M3Sense: Affect-Agnostic Multitask Representation Learning Using Multimodal Wearable Sensors

SIRAT SAMYOUN*, MD MOFIJUL ISLAM*, TARIQ IQBAL, and JOHN STANKOVIC, University of Virginia, USA

Modern smartwatches or wrist wearables having multiple physiological sensing modalities have emerged as a subtle way to detect different mental health conditions, such as anxiety, emotions, and stress. However, affect detection models depending on wrist sensors data often provide poor performance due to inconsistent or inaccurate signals and scarcity of labeled data representing a condition. Although learning representations based on the physiological similarities of the affective tasks offer a possibility to solve this problem, existing approaches fail to effectively generate representations that will work across these multiple tasks. Moreover, the problem becomes more challenging due to the large domain gap among these affective applications and the discrepancies among the multiple sensing modalities. We present M3Sense, a multi-task, multimodal representation learning framework that effectively learns the affect-agnostic physiological representations from limited labeled data and uses a novel domain alignment technique to utilize the unlabeled data from the other affective tasks to accurately detect these mental health conditions using wrist sensors only. We apply M3Sense to 3 mental health applications, and quantify the achieved performance boost compared to the state-of-the-art using extensive evaluations and ablation studies on publicly available and collected datasets. Moreover, we extensively investigate what combination of tasks and modalities aids in developing a robust Multitask Learning model for affect recognition. Our analysis shows that incorporating emotion detection in the learning models degrades the performance of anxiety and stress detection, whereas stress detection helps to boost the emotion detection performance. Our results also show that M3Sense provides consistent performance across all affective tasks and available modalities and also improves the performance of representation learning models on unseen affective tasks by 5% – 60%.

CCS Concepts: • **Computing methodologies** → *Multi-task learning*; • **Applied computing** → **Health informatics**; • **Human-centered computing** → *Ubiquitous and mobile computing theory, concepts and paradigms.*

Additional Key Words and Phrases: Multimodal Learning, Multitask Learning, Representation Learning, Affect Recognition, Domain Adaptation, Wearable Sensors, Mental Health, Health Informatics

ACM Reference Format:
Sirat Samyoun, Md Mofijul Islam, Tariq Iqbal, and John Stankovic. 2022. M3Sense: Affect-Agnostic Multitask Representation Learning Using Multimodal Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 73 (June 2022), 32 pages. https://doi.org/10.1145/3534600

1 INTRODUCTION

Every year nearly 1 billion people suffer from different forms of mental health related problems, such as stress, anxiety, and depression [70]. Given the chronic nature of these problems, this leads to increased suicides and morbidity and accounts for significant economic burden [70]. Most importantly, since the COVID-19 pandemic, there has been an alarming rise in these mental health related conditions worldwide [69, 70]. Accurate and

*Both authors contributed equally to this research.

Authors' address: Sirat Samyoun, samyoun@virginia.edu; Md Mofijul Islam, mi8uu@virginia.edu; Tariq Iqbal, tiqbal@virginia.edu; John Stankovic, stankovic@virginia.edu, University of Virginia, Charlottesville, Virginia, USA, 22903.

2474-9567/2022/6-ART73
https://doi.org/10.1145/3534600

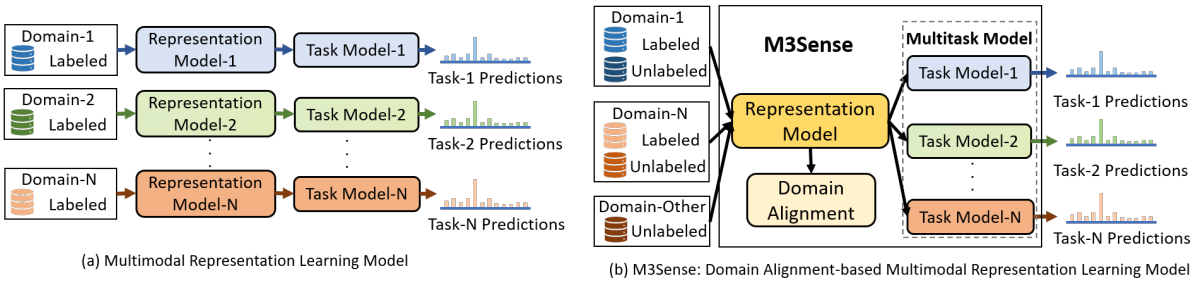


Fig. 1. (a) State-of-the-art multimodal representation models learn to extract task-specify representation to learn task models separately. These learning models require labeled data from a specific domain to train a particular task. (b) Our proposed Multitask Domain Alignment-based Multimodal Representation Learning Model (*M3Sense*) can use both labeled and unlabeled data to train the representation learning model to extract an affect-agnostic generalized representation. This domain alignment-based training method helps to transfer the model across tasks and domains.

early diagnosis of these conditions are crucial for better interventions and improved healthcare outcomes [69]. Several research works in affect recognition have focused on accurately detecting these conditions in the form of stress detection, anxiety detection, and emotion (happy, sad, angry, etc.) detection. Such solutions, also known as *affective applications*, are designed to detect mental health related conditions from the behavioral and physiological changes of a person by using facial expression captured via a video camera [95, 114], voice or speech captured via a microphone [63], or physiological sensing data from wearable sensors [72, 89], placed on the wrist, head or chest of the body.

Besides accurate detection, finding a subtle way to detect mental health conditions is important for better interventions. State-of-the-art works predominately use a video camera or a microphone [63, 98, 114] to detect these mental health conditions. However, placing these devices inside a room is highly privacy-invasive and involves substantial installation and processing costs [94, 103]. Thus, these approaches are not ubiquitous and limit usability in real-world settings. For example, a video camera or microphone based solution does not work when the user is outside. On the contrary, the wearable sensors are unobtrusive, privacy-preserving, and offer great ubiquity [25, 94]. Among the wearable sensors, the wrist-worn devices, including smartwatches, have emerged as a very popular and convenient way to monitor these mental health conditions [98].

Despite these significant benefits, *affective applications* suffer performance degradation when built using wrist physiological sensor data [7, 88, 106]. There are a couple of reasons behind this performance degradation. First, the wrist devices have a small form factor and they usually provide less accurate data compared to the chest or head based devices [90]. Although following the state-of-the-art works, we can use multimodal sensor data over unimodal data to improve the model robustness [29, 33, 34, 54, 78], but wrist sensors are heterogeneous in terms of sampling frequencies, and they primarily provide limited physiological signals with poor frequencies [23]. Thus, a learning model built with multiple wrist sensing modalities will struggle to detect the affective conditions accurately. Moreover, Islam and Iqbal [34] showed that the presence of different combinations of modalities often impacts the performance of a learning model. Therefore, to guarantee consistent performance for a task, a learning model needs to carefully attend to different modalities to extract robust multimodal representation.

Second, there is a scarcity of labeled data for mental health conditions [60, 98], because such conditions can only be annotated by experts, and in many cases, only a limited portion of the data represents a specific condition (e.g., stress, anxiety, or negative emotion) [86]. On the other hand, we can collect a huge quantity of unlabeled data with minimal human effort. However, as the state-of-the-art supervised learning approaches are predominately using labeled data, the unlabeled data has been under utilized. Moreover, the performance of the learning models

are highly dependent on the amount of training data. Thus, developing a learning framework which can utilize both the labeled data and unlabeled data will help to improve the performance of the model (Fig. 1), and will also reduce the human effort and the overhead to develop learning models for various affective applications.

Third, state-of-the-art multimodal representation learning models can not effectively extract generalized representations which are transferable across tasks. These learning models [9, 33, 34, 38, 44, 52, 54, 65] are trained to learn task-specific multimodal representations to improve the task performance which limits these models usability across tasks and domains (Fig. 1 (a)). Specifically, most of the learning models for affective applications are trained to learn affect-specific representations to improve the task performance. For example, a learning model for emotion recognition (a source task) is trained to learn the salient representation to recognize an emotion accurately. As a result, such learned representations will not effectively help to detect anxiety (target task), even though there is a physiological association between emotion and anxiety. Moreover, the performance of a model declines further if the source and target task domains are completely different [36]. For example, a model trained to learn stress-specific affective representation in the lab environment may be able to detect stress in a hospital environment, but it will not usually be effective in detecting anxiety in public speaking settings. Thus, a fundamental limitation of state-of-the-art multimodal representation learning models for affective applications is that the learned representations are not *affect-agnostic*, i.e, invariant to the affective applications. To the best of our knowledge, the state-of-the-art works do not consider the discrepancy between both task and domain characteristics to transfer a learning model trained on a source task to a target task from a different domain.

Fourth, although it is possible to develop a single and unified learning model to train multiple tasks, instead of developing individual models for multiple tasks, one task often dominates the learning process which leads to poor performance for other tasks [27, 28, 83, 97, 116]. This phenomenon is termed as negative knowledge transfer across tasks in the multitask learning model (MTL) [28, 97]. For example, Standley et al. showed that task-relatedness plays a crucial role in ensuring robust performance for all tasks in an MTL model [97]. Unrelated tasks can force an MTL model to learn over-generalized features leading to negative knowledge transfer across tasks [24, 27, 28, 31, 83, 97, 101, 116, 118]. Although several psychology studies showed the associations among affective tasks [3, 17, 25], task relatedness among the affective tasks has not been explored to develop a robust MTL model. Most importantly, the impact of various combinations of modalities and tasks has not been studied in developing an MTL model to ensure robust performance across tasks and domains.

Interestingly, several works in psychology have revealed that the different mental health conditions, such as, anxiety, stress, and emotions, have significant associations among each other [3, 17, 25]. For example, stress is considered as a primary indicator of anxiety disorders [18, 105]. Moreover, people during stressful periods experience different kinds of emotions [17, 20]. Often these mental health conditions trigger similar physiological response to the human body, such as, irregular heart rhythms, increased skin sensitivity, and chest pain [2, 17, 62]. Along this line, state-of-the-art works have shown that both stress and anxiety can be placed in the same region valence-arousal scale [11, 51, 112]. Despite these crucial insights of affective associations from several the psychology studies, no prior works consider the affective associations across tasks to extract generalized affect-agnostic representation to improve the task performance.

To address the above-mentioned challenges, we present a novel multimodal, multitask learning framework, called *M3Sense*. In *M3Sense*, we develop *Conditional Attention-based Multimodal Fusion* (*CAM*), where we fuse multimodal features conditioned on tasks by employing our proposed conditional attention model. To train our proposed multimodal representation learning model, we devise a Domain Alignment-based Multitask Learning Method by following the insights of the associations among the affective states from the aforementioned studies. This training method utilizes a novel *Domain Alignment Module* to align the distributions of multimodal representations across multiple tasks and domains. This method guides the representation learning model of *M3Sense* to effectively learn the affect-agnostic representations, which will work across tasks and domains.

Moreover, this method enables *M3Sense* to utilize both labeled and unlabeled data from both affective and non-affective domains to train a robust representation model in an end-to-end manner. Finally, the affect-agnostic representation is used to train multitask learning models to produce task-specific predictions.

We conducted extensive experimental evaluations to evaluate the performance of *M3Sense* on three affective tasks: anxiety, emotion, and stress detection. We also compared the performance of *M3Sense* with state-of-the-art multimodal representation learning models and handcrafted feature-based machine learning models. The experimental results suggest that *M3Sense* outperforms feature-based models and representation learning models by $9.6\% - 26.2\%$ and $0.6\% - 11.7\%$ across three evaluated tasks, respectively. Moreover, using ablation studies, we investigate and find out which task combinations lead to performance degradation for the affective tasks. For example, if we train *M3Sense* with anxiety and stress detection tasks, then the Top-1 accuracy of these tasks are 70.3% and 85.8%, respectively. However, if we introduce emotion in *M3Sense*, the Top-1 accuracy of anxiety and stress detection tasks degrades to 60.2% and 82.4%, respectively. Furthermore, we evaluated the impact of different combinations of modalities in affective tasks. The experimental evaluations suggest that *M3Sense* consistently outperforms across all the combinations of modalities. To the best of our knowledge, we are the first to investigate the impact of task and modality combinations in the multitask learning models.

Additionally, we evaluated the generalized representation learning capability of state-of-the-art learning models and *M3Sense* by pre-training a model for a set of tasks and fine-tune that model for an unseen task prediction with a few training samples. The experimental results suggest that the Top-1 accuracy of the state-of-the-art models degrades considerably on many unseen task-transfer settings across heterogeneous domains, whereas *M3Sense* can increase the accuracy of these models by approximately $5\% - 60\%$. Thus, *M3Sense* can help to train the representation learning models to extract affect-agnostic multimodal representations, which can be used for unseen tasks across heterogeneous domains. To the best of our knowledge, we are the first to develop a multimodal and multitask learning model which can be used across heterogeneous tasks and domains.

The key contributions of this work are:

- We develop a novel multitask learning framework, called *M3Sense*, which learns affect-agnostic representation from multimodal sensor data. *M3Sense* allows using both labeled and unlabeled data from heterogeneous affective and non-affective domains to train a learning model.
- We design a novel domain alignment-based training algorithm to train representation learning models to extract generalized representations, which can be used across multiple unseen tasks and domains.
- We extensively evaluate the performance of *M3Sense* in three different affective domains. Experimental results and ablation studies show that *M3Sense* outperforms handcrafted feature-based models and state-of-the-art representation learning models across all the evaluated tasks and domains based on three datasets.
- We present valuable insights on which combinations of tasks and modalities aid in developing a robust Multitask Learning (MTL) model. For example, stress detection helps to boost emotion detection performance in an MTL model, whereas emotion detection deteriorates the anxiety and stress detection performance.
- We show that the accuracy of state-of-the-art learning models degrades considerably on many unseen affective task transferring settings with different combination of modalities, whereas *M3Sense* improves the accuracy of these models by approximately $5\% - 60\%$.

2 RELATED WORK

Wearable-based affect recognition approaches: Wearable devices have emerged as the most privacy-preserving, unobtrusive, and ubiquitous way for affect recognition. Several past works [26, 93, 95] used EEG signals data collected from a head device, while the works in [58, 93] used different combinations of the physiological signals collected from the chest, such as ECG (or heart-rate), EMG signal, respiratory signals, and skin conductance. However, wearing a device on the chest or on the head is highly impractical for continuous monitoring, limiting

these solutions' effectiveness. On the contrary, several works [75, 76, 88, 100, 117] have used different combinations of the wrist physiological signals, such as the BVP/PPG, skin conductance, temperature, and fingertip oxygen saturation. However, these solutions require running independent models for detecting anxiety, emotion, and stress, which is a major hurdle for such wrist devices having limited battery and processing power.

Handcrafted feature-based traditional machine learning approaches: Several works in the literature have used traditional machine learning-based approaches for affective applications. Among these, the ensemble-based methods (e.g., Random Forest, Extra Trees) provided the best results [75, 88, 91], as such methods boost the performance by selecting the best features for the tasks (e.g., anxiety, emotion, stress). Among the other methods, the Support Vector Machine [100, 117] and Linear discriminatory Analysis [94] methods were popularly used. However, a significant limitation of all the traditional models is that they can not effectively capture the salient representations and the long-term dependencies among the sensing modalities. Therefore, they yield much lower performance than the deep models when built using wrist sensors.

Multimodal representation learning approaches: Multimodal representation learning approaches have produced state-of-the-art results for various applications [29, 78, 81], such as human activity recognition [34, 35, 38, 44, 54], gesture recognition [38, 115], video classification [16, 30, 40], image captioning [57, 110], and visual question answering [49, 56]. The wearable based affect recognition works have mostly used CNN [54, 79] and RNN [109][50][88] models to effectively learn the spatial and temporal representations. Moreover, attention-based mechanisms have been popularly integrated into such models [33, 54] to selectively focus on the task-specific important information present in the multiple modalities. However, there are several reasons why such approaches can not be effectively applied in multitask affect recognition. First, the representation learned by the models is not *affect-agnostic*. Therefore, these representations when used on an unseen task leads to poor performance. Second, the scarcity of expert-annotated data representing a condition (e.g., anxiety, emotion, stress) makes it difficult to handle the disparities among the modality distributions. Moreover, the wrist data signals are often of poor quality, which makes accurate detection challenging. Third, these approaches do not consider the associations among these affective domains, and therefore are incapable of building a unified framework for all domains.

Multitask learning approaches: Multitask learning (MTL) is a machine learning paradigm that aims to improve the performance of learning models by training multiple tasks jointly [83, 116]. Several multitask learning approaches have been proposed in the literature to improve the task performance by learning a shared generalized representation [24, 27, 31, 83, 101, 116, 118]. However, there are several reasons why the existing solutions can not be straightforwardly used for our problem. First, state-of-the-art MTL frameworks learn all the tasks simultaneously in a single model. Thus, all the tasks are available during the training phase. However, these MTL models are not designed to train for one set of tasks and use the learned representation to learn an unseen task. Second, the exiting MTL models assume the tasks are from related domains. However, the domain gap for the affective tasks is considerably high when the source and target domain tasks are entirely different distributions (e.g., from stress detection to anxiety detection), unlike same task under different environments (e.g., from detecting stress in the hospital environment to detecting stress in a home environment). Third, these MTL models do not employ any domain alignment technique, and thus, can not be applied to utilize labeled or unlabeled data from entirely different domains.

Representation learning approaches using unlabeled data: Utilizing unlabeled data have been widely studied in the literature [29, 53, 56, 66, 67, 77, 82], mostly in computer vision, speech recognition, and text mining fields. Such techniques learn the underlying representation from data by applying self-supervised or semi-supervised learning methods. However, in the affective domains, particularly the multitask and multimodal setting, no such techniques have been explored or evaluated. For example, these works can not be used to align unlabeled data from different modalities from the anxiety domain to the same in the stress domain or vice versa. *M3Sense* overcomes this significant limitation of the state-of-the-art, and can be applied to any affective domain using physiological signals in a multitask setting.

3 BACKGROUND

3.1 Affective Applications for Mental Health

Affect refers to the outward expression of the feeling of the mind, which is often considered as an umbrella term in mental health domains [92]. Affect recognition is an interdisciplinary field that deals with automatically recognizing and modeling different mental health conditions of the user [113]. In this paper, we focus on a set of affective applications from different aspects of mental health. We briefly discuss these applications below:

3.1.1 Stress detection: Stress is the state of being overwhelmed with physical or psychological pressure. It refers to the bodily changes induced by external events or conditions [25]. Stress leads to several health problems, such as, stroke, hypertension, depression [3]. Psychological stress can be detected using audio or visual modalities [114] or physiological sensor data [85, 89].

3.1.2 Anxiety detection: Anxiety is the reaction of human body to various adverse situations, which is accompanied by panic, fear, uneasiness. Scientists have often described anxiety as a reaction to stress, as they share very similar symptoms [25] [3]. Past works on anxiety detection have mostly used self-assessment screening [4], and physiological parameters based approaches [72, 117].

3.1.3 Emotion detection: Emotion is a state of mind that people experience in daily life. According to different theories of emotion [73, 84], there are *positive emotions* (happy and sad) as well as *negative emotions* (anger, fear, and frustration). Similar to stress and anxiety, existing emotion detection methods have used audio-visual modalities [114], and physiological sensor data [75, 76, 100].

3.2 Correlations among Affective Domains

Over the years, researchers in psychology have revealed interesting associations among stress, anxiety and emotions in different contexts. For example, the Circumplex model of affect, originally presented by Russel et al. [84], has shown that all human emotions can be interpreted using a scale of two dimension: valence and arousal. Valence refers to the positivity or negativity of an emotion, and arousal is a measure to the intensity or activation. Later on, psychologists have extended this model over the years, and found that anxiety and stress can be placed in the high valence and negative arousal region of this scale [11, 25, 51, 112], as they show very similar symptoms to many of the emotions. For example, anxiety is significantly associated with fear and panic [80], two negative emotions, while stress is often related to anger [68], another negative emotion. Moreover, works in [18, 105] observed that stress can positively indicate the symptoms of anxiety, while other studies showed that people feel a complex array of emotions during stress and anxiety periods [17, 19, 20]. Moreover, previous works [25] have justified the placement of stressed state in the Circumplex model of emotion, and also pointed out the similarities among different states of stress, anxiety and emotions.

While stress, anxiety and different emotions exhibit significant correlations in terms of behavioral and physiological response to the human body, they differ too. For example, stress is associated to increased heart rate variability (HRV) [41], while in many cases, anxiety is characterized by low heart rate variability [2]. This paper to exploit such remarkable associations and differences among these affective domains using physiological sensors.

3.3 Wearable-based Affect Recognition using Physiological Signals

The wearable devices, usually placed on the head, chest or wrist of the body, provide most ubiquitous, and privacy-preserving approach for affect recognition in the wild. For example, the Empatica E4 wristband [15] provides several physiological signals from the wrist, while the RespiBAN Professional device [74] provides several physiological signals from the chest. We briefly discuss such physiological modalities below.

- **EEG:** Electroencephalography (EEG) captures the electrical activity of the brain, and is associated with stress, anxiety, and emotions [26, 95]. It is captured from a wearable device placed on the head.
- **EDA:** Electrodermal activity (EDA) captures the electrical changes of the skin arising from the brain signals that are caused by any stimulation [46, 94]. EDA can be measured from the skin conductance electrodes present in both the wrist and the chest sensors.
- **ECG and BVP:** ECG refers to the Electrocardiogram signal measured from the chest and BVP represents the Blood Volume Pulse signal available from the wrist. Both ECG and BVP signals are used to calculate the heart-rate and the heart rate variability, which are vitals representing the mental health conditions [2, 41].
- **Respiration:** Previous studies found that human body respiration behavior changes (e.g., rapid breathing) due to affective states changes [107]. The respiratory signal can be captured from a chest device.
- **EMG:** Electromyography (EMG) refers to the electrical activity produced by the skeletal muscles. An EMG sensor is usually placed on the chest. Previous research have shown that different mental health disorders can lead to increased EMG in specific muscles of the body [58, 94].
- **Temperature:** Skin temperature can be measured by using wrist or chest sensors. Prior studies have shown the associations between skin temperature changes in response to emotions, stress or anxiety [94, 104].

While using a wrist device or a smartwatch is far more convenient than using a head or chest device in daily life settings, the wrist devices have very limited resources in terms of processing power and battery. Therefore, we aim to build a single wrist-based solution that will exploit the correlations among these domains, and will work across each of these affective tasks (e.g., stress detection, anxiety detection, emotion detection).

4 AFFECT-AGNOSTIC MULTIMODAL REPRESENTATION LEARNING

4.1 Problem Formulation

Based on the motivation and background, we formulate the key research goals of this paper, which are two-fold. First, we train a multimodal representation learning model to extract affect-agnostic representation by utilizing labeled ($D^L = (D^L_1, D^L_2, \ldots, D^L_{N^L})$) and unlabeled datasets ($D^U = (D^U_1, D^U_2, \ldots, D^U_{N^U})$) from heterogeneous domains. Second, using the affect-agnostic multimodal representation, we train a multitask model consisting of N affective tasks $T = (T_1, T_2, \ldots, T_N)$. Each data sample in a labeled or unlabeled dataset D^i consists of modalities M, where $M = (M_1, M_2, \ldots, M_K)$, and where K is the number of categories of wrist modalities. Each data sample comes from a domain, where a domain is represented by the context (e.g., public speaking, debating, or exercising) and the environment (e.g., real-life, laboratory) where the data acquisition was performed. Our goal is to accurately detect all these affective tasks by using the data from the modalities of M present in D^L and D^U, where each task T_i classifies each data sample to the task class labels.

4.2 Design Goals and Learning Framework Overview

We identify the key design goals for developing the proposed affect-agnostic representation learning framework:

- **Accurate affect recognition using multimodal wrist sensors:** The learning model should accurately detect multiple mental health conditions and should ensure ubiquitous usability by using wrist sensors only. To facilitate robust performance, it should utilize multimodal data across the affective tasks.
- **Bridging the gap among the affective tasks and domains:** The framework should be able to learn generalized representations from the data that will work well across heterogeneous affective tasks (e.g., detecting stress, emotion, and anxiety). Moreover, the performance should not degrade even if the trained solution is tested in other domains, i.e, different contexts and environments.
- **Minimize the burden on labeled data and utilize the unlabeled data:** The framework should be able to utilize the limited amount of labeled data available for each task from a particular domain. To ensure

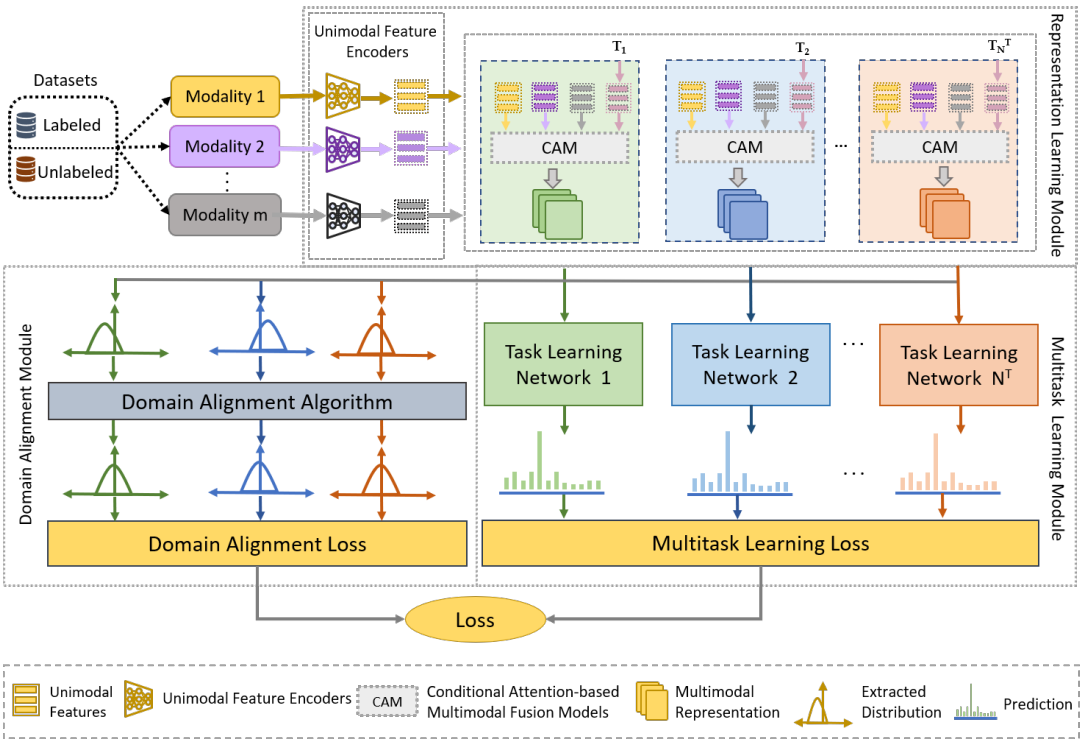


Fig. 2. Overview of our proposed affect-agnostic multitask representation learning framework, *M3Sense*. First, the *Representation Learning Module* uses different Unimodal Feature Encoders to extract unimodal representation from each modality independently. Similarly, it uses multiple Conditional Attention-based Multimodal Fusion Models to extract task-specific representations. *M3Sense* fuses these features to extract multimodal representation for each task. Second, the *Domain Alignment Module* extracts distributions from the multimodal representations for all tasks, which are aligned to guide the *Representation Learning Module* to learn affect-agnostic representation. Third, the *Multitask Learning Module* uses the extracted multimodal representation for multiple task prediction. Finally, *M3Sense* uses both the *Domain Alignment Loss* and *Multitask Learning Loss* to train multiple tasks with a single shared *Representation Learning Module* using both labeled and unlabeled data from heterogeneous tasks and domains simultaneously.

robust performance with limited labeled data, it should also have the capability to utilize the huge amount of unlabeled data from the heterogeneous tasks and domains.

- **Reduce the model size and complexity for wrist device efficiency:** To ensure ubiquitous usability of the learning model, a single representation learning model for multiple tasks is preferred instead of learning multiple models for each affective task. That way, we can compress the model size and reduce complexity to ensure smooth running performance on the resource-constrained wrist devices.

Based on these design goals, we develop *M3Sense*, a multitask, multimodal representation learning framework that effectively learns the affect-agnostic physiological representations to accurately detect affective conditions. The complete architecture of *M3Sense* is depicted in Fig. 2. *M3Sense* consists of three learning modules: *Representation Learning Module*, *Domain Alignment Module*, and *Multitask Learning Module*. We also design a Domain Alignment-based Multitask Training Algorithm to train these learning modules in an end-to-end manner.

- **Representation Learning Module:** This module learns the salient multimodal representations from the data for the affective tasks predictions. To achieve this, it sequentially applies the following two models.
  - The **Unimodal Feature Encoder (UFE)** extracts the modality-specific salient spatial-temporal representation by sequentially employing spatial and temporal feature encoders and unimodal attention modules. We have used different UFEs for different modalities to capture modality-specific feature distribution.
  - **Conditional Attention-based Multimodal Fusion (CAM)** uses a novel multimodal fusion model that extracts and fuses the task-specific salient representation using a conditional attention mechanism. Our proposed architecture enables *M3Sense* to learn multiple tasks simultaneously by reusing the representation learning module across tasks.
- **Domain Alignment Module**: To extract generalized multimodal representations, we have introduced a Domain Alignment Module. The reasoning behind incorporating this module in *M3Sense* is two fold. First, this module aligns the distributions of multimodal representations for multiple tasks to ensure the extraction of affect-agnostic representations. Second, this module enables *M3Sense* to utilize both labeled and unlabeled data from various domains to train a robust multimodal representation learning model.
- **Multitask Learning Module**: This module utilizes the extracted affect-agnostic generalized representation to learn the task-specific representations using a Task Learning Network (TLN) and produces predictions.

4.3 Representation Learning Module

M3Sense employs a *Representation Learning Module* to extract multimodal representations for each task $T_i \in T$, which consists of *Unimodal Feature Encoders (UFE)* and *Conditional Attention-based Multimodal Fusion Models (CAM)*. We have used different UFEs to extract modality-specific feature representation. Moreover, UFEs are shared among the tasks to transfer knowledge among the tasks. Similarly, *M3Sense* employs multiple CAMs to extract task-specific multimodal representations using our proposed conditional attention model:

$$X_{T_i} = E_{T_i}(E^u_m(X^r_m)) \quad , \quad T_i \in T \tag{1}$$

Here E_{T_i} is the task-specific multimodal feature encoder in *CAM* for task T_i and E^u is the shared unimodal feature encoder (u stands for unimodal and r stands for raw). Moreover, X_{T_i} is the output of representation learning module, which is a vector (tensor) of the extracted feature representation of task T_i.

4.3.1 Unimodal Feature Encoder (UFE):. State-of-the-art works have shown that each modality coming from the wearable sensors has unique physiological characteristics and distributions to detect these affective tasks [46, 59, 94]. To capture the diverse characteristic of the modalities, we design the *Unimodal Feature Encoders* by adopting a similar learning architecture proposed by Islam and Iqbal [33]. Each UFE employ a spatial-temporal feature encoder and a unimodal self-attention module to extract the unimodal representation from the data:

$$X^t_m = E^u_m(X^r_m) = E^a_m(E^t_m(E^s_m(X^r))) \quad , \quad m \in M \tag{2}$$

Here, E^u_m is the unimodal feature encoder for modality m, which consists of three sequential learning models: a spatial feature encoder (E^s_m), a spatial-temporal feature encoder (E^t_m), and a unimodal attention model (E^a_m) (u, s, t, and a stands for unimodal, spatial, spatial-temporal, and attention, respectively). We present the architecture of these learning modules in the subsequent section.

Spatial-temporal feature encoders: This part of UFE uses a combination of models to extract the spatial-temporal features from the raw sequential data $X^r = (X^r_1, X^r_2, \ldots, X^r_M)$. Even though the wrist sensors provide limited quantity of samples, this part of our solution ensures splitting data into subsequent windows and capturing the best spatial and temporal features present within the data windows. Here, $X^r_m = (x^r_{m,1}, x^r_{m,2}, \ldots, x^r_{m,L^r_m})$ and L^r_m is

the length of the raw feature sequence of modality m (r stands for raw, and raw feature is the sensor data before any processing or employing feature encoder).

First, the unimodal raw feature sequence is split with window size of S_m^w and stride size of S_m^{st} to produce segmented data, $X_m^{r'} = (x_{m,1}^{r'}, x_{m,2}^{r'}, \ldots, x_{m,1}^{r'}) \in R^{B \times L_m^s \times S_m^w \times S_m^r}$, where B is the Batch size, L_m^s is the total number of segment and S_m^r is the feature dimension of modality m (w stands for window and st stands for stride size).

Second, each segment is encoded using a spatial feature encoder (E_m^s) to capture the spatially encoded features sequence, $X_m^s = (x_{m,1}^s, x_{m,2}^s, \ldots, x_{m,1}^s) \in R^{B \times L_m^s \times S_m^s}$, here S_m^s is the dimension of the spatial feature (s stands for encoded spatial feature). As the spatial feature encoders extract features from each window independently and lack temporal features co-relation, these extracted features are referred to as spatial [33–35]. The primary reason for splitting the raw sensor data and extracting spatial features is to reduce the temporal feature dimension, as the temporal feature encoder, such as LSTM, suffers from extracting salient long-range features due to the vanishing gradient. This reduction of temporal feature dimension also reduce the model complexity. We have used a co-occurrence learning model [48] to design this encoder.

Third, we design a spatial-temporal feature encoder (E_m^t) to extract the spatial-temporal feature representation, $X_m^t = (x_{m,1}^t, x_{m,2}^t, \ldots, x_{m,1}^t) \in R^{B \times L_m^s \times S^u}$ from the encoded spatial feature X_m^s, where S^u is the spatial-temporal feature dimension (t stands for spatial-temporal feature, and u stands for unimodal). As recognizing affects involves capturing long-term feature correlations, we employed a Long Short-Term Memory (LSTM), a variant of recurrent neural network (RNN), to design this encoder. It must be noted that although the data from various affective domains are not temporally aligned among each other, this part of *M3Sense* automatically extracts the best temporal features that represent a specific affective condition. Past works in the literature have considered the minimum duration for affective changes as 5 seconds [71, 91]. In our solution, the window size S_m^w and batch size B are chosen in a way that captures the temporal features of a specific condition within this minimum duration which provides the best performance.

Unimodal self-attention model: Although the aforementioned feature encoders capture the long-range spatial-temporal features from raw sensor data, they can not effectively learn the sparse salient features from the unimodal feature sequence [5, 33, 34, 54, 61]. State-of-the-art sequence learning models [33, 34, 52, 54], specifically natural language representation models [5, 14, 49, 56, 61, 102], extensively used attention models to sparsely weight different parts of temporal features for extracting salient representations. Recently, several attention models have been proposed to extract salient representations from multimodal sensor data. For example, Long et al. proposed a lightweight attention model, called Keyless, to extract salient unimodal representations, which are then concatenated to produce a multimodal representation [54]. Moreover, Islam et al. proposed a multimodal attention model to prioritize the modalities for extracting multimodal representations [33]. Following the insights from these works, we design a unimodal attention model, E_m^a, that uses a self-attention mechanism to extract the unimodal representation, $X_m^a \in R^{B \times S^u}$, from encoded spatial-temporal features X_m^t in the following way:

$$X_m^a = E_m^a(X_{m,i}^t) = \sum_{i=1}^{L_m^s} \alpha_{m,i} X_{m,i}^t \quad , \quad m \in M \tag{3}$$

Here the attention weights $\alpha_{m,i}$ are calculated as follows,

$$\beta_{m,i} = W_m^{a^T} X_{m,i}^t \tag{4}$$

$$\alpha_{m,i} = \frac{exp(\beta_{m,i})}{\sum_i^{L_m^s} exp(\beta_{m,i})} \tag{5}$$

Here W_m^a is the modality-specific learnable parameters. Finally, the attended unimodal representations are combined to produce a unimodal feature representation set, $X^u = (X_1^a, X_2^a, \ldots, X_K^a) \in R^{B \times K \times S^u}$.

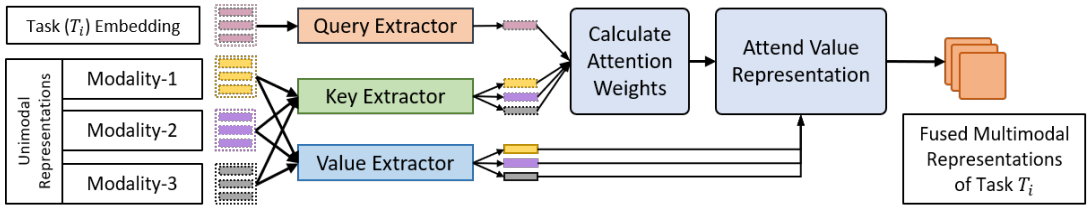


Fig. 3. *Conditional Attention-based Multimodal Fusion* (*CAM*). The CAM extracts key and value representations from the unimodal representations. CAM also extracts task (T_i) embedding, which is used to calculate the attention weights to calibrate value representations. The attention representation are fused (summed) to produce multimodal representations of task T_i.

4.3.2 Conditional Attention-based Multimodal Fusion (CAM):. *M3Sense* integrates a novel *Conditional Attention-based Multimodal Fusion* model, called CAM, that fuses the multimodal features to generate robust multimodal representations from the sensor data (Fig. 3). Unlike the task-specific state-of-the-art models, which require training from scratch, CAM allows *M3Sense* to train multiple tasks by reusing the same representation model. This architecture reduces the memory and computational complexity compared to the state-of-the-art models. CAM helps to prioritize the modalities and extracts task-specific salient multimodal representations using conditional attention mechanism:

$$X_{T_i} = E_{T_i}(X^u) \quad , \quad T_i \in T \tag{6}$$

First, CAM projects each one-hot task identifier vector T_i from a task set $T = T_1, T_2, \ldots, T_{N^T}$ to an embedding, Q_{T_i} (N^T is the total number of tasks and T stands for task):

$$Q_{T_i} = T_i W^q \tag{7}$$

Here W^q are the learnable parameters (q stands for query). Q_{T_i} represents the task representation which CAM uses to query the unimodal features, X^u, to extract task-specific salient representations.

Second, it embeds the unimodal features, X^u, to produce key (K^u) and value (V^u) feature vector representations in the following way:

$$K^u = X^u W^K \tag{8}$$
$$V^u = X^u W^V \tag{9}$$

Here, W^K and W^V are learnable parameters for key and vector representation projections, respectively.

Finally, for each task $T_i \in T$, CAM uses a task-specific representation, Q_{T_i}, as prior to query the key and value representations of unimodal features to fuse and extract the multimodal representation, X_{T_i}:

$$X'_{T_i} = \sigma\left(\frac{Q_{T_i} {K^u}^T}{\sqrt{D^u}}\right) V^u \qquad , T_i \in T \tag{10}$$
$$X_{T_i} = W^f_{T_i} X'_{T_i} \qquad , T_i \in T \tag{11}$$

Here, $W^f_{T_i}$ is a learnable projection parameter to project fused multimodal representation, X_{T_i}. We use this multimodal representation in the subsequent parts of *M3Sense*.

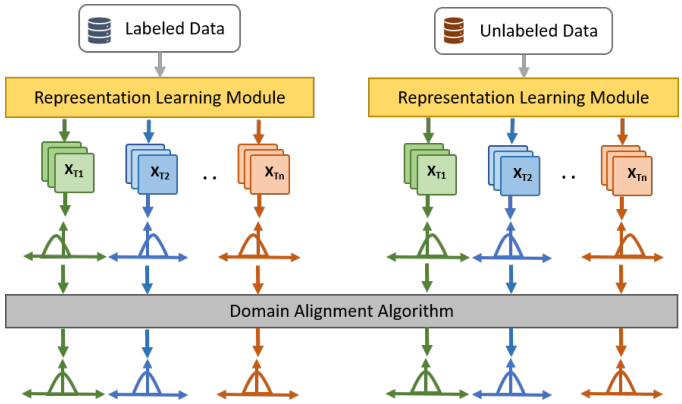


Fig. 4. Distribution-based *Domain Alignment Module* in *M3Sense*. It extracts distributions of multimodal representations for each task on labeled and unlabeled training samples. Our proposed *Domain Alignment Algorithm* aligns these distributions to guide the model during the training phase to learn affect-agnostic multimodal representation.

4.4 Domain Alignment Module

M3Sense uses *CAM* to learn task-specific salient multimodal representations from data. However, these representations are not affect-agnostic, and therefore not generalized across tasks. Thus, these representations are not beneficial for unseen tasks, which were not present during the training phase. To extract generalized multimodal representations, we design a *Domain Alignment Module* in *M3Sense*, which utilizes the distributions of multimodal representations for all tasks $T_i \in T$ across domains (Fig. 4). Specifically, it aligns the distribution across heterogeneous domains in a multitask setting (e.g., aligning a modality's distribution from anxiety and emotion domains to stress domains), which is the novelty of this module. Moreover, this module is capable of utilizing both labeled datasets ($D^L = (D^L_1, D^L_1, \ldots, D^L_{N^L})$) and unlabeled datasets ($D^U = (D^U_1, D^U_1, \ldots, D^U_{N^U})$) from heterogeneous domains (L stands for labeled, and U stands for unlabeled data) during training. Here, N^L and N^U are the numbers of labeled and unlabeled domains.

First, the *Domain Alignment Module* uses the extracted multimodal representation for each task from the previous step (Section 4.3, Eq. 1), X_{T_i}, to produce a posterior distribution for each task representation:

$$\alpha_{T_i} \sim (\mu_{T_i}, \sigma_{T_i}) = q_\theta((\mu_{T_i}, \sigma_{T_i})|E_{T_i}(E^u(X^r))) \quad , \quad T_i \in T, \;\; X^r \in D = (D^L \cup D^U) \tag{12}$$

Finally, it calculates the KL-divergence loss between the posterior distributions of multimodal representations for all tasks and a referenced prior distribution:

$$L_{align} = \sum_{X^r r \in D} \sum_{T_i \in T} D_{KL}\left[q_\theta(\alpha_{T_i}|E_{T_i}(E^u(X^r)))||p(\beta)\right] \quad , \quad X^r \in D = (D^L \cup D^U) \tag{13}$$

Here, $p(\beta)$ is the prior reference distribution, with respect to which the extracted distribution is aligned. We model $p(\beta)$ as a Normal distribution, $\mathcal{N}(\mu_\beta, \sigma_\beta)$ with zero mean and unit standard deviation, following the re-parameterization trick proposed by Kingma and Welling [42]. Prior works on variational inference have used normal distribution as a prior as it allows the use of the re-parameterization trick and provides an analytical evaluation of the KL-divergence objective [6, 13, 42]. We use a distribution recognition neural network q with parameters θ to obtain μ_{T_i} and σ_{T_i}. In the re-parameterization trick, a random variable $\epsilon \sim \mathcal{N}(0, 1)$ is sampled

Algorithm 1: Domain Alignment-based Multitask Training Method

Input: T: Task list, X^r: Raw features, M: Modalities, D^L: Labeled datasets, D^U: Unlabeled datasets
Output: Affect-agnostic learning model

for $epoch \leftarrow 1$ *to* N^e **do**
  $L_{align}, L_{multitask} \leftarrow 0, 0$ ▷ Initialize the domain alignment and multitask learning losses to zero
  for $X^r \in D^L$ **do**
    ▷ Sample a batch of data from the labeled datasets
    for $T_i \in T$ **do**
      $X_{T_i} = E_{T_i}(E^u(X^r_m))$ ▷ Extract multimodal representation for each task (Eq. 1)
      $\alpha_{T_i} \sim (\mu_{T_i}, \sigma_{T_i}) = q_\theta((\mu_{T_i}, \sigma_{T_i})|E_{T_i}(E^u(X^r)))$ ▷ Encode posterior distribution (Eq.12)
      $L_{align} \leftarrow L_{align} + D_{KL}\left[q_\theta(\alpha_{T_i}|E_{T_i}(E^u(X^r)))||p(\beta)\right]$ ▷ Calculate domain alignment loss (Eq. 13)
      $\hat{y}_{T_i} = F_{T_i}(X_{T_i})$ ▷ Produce the task prediction (Eq. 14)
      $L_{multitask} \leftarrow L_{multitask} + \frac{1}{B}\sum_{j=1}^{B} y_{T_i} \log \hat{y}_{T_i}$ ▷ Calculate the multitask prediction loss (Eq. 15)
    end
  end
  for $X^r \in D^U$ **do**
    ▷ Sample a batch of data from the unlabeled datasets
    for $T_i \in T$ **do**
      $X_{T_i} = E_{T_i}(E^u(X^r_m))$ ▷ Extract multimodal representation for each task (Eq. 1)
      $\alpha_{T_i} \sim (\mu_{T_i}, \sigma_{T_i}) = q_\theta((\mu_{T_i}, \sigma_{T_i})|E_{T_i}(E^u(X^r)))$ ▷ Encode posterior distribution (Eq.12)
      $L_{align} \leftarrow L_{align} + D_{KL}\left[q_\theta(\alpha_{T_i}|E_{T_i}(E^u(X^r)))||p(\beta)\right]$ ▷ Calculate domain alignment loss (Eq. 13)
    end
  end
  $L = L_{multitask} + \gamma_{align} \times L_{align}$ ▷ Calculate training loss (Eq. 16)
  Backpropagate the learning model to minimize the training loss L
end
return Affect-agnostic learning model

and multiplied by the μ_{T_i} and σ_{T_i}. The recognition neural network with the re-parameterization trick allows the end-to-end training of the representation learning model of *M3Sense* and back propagate through the distributions.

4.5 Multitask Learning Module

The third and final component of *M3Sense* is a *Multitask Learning Module*. It utilizes the learned affect-agnostic representation, X_{T_i} to produce the prediction of each task $T_i \in T$. This module integrates a *Task Learning Network* (TLN) that uses a neural network having fully-connected layers followed by a Softmax activation to predict the classification probability for each task:

$$\hat{y}_{T_i} = F_{T_i}(X_{T_i}) \quad , \quad X_{T_i} = E_{T_i}(E^u(X^r)), \;\; T_i \in T, \;\; X^r \in D^L \tag{14}$$

Here, F_{T_i} is the *Task Learning Network* for task T_i. Finally, we use the predictions, $\hat{y}_{T_i}$, for all tasks to calculate the multitask learning loss $L_{multitask}$, where y_{T_i} is the ground-truth label, and B is the batch size.

$$L_{multitask}(y, \hat{y}) = \frac{1}{B}\sum_{j=1}^{B}\sum_{T_i \in T} y_{T_i} \log \hat{y}_{T_i} \tag{15}$$

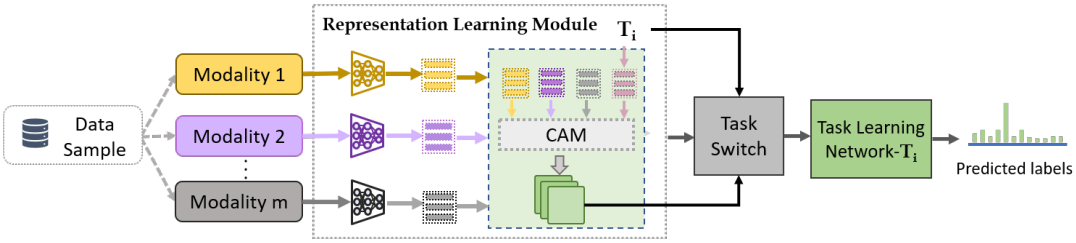


Fig. 5. Inference pipeline of *M3Sense*. *M3Sense* uses *Representation Learning Module* to extract multimodal representation for a particular task. *Task Switch Module* activates a particular *Task Learning Network* to produce predictions for the intended task. However, we can also extract representations and produce predictions for multiple tasks in parallel by skipping the *Task Switch Module* in *M3Sense*.

4.6 Domain Alignment-based Multitask Training Algorithm

Several state-of-the-art works have designed multimodal representation learning models to train a specific task from a particular domain [16, 30, 33, 34, 38, 44, 54, 108]. However, by design, the learned representations of these models are domain-specific, and therefore, can not effectively transfer knowledge from a task T_i in the domain D_i to another task T_j in the domain D_j.

To overcome the above-mentioned issue, we devise a novel algorithm that extracts the affect-agnostic multimodal representation for multiple tasks from heterogeneous domains. The representation learned using this algorithm in one domain (e.g., emotions while debating in a laboratory environment) is generalized and can be effectively used to train multiple tasks across a heterogeneous domain (e.g., anxiety while public speaking in a classroom environment). The procedure of our proposed training method is presented in Algorithm 1. This algorithm consists of two steps. First, it extracts the distribution from the fused multimodal representation to calculate the domain alignment loss, L_{align} (Section 4.4, Eq. 13). This loss guides the model to extract affect-agnostic representation across tasks and domains from both the labeled and unlabeled data. Second, *M3Sense* utilizes the *Multitask Learning Module* and labeled data samples to calculate the task-specific prediction loss, $L_{multitask}$ (Section 4.5, Eq. 15). This loss is used to train the *Task Learning Network* to produce predictions. Finally, the algorithm combines these two losses, and trains the learning model in an end-to-end manner to extract affect-agnostic multimodal representation and improve the task performance:

$$L = L_{multitask} + \gamma_{align} \times L_{align} \tag{16}$$

Here, γ_{align} is the weight to the domain alignment loss, which are chosen based on the network performance. We sample a batch of data from either labeled or unlabeled data samples during training. If the training batch of data contains unlabeled data, then we only calculate L_{align} loss and back-propagate to update the model parameters. Otherwise, we calculate both the L_{align} and $L_{multitask}$ loss and update the model accordingly. Domain alignment loss aims to extract similar representations for all the tasks for a particular domain. This task-based supervision allows utilizing the unlabeled data without aligning the labels across tasks. Testing data split only contains labeled data samples to evaluate the model performance for the affect recognition tasks.

4.7 Task Inference in *M3Sense*

As mentioned before, *M3Sense* utilizes three learning modules to train the representation learning model for multiple tasks using both labeled and unlabeled data. However, all these modules are not needed in the inference phase. The inference architecture of *M3Sense* is depicted in Fig. 5, which involves three modifications compared

Table 1. Training and evaluation datasets.

Datasets	Application	Labels used	Modalities used	Subjects
VerBIO [111]	Anxiety	High or no anxiety	EDA, BVP, TEMP	55
WESAD [91]	Stress	Stress, Amusement, Baseline	EDA, BVP, TEMP	15
K-EmoCon [71]	Emotion	High valence and arousal, others	EDA, BVP, TEMP	32
MMSDN [32]	Stress, Mental workload	-	EDA, BVP, TEMP	15
MAYA	Physical Activity	-	EDA, BVP, TEMP	12

to the training architecture. First, the *Domain Alignment Module* is removed, as we no longer need to align the representation across domains and tasks for the predictions. Second, the multimodal representation is for a particular task is extracted in the *Representation Learning Module*. However, if predictions for multiple tasks are desired, it extracts the representations for multiple tasks in parallel. Third, a *Task Switch Module* is introduced, which activates a particular *Task Learning Network* to generate task-specific predictions. During inference, *M3Sense* first takes data samples from the wrist sensors and extracts a multimodal representation for a particular task T_i. Next, it uses the task switch to activate a particular *Task Learning Network* to produce predictions for the intended task. However, if predictions for all the tasks is desired, *M3Sense* skips the *Task Switch Module*. This design offers flexibility to choose tasks for inference. Moreover, as *M3Sense* is able to *Task Learning Network* for multiple tasks together, it reduces the inference time too.

5 EXPERIMENTAL SETUP

5.1 Datasets

We have used five different datasets to train and evaluate *M3Sense*: VerBIO [111], WESAD [91], K-EmoCon [71], MMSDN [32], and a new collected dataset, called MAYA. Among these, MMSDN and MAYA were used as unlabeled datasets for training only, the others were used for both training and evaluation. We summarize the datasets in Table 1.

5.1.1 Anxiety Dataset: VerBIO dataset [111] is a multimodal bio-behavioral dataset that explores an individuals' anxiety during public speaking in real-life environments. This dataset provides the audio recordings and the physiological signals captured using a wristband from total 55 participants while performing class presentations. The participants were given self-report questionnaires to obtain their state-based anxiety during presentation and personality-based anxiety using the popular STAI (State and Trait Anxiety) scale [96, 112].

5.1.2 Stress Dataset: WESAD [91] is a stress and affect detection dataset that provides physiological data from the wearable sensors placed on the wrist and the chest from 15 participants. The participants underwent the Trier Social Stress Test (TSST) [43] consisting of mental load tasks as well as the neutral, amusement and recovery phases during the study. Next, self-report questionnaires were used to obtain the ground-truth annotations. Overall, the dataset provides three kinds of labels, namely stress, amusement, and baseline.

5.1.3 Emotion Dataset: K-EmoCon [71] is publicly available dataset for emotion detection in natural conversations. This dataset contains synchronised physiological signals recorded from the chest, wrist and the head sensors from 32 participants, during 16 paired debate sessions. The ground-truth annotations are provided on a valence-arousal scale, a popular model of emotion [84], along with the ratings for different emotion categories. The annotations were provided by the participants themselves, as well as by external raters.

5.1.4 Unlabeled Dataset - 1: MMSDN dataset, presented in [32] is a multi-modal affective domain dataset which aimed for detecting the stressful situations and mental workload of overall 15 nurses in a hospital. The acquisition

of the dataset was performed in a natural working environment during their long working shifts during the COVID-19 pandemic. Following the shifts, the participants filled out a survey, where they self-reported the levels of different kinds of stress, mental workload, and physical workloads during the shifts.

5.1.5 Unlabeled Dataset - 2: As mentioned before, *M3Sense* integrates a *Domain Alignment Module* which is able to utilize the unlabeled data from the other affective domains as well as from the non-affective domains that involves any changes of physiological response to the body. Previous works in the literature have shown significant correlations of physiological parameters changes during physical exercise [8, 99]. Inspired by these works, we collect a new dataset, named MAYA from a completely different and non-affective domain that contains the multimodal physiological data for different physical activities in daily life from 12 participants.

5.2 MAYA Dataset Collection and Study Protocol:

For this study, we chose a set of basic arm-based exercises, which are commonly recommended by experts for healthy individuals as well as for patients to ensure arm mobility in daily life, such as, *Shoulder Flexion*, *Shoulder Abduction*, *Wrist Flexion*, *Wrist Extension*, *Elbow Flexion*. Each exercise consists of a set of steps which involves movement of different limbs of the arm (e.g., elbow, shoulder or wrist). Data was collected using the Empatica E4 wristband. All the sessions were recorded by a video camera to obtain the ground-truth annotations. Overall, 12 subjects (57% men, 43% women) were included. The participants were healthy individuals, aged 26-39 years, with an average age of 33 years. The study protocol involved 3 phases: the *baseline phase*, the *exercise phase*, and the *resting phase*. We discuss each of these phases below, which are also demonstrated in Fig. 6:

5.2.1 Baseline phase: First the participants were asked to wear the watch in their own ways. Then they were shown 5 instructional videos showing each of the physical exercises. The videos were chosen by physical therapists and these demonstrated how to perform each step of the exercise correctly. Afterwards, the participants performed 2 practice sessions to make themselves comfortable with the exercises. On average, this phase took 20 minutes.

5.2.2 Physical exercise phase: The baseline phase helped the participants to achieve a physiological baseline. Next, each participant began performing the exercise sessions. During a session, each participant performed 5 exercises, one after another. Overall, each participant performed 10 sessions of each exercise. The average duration length for each participants exercise phase was 7 minutes.

5.2.3 Resting phase: In this phase, the participants were allowed to rest for 15 minutes to get back to their usual physiological state. Then each participant was provided self-report questionnaires that included three categories of questions. First, they were asked to rate the physical workload level and the intensity level on a scale of 1 to 10. Second, they were asked to report if they faced any pain or inconveniences during the exercises. Third, they were also asked to rate different phases of the data collection.

5.3 Data Preparation

Since *M3Sense* is designed to be a single learning framework for multiple mental health applications, the datasets were combined and processed in a similar manner. First the data from different modalities were split into several windows, and then synchronized into combined files. Each row of a combined file indicates a data window that is assigned a class label having a value between 0 to the number of classes for the affective application, an identifier for the application, and another identifier for the participant. The window length was chosen as 20 seconds, based on the previous works [10, 25, 94]. The signals from the wrist devices are often accompanied by noise and motion artifacts, mostly due to attachment of the sensors. To handle such artifacts from the BVP(PPG), EDA and

Fig. 6. Three phases of the MAYA dataset collection and study protocol: (a) *Baseline phase*- the participant is preparing for the exercise phase and watching the video of an exercise. (b) *Exercise phase*- the participant is performing an exercise. (c) *Resting phase*- the participant resting after a session.

TEMP signals, and thus to improve the quality of these signals, we applied a combination of a low pass filter and a finite impulse response filter, similar to the techniques described in [88].

For the VerbIO dataset, we utilized the sessions that involved data collection in real-life settings in the context of public-speaking anxiety. Following previous work [39], we assigned two labels for anxiety from the STAI scale, one for no or low anxiety (scores between $20-37$) and high anxiety (scores between 37–higher). For the WESAD dataset, it already contained labels from 0 to 3, no additional processing was required. For the K-EmoCon dataset, we assigned the labels based on the valence and arousal scale ratings. Label 1 was assigned for the data samples having high valence (value greater than 2) and high arousal (value greater than 2), label 0 was assigned otherwise. For the unlabeled MMSDN dataset, we only used data from one randomly chosen session for each participant. For the MAYA dataset, we utilized the data for the physical exercise phase only, as this phase represents the physiological changes. For the unlabeled data samples, we used -1 as labels to avoid conflicts.

Following this step, two sets of data were prepared from each of the combined data files for a modality: a training-validation set (35% of the overall samples), and a testing set (55% of the overall samples). The remaining data samples were used as unlabeled, with a label -1. The training-validation set was split into training and validation sets with a split of $75\% - 25\%$. To ensure fair evaluation, a stratified splitting technique was followed that ensured that data samples for each class for each application appeared in these sets in the same ratio.

5.4 Implementation Details of Learning Models

5.4.1 Training Architecture of M3Sense : To implement and train *M3Sense* using the datasets, the data from the wrist physiological modalities were segmented with a segment size of 5 and a stride size of 5. Following the implementation of *Unimodal Feature Encoder* (UFE) proposed by [33], we implemented that of *M3Sense*. We used a co-occurrence learning model [48] to implement the spatial encoder. The implementation consists of a two-layer CNN (Convolutional Neural Network). The CNN has 64 and 32 channels with the kernel sizes of (1×1) and (3×3), respectively. In CNN, the role of kernels is to pool features at a different level of abstraction. Small kernels (1×1) can pool local features and large kernels (3×3) can pool global features. The feature embedding size of each encoded spatial unimodal feature for each modality was 128. Following this step, we applied batch normalization to standardize the input layers, ReLU-activation to allow non-linear activation in the learning model, and dropout layers (having a probability of 0.3) to regularize the learning model and prevent overfitting during the training. These values were chosen based on our experimental evaluation and the implementation details of prior works on representation learning [33, 34, 48] to ensure reproducibility of the experimental evaluations.

To implement the temporal feature extraction model, we used an LSTM (Long Short Term Memory) network with a hidden feature size of 128. ReLU-activation and dropout layers (with a probability of 0.1) were used. To implement the self-attention based module, a one dimensional Convolutional layer was utilized, along with

batch normalization, ReLU-activation, and dropout layers (having a probability of 0.3). To implement the CAM module, we used our proposed conditional attention-based multimodal fusion model. Following the fusion, batch normalization, ReLU-activation, and dropout layers (having a probability of 0.3) were applied. Lastly, these output features were passed through two fully connected layers, having a ReLU activation. To implement the *Task Learning Network*, we used a fully connected neural network followed by ReLU-activation. The resulting representations were passed through a Softmax layer to produce class labels probabilities for a particular affective task, which provides $L_{multitask}$ using Cross Entropy Loss. Additionally, following the previous works [6, 13, 42], we used the Normal distribution as a reference to calculate the domain alignment loss L_{align}. We empirically set the domain alignment loss weight $\gamma_{align} = 0.3$, which helped to train the model.

5.4.2 Training Architecture of Baseline Models: For performance comparison, we implemented two sets of baseline models. We discuss the implementation and details of these models below.

Handcrafted features-based machine learning models: Previous works on affect recognition have successfully applied different traditional machine learning methods [75, 88, 91, 100]. We implemented 6 most commonly used classifiers among these works for performance comparison. These are, *Random Forest*, *Decision Tree*, *Extra Trees*, *Linear Regression*, *Linear discriminatory Analysis*, and *Support Vector Machine*. The tree-based models (e.g., *Random Forest*, *Extra Trees*) were implemented with 100 number of estimator trees, and with Gini impurity based information gain. The *Decision Tree* implementation also used Gini-based gains, and followed the best splitting strategy at each node. The *Linear Regression* implementation used a linear classifier with L2-regularization, while the *Support Vector Machine* implementation used a classifier with the Radial Basis Function (RBF) kernel, based on previous works [100, 117]. For the *Linear discriminatory Analysis* implementation, the singular value decomposition strategy was used. It must be noted that these parameters and options provided the best results for the training environment, and were chosen accordingly. For all of these implementations, the physiological signals were split into windows of size 20 seconds (described in Section 5.3). Next we computed different handcrafted statistical features (e.g., Mean, Standard Deviation, Maximum, Minimum) from the windows. We used a 10-fold cross validation strategy to avoid overfitting by the models. The average scores were chosen across the runs.

Deep Multimodal Representation Learning Models: We compared the performance of *M3Sense* with two baselines and two state-of-the-art multimodal representation learning models, which are as follows.

- ***Non-Attention:*** This baseline model uses a *Unimodal Feature Encoder*, similar to the learning architecture of *M3Sense* (Section 4.3.1), except the self-attention learning model was removed. The spatial-temporal feature encoder was kept to extract unimodal representation from data. Finally, the extracted representations are summed to produce a fused multimodal representation, which is used for learning a specific task.
- ***Multimodal-Attention:*** Similar to the Non-Attention baseline model, this baseline model uses a spatial-temporal feature encoder without a unimodal attention model. However, it utilizes a multimodal attention model, adopted from the learning architecture of Keyless [54]. The implementation consists of a 1-D CNN model that calculates the attention weights that are used to attend the feature from different modalities to produce a multimodal representation, which is used for a particular task learning.
- ***Keyless* [54]:** Keyless is a state-of-the-art multimodal representation learning approach that uses a light-weight attention model to extract unimodal features [54]. We implement this model by employing a *Unimodal Feature Encoder* having a spatial-temporal feature encoder and a self-attention model. It follows a similar architecture to the UFE of *M3Sense*.
- ***HAMLET* [33]:** HAMLET is another state-of-the-art multimodal representation learning model for a single task learning, which employs a unimodal attention model to extract salient unimodal feature and then an attention approach to fuse multimodal representation. It also utilizes a *Unimodal Feature Encoder* architecture, similar to that of *M3Sense*. HAMLET leveraged the transformer-style self-attention model [102] in designing both unimodal and multimodal attention model.

5.4.3 M3Sense Variants: We develop three variants of *M3Sense*, based on the two state-of-the-art models, and a baseline model. We replace the CAM in *Representation Learning Module* of *M3Sense* with three baseline models: Non-Attention, Keyless [54], and HAMLET [33]. Unlike the baseline models, the *M3Sense* variants share the multimodal representation across the tasks using the *Multitask Learning Module*, and also use the Domain Alignment Algorithm to generate affect-agnostic representation to improve the task performance.

- ***M3Sense(Non-Attention):*** In this variant, we remove the unimodal self-attention model from the *Representation Learning Module* of *M3Sense* (Section 4.3) and summed the extracted spatial-temporal feature to produce unimodal representation. Moreover, we remove the CAM from *M3Sense* and summed the unimodal representation to produce multimodal representation representation. This fused representation is used in all task-specific networks of *M3Sense* to produce multiple task predictions.
- ***M3Sense(Keyless):*** In this variant, unimodal feature encoder implementation is similar to the *Unimodal Feature Encoder* of *M3Sense*. However, CAM is removed from *M3Sense*, and the extracted unimodal representations are summed and used for task learning.
- ***M3Sense(HAMLET):*** This variant utilizes a *Unimodal Feature Encoder* is similar to that of *M3Sense*. However, we replace CAM in *M3Sense* with multimodal attention based fusion approach from HAMLET [33]. This approach fuses the multimodal representation which is used by *Task Learning Network* of *M3Sense*. Similar to the other baseline variant of *M3Sense*, each *Task Learning Network* shared the same fused representation.

5.5 Training Environment

We used PyTorch deep learning framework to implement the learning models of *M3Sense* and baseline models. An Adam optimizer with weight regularization and cosine annealing warm restarts [55] were used to train all the learning models. The initial leaning rate was set to $3e^{-4}$. The cycle length (T_0) and the cycle multiplier (T_{mult}) were set to 30 and 2, respectively. As we are training the learning models with multiple modalities for multiple tasks on GPUs with limited memory, we set the batch size 2. We trained all the learning models, including *M3Sense* and baselines, for 65 epochs. To ensure reproducibility, we used a fixed random seed (333) in the PyTorch-Lightning framework. Finally, the models were trained in a distributed manner on a GPU cluster environment with each each cluster node having $1-4$ GPUs from the following set of GPU models: $P100$, $V100$, $RTX-2080$, and $RTX-6000$.

6 RESULTS AND DISCUSSION

6.1 Comparison with Multimodal Learning Models

We compared the performance of *M3Sense* with the aforementioned baseline models. The experimental results are presented in Table 2 and described below.

6.1.1 Results: The experimental results suggest that *M3Sense* outperformed all the features-based machine learning models and the state-of-the-art multimodal representation models across all the tasks by achieving the highest Top-1 accuracy in anxiety: 71.5%, emotion: 75.2%, and stress: 87.0% detection tasks (Table 2). For the anxiety detection task, the best performing variants of *M3Sense* (*M3Sense*(Non-Attention)) outperformed the best performing feature-based models (*Extra Trees*) and the multimodal representation model (*HAMLET* [33]) by 9.6% and 1.3%, respectively. For the emotion recognition task, the best performing models of *M3Sense* (*M3Sense*(CAM)) outperformed the best performing feature-based models (*Support Vector Machine*) and the multimodal representation model (*HAMLET* [33]) by 26.2% and 11.7%, respectively. For the stress detection task, the best performing variants of *M3Sense* (*M3Sense(Keyless)*) outperformed the best performing feature-based models (*Extra Trees*) and the multimodal representation learning model (*HAMLET* [33]) by 15.5% and 0.6%, respectively.

Table 2. Performance comparison of multimodal learning models

Approach	Learning Models	Task (Top-1 Accuracy (%))		
		Anxiety	Emotion	Stress
Handcrafted Feature-Based Machine Learning	Random Forest	60.0	46.9	69.6
	Decision Tree	50.5	43.0	61.8
	Extra Trees	61.9	43.9	71.5
	Linear discriminatory Analysis	37.6	45.1	66.1
	Linear Regression	40.0	48.1	65.3
	Support Vector Machine	48.0	49.0	67.2
Deep Multimodal Representation Learning	Non-Attention	60.3	62.0	82.8
	Multimodal Attention	60.7	63.5	86.1
	Keyless [54]	64.0	63.5	82.8
	HAMLET [33]	70.2	63.5	86.4
M3Sense	M3Sense (Non-Attention)	**71.5**	75.2	81.5
	M3Sense (Keyless)	66.7	75.2	**87.0**
	M3Sense (HAMLET)	71.5	75.2	81.5
	M3Sense (*CAM*)	70.3	**75.2**	85.7

Are handcrafted feature-based models suitable for affective tasks? The results in Table 2 suggest that handcrafted feature-based machine learning models can not extract task-specific salient features for the affective tasks. There is a considerable performance gap between the best-performing machine learning model and the deep multimodal representation learning model. The reasoning behind this performance gap is such feature-based models depend on the manual selection of features which do not help to train a generalized model to achieve considerably better performance on the unseen data samples. Moreover, these models can not effectively capture the temporal and spatial correlations in the data that represents these affective conditions.

Are the deep models and the attention approach helpful? Our results strongly indicate that all the deep multimodal representation learning models, including *M3Sense*, outperformed the handcrafted feature-based machine learning models by a good margin across all the tasks. Our explanation is these deep models extract generalized feature representation for a particular task, which leads to better performance on unseen data samples. However, the performance degrades, if an attention method is not included in the architecture. For example, the Non-Attention model, which does not use the attention method under-performs than the other baseline models across all the tasks. Thus, the learning model architecture plays a crucial role in improving task performance. Additionally, to the best of our knowledge, we are the first to conduct ablation studies to evaluate the impact of various attention mechanisms for affective tasks. The results clearly show that for all attention mechanisms, when *M3Sense* is applied to the representation learning models, the task performance improves consistently. Thus, *M3Sense* provides a generalized learning framework that helps to improve the performance of multimodal learning models, regardless of which attention mechanism is used for multimodal fusion.

Can *M3Sense* further improve the performance? The results (Table 2) show that our framework, *M3Sense* outperforms both the deep multimodal representation learning models and the handcrafted features-based models across all the affective tasks. In particular, the performance of the deep models is considerably lower for some tasks (e.g., 62.0% − 63.5% accuracy in the emotion detection task), while all variants of *M3Sense* performs consistently well (e.g., 75.2%). We also observe that the Non-Attention learning model does not outperform the handcrafted features-based models for anxiety detection, however, *M3Sense* can guide Non-Attention model to learn generalized representations and improve the performance across all the tasks. Notably, the Non-Attention

Table 4. Impact of task and domain variations in *M3Sense*. We evaluate which affective tasks should be learned together in multitask learning model.

Task Combinations	Context Combinations	Environment Combinations	Tasks (Top-1 Accuracy (%))		
			Anxiety	Emotion	Stress
Anxiety Emotion Stress	Presentation Debate Mental load tasks	Real-life Real-life Laboratory	60.2	**75.2**	82.7
Anxiety Emotion	Presentation Debate	Real-life Real-life	63.5	63.5	-
Anxiety Stress	Presentation Mental load tasks	Real-life Laboratory	**70.3**	-	**85.8**
Emotion Stress	Debate Mental load tasks	Real-life Laboratory	-	75.2	82.4

model with *M3Sense* outperforms baseline learning models for anxiety detection. Moreover, *M3Sense* helps to improve all the evaluated baseline models across all the tasks by achieving the highest top-1 accuracy. The reasoning behind this performance improvement is that *M3Sense* uses our proposed domain alignment approach that guides the learning model to extract affect-agnostic representation, which benefits all tasks. Additionally, it utilizes both the labeled and unlabeled data which helps the learning models to learn generalized representations across tasks and domains to ensure robust performance in the multitask learning setting.

Moreover, as mentioned before we have used only wrist sensors with a very low sampling frequency to evaluate the multimodal learning models. For example, for a stressful condition having a duration 60 seconds, the Empatica E4 wristband [15] will provide $4Hz \times 60 = 240$ samples of EDA only, while a respiBAN chest strap [74] will provide $700Hz \times 60 = 42000$ samples. Our results showing superior performance over the state-of-the-art representation learning models suggest that the combination of CAM and the *Domain Alignment Module* helps to extract the useful information from these fewer samples from wrist devices.

6.2 Which Affective Tasks Should be Learned Together in Multitask Learning Model?

We experimentally analyze what combinations of tasks and domains are suitable for the affective applications in the multitask learning settings.We trained separate multitask models with various combinations of tasks, while keeping the modality combination same: BVP, EDA, and TEMP. We developed three baselines multitask learning models by incorporating state-of-the-art multimodal representation learning models into *M3Sense*. Table 4 presents the results. We discuss our findings from the results below.

Detecting stress facilitates emotion detection regardless of domains: The experimental results suggest that incorporating stress detection in the multitask learning model helps to improve the performance of emotion detection, regardless of the context and environment. For example, for the emotion recognition task, a combination of anxiety and emotion provides 63.5% Top-1 accuracy, while the introduction of stress improves the accuracy to 75.2%. Thus, stress detection in a multitask learning model helps to regularize the representations to improve the performance of emotion detection. It must be noted that these results show the relevance with the findings of psychology studies that stress is usually considered as a cause, while emotions often result from stress [3, 19]. We also notice that the performance of our models slightly changes with the change of the context or environment; rather, it is dependent on the variations of tasks. For example, stress data were collected in the laboratory in the context of mentally stressful tasks; however, the learned generalized representations are used for emotion detection, which was performed in different real-life conversational contexts.

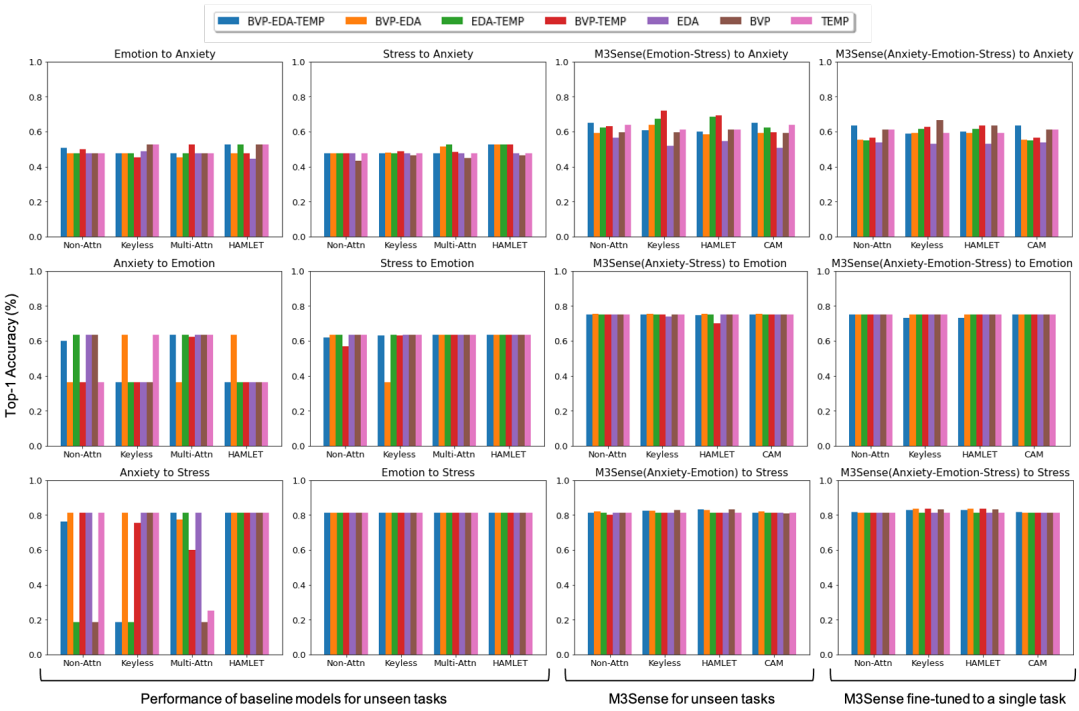


Fig. 7. Experimental evaluations of state-of-the-art multimodal representation learning model and *M3Sense* for unseen tasks. Here, *M3Sense([Task-Set]) to [Task]* in the figure title denotes that *M3Sense* is pre-trained for *[Task-Set]* and fine-tuned for *[Task]*. (Non-Attn: Non-Attention, Multi-Attn: Multimodal-Attention)

Emotion detection degrades the performance of other tasks: Interestingly, our experimental evaluations indicate that incorporating an emotion detection task in a multitask learning model degrades the performance of other tasks considerably. However, the emotion detection performance is leveraged from the other tasks representations. For example, a multitask model with anxiety and stress detection tasks achieves the highest Top-1 accuracy of 70.3% and 85.8% for anxiety and stress detection tasks, respectively. However, incorporating emotion in that model reduces the accuracy of anxiety and stress detection tasks 10.1% and 3.1%, respectively. The reasoning behind these performance degradations is that emotion detection is much more complicated than stress and anxiety because of a wide variety of emotions. Although both stress and anxiety are located in the negative valence and high arousal region of the circumplex model [25, 84], the other regions of the model have many other emotions, which are not beneficial to learn representations for stress or anxiety detection tasks. As a result, the emotion detection task dominates the multitask learning and forces the models to learn over-generalized representation, which leads to performance degradation for other tasks. Thus, we should not use emotion with stress, and anxiety detection tasks in an affective multitask learning model. However, incorporating anxiety and stress detection will help to improve the performance of emotion detection.

6.3 Generalizability of Representation Learning Models

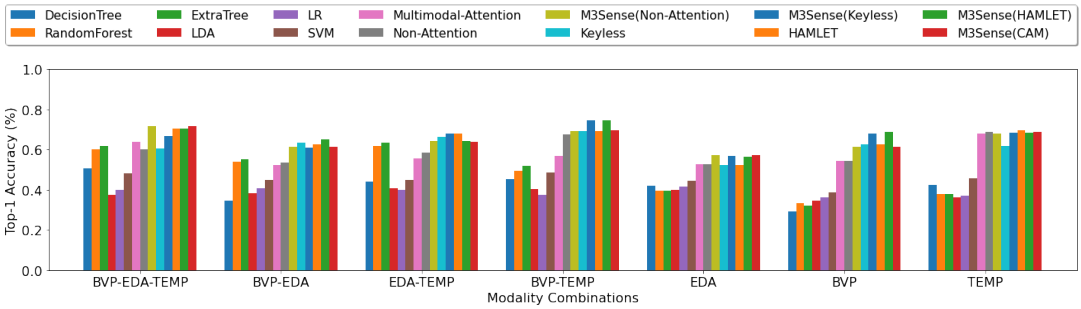
We investigated whether a representation learned for one task can be used for another unseen task prediction (a task which was not present during training), and thus reduce the dependency on a huge labeled dataset to train a model for the target task. We conduct this experimentation in three phases:

- **Transfer baseline models across tasks:** In this phase, we investigated whether a model can learn a generalized representation that can be used to train a model for an unseen task with limited training data. We trained the baseline models for emotion, anxiety, and stress tasks separately, called source tasks. After that, we froze the *Representation Learning Module* of these models and replace only the *Task Learning Network* of the source task with a model for the target task. Finally, we used this modified model and fine-tuned it with a small labeled dataset for the unseen task.
- **Transfer multitask model (*M3Sense*) across tasks:** The goal of this experimentation is to investigate whether the learned representation by *M3Sense* generalizes well to learn unseen tasks with a few labeled training samples. We pre-trained *M3Sense* with a pair of tasks and then froze the *Representation Learning Module* of the model. After that, the *Task Learning Network* part was replaced with a model for the unseen target task, and the modified model was with a small labeled dataset for the unseen task.
- **Fine-tune a multitask model (*M3Sense*) to a single task:** In this phase, we investigated the impact of fine-tuning the multitask learning model of *M3Sense* to a single task model. This experimentation aims to investigate whether fine-tuning a multitask model can lose the generalizability of learned representation across tasks. We followed the similar procedure of prior experiments. In this case, the model was pre-trained with all the tasks (anxiety, emotion, and stress) and then fine-tuned for a specific target task.

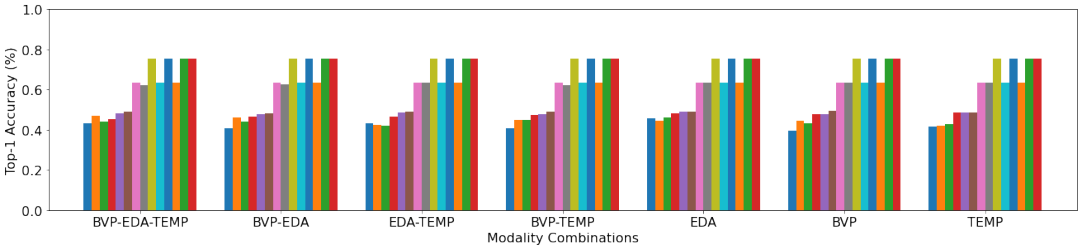
Are the state-of-the-art models capable of handling unseen tasks? The experimental results in Fig. 7 suggest that the performance of state-of-the-art multimodal learning models for unseen task degrades considerably for some combination of modalities. For example, if we transfer Keyless [54] and HAMLET [33] models from anxiety detection to emotion recognition, then the Top-1 accuracy degrades to less than 40% for the following combination of modalities: BVP-EDA-TEMP, EDA-TEMP, BVP-TEMP, EDA, BVP. The capability of transferring a state-of-the-art multimodal learning model across tasks is highly dependent on the combination of modalities and the pair of source and target tasks. The reasoning behind this performance degradation for unseen task is that these multimodal learning models extract task-specific representation which are not generalized across tasks. Thus, these models are not effective for unseen task learning. Additionally, given the fact that the commercially available wrist devices come with different available sensors combinations, the state-of-the-art models when trained for one device will struggle to perform well in new devices having different modalities.

Can *M3Sense* improve the performance of models on unseen tasks? The experimental results in Fig. 7 suggest that if we trained *M3Sense* with a source task set and fine-tune that model for an unseen task, then the performance of the fine-tuned model stays the same or slightly reduced compared to a model which is trained with all the source tasks and the unseen task. For example, if we transfer *M3Sense*(Anxiety-Stress) model to learn the emotion recognition task, then the Top-1 accuracy of the emotion recognition task stays the same (*M3Sense*(Anxiety-Stress) denotes that *M3Sense* is trained for anxiety and stress tasks). However, if we fine-tune *M3Sense*(Emotion-Stress) model to learn anxiety detection tasks with a few training samples, then the performance of anxiety detection reduces slightly. Although the performance for an unseen task in the fine-tuned model of *M3Sense* degrades slightly in some settings, this fine-tuned model of *M3Sense* outperforms all the evaluated state-of-the-art model performance to learn an unseen task. The reasoning behind this consistent performance to learn an unseen task is that *M3Sense* can learn affect-agnostic representation by using the *Domain Alignment Module* which being generalized is effective for unseen tasks.

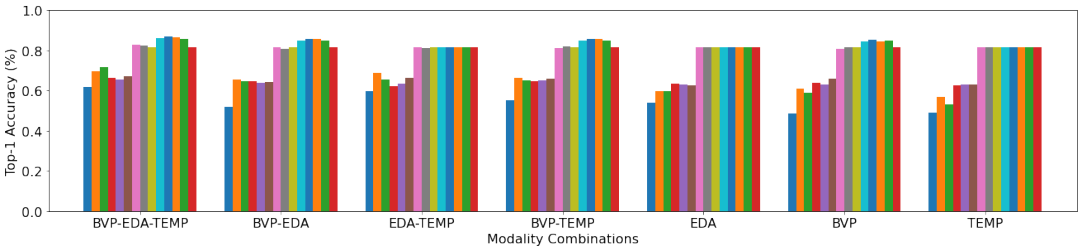
Does fine-tuning *M3Sense* for a single affective task help? Experimental results in Fig. 7 suggest that if we train *M3Sense* with all the tasks and then fine-tuning for a target task, the accuracy degrades slightly. Despite



(a) Anxiety detection models.



(b) Emotion detection models.



(c) Stress detection models.

Fig. 8. Impact of modality combinations in handcrafted feature-based machine learning models, multimodal representation learning models, and the variants of *M3Sense* for affect recognition.

this, the fine-tuned model still outperforms the task-specific baselines and state-of-the-art models. The reasoning behind this slight performance degradation is that the fine-tuned model learns task-specific representation, which leads to the loss of generalizability of the learned representation. However, it does not degrade considerably, as we only fine-tuned the task learning model part.

6.4 Ablation Studies

6.4.1 Impact of Modalities on Task Performance: Previous works [23, 45] have found that using multimodal data over unimodal data improves the representation learning performance. However, no prior works on affect recognition have demonstrated whether all or specific combinations of modalities can work better for detecting these tasks. To bridge this gap of literature, we do an ablation study using different modality combinations. All 3 categories of compared methods were used, i.e., the feature-based traditional models, the deep multimodal

Table 5. Impact of utilizing unlabeled data through the domain alignment module in multitask learning.

Training Method	Learning Models	Task (Top-1 Accuracy (%))		
		Anxiety	Emotion	Stress
Without unlabeled data	*M3Sense* (Non-Attention)	54.9	75.2	81.5
	M3Sense (HAMLET)	52.7	75.2	83.6
	M3Sense (Keyless)	55.0	75.2	81.5
	M3Sense (*CAM*)	51.4	75.2	81.7
With unlabeled data	*M3Sense* (Non-Attention)	**71.5**	75.2	81.5
	M3Sense (Keyless)	66.7	75.2	**87.0**
	M3Sense (HAMLET)	71.5	75.2	81.5
	M3Sense (*CAM*)	70.3	75.2	85.7

representation learning models, and the variants of *M3Sense*. Fig. 8 demonstrates the results, from which we make the following observations.

Unimodal or multimodal sensors for affect recognition? The results (Fig. 8) show that whether to use unimodal or multimodal sensors for better performance- it comes down to the individual affective tasks. For example, for anxiety detection, the best result obtained using a combination of the BVP and TEMP modalities was 74.7% (using *M3Sense*(HAMLET)), while the same model using all modalities provided 70.3%. Moreover, for anxiety, using the single modality TEMP achieved performance (69.4%) nearly as good as using all modalities using the same model (70.2%). It shows clinical relevance with previous works that have shown that temperature (TEMP) is a strong indicator of anxiety [3, 64]. Overall, our findings suggests that a combination of EDA-BVP-TEMP yielded the best accuracy performance for emotion (75.2% using *M3Sense*(HAMLET)) and stress (86.4% using HAMLET), while a combination for BVP and TEMP modalities provided the best results for anxiety (74.7% using *M3Sense*(HAMLET))). Additionally, *M3Sense* outperforms the feature-based machine learning models and state-of-the-art multimodal representation learning models in all combination of modalities.

A solution with consistent performance is needed: A major problem with the commercially available wrist devices or smartwatches is they come with a variety of embedded sensors, For example, the latest version of the Apple watch (Series 7) does not provide an temperature sensor, while that for the Samsung Galaxy (Series 4) does not include a EDA sensor [12]. Therefore, we need a learning model that will provide consistent performance regardless of which sensing modalities are available. From the results, we see that all variants of *M3Sense*, our solution, provides consistent performance across all sensors combinations, where the feature-based traditional models, or the state-of-the-art models often do a poor job. For example, for emotion detection, if the EDA sensor is absent, the best performance achieved by using the remaining sensors with the feature-based traditional models is 49% accuracy (using Support Vector Machine), and that with the deep representation models is 63.5% accuracy (using HAMLET), and while our solution using the same model (*M3Sense* (HAMLET)) produces 75.2% accuracy.

6.4.2 Impact of utilizing unlabeled data in M3Sense: In addition to using labeled data from the same domain for learning the representation as the state-of-the-art does, *M3Sense* is capable of utilizing unlabeled data from heterogeneous domains through the domain alignment module. To show the effectiveness of this part of our solution, we conducted experiments with the variants of *M3Sense* under different settings. First, we trained the variant of *M3Sense* by using only labeled data. As we did not use unlabeled data, only multitask learning losses were used to train the models. Second, we used both labeled and unlabeled data and utilized the multitask learning losses and domain alignment loss to align the distribution of the representation. We present the results in Table 5. We make the following interpretations from the results.

Table 6. Comparison of mode size (number of parameters in thousands) of different learning models

Learning Models	Model Size			Combined Model Size
	Anxiety	Emotion	Stress	
Non-Attention	1110 K	1110 K	1110 K	3330 K
Multimodal Attention	1112 K	1112 K	1112 K	3336 K
Keyless [54]	1113 K	1113 K	1113 K	3339 K
HAMLET [33]	1115 K	1115 K	1115 K	3345 K
M3Sense (Non-Attention)	–	–	–	2244 K
M3Sense (Keyless)	–	–	–	2248 K
M3Sense (HAMLET)	–	–	–	2250 K
M3Sense (*CAM*)	–	–	–	2251 K

Does unlabeled data help in affect recognition? The experimental results in Table 5 suggest that the addition of unlabeled data helps to improve the accuracy of anxiety and stress detection tasks. Among all the tasks, the anxiety detection task achieved the highest performance improvement of 16.6% in Top-1 accuracy, while the stress detection task achieves an accuracy gain of 3.6%. The reasoning behind this performance improvement is that domain alignment module enables *M3Sense* to effectively utilize unlabeled data to learn affect-agnostic representations from multiple heterogeneous domains. Our explanation is that the generalized representation learned from the unlabeled data mostly provide additional salient information for the stress and anxiety detection task. Moreover, the unlabeled datasets we have used in this work are from the mental and physical workload domains (Section 5.1.4 and 5.1.5). Previous works in the literature have shown significant correlations among these domains with stress and anxiety [1, 32, 47]. On the contrary, compared to stress and anxiety detection, emotion detection is a more complex and difficult task, as it involves a variety of emotions placed in the circumplex model [84]. Thus, these results show clinical relevance with the aforementioned works in the literature. We also note that the overall experiments show the prospect of unlabeled data to leverage the affective tasks performance in multitask settings. Although the utilizing unlabeled data have been explored in the literature [29, 53, 56, 66, 67, 77, 82], our *Domain Alignment Module* aligns the unlabeled data distribution across heterogeneous domains in a multitask setting (e.g., aligning a unlabeled data of a modality from anxiety and emotion domains to stress domain), which is the main benefit of this module over these existing works.

6.5 Comparison of Learning Model Complexity (Space and Time)

We analyzed the space and time complexity for different deep multimodal representation learning models and the variants of our proposed multitask learning framework, *M3Sense*. We conducted these analyses by executing a batch of size 2 with multimodal data samples on a $P100$ GPU node in a cluster computing environment. This computing node has $60GB$ memory and 10 CPUs. The analyses of space (number of parameters in a model) and time (batch of data samples execution time) model complexity are presented in Table 6 and 7, respectively.

Can *M3Sense* reduce the model size? The results in Table 6 suggest that each deep multimodal learning model has the approximately same number of parameters ($1112K$) in a single task setting. For three affective tasks (anxiety, emotion, and stress) with three modalities (BVP, EDA, TEMP), the combined model size is approximately $3337K$. Our proposed multitask learning framework *M3Sense* learns all the tasks using a single model, unlike a separate learning model for each task. The number of parameters of *M3Sense* with baseline multimodal learning models is approximately $2248K$, which is approximately 33% lower than the combined model size of baseline learning models for multiple tasks. The reasoning behind this space reduction is that *M3Sense* uses a single and generalized representation learning model for all the tasks compared to the baseline learning models, which use

Table 7. Comparison of execution of time (Millisecond) per batch size of 2 for different learning models

Method	Execution Time (Millisecond)			Combined Execution Time (Millisecond)
	Anxiety	Emotion	Stress	
Non-Attention	142	141	143	426
Multimodal Attention	160	159	158	477
Keyless [54]	173	170	174	517
HAMLET [33]	195	192	196	583
M3Sense (Non-Attention)	–	–	–	**172**
M3Sense (Keyless)	–	–	–	191
M3Sense (HAMLET)	–	–	–	212
M3Sense (*CAM*)	–	–	–	203

a separate model for each task. Thus, in the multitask learning setting, *M3Sense* reduces the combined model size of state-of-the-art single task-based learning models without compromising the performance.

Can *M3Sense* reduce the execution time? The results in Table 7 suggest that each baseline multimodal learning model requires less than 200*ms* computing time to execute a batch of size 2 with three modalities (BVP, EDA, TEMP). These baseline models require more than 400*ms* computing time to infer the prediction for all tasks. On the other hand, *M3Sense* can reduce this multitask inference time to approximately 200*ms*, which 50% reduction compared to the baseline learning models.

7 BROADER IMPACT

Our research shows that multitask learning (MTL) can be a new perspective in the mental health domains research by utilizing the significant associations among these domains. Our framework, *M3Sense* provides a generalized and scalable platform that can be integrated into any wearable-based affection detection pipeline. Future work would be interesting to extend this research to other mental health domains, such as depression and mental workload, and to find out which affective tasks should be learned together to guarantee robust performance across all domains. Moreover, it must be noted that our multitasking framework considers each of the mental health domains as individual affective tasks (e.g., stress, anxiety, emotion), given the correlations among these domains found from the literature (explained in Section 3), and it uses the individual rating scales of the domain for ground truth labels, such as, the valence-arousal scale of emotion, STAI scale of anxiety. The other rating scales from these domains can also be utilized in a similar manner, which we leave as a future work. Future works also can manifest the ability of MTL to develop personalized models that can account for subjective differences based on potential bio-markers (e.g., age, gender, and personality).

Additionally, this research shows the potential usability of unlabeled data from heterogeneous domains, which can be instrumental in reducing the dependency on expert-annotated data for these mental health conditions. Moreover, since the COVID-19 pandemic, there has been a massive surge in the mental wellness and wearable industry's growth [21, 22]. Companies like Apple, Samsung, and Garmin are enhancing new physiological capabilities and mental wellness features to their latest smartwatches [37, 87]. Our research gives the impression that a single model can detect various mental health conditions using different embedded sensors from these watches with consistent performance. Our proposed learning framework can reduce the model size and execution time, which will be crucial for these resource-constrained devices. It also opens us with many possibilities in building intelligent cognitive assistants on smartwatches, including early assessment of multiple mental health conditions and providing preventive interventions by utilizing its interaction capabilities.

8 CONCLUSION

In this work, we developed a novel multitask learning framework, called *M3Sense*, to learn affect-agnostic multimodal representations for affect recognition tasks. Moreover, we designed a novel domain alignment algorithm to train multitask models to learn generalized representation using limited labeled and a huge amount of unlabeled data. Our extensive experimental evaluations and ablation studies suggest that *M3Sense* outperforms all the evaluated handcrafted feature-based machine learning models and state-of-the-art multimodal representation learning models across all the evaluated tasks and domains. Moreover, we evaluated the impact of various task combinations and modalities in *M3Sense*, which provides valuable insights to design a multitask learning model for affect recognition tasks. Additionally, *M3Sense* is a unified learning model that can help to improve the performance of the state-of-the-art representation learning model by 5% – 60% for several unseen tasks with various input modalities. Moreover, the findings from our experimental evaluations can be used to develop robust multitask learning models for various mental health applications.

REFERENCES

[1] Norah H Alsuraykh, Max L Wilson, Paul Tennent, and Sarah Sharples. 2019. How stress and mental workload are connected. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 371–376.

[2] Anxiety and Heart Disease. 2020. John Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/anxiety-and-heart-disease.

[3] American Psychological Association. Accessed:2021. What's the difference between stress and anxiety? https://www.apa.org/topics/stress/anxiety-difference.

[4] PR Aylard, JH Gooding, PJ McKenna, and RP Snaith. 1987. A validation study of three anxiety and depression self-assessment scales. *Journal of psychosomatic research* 31, 2 (1987), 261–268.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[6] Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. 2017. Variational attention for sequence-to-sequence models. *arXiv preprint arXiv:1712.08207* (2017).

[7] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine* 3, 1 (2020), 1–9.

[8] Deborah Anne Burton, Keith Stokes, and George M Hall. 2004. Physiological effects of exercise. *Continuing Education in Anaesthesia Critical Care & Pain* 4, 6 (2004), 185–188.

[9] C. Chen, R. Jafari, and N. Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE ICIP*. 168–172.

[10] Youngjun Cho, Simon J Julier, and Nadia Bianchi-Berthouze. 2019. Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR mental health* 6, 4 (2019), e10140.

[11] Sven-Åke Christianson. 1992. Emotional stress and eyewitness memory: a critical review. *Psychological bulletin* 112, 2 (1992), 284.

[12] CNET. Accessed:2021. Apple Watch 7 vs. Samsung Galaxy Watch 4: All the major differences between smartwatch rivals. www.cnet.com/tech/mobile/apple-watch-7-vs-samsung-galaxy-watch-4-all-the-major-differences-between-smartwatch-rivals/.

[13] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*. 9712–9724.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Empatica. Accessed:2021. Empatica E4 Wristband. https://www.empatica.com/research/e4/.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *CVPR*.

[17] Pamela J Feldman, Sheldon Cohen, Stephen J Lepore, Karen A Matthews, Thomas W Kamarck, and Anna L Marsland. 1999. Negative emotions and acute physiological responses to stress. *Annals of Behavioral Medicine* 21, 3 (1999), 216–222.

[18] Nancy Fiedler, Robert Laumbach, Kathie Kelly-McNeil, Paul Lioy, Zhi-Hua Fan, Junfeng Zhang, John Ottenweller, Pamela Ohman-Strickland, and Howard Kipen. 2005. Health effects of a mixture of indoor air volatile organics, their ozone oxidation products, and stress. *Environmental health perspectives* 113, 11 (2005), 1542–1548.

[19] Susan Folkman. 1997. Positive psychological states and coping with severe stress. *Social science & medicine* 45, 8 (1997), 1207–1221.

[20] Susan Folkman. 2008. The case for positive emotions in the stress process. *Anxiety, stress, and coping* 21, 1 (2008), 3–14.

[21] Forbes. Accessed:2021. Smart Wearables Market To Double By 2022: $27 Billion Industry Forecast. www.forbes.com/sites/paullamkin/2018/10/23/smart-wearables-market-to-double-by-2022-27-billion-industry-forecast.

[22] Forbes. Accessed:2021. Venture Funding For Mental Health Startups Hits Record High As Anxiety, Depression Skyrocket. www.forbes.com/sites/katiejennings/2021/06/07/venture-funding-for-mental-health-startups-hits-record-high-as-anxiety-depression-skyrocket/.
[23] Andrea E Frank, Alyssa Kubota, and Laurel D Riek. 2019. Wearable activity recognition for robust human-robot teaming in safety-critical environments via hybrid neural networks. In *IROS*. IEEE, 449–454.
[24] Christian Gagné. 2019. A Principled Approach for Learning Task Similarity in Multitask Learning. In *IJCAI*.
[25] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. 2019. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing* (2019).
[26] Giorgos Giannakakis, Dimitris Grigoriadis, and Manolis Tsiknakis. 2015. Detection of stress/anxiety state from EEG features during video watching. In *2015 International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6034–6037.
[27] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. 2018. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 270–287.
[28] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. 2020. Learning to branch for multi-task learning. In *International Conference on Machine Learning*. PMLR, 3854–3863.
[29] W. Guo, J. Wang, and S. Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access* 7 (2019), 63373–63394.
[30] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *EMNLP*.
[31] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587* (2016).
[32] Seyedmajid Hosseini, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, Kenneth Cochran, and Raju Gottumukkala. 2021. A multi-modal sensor dataset for continuous stress detection of nurses in a hospital. *arXiv* (2021).
[33] Md Mofijul Islam and Tariq Iqbal. 2020. HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 10285–10292.
[34] Md Mofijul Islam and Tariq Iqbal. 2021. Multi-GAT: A Graphical Attention-based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. In *IEEE Robotics and Automation Letters (RA-L)*.
[35] Md Mofijul Islam and Tariq Iqbal. 2022. MuMu: Cooperative Multitask Learning-based Guided Multimodal Fusion. In *AAAI*.
[36] Md Tamzeed Islam and Shahriar Nirjon. 2021. Sound-Adapter: Multi-Source Domain Adaptation for Acoustic Classification Through Domain Discovery. In *Proceedings of the International Conference on Information Processing in Sensor Networks*. 176–190.
[37] Wall Street Journal. Accessed:2021. Apple Is Working on iPhone Features to Help Detect Depression, Cognitive Decline. https://www.wsj.com/articles/apple-wants-iphones-to-help-detect-depression-cognitive-decline-sources-say-11632216601.
[38] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. 2020. MMTM: Multimodal Transfer Module for CNN Fusion. In *CVPR*.
[39] Ozcan Kayikcioglu, Sinan Bilgin, Goktug Seymenoglu, and Artuner Deveci. 2017. State and trait anxiety scores of patients receiving intravitreal injections. *Biomedicine hub* 2, 2 (2017), 1–5.
[40] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*. 5492–5501.
[41] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. 2018. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation* 15, 3 (2018), 235.
[42] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[43] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.
[44] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *ICCV*. 8658–8667.
[45] Alyssa Kubota, Tariq Iqbal, Julie A Shah, and Laurel D Riek. 2019. Activity recognition in manufacturing: The roles of motion capture and sEMG+ inertial wearables in detecting fine vs. gross motion. In *ICRA*. IEEE.
[46] Richard S Lazarus, Joseph C Speisman, and Arnold M Mordkoff. 1963. The relationship between autonomic indicators of psychological stress: Heart rate and skin conductance. *Psychosomatic medicine* (1963).
[47] Ying Lean and Fu Shan. 2012. Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries* 22, 3 (2012), 177–187.
[48] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *IJCAI* (2018).
[49] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *NeurIPS*.
[50] Russell Li and Zhandong Liu. 2020. Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making* 20, 11 (2020), 1–10.

[51] Alexandros Liapis, Christos Katsanos, Dimitris G Sotiropoulos, Nikos Karousos, and Michalis Xenos. 2017. Stress in interactive applications: analysis of the valence-arousal space based on physiological signals and self-reported data. *Multimedia Tools and Applications* 76, 4 (2017), 5051–5071.

[52] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, and P. Liu. 2019. Action Recognition Based on 3D Skeleton and RGB Frame Fusion. In *IROS*. 258–264.

[53] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2019), 1862–1878.

[54] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. 2018. Multimodal keyless attention fusion for video classification. In *AAAI*.

[55] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*.

[56] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

[57] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*. 375–383.

[58] Rosan Luijcks, Hermie J Hermens, Lonneke Bodar, Catherine J Vossen, Jim van Os, and Richel Lousberg. 2014. Experimentally induced stress validated by EMG activity. *PloS one* 9, 4 (2014), e95215.

[59] Wenjie Luo, Zhenyu Yan, Qun Song, and Rui Tan. 2021. PhyAug: Physics-Directed Data Augmentation for Deep Sensing Model Transfer in Cyber-Physical Systems. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*. 31–46.

[60] Yun Luo, Li-Zhen Zhu, Zi-Yu Wan, and Bao-Liang Lu. 2020. Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *Journal of Neural Engineering* 17, 5 (2020), 056021.

[61] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

[62] Linda Mah, Claudia Szabuniewicz, and Alexandra J Fiocco. 2016. Can anxiety damage the brain? *Current opinion in psychiatry* 29, 1 (2016), 56–63.

[63] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. 2010. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications* 49, 2 (2010), 277–297.

[64] Elba Mauriz, Sandra Caloca-Amber, and Ana M Vázquez-Casares. 2020. Effect of Facial Skin Temperature on the Perception of Anxiety: A Pilot Study. In *Healthcare*, Vol. 8. Multidisciplinary Digital Publishing Institute, 206.

[65] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *ACM ISWC* (Maui, Hawaii). 158–165.

[66] Michael Neumann and Ngoc Thang Vu. 2019. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7390–7394.

[67] Kanal Paul Nigam. 2001. *Using unlabeled data to improve text classification.* Carnegie Mellon University.

[68] Raymond W Novaco. 1978. Anger and coping with stress. In *Cognitive behavior therapy*. Springer, 135–173.

[69] National Institute of Mental Health. 2021. COVID-19 and Mental Health. https://www.nimh.nih.gov/about/director/messages/2021/one-year-in-covid-19-and-mental-health.

[70] World Health Organization. 2020. Mental Health. https://www.who.int/news/item/27-08-2020-world-mental-health-day-an-opportunity-to-kick-start-a-massive-scale-up-in-investment-in-mental-health.

[71] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 293.

[72] Rosalind W Picard and Jennifer Healey. 1997. Affective wearables. *Personal technologies* 1, 4 (1997), 231–240.

[73] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[74] PLUX. Accessed: 2021. respiBAN Professional. https://biosignalsplux.com/products/wearables/respiban-pro.html.

[75] Juan Carlos Quiroz, Elena Geangu, and Min Hooi Yong. 2018. Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR mental health* 5, 3 (2018), e10153.

[76] Martin Ragot, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. 2017. Emotion recognition using physiological signals: laboratory vs. wearable sensors. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 15–22.

[77] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*. 759–766.

[78] D. Ramachandram and G. W. Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108.

[79] Nafiul Rashid, Luke Chen, Manik Dautta, Abel Jimenez, Peter Tseng, and Mohammad Abdullah Al Faruque. 2021. Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor. *arXiv preprint arXiv:2108.03166* (2021).

[80] Steven Reiss. 1991. Expectancy model of fear, anxiety, and panic. *Clinical psychology review* 11, 2 (1991), 141–153.
[81] Alina Roitberg, Nikhil Somani, Alexander Perzylo, Markus Rickert, and Alois Knoll. 2015. Multimodal Human Activity Recognition for Industrial Manufacturing Processes in Robotic Workcells. In *ICMI*.
[82] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery* 13, 6 (2018), 925–933.
[83] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).
[84] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[85] Mario Salai, István Vassányi, and István Kósa. 2016. Stress detection using low cost heart rate sensors. *Journal of healthcare engineering* 2016 (2016).
[86] Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 2 (2018), 1–26.
[87] Samsung. Accessed:2021. Samsung Teams up with Calm to Provide Better Mindfulness and Wellness Experiences. https://news.samsung.com/global/samsung-teams-up-with-calm-to-provide-better-mindfulness-and-wellness-experiences.
[88] Sirat Samyoun, Abu Sayeed Mondol, and John A Stankovic. 2020. Stress detection via sensor translation. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 19–26.
[89] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676.
[90] Francesco Sartor, Jos Gelissen, Ralph Van Dinther, David Roovers, Gabriele B Papini, and Giuseppe Coppola. 2018. Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients. *BMC Sports Science, Medicine and Rehabilitation* 10, 1 (2018), 10.
[91] 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 2018 on ICMI*. ACM, 400–408.
[92] Simone Schnall. 2010. Affect, mood and emotions. *Social and emotional aspect of learning* (2010), 59–64.
[93] Dongmin Shin, Dongil Shin, and Dongkyoo Shin. 2017. Development of emotion recognition interface using complex EEG/ECG bio-signal for interactive contents. *Multimedia Tools and Applications* 76, 9 (2017), 11449–11470.
[94] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.
[95] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2015), 17–28.
[96] Charles D Spielberger. 1983. State-trait anxiety inventory for adults. (1983).
[97] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning?. In *International Conference on Machine Learning*. PMLR, 9120–9132.
[98] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry* 10, 1 (2020), 1–26.
[99] Phillip D Tomporowski and Norman R Ellis. 1986. Effects of exercise on cognitive processes: A review. *Psychological bulletin* (1986).
[100] Goran Udovičić, Jurica Đerek, Mladen Russo, and Marjan Sikora. 2017. Wearable emotion recognition system based on GSR and PPG signals. In *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*. 53–59.
[101] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2020. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379* (2020).
[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017), 5999–6009.
[103] Sergio A Velastin. 2009. CCTV video analytics: Recent advances and limitations. In *International Visual Informatics Conference*. Springer.
[104] Christiaan H Vinkers, Renske Penning, Juliane Hellhammer, Joris C Verster, John HGM Klaessens, Berend Olivier, and Cor J Kalkman. 2013. The effect of stress on core and peripheral body temperature in humans. *Stress* 16, 5 (2013), 520–530.
[105] Ajai Vyas and Sumantra Chattarji. 2004. Modulation of different states of anxiety-like behavior by chronic stress. *Behavioral neuroscience* 118, 6 (2004), 1450.
[106] Robert Wang, Gordon Blackburn, Milind Desai, Dermot Phelan, Lauren Gillinov, Penny Houghtaling, and Marc Gillinov. 2017. Accuracy of wrist-worn heart rate monitors. *Jama cardiology* 2, 1 (2017), 104–106.
[107] Devy Widjaja, Michele Orini, Elke Vlemincx, and Sabine Van Huffel. 2013. Cardiorespiratory dynamic response to mental stress: a multivariate time-frequency analysis. *Computational and mathematical methods in medicine* 2013 (2013).
[108] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. 2020. Audiovisual SlowFast Networks for Video Recognition. *arXiv preprint arXiv:2001.08740* (2020).

[109] Xiaofen Xing, Zhenqi Li, Tianyuan Xu, Lin Shu, Bin Hu, and Xiangmin Xu. 2019. SAE+ LSTM: A New framework for emotion recognition from multi-channel EEG. *Frontiers in neurorobotics* 13 (2019), 37.
[110] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 2048–2057.
[111] Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir Behzadan, and Theodora Chaspari. 2020. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing* (2020).
[112] Jung-Yi Yoo and Jang-Han Lee. 2015. The effects of valence and arousal on time perception in individuals with social anxiety. *Frontiers in psychology* 6 (2015), 1208.
[113] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2008), 39–58.
[114] Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson. 2007. Audio-visual affect recognition. *IEEE Transactions on multimedia* 9, 2 (2007), 424–428.
[115] Y. Zhang, C. Cao, J. Cheng, and H. Lu. 2018. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Transactions on Multimedia* 20, 5 (2018), 1038–1050.
[116] Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017).
[117] Yali Zheng, Tracy CH Wong, Billy HK Leung, and Carmen CY Poon. 2016. Unobtrusive and multimodal wearable sensing to quantify anxiety. *IEEE Sensors Journal* 16, 10 (2016), 3689–3696.
[118] Fan Zhou, Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. 2020. Task Similarity Estimation Through Adversarial Multitask Neural Network. *IEEE Transactions on Neural Networks and Learning Systems* (2020).