state categories across datasets. Consistent with baseline methodologies (e.g., [8, 29]), our primary evaluation used Leave-One-Subject-Out (LOSO) cross-validation where $P-1$ participants' data trains the model, with the P^{th} subject's data reserved for testing - repeated for all P participants. For the ablation study, we randomly segment the data collection subject-wise in a $3:1:1$ ratio, a test scenario that imposes a stricter evaluation condition as it uses data collected from only 60% of subjects for model training (contrasting with LOSO's progressive utilization of all but one participant per iteration), 20% validation, and the remaining 20% subjects for testing. The tables report the average categorical prediction accuracy and the average F1 score computed from the set of P iterations.

Comparative Study The proposed *ActDiffNet* is evaluated against multiple state-of-the-art baseline models, which include: SimpDCNN [24]; SigRep [8]; SSL[35]; MULT[32]; CorrNet [36]; and Sense & Learn framework (S&T)[25]. Table 2 reports the comparative performance of the proposed *ActDiffNet* against several state-of-the-art methods. Due to significant class imbalances present in both datasets, reporting only the Accuracy scores is not enough. There- fore, in the table, we report the performance details using both the Accuracy and the F1 scores. As observed in the table, compared to SSL [35], the best of the existing baseline methods, *ActDiffNet* consistently demonstrates outperformance across several experiment settings. In particular, *ActDiffNet* reports around 1.5% (and 2%) improved F1-score in Valence-3 (and Arousal-3) task in CASE dataset and around 3% (and 2%) improved *Accuracy* (and *F1* scores) in WESAD dataset.

5 Ablation Study:

To evaluate the robustness of *ActDiffNet* in handling a limited data environment, in this set of experiments, the initial version of the multisensor classifier is trained using only a smaller training set that contains 20% less samples compared to the existing methods used as baselines for this work. However, as observed in Figure 2, this initial classifier quickly upgrades itself via targeted finetuning on a small number of 'hard' samples. For example, compared to the best baseline SigRep [8], in Emotion-3 task, by using only 1200 additional synthesized samples (which in size forms approximately 2.25% of the total training collection used by the existing literature), the upgraded model reports a competitive state-of-the-art performance of accuracy (and F1-Score) of 78.18 (77.08). Finally, with 2000 additional synthesized samples (which in size forms approximately 3.8% of the total training collection used by the existing literature), the proposed *ActDiffNet* attains a competitive accuracy (and F1-Score) of 81.1% (79.1). A similar performance trend is also observed in Valence-3 and Arousal-3 tasks, where the model requires 900 (which in size forms approximately 1.67% of the total training collection used by the existing literature) samples to report a performance equivalent to its best-performing baseline SSL[35]. Finally again, with 2000 additional synthesized samples, the proposed *ActDiffNet* attains competitive performance metrics, which exceeds SSL[35], the best-performing baseline's performance by around $2-3\%$.

6 Conclusion

This paper presents an effective active learning-based framework for multisensor affective state recognition that effectively synthesizes minority patterns via sensor-specific diffusion modules, conditioned on a text-based description highlighting individual's personal and other exterior context dynamics. The proposed *ActDiffNet* achieves state-of-the-art performance by significantly outperforming existing models, while requiring a considerably smaller data collection

for training. Despite these advancements, due to its targeted finetuning on the identified 'hard samples' there is a risk of bias in the model's interpretation of these patterns. Future work includes adaptive thresholding and signal diversity analysis for enhanced system robustness.

7 Future Work

While our current approach demonstrates strong performance, there remains significant room for improvement. In future work, we plan to explore the integration of large language models (LLMs) for the fair and adaptive synthesis of biological signals. This will enable the model to better capture underrepresented and rare data patterns. The main goal is improving generalizability and robustness across diverse subjects and conditions. Additionally, we aim to investigate advanced self-supervised learning techniques and multi-modal fusion strategies to further enhance emotion recognition accuracy, particularly in low-resource or imbalanced settings.

References

[1] Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6090–6100.

[2] Kleanthis Avramidis, Dominika Kunc, Bartosz Perz, Kranti Adsul, Tiantian Feng, Przemysław Kazienko, Stanisław Saganowski, and Shrikanth Narayanan. 2024. Scaling representation learning from ubiquitous ecg with state-space models. *IEEE Journal of Biomedical and Health Informatics* (2024).

[3] Anubhav Bhatti, Behnam Behinaein, Paul Hungler, and Ali Etemad. 2024. Attx: Attentive cross-connections for fusion of wearable signals in emotion recognition. *ACM Transactions on Computing for Healthcare* 5, 3 (2024), 1–24.

[4] João Cálem, Catarina Moreira, and Joaquim Jorge. 2024. Intelligent systems in healthcare: A systematic survey of explainable user interfaces. *Computers in Biology and Medicine* 180 (2024), 108908.

[5] Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. 2024. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In *International Conference on Human-Computer Interaction (HCII)*.

[6] Manas Dave and Neil Patel. 2023. Artificial intelligence in healthcare and education. *British dental journal* 234, 10 (2023), 761–764.

[7] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.

[8] Vipula Dissanayake, Sachith Seneviratne, Rajib Rana, Elliott Wen, Tharindu Kaluarachchi, and Suranga Nanayakkara. 2022. Sigrep: Toward robust wearable emotion recognition with contrastive representation learning. *IEEE Access* 10 (2022), 18105–18120.

[9] Cunhang Fan, Heng Xie, Jianhua Tao, Yongwei Li, Guanxiong Pei, Taihao Li, and Zhao Lv. 2024. ICaps-ResLSTM: Improved capsule network and residual LSTM for EEG emotion recognition. *Biomedical Signal Processing and Control* 87 (2024), 105422.

[10] Hany Ferdinando, Tapio Seppänen, and Esko Alasaarela. 2016. Comparing features from ECG pattern and HRV analysis for emotion recognition system. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 1–6.

[11] Ming Guo, Wenrui Li, Chao Wang, Yuxin Ge, and Chongjun Wang. 2024. Smile: Spiking Multi-Modal Interactive Label-Guided Enhancement Network for Emotion Recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[12] Ean-Gyu Han, Tae-Koo Kang, and Myo-Taeg Lim. 2023. Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features. *Electronics* 12, 13 (2023), 2933.

[13] Fazheng Hou, Junjie Liu, Zhongli Bai, Zhiyi Yang, Jiayin Liu, Qiang Gao, and Yu Song. 2023. EEG-based emotion recognition for hearing impaired and normal individuals with residual feature pyramids network based on time–frequency–spatial features. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–11.

[14] Lam Huynh, Tri Nguyen, Thu Nguyen, Susanna Pirttikangas, and Pekka Siirtola. 2021. Stressnas: Affect state and stress detection using neural architecture search. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 121–125.

[15] Md Rabiul Islam, Mohammad Ali Moni, Md Milon Islam, Md Rashed-Al-Mahfuz, Md Saiful Islam, Md Kamrul Hasan, Md Sabir Hossain, Mohiuddin Ahmad, Shahadat Uddin, Akm Azad, et al. 2021. Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques. *IEEE Access* 9 (2021), 94601–94624.

[16] Karl LaFleur, Kaitlin Cassady, Alexander Doud, Kaleb Shades, Eitan Rogin, and Bin He. 2013. Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface. *Journal of neural engineering* 10, 4 (2013), 046003.

[17] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 47–54.

[18] Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and CKM Lee. 2023. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer Science Review* 48 (2023), 100545.

[19] Hongqi Li, Xiaoya Li, and José del R Millán. 2024. Noninvasive EEG-Based Intelligent Mobile Robots: A Systematic Review. *IEEE Transactions on Automation Science and Engineering* (2024).

[20] Joe Li, Peter Washington, et al. 2024. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: machine learning study. *JMIR AI* 3, 1 (2024), e52171.

[21] Xiaomin Li, Mykhailo Sakevych, Gentry Atkinson, and Vangelis Metsis. 2024. BioDiffusion: A Versatile Diffusion Model for Biomedical Signal Synthesis. *arXiv e-prints* (2024), arXiv–2401.

[22] Van-Tu Ninh, Manh-Duy Nguyen, Sinéad Smyth, Minh-Triet Tran, Graham Healy, Binh T Nguyen, and Cathal Gurrin. 2022. An improved subject-independent stress detection model applied to consumer-grade wearable devices. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 907–919.

[23] Guanxiong Pei, Qian Shang, Shizhen Hua, Taihao Li, and Jia Jin. 2024. EEG-based affective computing in virtual reality with a balancing of the computational efficiency and recognition accuracy. *Computers in Human Behavior* 152 (2024), 108085.

[24] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[25] Aaqib Saeed, Victor Ungureanu, and Beat Gfeller. 2021. Sense and learn: Self-supervision for omnipresent sensors. *Machine Learning with Applications* 6 (2021), 100152.

[26] Stanisław Saganowski, Bartosz Perz, Adam G Polak, and Przemysław Kazienko. 2022. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing* 14, 3 (2022), 1876–1897.

[27] SKB Sangeetha, Rajeswari Rajesh Immanuel, Sandeep Kumar Mathivanan, Jaehyuk Cho, and Sathishkumar Veerappampalayam Easwaramoorthy. 2024. An Empirical Analysis of Multimodal Affective Computing Approaches for Advancing Emotional Intelligence in Artificial Intelligence for Healthcare. *IEEE Access* (2024).

[28] Pritam Sarkar and Ali Etemad. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1541–1554.

[29] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.

[30] Karan Sharma, Claudio Castellini, Egon L Van Den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* 6, 1 (2019), 196.

[31] Fazli Subhan, Alina Mirza, Mazliham Bin Mohd Su'ud, Muhammad Mansoor Alam, Shibli Nisar, Usman Habib, and Muhammad Zubair Iqbal. 2023. AI-enabled wearable medical internet of things in healthcare system: A survey. *Applied Sciences* 13, 3 (2023), 1394.

[32] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[33] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* 83 (2022), 19–52.

[34] Chenwei Wu and Cheng Ding. 2024. Self-Supervised Learning for Biomedical Signal Processing: A Systematic Review on ECG and PPG Signals. *medRxiv* (2024), 2024–09.

[35] Yujin Wu, Mohamed Daoudi, and Ali Amad. 2023. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Transactions on Affective Computing* 15, 1 (2023), 157–172.

[36] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. Corrnet: Fine-grained emotion recognition for video watching using wearable physiological sensors. *Sensors* 21, 1 (2020), 52.

[37] M Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios J Hadjileontiadis, and Ahsan Khandoker. 2022. LSTM-modeling of emotion recognition using peripheral physiological signals in naturalistic conversations. *IEEE Journal of Biomedical and Health Informatics* 27, 2 (2022), 912–923.