Neural Computing and Applications (2023) 35:24435–24454
https://doi.org/10.1007/s00521-023-09036-4

ORIGINAL ARTICLE

Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress

Gulin Dogan[1] · Fatma Patlar Akbulut[2]

Received: 6 December 2022 / Accepted: 6 September 2023 / Published online: 3 October 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
Mental stress is a significant risk factor for several maladies and can negatively impact a person's quality of life, including their work and personal relationships. Traditional methods of detecting mental stress through interviews and questionnaires may not capture individuals' instantaneous emotional responses. In this study, the method of experience sampling was used to analyze the participants' immediate affective responses, which provides a more comprehensive and dynamic understanding of the participants' experiences. WorkStress3D dataset was compiled using information gathered from 20 participants for three distinct modalities. During an average of one week, 175 h of data containing physiological signals such as BVP, EDA, and body temperature, as well as facial expressions and auditory data, were collected from a single subject. We present a novel fusion model that uses double-early fusion approaches to combine data from multiple modalities. The model's F1 score of 0.94 with a loss of 0.18 is very encouraging, showing that it can accurately identify and classify varying degrees of stress. Furthermore, we investigate the utilization of transfer learning techniques to improve the efficacy of our stress detection system. Despite our efforts, we were unable to attain better results than the fusion model. Transfer learning resulted in an accuracy of 0.93 and a loss of 0.17, illustrating the difficulty of adapting pre-trained models to the task of stress analysis. The results we obtained emphasize the significance of multi-modal fusion in stress detection and the importance of selecting the most suitable model architecture for the given task. The proposed fusion model demonstrates its potential for achieving an accurate and robust classification of stress. This research contributes to the field of stress analysis and contributes to the development of effective models for stress detection.

Keywords Stress detection · Sequential and non-sequential model · Fine-tuning · Multi-modality

1 Introduction

Modern society has become inseparable from stress, which makes it nearly impossible to prevent stress factors. In every aspect of existence, people are exposed to different types and levels of stress. Stress is an internal experience that occurs as a result of a person leaving their comfort zone and undergoing a change in their daily routine. It renders a person incapable of coping with threats to their physical, emotional, or mental health, and can even result in the development of chronic diseases. According to the World Health Organization [1], stress-related disorders have emerged, including chronic stress caused by malfunctions, decreased well-being, increased disease rates, fatigue syndrome, and depression.

Several tests and questionnaires, such as the perceived stress scale (PSS) [2] and the stress response inventory (SRI) [3], can be used to assess mental stress. However, the signs of stress can only be examined after the individual has experienced stress [4], and there are only subjective solutions available, delaying the ultimate result. It can even be incorrectly evaluated if the individual does not recognize that they are under stress. To ensure a correct diagnosis, it is essential to look beyond the person's words and

Gulin Dogan and Fatma Patlar Akbulut have contributed equally to this work.

✉ Fatma Patlar Akbulut
f.patlar@iku.edu.tr

Gulin Dogan
1900000822@stu.iku.edu.tr

[1] Department of Computer Engineering, Istanbul Kultur University, Istanbul 34158, Turkey

[2] Department of Software Engineering, Istanbul Kultur University, Istanbul 34158, Turkey

provide a more objective evaluation method. In such situations, it is possible to study an individual's unconscious physiological signals [5–8] or to make a more precise evaluation by combining them with facial images [9] and a person's voice. The most significant acquisition is the use of methods, such as experience sampling (ES), that enable you to detect the event at the time of the incident, as opposed to analysis performed after the incident has occurred.

The experience sampling method is one of the most effective applications for analyzing daily life. The method of experience sampling permits the use of multiple sensors, recording channels, and application-based evaluations. This technique is also known as an "ecological instant assessment," which is a journaling technique that evaluates the current environment and mental state. Using this method with technological devices enables the capture of instantaneous experiences in real-time by randomly [10] or at regular intervals posing queries to participants. Its ability to be incorporated into everyday life also expands its field of application: from tracing tobacco, alcohol, cannabis, or opium addiction [11] to chronic pain treatment [12], cognitive impairments, and susceptibility to stress [13], fatigue self-management programs, and enhancing the quality of life by minimizing fatigue [14]. It has been utilized frequently in recent medical assessments [15].

The widespread use of smart devices facilitated by emerging technology has transformed the way we interact with technology and opened up new avenues for understanding and addressing the complex phenomenon of environmental stress [16, 17]. Smartphones, smartwatches, and fitness trackers have become an integral part of our daily existence, providing unprecedented opportunities to monitor and analyze various aspects of the human experience. Wearable devices equipped with sophisticated sensors have made it possible to collect a vast quantity of personal data that exceeds traditional self-reporting measures. These devices are capable of capturing physiological signals such as heart rate, heart rate variability, electrodermal activity, and sleep patterns, providing objective and real-time insights into the physical and emotional states of individuals. In addition, they enable the monitoring of emotion and cognition via techniques such as facial expression analysis, speech recognition, and activity tracking, thereby providing a multidimensional understanding of the subjective experiences of individuals. The availability of such abundant and context-specific data has sparked interest in the use of intelligent devices to detect and prevent ecological stress. Stress, a pervasive and debilitating condition that affects individuals in numerous spheres of life, poses significant obstacles to both individual and societal productivity. By leveraging wearable devices and the wealth of data they provide, researchers and practitioners can gain a deeper understanding of the triggers, patterns, and consequences of stress, ultimately leading to the development of effective stress management and prevention strategies.

In addition to the opportunities presented by wearable devices and stress monitoring, there are crucial factors to consider. Privacy and data security are of the utmost importance, as the collection of personal information raises questions regarding data ownership, consent, and responsible data management practices. For widespread acceptance and ethical implementation of these technologies, it is essential to strike a balance between the benefits of data-driven stress detection and the preservation of individuals' privacy rights. In view of these opportunities and challenges, the purpose of this study is to investigate the potential of smart devices and wearable technology for stress detection and management in ecological environments. Using the rich and diverse data collected from ubiquitous devices, including physiological signals, emotion recognition, and contextual information, we aim to develop a comprehensive framework that integrates these multiple modalities for accurate, real-time stress detection. Through this research, we hope to contribute to the expanding corpus of knowledge on stress detection and prevention and pave the way for innovative and individualized interventions that enhance the well-being of individuals in the face of everyday stressors.

The specific objective of this study is to develop a method for early detection of mental stress in the workplace as a whole. This method eliminates the need to induce stress artificially in participants. By responding to specific inquiries at designated times throughout the day, participants are asked to convey their mood. In addition, participants are required to input data into a mobile application that assesses heart rate variability (HRV), a well-established physiological indicator of stress. The application's collected data are then analyzed using machine learning algorithms to accurately detect and predict stress levels. By focusing on workplace stress detection, this study seeks to provide valuable insights into identifying and addressing stress in occupational settings, with the ultimate objective of developing effective stress management and prevention strategies. The contributions of the study to the literature are listed below:

- A fusion dataset that includes multi-sensor signals, facial images, and audio recordings has been painstakingly developed utilizing data from study participants. This large and varied dataset can be used to facilitate future studies and developments in stress detection techniques. The fusion database is used as a resource for testing and benchmarking cutting-edge models and algorithms.

- An improved deep learning architecture was developed for the identification of the seven basic emotions: anger, disgust, fear, happiness, sorrow, surprise, and apathy. This architecture provides a sophisticated framework for reliably identifying and categorizing stress-related emotions by integrating image and signal-based analysis. Building this framework helps us comprehend the nuanced nature of our reactions to stress.
- The study compares and contrasts different deep learning architectures and uses multi-sensor data to determine which is most effective for detecting stress. The study finds the best methods of stress detection and evaluation by comparing and contrasting various methods and algorithms. In particular, the investigation demonstrates that blood volume pulse (BVP) and electrodermal activity (EDA) signals are very applicable in stress detection. These results shed light on the most appropriate methods for precise stress assessment and hence contribute to the optimization of stress detection models and algorithms.
- By combining the aforementioned contributions, this study makes important strides forward in our understanding of how to recognize and prevent stress. The research improves our understanding of the underlying mechanisms of stress and provides valuable insights into effective strategies for detecting, managing, and preventing stress in a variety of contexts by leveraging the fusion database, sophisticated deep learning architecture, and comparative analysis.

2 Related work

Advances in stress detection have been made possible by new modalities and the application of machine learning methods. We discuss previous research that has investigated stress detection through a variety of channels, such as physiological signals, audio-based methods, and facial expressions. We also explore the feature extraction and selection, machine learning models, and data fusion and integration that are used in stress detection as part of the machine learning and data analysis methodologies. By reviewing the relevant literature, we can learn about the various strategies that have been used for the problem of stress detection.

2.1 Modalities for stress detection

2.1.1 Physiological signals

Physiological signals of stress reactions have been appreciated for quite some time. These readings are objective indicators of the body's physiological processes and reactions to stress. The electrocardiogram (ECG), electrodermal activity (EDA), electromyography (EMG), respiration, and electroencephalogram (EEG) are only a few of the physiological signals that have been widely explored for stress detection. The ECG is used to monitor the heart's electrical activity [18] and get knowledge about the heart's inner workings. Heart rate variability (HRV), the quantification of differences in the time intervals between consecutive heartbeats [19], is one feature that may be extracted from ECG data through analysis. An index of autonomic nervous system function and emotional stability, HRV has been linked to stress levels. Galvanic skin response (GSR) or EDA assesses the skin's electrical conductivity. Sweat gland activity affects EDA and is controlled by the sympathetic nervous system. Changes in skin conductance occur as a result of increased sympathetic activity under stress. Emotional arousal and stress levels can be deduced from an examination of EDA signals. EMG records the electrical potential changes caused by contracting muscles. Muscle tension and activation can be reflected in EMG readings, both of which are linked to the body's stress response. When people are under stress, they tend to tense up and do more with their muscles. It is possible to learn a great deal about stress-related facial expressions and bruxism (teeth grinding) by analyzing EMG signals from specific muscle groups, such as the forehead or jaw. One such vital physiological indication that may provide useful information for identifying stress is respiration. Respiratory signals record details about breathing, including the rate, depth, and variability of breaths. When under stress, the body alters its breathing to better prepare for fight-or-flight reactions. Abnormalities in breathing patterns and the physiological responses to stress can be uncovered by analyzing respiratory signals. The electrical activity of the brain can be measured with a physiological signal called an EEG. Electrodes are placed on the scalp to record electrical activity in the brain in the form of an EEG signal. Cognitive and emotional processes, such as the stress reaction, can be understood by analysis of EEG data. Different mental states and stress levels are related to different frequency bands in the EEG spectral distribution, including alpha, beta, theta, and delta waves. Researchers can examine stress-related cognitive and affective states, as well as their brain correlates, by analyzing EEG data. Several signal processing methods are used to make sense of physiological signals. Frequency or time-domain properties can be extracted from physiological signals using signal processing techniques including the Fourier transform, the wavelet transform, and time-frequency analysis. Using these qualities, stress-related patterns can be identified by capturing the signals' underlying dynamics and properties.

2.1.2 Audio-based approaches

Stress levels can be inferred from an individual's speech and voice characteristics [20] using audio-based methods. Emotional states, such as stress, can be gleaned from one's voice through the application of various methods of feature extraction. Prosody [21] is one of the most important elements used to analyze audio data for stress detection. Prosody is the term used to describe how speakers alter their voices, rhythms, and rhythms to enhance their messages. It's common for people to raise their pitch, enhance their intensity, and speak more quickly when under stress. Signal processing methods like pitch analysis, energy estimate, and speech rate calculation can be used to quantify these prosodic shifts. Speech is another crucial characteristic that can be determined from audio waves. Stress levels can be deduced from a person's word choice and the way they express themselves verbally [22]. Sentiment analysis, keyword extraction, and semantic analysis are just some of the NLP methods that have been investigated for application with speech content. Words or language patterns that are indicators of stress or emotional distress can be isolated with the use of these methods.

Audio signals can convey information not just through the prosody and content of speech, but also through nonverbal vocal cues. Laughter sighs, hesitance, and variations in breathing patterns are all examples of nonverbal cues that can express emotion. Long pauses, sighs, and erratic breathing patterns may all be signs of stress. The development of stress detection algorithms can benefit from the analysis of these nonverbal indicators in audio recordings. Audio-based stress detection relies heavily on machine learning methods. They can be taught using audio data in which people's stress levels have been marked or self-reported. Support vector machines (SVM), hidden Markov models (HMM), and deep learning models (e.g., recurrent neural networks or convolutional neural networks) are only some of the machine learning techniques that have been used to categorize audio signals into stress levels [23]. Combining data from many audio aspects or modalities is one way to enhance the precision of stress detection based on audio. A more complete picture of a person's stress level, for instance, can be obtained by integrating prosodic traits with speech content or nonverbal signs. It's important to recognize the obstacles that audio-based stress detection methods must overcome. The quality of the data used to train stress detection models can be negatively impacted by factors such as environmental noise, speaker unpredictability, and language variations. In order to reliably and accurately detect stress from audio signals, robust feature extraction techniques and algorithms are required.

2.1.3 Facial expressions

Expressions on a person's face can tell you a lot about how they're feeling, including whether or not they're stressed. Capturing and deciphering facial movements and muscle activations to infer stress levels is the goal of facial expression analysis [24]. The use of computer vision algorithms to identify and follow the motion of facial features is an important part of facial expression analysis. It is feasible to extract facial features that reflect distinct emotional states, including stress, by examining the positions and movements of face landmarks such as the eyebrows, eyes, nose, and lips [25]. Methods including optical flow analysis [26], geometric modeling and facial landmark detection are frequently used in this setting. Analysis of facial expressions is commonly performed using the facial action coding system (FACS). Action units (AUs), which are used to code and measure face motions, are linked to individual muscle activations. Some permutations of AUs are more predictive of particular emotional expressions, such as those associated with stress than others. Researchers can deduce a person's stress levels from the detection and analysis of these AUs.

Convolutional neural networks (CNNs) and other deep learning advancements have significantly boosted the accuracy of facial expression interpretation. Emotion detection accuracy is greatly improved by CNN-based models' ability to learn to automatically extract discriminative characteristics from facial photos. These models are able to generalize well to new data since they are trained on big labeled datasets of facial expressions. Dynamic facial expressions, in addition to static ones, can reveal a lot about a person's stress levels. Micro-expressions, or fleeting facial expressions, are faint and transient, but they might betray hidden emotions or tension. These micro-expressions are captured with high-speed cameras or specialized equipment and then detected and interpreted with cutting-edge analysis methods. Significant progress in stress detection has been made by the combination of facial expression analysis and machine learning methods. Accurate and reliable stress detection systems have also been demonstrated to be possible with the help of ensemble approaches, which combine many classifiers or models.

Facial expression-based stress detection can benefit greatly from fusion methods. More information about a person's stress level can be gleaned through a combination of facial expression elements with physiological data and audio-based indicators. Stress can be detected by facial expressions, but this task is complicated by factors such as individual differences in expression, lighting, and occlusions, and the requirement for huge annotated datasets for deep learning model training. To overcome these obstacles and enhance the accuracy and generalizability of facial

expression-based stress detection systems, robust facial expression analysis techniques, such as preprocessing procedures, feature selection, and model training strategies, are necessary.

2.2 Machine learning and data analysis techniques for stress detection

2.2.1 Feature extraction and selection

Extracting and selecting significant characteristics from data collected via several modalities is crucial for efficient stress detection. The goal of stress-related feature extraction methods is to collect data that is meaningful and indicative of an individual's stress levels. Heart rate variability (HRV), skin conductance response (SCR), respiratory rate, and blood pressure are only a few of the properties that can be derived from physiological inputs. These characteristics shed light on the workings of the autonomic nervous system and the physiological alterations that occur in response to stress.

A number of acoustic parameters, such as pitch, intensity, speech rate, spectral properties, and voice quality, can be extracted from voice recordings and used for stress detection. Emotional and stress states can be detected by these observable differences in how people speak, sound, and use prosody. Using features taken from images or recordings of the face, stress can be detected. Various face expressions can be indicated by a combination of facial muscle movements represented by facial action units (AUs). Furthermore, information about a person's stress levels can be gleaned from their look, including geometric features like the distances between face landmarks and appearance-based features like texture patterns and color distributions. To improve the efficacy of stress detection models, researchers have developed feature selection methods to zero in on the most informative and discriminative features. The objective is to optimize the discriminative ability of the stress detection system while simultaneously decreasing the number of dimensions involved. Recursive feature elimination (RFE) and sequential forward/backward selection are two examples of wrapper methods that can be used with other approaches such as correlation analysis and mutual information. Time-varying and frequency-specific stress-related patterns in physiological and acoustic data can be revealed by employing advanced feature extraction techniques like wavelet transformations, empirical mode decomposition (EMD), and time-frequency analysis via spectrograms or wavelet scalograms.

When choosing features for stress detection, it is crucial to take into account the modality-specific traits and physiological relevance. The selection of significant traits can be guided by domain expertise and a prior understanding of stress's physiological and psychological elements. In addition, real-time processing, computational confusion, and interpretability are all factors that should guide the selection of features for a stress detection system. The dimensionality of the data can be reduced and the necessary information extracted from the data by combining effective feature selection methods with appropriate feature extraction techniques. This makes it easier to create models for detecting stress that is both effective and efficient, which in turn leads to more precise and trustworthy assessments of stress in a variety of contexts.

2.2.2 Machine learning models

Because they allow the creation of prediction models that can learn patterns and relationships from the collected variables, machine learning techniques play a significant role in stress detection. Stress detection tasks have been used with supervised, unsupervised, and semi-supervised machine learning techniques. Stress detection typically employs supervised learning methods due to the availability of a labeled dataset for model training. Support vector machines (SVM), random forests, decision trees, and neural networks are only a few of the classification techniques that have shown great promise in the classification of stress. These algorithms study the retrieved data and learn to categorize stress levels and make predictions about future stress conditions. For supervised learning models to work, a training dataset that accurately portrays the spectrum of stresses is necessary. When labeled data are scarce or nonexistent, unsupervised learning methods are used. On the basis of the collected features, clustering techniques like K-means, hierarchical clustering, and gaussian mixture models can classify stress patterns into groups with high similarity. Through the use of unsupervised learning, we can get insight into the variability of stress responses across persons and contexts. To train stress detection models, semi-supervised learning uses both labeled and unlabeled data. To drive the learning process, this method makes use of the small amount of labeled data available, while the vast amount of unlabeled data is used to capture a more comprehensive picture of stress patterns. In order to make the most of labeled and unlabeled data, several methods have been investigated for use in stress detection. These include self-training, co-training, and multi-view learning.

In recent years, there has been a lot of interest in using deep learning models for the detection of stress, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Convolutional neural networks (CNNs) are well-suited for facial expression-based stress detection due to their proficiency in extracting spatial characteristics

from pictures or spectrograms. However, RNNs excel in sequential data processing and have been used effectively in both audio and physiologically-based stress detection tasks. Stress detection research has also looked into transfer learning, a method that uses already-trained models on huge datasets. Even when there is a dearth of stress-specific data, stress detection systems can take advantage of the information obtained on comparable tasks or datasets by fine-tuning pre-trained models. Considerations such as data type, availability of labeled or unlabeled samples, model confusion, and human interpretability all play a role in deciding which machine learning model to use. To achieve reliable and effective stress detection, researchers must carefully pick and assess the right machine-learning algorithms based on these factors.

2.2.3 Data fusion and integration

For a more complete picture of an individual's stress level, it's helpful to combine data from several different modalities or sources. Data fusion and integration methods aim to increase stress detection performance by capitalizing on the synergistic benefits of many modalities [27]. Two main groups of methods exist for fusing data: early fusion and late fusion [28]. Raw or preprocessed characteristics from several modalities are fused early on to form a unified representation. It is possible to build a fused feature vector by concatenating or combining the features of different types of data, such as physiological signals, audio features, and facial expression features. Using early fusion methods, the model can pick up on complicated interactions between modalities and learn from integrated data. Data with various dimensions and sizes across modalities may present difficulties for early fusion. On the other hand, late fusion entails developing distinct models for each modality and integrating their forecasts at a later time. Voting, weighted averaging, and stacking are all viable options for combining the forecasts. By modeling each modality independently, late fusion approaches make it possible to include properties that are unique to each modality. In addition, late fusion can independently deal with modalities that have various dimensions and sizes. However, the interactions and dependencies between modalities in the early stages of the model may be missed by a late fusion.

Multi-modal fusion approaches [27], such as decision-level fusion and feature-level fusion, have also been investigated for use in stress detection, in addition to early and late fusion. The output decisions or scores from many modality-specific models are fused at the decision level to provide a single prediction. The predictions can be aggregated using a variety of fusion procedures, such as majority voting, weighted voting, and fusion based on fuzzy logic. However, feature-level fusion involves merging features from different modalities through statistical procedures like mean, median, or concatenation. The goal of feature-level fusion is to extract valuable, modality-specific information that can then be fed into the model.

Moreover, data integration methods take into account the incorporation of additional relevant data sources, such as contextual information or self-reported data, in addition to the fusion of several modalities. A person's stress reaction can be affected by their surroundings, their level of activity, the time of day, and their social contacts. Understanding stress in certain settings can be improved by combining contextual knowledge with multi-modal data. Subjective information regarding an individual's perceived stress levels and experiences can be gleaned from self-reported data gathered via questionnaires or diaries. A more complete picture of an individual's stress level can be obtained by combining self-reported data with objective measurements such as physiological, auditory, or facial expression data. Modality features, data availability, data source confusion, stress detection job difficulty, and computational resources are all important considerations when deciding on a data fusion and integration technique. When designing a stress detection system, scientists must weigh the benefits and drawbacks of various fusion methods to settle on the most appropriate strategy.

3 Methodology

In this research, we propose a comprehensive strategy for stress detection by fusing multiple modalities using deep learning architectures. In our approach, electrodermal activity, blood volume pressure, skin temperature, accelerometer, speech, and facial expressions are considered as modalities. Each modality is subjected to a rigorous preprocessing procedure to guarantee data quality and dependability. Early fusion occurs in two phases during the fusion process. First, the biosignal modalities (EDA, BVP, TEMP, and ACC) are combined to capture the temporal dynamics of physiological stress-responsive signals. The combined biosignal modality is then combined with facial expressions and audio waves to generate a comprehensive picture of stress levels. These fusion structures facilitate accurate and dependable detection of stress by capturing and interpreting pertinent information from multiple modalities. Figure 1 depicts the architecture and procedure of our multi-modality stress detection system, illustrating the techniques utilized in this investigation.

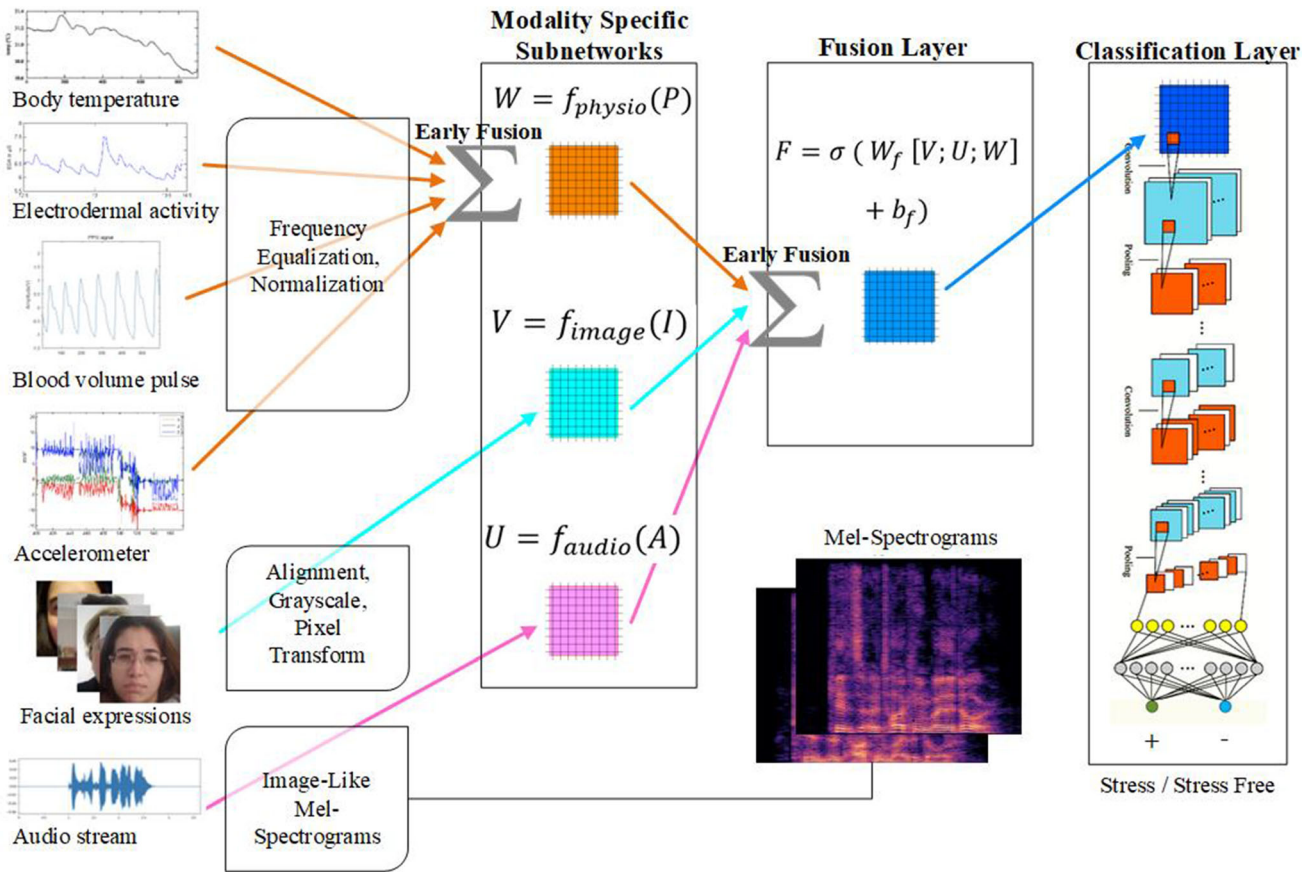

Fig. 1 The general pipeline of the proposed model for automatic stress recognition using facial expressions, audio, and physiological signals

3.1 Experience sampling mobile application development

The development of the mobile ES software necessitated a comprehensive strategy in order to produce a robust data collection instrument for our stress research. For efficient communication between the mobile application and the remote server, the software architecture utilized industry-standard protocols such as RESTful APIs. To collect real-time stress data, the ES software utilized sophisticated scheduling algorithms that allowed participants to report their stress levels and experiences at predetermined intervals or in response to specific events. The scheduling module managed the timing and frequency of prompts intelligently, taking into account variables such as participant availability, contextual relevance, and user preferences. In addition, the software was integrated with the device's native notification system, allowing for timely, non-intrusive prompts to optimize participant engagement while minimizing disruption to their daily activities. To assure data integrity and reduce recall bias, the ES software implemented secure and dependable mechanisms for data logging. User responses and contextual data were instantaneously encrypted and transmitted to the backend server. Data storage complied with stringent privacy and security protocols, and participants provided explicit consent for data collection and use. In addition, the software included data validation mechanisms to detect and manage incorrect or inconsistent inputs, ensuring the accuracy and reliability of the collected data. During the development process, compatibility and scalability played a significant role. The ES software was developed to support prominent mobile operating systems such as iOS and Android. The codebase adhered to industry standards and made use of modular and extensible architectures, allowing for future expansions and modifications. The backend infrastructure of the software was designed to manage concurrent user requests, employ load-balancing techniques, and scale horizontally to accommodate increasing participant numbers without compromising performance or data integrity.

After the application prototype was created, the type of ES application to be developed was determined. Our study's ES technique was intended to be a random half-hourly schedule with seven alerts each day between 10 a.m. and 8 p.m. Six times throughout the day, participants' smartphones would alert them with a link to an immediate survey. Electrodermal activity, heart rate, blood pressure, and skin temperature were all measured and recorded

concurrently for 15 min. Participants were also asked to document their own emotional experiences through photographs and audio recordings. No artificial stress was created for this study; rather, researchers looked for and graded stress in the context of everyday work. Stress in the workplace was studied by collecting data from a wide range of professions in an effort to produce a representative cross-section of society. The inclusion of both high- and low-stress professions enriched our understanding of stress in its many forms.

We determined which questions to be asked after determining the frequency of use. The structure and question order of the questionnaire were meticulously designed to encourage neutral and objective responses. The responses of participants and research outcomes may be influenced by factors such as the order and structure of the questions posed. In contrast to questions with predetermined response options, such as the Likert scale, open-ended questions enable respondents to respond on their own terms. While responses to open-ended inquiries tend to be more in-depth, the rate of non-responses may be higher. For grading purposes, the seven-point Likert scale is extensively employed. In this research, Likert scale questionnaires utilized both 4- and 5-point scales. The research utilized a number of Likert-scale questionnaires and a demographics questionnaire. Age, gender, occupation, marital status, smoking status, and medication use were among the eleven queries included in the questionnaire's demographic section. Six items comprised the brief questionnaire that was completed alongside physiological signals:

1. How do you currently feel? *Options: Happy, Unhappy, Good, Bad, Neutral, Relaxed, Satisfied, Energetic, Excited, Tired, Nervous, Sad, Angry, Worried, Lonely, Guilty, Sick, and Other*
2. What are you doing at the moment? *Options: Work, rest, food/drink, cleaning, sports, hobby, awareness work, listening to music, watching videos, free time, and other*
3. Is anyone else with you? *Options: Manager, Owner, Colleague, Family Members, Partner, Friend, Strangers, Pets, and Nobody*
4. Is there someone or something currently troubling you? *Options: Yes, No and I Don't Know*
5. Would you prefer to be somewhere else right now? *Options: Yes No and I Don't Know*
6. Do you have the energy to complete additional tasks today? *Options: Yes No and I Don't Know*

The general stress test, the PANAS brief scale, and other surveys all contributed to the investigation of the emotional states of study participants. The general stress test derives a stress score from the responses of participants to twenty scenarios of normal, ordinary life. It asks questions such as "Would you attempt to cross the street when the red light is about to turn on?" and rates your likelihood of doing so based on how close the light is to turning red. The PANAS scale is a concise, five-point Likert scale questionnaire used to assess positive and negative emotions. These surveys were conducted for statistical purposes and to gain a better understanding of the attitudes and behaviors of the sample population. The questionnaire's questions are enumerated below.

1. Are you attempting to juggle multiple jobs in a short period of time?
2. Do you grow impatient in cases of business disruptions or delays?
3. Do you always feel like you have to win in the games you play, even if it's for fun?
4. Do you try to cross the traffic light with your car when the red light is about to turn on?
5. Even if you need help with something you do, would you refrain from asking?
6. Do you always feel the need to earn the admiration of others and be respected?
7. Do you always criticize the way others do their job?
8. Do you frequently look at your watch or a clock?
9. Do you ever have excessive ambitions to improve your achievements and position?
10. Do you get the idea that time is not enough for you?
11. Do you have the habit of doing more than one task at once?
12. Do you often feel nervous or angry?
13. Do you find it difficult to find time for yourself and your hobbies?
14. Do you have a tendency to talk quickly or speed up conversations?
15. Do you consider yourself a difficult person to get along with?
16. Do your friends or relatives say that it's hard to get along with you?
17. Do you have a tendency to be involved in more than one project?
18. Do you often set deadlines for finishing your work?
19. Do you feel guilty when you take time off to rest or sit idle?
20. Do you ever put too much responsibility on yourself?

As a final step, different modalities for data collection have been established: (1) survey data on general stress positive–negative emotions, and real-time mood detection; (2) facial expressions captured in real time; (3) audio signals; and (4) physiological signal data (Fig. 2). Participants were required to create an account in order to continuously capture data for seven days. They were presented with a screen requesting demographic information after

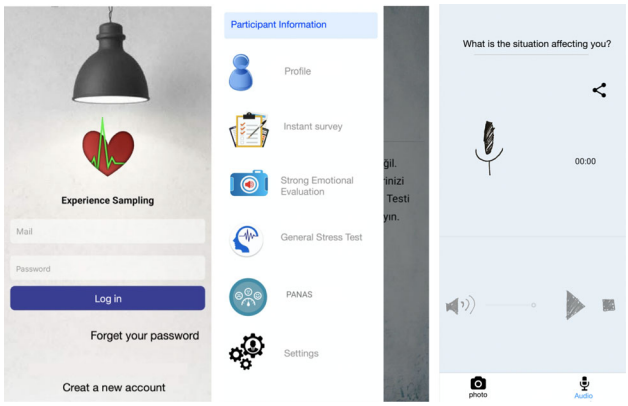

Fig. 2 Sample views of mobile app: **a** home screen, **b** navigation, and **c** audio signal recording screen

registering in. User profiles, instantaneous surveys, extreme emotional experiences, the General Stress Test, the PANAS scale, and application preferences were all available. The participants' demographic information was collected beforehand, and they then completed a 20-question survey titled the general stress test to assess their emotional reactivity to commonplace events. This standardized measure of stress utilized a four-point Likert scale with the following response options: "Never," "Sometimes," "Often," and "Always." After completing the survey, participants were provided with comprehensive feedback regarding their performance. When experiencing extreme emotions, participants captured both video and audio. Participants shot photographs expressing their current mood and discussed the experience verbally. The PANAS scale required participants to rate their level of five positive and five negative emotions on a Likert scale ranging from 0 to 5. The spectrum of emotions examined included initiative, concentration, inspiration, vigilance, anxiety, rage, melancholy, resentment, and humiliation. Before sending notifications, using their cameras, or recording their audio during emotional highs and lows, users were required to register for the application. Six notifications were sent throughout the course of the day to remind participants to review their instant self-reporting progress.

In addition to requiring participants to answer specific questions, the system also allows them to document photographs and sounds during times of extreme emotions, enabling the collection of biometric data quickly.

3.1.1 Dataset of Workplace Stress in 3 Dimension(WorkStress3D)

Workplace stress in three dimensions (WorkStress3D) is the name of the dataset we gathered, which consists of three modalities: biosignals, facial expressions, and speech signals. Throughout the research, participants wore a wrist-worn Empatica E4 sensor device. They were given questionnaires to complete and asked to disclose their physiological data collected over a seven-day period. The smartphone captured survey data and visual/auditory aspects, whereas the wrist-worn sensor device recorded physiological information including electrodermal activity, heart rate, blood volume pressure, skin temperature, and accelerometer readings. With the Experience Sampling (ES) application deployed on the mobile devices of participants, data administration was simplified. The survey questions, which consisted of six concise and direct inquiries, were crafted with great care to accurately capture respondents' sentiments. Participants were also encouraged to photograph and record audio recordings of their experiences with intense emotions. Four times a day for a maximum of 15 min each, physiological indicators were recorded. The wearable sensor device collected physiological data continuously for 150 min during each recording session, while the mobile application collected immediate survey data in real-time.

Twenty participants were included in the research. The participants included a computer engineer, a research assistant, a judge, a lawyer, a doctor, a marketer, an entrepreneur, a professor, a secretary, and a self-employed individual. There were approximately 35% women and 65% males, with a mean age of 38.7. Seventy percent of the participants were employed in the private sector, compared to thirty percent in the public sector. Twenty-five percent of the participants smoked, while 75% did not. Eighty percent of the participants in the study reported being in excellent health, while twenty percent reported having a medical condition. In addition to minor ailments and injuries, eczema and a heart condition were also mentioned. One participant bowed out prior to the conclusion of the study and was therefore not included in the final analysis.

In addition to the data acquired from the participants, the study used multiple public datasets to train the model. These datasets included Cohn-Kanade (CK), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS), and facial expression recognition 2013 (Fer2013). The audio datasets contained recordings of human utterances from various sources, while the facial expression datasets contained tagged photographs of people's facial expressions.

3.2 Proposed Deep Multi-modality Fusion Model

We present a deep multi-modality fusion model that combines the pre-processed image, audio signal, and physiological data to efficiently harness the information from many modalities and accomplish robust emotion

recognition. The model's goal is to accurately anticipate emotional states by capturing the supplementary properties of each modality and capitalizing on their synergistic benefits. In order to extract meaningful representations from raw data, the suggested fusion model uses a deep learning architecture.

3.3 Prepocessing of WorkStress3D Dataset

The WorkStress3D dataset was created by combining data from multiple sources to leverage the benefits of multiple modalities and enhance comprehension of the emotional states of participants. By integrating and extracting pertinent information, the purpose of the fusion dataset was to improve the data's performance and classification precision. Multiple iterations of data consolidation, reduction, and fusion were required to produce the final fusion dataset.

1. **Visual Data Pre-processing**: The original 1152 x 2048 pixels and RGB hues were lost in the preprocessing of the collected face expression images. It was necessary to separate the actual facial expression from the background in order to facilitate feature extraction and subsequent image processing. This was accomplished through face alignment, normalization, and augmentation. Additionally, all photographs were scaled down to the lowest resolution feasible for the study.
2. **Audio Signal Pre-processing**: The spectrum characteristics of the collected audio signals were investigated. Using a binary scale, we extracted power spectra and discovered that mel spectra capture the most significant auditory details. Following these procedures, the audio data were more accurately represented, which facilitated further analysis and fusion.
3. **Physiological Signal Pre-processing**: The physiological signals of participants were sampled at varying frequencies using the wrist-worn Empatica E4 sensor device. To resolve this issue, a downsampling technique was used to make the frequencies of the signals more uniform. Downsampling was chosen to achieve a balance between computing efficiency and accuracy. By downsampling all physiological inputs to 4 Hz, consistency and effective fusion were attained.
4. **Feature Transformation**: Quadratic features were subjected to polynomial feature transformations to capture higher-order correlations and enhance the representation of physiological signals. By integrating the physiological signals' raw data, additional features were generated. The objective was to more precisely depict the physiological responses of participants and to capture nonlinear interactions.

After we complete the preprocessing, we have labeled the dataset. The labeling process involved trained human annotators who carefully classified each data instance as either stressful or stress-free based on established criteria. The WorkStress3D dataset was generated by combining processed visual information, auditory signals, and physiological data. Using this dataset, researchers were able to combine and combine modalities to gain a more complete understanding of the emotions of individuals. The dataset served as the basis for additional modeling and analysis, enabling the development of efficient emotion recognition and classification algorithms.

3.3.1 Modality-Specific Subnetworks

The proposed deep multi-modality fusion model relies heavily on the specialized subnetworks for each input modality, which are responsible for extracting modality-specific features. As a result, the model can learn modality-specific representations that are particularly well-suited for emotion recognition based on the image, audio, and physiological inputs.

A function $f_{\text{image}}(\mathbf{I})$, where $\mathbf{I}$ is the input face image, stands in for the image subnetwork. To extract basic visual characteristics, we use a CNN architecture that has been pre-trained on a large-scale image dataset, such VGGNet or ResNet, when we fine-tuned the CNN layers for emotion recognition to tailor the learned characteristics to the problem at hand. Here is how we can write down the image subnetwork:

$$\mathbf{V} = f_{\text{image}}(\mathbf{I}) \tag{1}$$

where $\mathbf{V}$ stands for the image characteristics that are unique to a given modality.

The audio subnetwork is represented by the notation $f_{\text{audio}}(\mathbf{A})$, where $\mathbf{A}$ is the audio signal that is being input. We used specialized audio analysis networks trained to detect temporal dynamics and auditory patterns, such as RNNs and CNNs. The audio subnetwork can be written in the form of:

$$\mathbf{U} = f_{\text{audio}}(\mathbf{A}) \tag{2}$$

where $\mathbf{U}$ stands for the acoustic characteristics unique to the modality in question.

Similarly, the input physiological signals are denoted by $\mathbf{P}$ in the representation of the physiological subnetwork, $f_{\text{physio}}(\mathbf{P})$. In order to describe the temporal dynamics and spatial patterns in the physiological data, we used either RNNs or CNNs. The physiological subnetwork can be written as follows using a convolutional network:

$$\mathbf{W} = f_{\text{physio}}(\mathbf{P}) \tag{3}$$

where **W** stands for physiological characteristics unique to the modality in question.

Subnetworks tailored to each modality are trained and optimized using loss functions and optimization techniques. The individual networks are trained to separate out modality-specific characteristics. In this way, the model is guaranteed to accurately collect the specific data associated with each modality, paving the way for a multifaceted comprehension of emotional states.

In conclusion, the **V**, **U**, and **W** features are the outputs of the modality-specific subnetworks, which extract modality-specific features from the input data. These characteristics serve as rich representations that aid in the fusing process of our proposed deep multi-modality fusion model, which in turn makes it easier to gain a holistic comprehension of human emotions.

3.3.2 Fusion Layers

We suggested deep multi-modality fusion model uses a fusion layer to bring together information from several sources, such as images, speech, and physiological data, into a single cohesive representation. Image, sound, and physiological characteristics that are unique to each modality will be denoted by **V**, **U**, and **W**. The fusion layer combines all of these properties into one unified representation, denoted by the notation **F**. With a fully connected layer and a nonlinear activation function, the fusion layer can be realized. To express the fusion process mathematically, we have:

$$\mathbf{F} = \sigma(\mathbf{W}_f[\mathbf{V};\mathbf{U};\mathbf{W}] + \mathbf{b}_f) \tag{4}$$

where $\mathbf{W}_f$ is the weight matrix and $\mathbf{b}_f$ is the bias vector of the fusion layer. The union of the modality-specific characteristics along the feature dimension is represented by the notation $[\mathbf{V};\mathbf{U};\mathbf{W}]$. An example of a nonlinear activation function is the sigmoid or rectified linear unit (ReLU), which is represented by the function $\sigma(\cdot)$. The model is able to capture the intricate interactions and dependencies across modalities because of the fusion layer's ability to incorporate complementary information from numerous modalities. The picture, audio, and physiological subnetworks' feature extractions are fused at the fusion layer to produce a representation that is representative of a holistic understanding of emotions.

The fusion layer's $\mathbf{W}_f$ and $\mathbf{b}_f$ parameters are trained to find the best possible fusion weights for the emotion recognition task as a whole. The fusion layer connects modality-specific data to a comprehensive emotion prediction. As has been demonstrated, a fully connected layer with a nonlinear activation function is used in the fusion layer to integrate the **V**, **U**, and **W** features that are unique to each modality. The model's ability to accurately predict feelings is due, in part, to the fact that the fused representation**F** captures the comprehensive information from many modalities.

3.3.3 Classification layer

Predicting emotions using the fused representation **F** acquired from the fusion layer is the job of the classification layer in the proposed deep multi-modality fusion model. This layer uses a softmax activation function after a completely connected base layer. Let us call the set of probabilities associated with each emotion category that has been predicted**Y**. The mathematical definition of the categorization layer is as follows:

$$\mathbf{Y} = softmax(\mathbf{W}_c\mathbf{F} + \mathbf{b}_c) \tag{5}$$

where $\mathbf{W}_c$ is the weight matrix of the classification layer and$\mathbf{b}_c$ is the bias vector. Matrix multiplication, represented by the symbol $\cdot$.

The predicted probability of emotions is meaningful and interpretable since the softmax activation function assures that they always add up to 1. This term means:

$$softmax(x) = \frac{exp(x)}{\sum_{i=1}^{C} exp(x_i)} \tag{6}$$

where **x** is a real-valued score vector and C is the total number of emotion categories.

The loss between the predicted probability **Y** and the ground truth emotion labels is minimized by training with optimum values for the classification layer parameters $\mathbf{W}_c$ and $\mathbf{b}_c$. The classification layer is the last step of the deep multi-modality fusion model, and it is responsible for converting the fused representation **F** into emotion probabilities. The model applies the softmax function to each class of emotions, giving each class a probability that represents how likely the input sample is to belong to that class.

As all the above points have demonstrated, predicting the emotion probabilities **Y** from the fused representation **F** is the job of the classification layer, which employs a fully connected layer followed by the softmax activation function. The probabilities are normalized and easily understood thanks to the softmax function. During training, the model's performance in emotion classification is optimized by learning the values of the parameters $\mathbf{W}_c$ and $\mathbf{b}_c$.

4 Experimental results

The experimental results section of this study investigates emotional responses by analyzing biosignals collected from multiple sensors as well as images and audio signals

obtained through the mobile application. Using a comprehensive list of seven emotion labels, the video and audio data were methodically analyzed to capture a wide range of emotional states. Positive emotions were used to symbolize stress-free situations, while negative emotions were used to represent stressful situations. In order to create a labeled dataset for stress prediction, the seven-day experiment survey responses of the participants were considered. The survey responses served as a valuable resource for distinguishing stress-related emotional manifestations from physiological signals. Specifically, the positive emotional expressions were labeled as "stress-free," signifying emotional well-being, whereas the negative emotional expressions were labeled as "stressful," indicating instances of elevated stress. This study sought to develop a comprehensive comprehension of the relationship between emotional responses and stress levels by meticulously labeling and classifying the emotional expressions captured from the collected data. In all experiments, datasets were separated into training and test sets of 80% and 20%, respectively. This analysis provides the basis for the subsequent evaluation and performance evaluation of the proposed deep multi-modality fusion model.

4.1 Survey findings and implications

We began by conducting a series of assessments with the goal of better comprehending how the participants' baseline circumstances were affected by their surrounding environment. We present the results of our analysis regarding the activities that participants engage in when experiencing negative emotions such as anger, fear, rage, anxiety, unhappiness, or sadness in order to gain a deeper understanding of the relationship between negative emotions, stress factors, and individual experiences. By analyzing the activities that participants engaged in during instances of negative emotions, we intended to identify key triggers and contexts that contribute to the occurrence of such emotions, particularly in work-related and social contexts. According to the results of the survey, the vast majority of respondents' unpleasant emotions (62%) occurred during job activities, followed by conversations with coworkers (19%), eating/drinking (9%), interacting on social media (4%), and doing chores (3%). Understanding these triggers can aid in the development of targeted interventions and strategies to reduce negative emotions and enhance well-being in these situations.

The most stressful experiences of the participants were analyzed to see if there were any commonalities that could shed light on the sources of stress in people's lives. The survey found that 71% of people who try to beat the signal at a crosswalk when it's about to turn red are under stress. A significant rate of association (77%) was also found between stress and a tendency toward rapid speech and accelerated discourse. In addition, 76 percent of respondents identified as "someone who is difficult to get along with," a trait that has been shown to increase the likelihood of experiencing stress. Physical fitness may have an effect on stress levels, as seen by the 84% association found between the individuals' body mass index and stress.

By comparing respondents' PANAS-reported feelings to their total stress levels, we were able to delve deeper into the psychological aspects of stress. The poll results (Table 1) showed a 0.97 correlation between being unhappy and being stressed. Similarly, there was a 0.96 association between shame and stress. The association between vigilance and stress was calculated to be 0.82, which is statistically significant.

As a whole, the investigation of survey results, combined with the descriptive results and numerical evaluation measures, provides a comprehensive picture of the relationship between negative emotions, stress factors, and individual experiences, with valuable implications for designing targeted interventions, promoting well-being, and nurturing psychological resilience in work-related and social contexts.

4.2 Stress modeling via individual data source

In this paper, we report on the findings of our separate studies analyzing facial expressions and audio data to determine the relevance of each of these types of data for the prediction of stress. The purpose of this experiment is to evaluate the efficacy of various data sets in terms of producing accurate forecasts of stress. For the analysis, in particular, CNN- and RNN-based deep learning models are utilized.

Table 1 The relationship between stress and emotions as measured by the Spearman Correlation between PANAS scale parameters and general stress scores

Emotions	Correlation value
Active	0.04
Determined	0.54
Attention	0.57
Inspired	0.63
Vigilance	0.82
Fear	0.05
Angry	0.09
Unhappiness	0.97
Hostility	0.22
Shame	0.96

4.2.1 Face expressions as an indicator of stress

Expressions on the face are an excellent indicator of emotional states, including stress. Figure 3 shows a sample cross-section of facial expressions collected from participants. We used a CNN model that was trained to classify stress levels based on facial expression data in order to evaluate the efficacy of facial expressions in stress prediction. Prior to classifying stress based on facial expressions, a succession of preprocessing steps was carried out. The facial expression data were converted to pixel representations and normalized to ensure consistency and eliminate any possible biases.

Figure 4 depicts the classification outcomes of the deep learning model for stress prediction based on facial expressions.

The WorkStress3D dataset containing 1375 stress-free and 1677 strained facial expressions was utilized to generate the confusion matrix. 1106 (80.43%) of the stress-free expressions were correctly identified as stress-free, while 269 (19.57%) were incorrectly predicted as anxious. 1542 (91.95%) of the strained facial expressions were correctly identified as stressed, while 135 (8.05%) were incorrectly identified as stress-free. Additionally, Table 2 shows the performance of the CNN model, which indicates high precision, recall, and an average F1 score of 85% for stress prediction based on facial expressions.

When studied with deep learning strategies, the findings suggest that facial expressions have the potential to serve as a trustworthy indicator for the prediction of stress, achieving high levels of accuracy and performance.

4.2.2 Speech signal as an indicator of stress

One of our experiments involved detecting stress through speech by analyzing audio signals to identify and classify individuals' stress levels. This method transforms raw audio data into pixel representations with Mel-spectrograms, which provide a visual representation of the audio

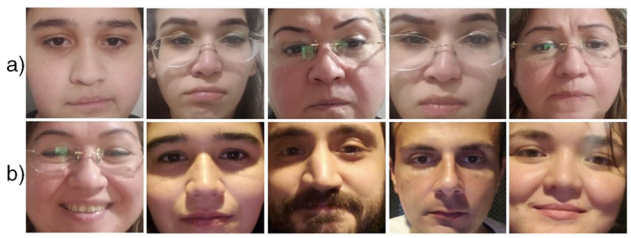

Fig. 3 Sample images from the WorkStress3D dataset: **a** facial expressions displaying signs of stress and tension, including furrowed brows and tense muscles, indicating elevated levels of stress, **b** facial expressions reflect a relaxed and tranquil state with a pleasant demeanor and the absence of stress-related characteristics, indicating low levels of stress

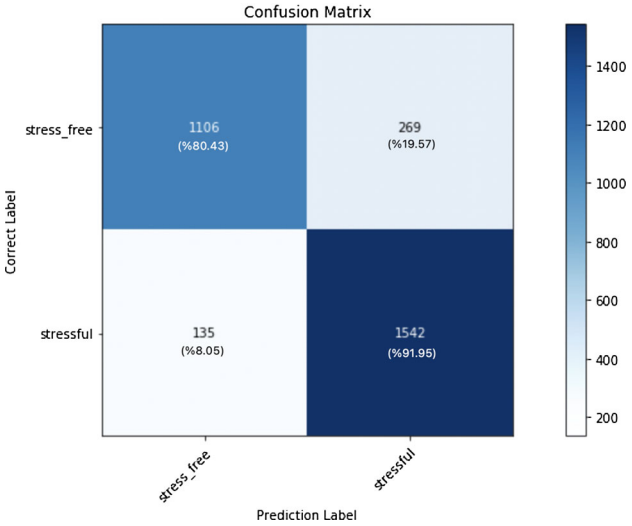

Fig. 4 Confusion matrix for estimating stress from facial expressions

signal. Figure 5 illustrates the waveforms and power spectrum of stressed audio.

The visual representations enable us to analyze the acoustic characteristics of stressful audio samples. It ensures that noises that are equally spaced on the scale are perceived as being equally spaced by humans. Using the Mel Scale, speech processing can capture the subtleties of how humans interpret sounds. Mel Spectrograms are visual representations of audio on the Mel Scale as opposed to the frequency domain. They enable us to visualize how distinct shapes and patterns are assumed by various sounds. These Mel Spectrograms are transformed into Mel Frequency Cepstral Coefficients (MFCCs) for stress detection in order to capture temporal dynamics between audio frames.

To facilitate stress detection, it is necessary to generate Mel-spectrograms that resemble images. This involves producing Mel-spectrogram segments that resemble RGB images and meet the input specifications of image CNNs. The audio signal is converted into three channels representing the static, delta, and delta-delta characteristics of Mel-spectrogram segments, with dimensions suitable for CNN model input. Figure 6 depicts the confusion matrix generated by the deep learning model for stress prediction using audio data.

Table 2 Performance of the CNN model for facial expression-based stress prediction

Stress situation	Precision	Recall	F1 score
Stress free	0.89	0.80	0.84
Stressful	0.85	0.91	0.87

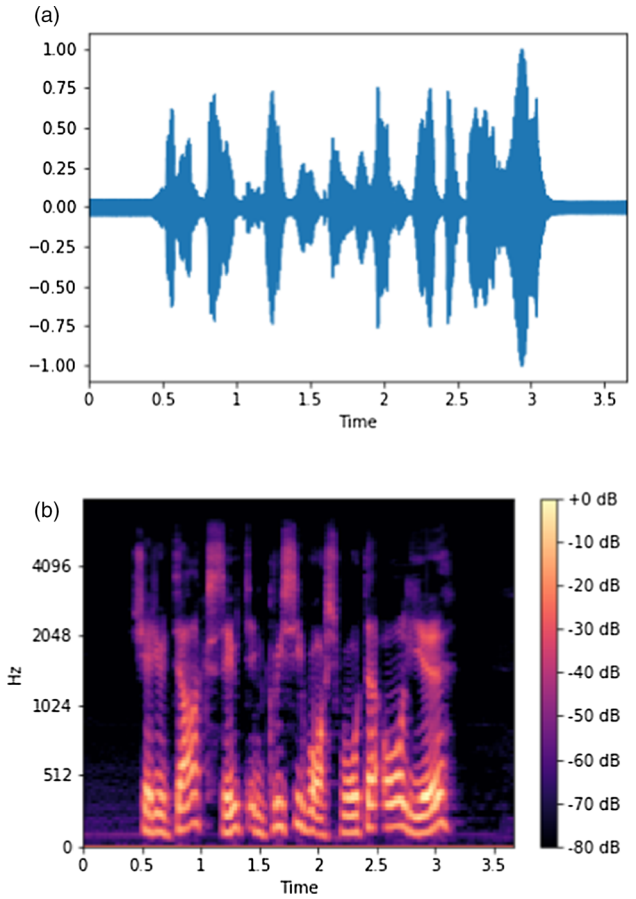

Fig. 5 Sample views of **a** the stressful audio signal and **b** the signal's power spectrum

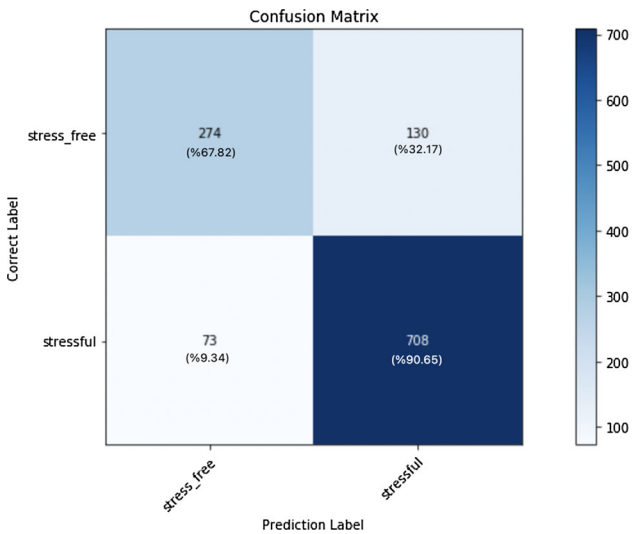

Fig. 6 Confusion matrix for estimating stress from audio data

The dataset containing 404 stress-free and 781 stressful audio samples was used to generate the confusion matrix. 274 (67.82%) of the stress-free samples were correctly classified as stress-free, whereas 130 (32.17%) were incorrectly predicted as anxious. In the case of stressed audio samples, 707 (90.65%) were correctly identified as stressed, whereas 73 (9.34%) were incorrectly identified as stress-free. The performance of the CNN model is summarized in Table 3, which demonstrates high precision, recall, and an average F1 score of 79% for stress prediction based on audio data.

The results demonstrate the efficacy of audio data analysis for predicting stress. We can accurately identify stress levels by analyzing the acoustic characteristics of speech and other audio signals. This provides vital insight into the emotional states of individuals and enables the development of targeted interventions and stress management support systems.

4.2.3 Biosignal as an indicator of stress

The aim of the investigation was to discover how the participants' sympathetic nervous systems responded to stress. The obtained physiological data is divided into two groups, "stressed" and "stress-free,". Under these conditions, the distributions of physiological signals show that 65.9% of the data is related to stress, whereas 34.1% represents stress-free states. It should be noted that although environmental conditions and the position of the device electrodes during measurement could have an effect, it is not considered significant in this investigation. EDA has a sample rate of 4 Hz and is one of the physiological signals of interest. There are three primary aspects of EDA: skin conductivity, phasic, and tonic components. To discriminate between tonic and phasic characteristics, a threshold of about 0.05 ms is applied to the skin conductivity reading. Phasic parameters are those that happen in response to stimuli, while tonic parameters describe ongoing electrical processes. Changes in these factors are thought to be indicative of arousal states. Examples of EDA signals in stressful and non-stressful conditions are shown in Fig. 7.

Data on skin temperature, recorded at a constant 4 Hz and expressed in degrees Celsius, show higher variation under stress than under normal circumstances. Figure 8 depicts the range of skin temperatures experienced by a person in a variety of conditions.

Table 3 CNN model performances for audio data

Stress situation	Precision	Recall	F1 score
Stress free	0.78	0.67	0.72
Stressful	0.84	0.90	0.86

Blood pressure data is provided by the BVP signal, which is sampled at a fixed rate of 64 Hz. BVP has greater frequency and amplitude under stress compared to non-stressful conditions, as seen by the signals displayed in Fig. 9 acquired by the Photoplethysmography (PPG) sensor.

In addition, the study makes use of accelerometer data collected at a constant rate of 64 Hz to measure the continuous gravitational force (g) in all three spatial dimensions (x, y, and z). The stress-free and stressful accelerometer signals are shown in Fig. 10.

All the gathered physiological signals are normalized and modeled across 15- , 30- , and 60-second intervals. In this study, we evaluate four different neural network models for stress detection throughout these intervals. The test results for each model and data frame combination are shown in Table 4.

According to the statistics in Table 4, the non-sequential DNN model produces the best performance for all window sizes. This model's superior performance can be attributed to its parallel design. Figure 11 shows that the best results come from using a time window of 60 s.

The performance of triple signal combinations in terms of stress analysis is evaluated using the optimal model and physiological signal windows. The validation accuracy of the non-sequential DNN model is presented in Table 5. The accuracy is shown for triple and binary combinations of physiological inputs measured over a period of 60 s.

According to the data presented in Table 5, the signal combination of EDA and BVP is the most efficient one for reaching a high level of performance. When compared to other combinations of physiological signals, the performance of this combination of signals is significantly higher when modeled in binary form. The fundamental purpose of this research is to construct a model for the early detection of stress in order to address issues related to the workplace, including stress and depression brought on by negative emotions. The identification of stress can be a first step toward its management, and it also has the potential to lead to research on the avoidance of stress.

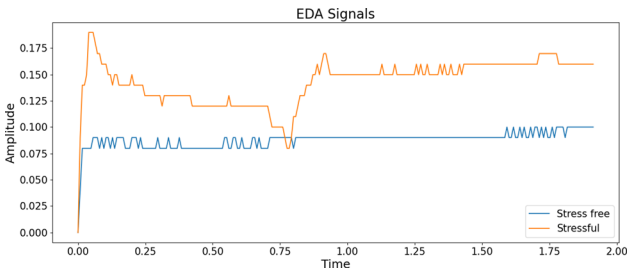

Fig. 7 Sample from a raw signal of EDA that gathered under stress-free and stressful situations

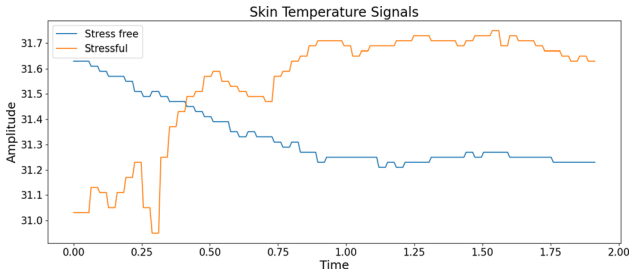

Fig. 8 Sample from a skin temperature signal that gathered under stress-free positive emotions and stressful situations

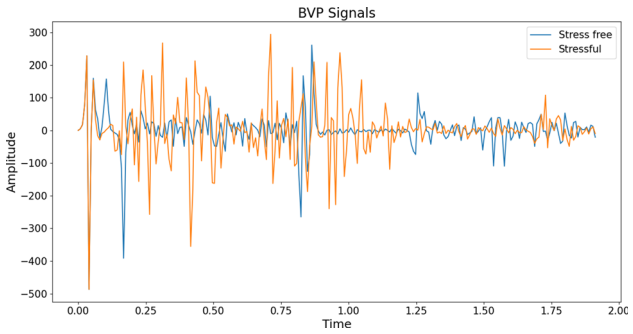

Fig. 9 BVP signal sample for stress-free positive emotions and stressful situations

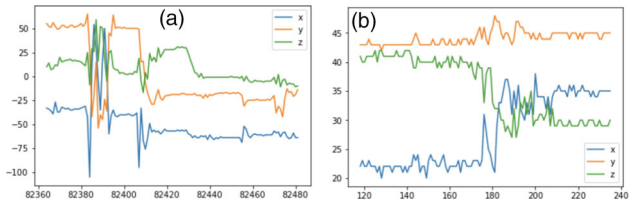

Fig. 10 Sample view from accelerometer signal for **a** stress-free positive emotions and **b** stressful situations

4.3 Stress modeling with multi-modal fusion

We have used the early fusion technique to create a complex fusion model in order to take advantage of the diverse and complementary data available from many sources. This approach incorporates acoustic data, biosignals, and facial expressions into a unified framework for stress prediction. To improve the model's stress classification accuracy and robustness, we fuse these different modalities early on to allow for joint analysis and feature extraction. Facial expressions, audio data, and biosignals can potentially be readily combined through the fusion model's well-thought-out architecture. The early fusion technique, which allows the model to merge the input from many modalities at an early stage, is key to this architecture. The model is able to identify and capture complex patterns and correlations between various modalities and stress levels because of the

Table 4 Comparison of testing accuracy and loss results for different data frames and network architectures

	15 sec frames	30 sec frames	60 sec frames
Models	Acc - Loss	Acc - Loss	Acc - Loss
Sequential DNN	0.89–0.27	0.91–0.21	0.91–0.21
Non sequential DNN	0.91–0.23	0.93–0.15	0.94–0.18
Sequential LSTM	0.72–0.53	0.85–0.34	0.87–0.28
Non sequential Bi-LSTM	0.84–0.36	0.88–0.28	0.86–0.32
Sequential GRU	0.86–0.33	0.91–0.21	0.92–0.18

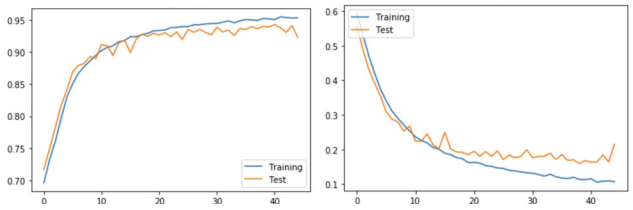

Fig. 11 Accuracy and loss results in the context of the 60-second time window

Table 5 Non-sequential DNN model performances for triple and binary combinations of 60 s long physiological signals

Signal Combinations	Val. Acc.
EDA, BVP, TEMP	68
BVP, TEMP, ACC	63
EDA, TEMP, ACC	66
EDA, BVP, ACC	67
EDA, BVP	70
BVP, TEMP	69
EDA, TEMP	70
TEMP, ACC	65
EDA, ACC	67
BVP, ACC	70

integrated representation that is sent down through the layers.

In particular, we used an early fusion method to incorporate biosignals as the major modality into the overall fusion process. In the first phase of our preliminary fusion strategy, we merged the biosignal data sets together. We generated a unified picture of the individual's physiological state under stress by combining signals from the individual's EDA, BVP, skin temperature, and accelerometer data. We were able to capture the interaction and synergistic effects of many biosignals in stress detection by fusing them early on. Next, we incorporated the remaining modalities, including speech and facial emotions, into the early fusion process by creating a multi-channel feature map. By combining the biosignals with these other sources of information, we were able to create a multi-modal depiction of stress that was both physiological and behavioral in nature.

We have done extensive experiments using the training set to systematically evaluate the efficacy of our fusion model. Eighty percent of the multi-modal data was used to train the fusion model, while the remaining twenty percent was used as a test dataset for careful evaluation of the model's efficacy. We use a 10-fold cross-validation method to achieve accurate performance evaluation using physiological signals. In the training phase, we utilize the use of the sparse categorical cross-entropy loss function, which is useful for classification tasks that involve a number of different classes. Seven distinct layers comprise the well-defined architecture of the CNN utilized in this investigation. These layers consist of six convolutional layers and one completely connected layer, contributing to the robustness and expressiveness of the network. In the initial convolutional layer (C1), 64 particles with a square shape and 5x5 dimensions are utilized. These kernels are convolved with the input data using the same 2x2 pixel padding and stride configuration. The succeeding convolutional layer, C2, consists of 64 5x5 filters with a 2x2 stride setting. The C3 and C4 layers inherit the same number of filters, stride, buffering, and kernel size as the C2 layer, but with 128 more kernels. The C5 and C6 layers that follow introduce 256 kernels of size 3x3 with a stride setting of 2x2 pixels. The final convolutional layer is followed by a flattening layer, which reformats the output into a vector format to facilitate its integration with the next fully connected layer. The 128 neurons in the fully connected layer contribute to the extraction of higher-level features. Notably, the final entirely connected layer has the same number of neurons as the classes being classified, ensuring a proper mapping of features to target classes. We also fine-tuned the following parameters: learning rates (e.g., 0.001, 0.01), batch sizes (e.g., 16, 32), and regularization techniques (e.g., L1 regularization, dropout).

In order to achieve an optimal value for this loss function, the models are trained, which ultimately enables a more accurate classification of the various levels of stress. In terms of stress prediction, our fusion model performs exceptionally well, as evidenced by its high levels of accuracy, precision, recall, and F1 score. These metrics are

strong indicators of the model's ability to distinguish between stress-free and stressful circumstances. The fusion model has shown exceptional proficiency in capturing the numerous linkages and nuances contained in the multimodal data, with an average F1 score of 0.94 which can be shown in Fig. 12.

The proposed multimodel fusion model's stellar results demonstrate the value of using early fusion to combine data from multiple sources. The model goes beyond the limitations of individual data sources to provide a more complete picture of an individual's emotional state by capitalizing on the complimentary nature of facial expressions, audio data, and biosignals. According to this all-encompassing knowledge, the proposed model can reliably anticipate stresses with high precision. In addition to advancing stress analysis as a whole, the proposed fusion model's successful implementation paves the way for the creation of very complex systems and applications in the real world. Individuals can be equipped with the tools they need to effectively manage and reduce stress through the use of multi-modal data analysis to the design of targeted treatments and supportive infrastructures. The potential for this innovative strategy to revolutionize stress management practices, improve people's overall health, and cultivate psychological resilience is enormous.

4.4 Impact of transfer learning on stress modeling

In our pursuit of accurate stress modeling, we initially investigated the use of state-of-the-art pre-trained models, including VGG16 and ResNet, as the base models. These models, which have been pre-trained on massive datasets for image recognition tasks, are renowned for their potent feature extraction capabilities. However, despite their success in image-related tasks, applying these pre-trained models directly to our biosignal data for modeling stress yielded suboptimal results, with accuracy rates scarcely exceeding 60%. This result suggested that the generic features learned by these models may not effectively capture the unique patterns and subtleties present in stress-related biosignals. We turned our focus to custom transfer learning in order to address this difficulty. Our goal was to transfer the knowledge gained by models trained on audio and image data, which are more closely related to biosignal analysis, to our biosignal stress modeling task since it captures together during experiments. We utilized a two-step procedure to accomplish this. Initially, we selected the model that had been trained on audio and image data with the CNN model. This pre-trained model had learned to extract pertinent facial expression clues and auditory features, such as prosody, intonation, and spectral characteristics, which could potentially capture crucial information about a speaker's emotional state during speech. Using the biosignal dataset, we then fine-tuned the previously trained model that can be seen in Fig. 13. This required retraining specific layers of the model while leaving the earlier layers unchanged, thereby preserving the learned representations from speech and facial images. We intended to adapt the pre-trained model to the unique characteristics of stress-related biosignals in this manner.

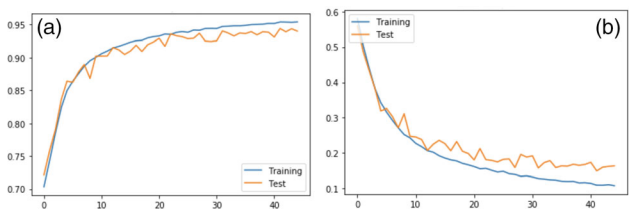

Fig. 12 Performance of the proposed multi-model fusion model: **a** accuracy and **b** loss curve

Table 6 demonstrates the beneficial effects of artificial neural network models for stress detection based on biological signals. The table summarizes, for various data window lengths for physiological signals, the success and loss rates attained by various models.

Three distinct window frames are used to evaluate the models: 15 s, 30 s, and 60 s. Across all frames, the non-sequential DNN consistently demonstrates greater accuracy and lower loss rates than the other evaluated models. Notably, the non-sequential DNN attains accuracy values in the range of 0.92 to 0.93 and loss values in the range of 0.21 to 0.17, indicating its robust performance in stress detection. In comparison, the sequential DNN, sequential LSTM, non-sequential Bidirectional LSTM, and sequential GRU models exhibit lower accuracy and greater loss rates. These models obtain accuracy values ranging from 0.77 to 0.93 and loss values ranging from 0.19 to 0.46, indicating that their stress detection capabilities have room for improvement.

We aimed to improve biosignal-based stress recognition by leveraging the feature extraction capabilities of the pre-trained model through transfer learning. By incorporating the learned representations and patterns from audio and video modalities, we sought to improve our biosignal stress modeling framework. These results highlight the promise of transfer learning as a beneficial strategy in stress modeling since it allows fusion structures to draw on the expertise of pre-trained models to enhance their sensitivity and specificity in stress detection. More work is needed to fully understand and improve transfer learning's potential for improving stress detection over a wide range of window lengths and application domains.

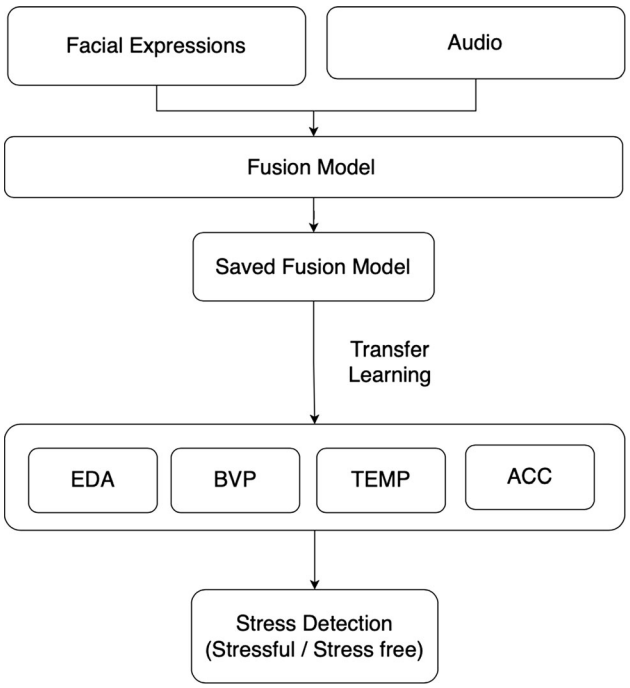

Fig. 13 Transfer learning technique with facial expressions, speech data and biosignal

5 Discussion

5.1 Main findings

The discussion of the key findings of our study delves deeply into the intricate complexities of multi-modality integration for stress detection. Our research uncovers the profound impact of integrating multiple modalities, such as biosignals, facial expressions, and sound waves, on the accuracy, robustness, and reliability of stress detection models through meticulous experimentation and in-depth analysis. This study demonstrates the remarkable efficacy of the double early fusion method, which combines biosignals, facial expressions, and audio signals simultaneously. By orchestrating this fusion with a sophisticated 2D CNN architecture, we successfully synchronize the rich visual and auditory stimuli that encompass both physiological responses and emotional manifestations. This fusion of diverse modalities generates a comprehensive and multidimensional representation of stress levels, thereby significantly enhancing the discriminative power of the stress detection system.

Moreover, our research on the inherent temporal dynamics of physiological signals reveals invaluable insights into stress-related events. Utilizing the capabilities of time series models, we decode the biosignals' intricate patterns, fluctuations, and temporal dependencies. By meticulously examining the temporal aspect, we reveal the subtle nuances of stress progression and dynamics, thereby enabling a granular and subtle analysis of stress over time. Intricate interactions between deep learning architectures and multi-modal data integration provide additional support for the impressive results we obtained. The potent combination of CNN and LSTM models demonstrates their exceptional capacity to capture and decipher intricate relationships and patterns embedded within multi-modal data. The multilayered neural networks effectively exploit the heterogeneous information present in biosignals, facial expressions, and sound waves by adapting to the various data sources. This adaptability enables the models to extract and exploit latent characteristics and correlations, resulting in enhanced stress detection capabilities.

5.2 Threats to validity

Potential threats to validity that could affect the reliability and generalizability of the findings should be taken into account when evaluating the results of any research project [29]. Research is said to have high validity if it reliably measures the variables of interest and if the findings can be extrapolated to other populations or settings. Here, we discuss about the risks to internal and external validity that our study of experience sampling for identifying mental stress faced. Our capacity to interpret and apply our results is enhanced by our attention to and consideration of these limitations.

5.2.1 Internal validity

The study may be limited in its applicability because of sampling error caused by the relatively small amount of participants. Selection biases, such as those based on

Table 6 Effects of transfer learning technique on fusion structures

	15 sec frames	30 sec frames	60 sec frames
Models	Acc - Loss	Acc - Loss	Acc - Loss
Sequential DNN	0.89–0.26	0.92–0.21	0.91–0.22
Non sequential DNN	0.92–0.21	0.92–0.19	0.93–0.17
Sequential LSTM	0.77–0.46	0.81–0.40	0.83–0.36
Non sequential Bi-LSTM	0.86–0.33	0.89–0.26	0.91–0.22
Sequential GRU	0.87–0.31	0.92–0.20	0.93–0.19

demographics or other traits, could potentially influence the findings. Confounding variables or demand characteristics may influence participant responses and affect the validity of the results of the experimental design of the study, including the length of the experiments and the specific tasks assigned to participants, which is not carefully considered. Measurement reliability concerns the potential for measurement errors or noise to affect the results of the study due to variations in the accuracy and reliability of the physiological sensors used to gather data, such as BVP, EDA, and BT, as well as the image and auditory data.

5.2.2 External validity

It is possible that the results of this study cannot be extrapolated to other people or settings. The people who took part in this study were chosen according to certain criteria, thus their opinions might not reflect those of the general public. Although data is obtained in a natural setting with participants going about their regular workday, not all participants can reflect jobs with the same level of stress as white-collar workers. From the perspective of ecological validity, it is possible that the complexity and diversity of real-world mental stress circumstances were not captured by the use of experience sampling and the specific tasks assigned to participants. Only job-related stress was examined. It is possible that the ecological validity of the findings and their application in natural settings for everyday stress only are constrained by the fact that the participants selected for the experimental setup did not experience particularly acute stress.

6 Conclusions

The objective of this research was to examine the feasibility of individual stress modeling by fusing information from multiple sources (facial expressions, speech, and biosignals) into a single model utilizing early fusion. The research has led to some significant discoveries and additions to the discipline of stress analysis. In the first place, we evaluated the multi-model fusion structure and discovered that fusing facial expressions, speech, and biosignals together increased stress classification performance over using any one modality alone. The proposed fusion model demonstrates its remarkable precision and accuracy in detecting and classifying stress levels, with an F1 score of 0.94 and a low loss of 0.18, exceeding the results of single modalities. Classifying stresses accurately also required picking the right time span in the sequential data. We discovered that a 30-second time window was the most effective, followed by 15- and 60-second ones. Given these results, a moderate time window duration appears to be the sweet spot for recording essential physiological and behavioral patterns while avoiding the loss of crucial temporal information. The proposed research emphasizes the relevance of considering numerous data sources and ideal frame size, as well as the efficacy of a multi-model fusion strategy, and so on. In addition to the fusion model, we conducted a stress prediction experiment using transfer learning, but we were unable to surpass the performance of the fusion model in stress detection despite utilizing transfer learning techniques. In spite of this, we have reduced the loss score by 0.01 points. Transfer learning yielded an accuracy of 0.93 and a loss of 0.17, which were inferior to the fusion model's remarkable results. Although transfer learning has been shown to be effective in numerous domains, it appears that stress analysis requires a more specialized approach. Despite this, the obtained results demonstrate a high level of accuracy and emphasize the significance of selecting the most appropriate model architecture for a given task. To explore alternative strategies and potentially build upon these findings, additional research and refinement are necessary. These results add to the growing body of knowledge on stress analysis and provide clues for how to improve existing stress detection programs.

Acknowledgements This work was supported by the Scientific Research Projects Coordination Unit of Istanbul Kultur University with project number: IKU-BAP2012. Within the scope of the study, data were collected with the permission of the ethics committee of Istanbul Kultur University, with the decision dated 20.05.2020 and numbered 2020.29.

Data availability The WorkStress3D dataset generated during and/or analyzed during the current study is available in the Mendeley repository https://data.mendeley.com/datasets/t93xcwm75r/5.

Declarations

Conflict of Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Jacobs N, Myin-Germeys Inez, Cathérine Derom P, Delespaul J Van, Os, and NA Nicolson, (2007) A momentary assessment study of the relationship between affective and adrenocortical stress responses in daily life. Biol Psychol 74(1):60–66
2. Cohen Sheldon, Kamarck Tom, Mermelstein Robin et al (1994) Perceived stress scale. Measur Stress Guide Health Soc Sci 10(2):1–2
3. Koh KB, Park JK, Kim CH (2000) Development of the stress response inventory. J Korean Neuropsychiatric Assoc 39(4):707–719
4. Dogan G, Akbulut FP, Catal C, Mishra A (2022) Stress detection using experience sampling: a systematic mapping study. Int J Environ Res Public Health 19(9):5693

5. Akbulut FP, Ikitimur B, Akan A (2020) Wearable sensor-based evaluation of psychosocial stress in patients with metabolic syndrome. Artific Intell Med 104:101824
6. Fatma Patlar Akbulut, Harry G Perros, and Muhammad Shahzad. Bimodal affect recognition based on autoregressive hidden markov models from physiological signals. Computer Methods and Programs in Biomedicine, 195:105571, 2020
7. Akbulut FP (2022) Evaluating the effects of the autonomic nervous system and sympathetic activity on emotional states. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi 21(41):156–169
8. Derdiyok S, Akbulut FP (2023) Biosignal based emotion-oriented video summarization. Multimed Syst 29(3):1513–1526
9. Yildirim E, Akbulut FP, Catal C (2023) Analysis of facial emotion expression in eating occasions using deep learning. Multimed Tools Appl 82:31659–31671
10. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. Annu Rev Clin Psychol 4:1–32
11. Serre Fuschia, Fatseas Melina, Debrabant Romain, Alexandre Jean-Marc, Auriacombe Marc, Swendsen Joel (2012) Ecological momentary assessment in alcohol, tobacco, cannabis and opiate dependence: a comparison of feasibility and validity. Drug Alcohol Depend 126(1–2):118–123
12. Abraham AD, Leung EJY, Wong BA, Rivera ZMG, Kruse LC, Clark JJ, Land BB (2020) Orally consumed cannabinoids provide long-lasting relief of allodynia in a mouse model of chronic neuropathic pain. Neuropsychopharmacology 45(7):1105–1114
13. Myin-Germeys I, Krabbendam L, Jolles J, Delespaul PA, van Os J (2002) Are cognitive impairments associated with sensitivity to stress in schizophrenia? an experience sampling study. Am J Psychiatry 159(3):443–449
14. Peters Stefan, Wilkinson Amanda, Mulligan Hilda (2019) Views of healthcare professionals on training for and delivery of a fatigue self-management program for persons with multiple sclerosis. Disabil Rehabil 41(23):2792–2798
15. Brys ADH, Di Stasio E, Lenaert B, Sanguinetti M, Picca A, Calvani R, Marzetti E, Gambaro G, Bossola M (2020) Serum interleukin-6 and endotoxin levels and their relationship with fatigue and depressive symptoms in patients on chronic haemodialysis. Cytokine 125:154823
16. Nalepa GJ, Kutt K, Giżycka B, Jemioło P, Bobek S (2019) Analysis and use of the emotional context with wearable devices for games and intelligent assistants. Sensors 19(11):2509
17. Setz Cornelia, Arnrich Bert, Schumm Johannes, La Marca Roberto, Tröster Gerhard, Ehlert Ulrike (2009) Discriminating stress from cognitive load using a wearable eda device. IEEE Transactions on information technology in biomedicine 14(2):410–417
18. Mohaddes F, da Silva RL, Akbulut FP, Zhou Y, Tanneeru A, Lobaton E, Lee B, Misra V (2020) A pipeline for adaptive filtering and transformation of noisy left-arm ECG to its surrogate chest signal. Electronics 9(5):866
19. Akbulut FP, Akan A (2018) A smart wearable system for short-term cardiovascular risk assessment with emotional dynamics. Measurement 128:237–246
20. Rothkrantz LJM, Wiggers P, Van Wees JWA, van Vark RJ (2004) Voice stress analysis. In: International conference on text, speech and dialogue, pp 449–456. Springer
21. Leung Y, Oates J, Chan SP (2018) Voice, articulation, and prosody contribute to listener perceptions of speaker gender: a systematic review and meta-analysis. J Speech Lang Hearing Res 61(2):266–297
22. Pennebaker JW (1993) Putting stress into words: health, linguistic, and therapeutic implications. Behav Res Therapy 31(6):539–548
23. Madhavi I, Chamishka S, Nawaratne R, Nanayakkara V, Alahakoon D, De Silva D (2020) A deep learning approach for work related stress detection from audio streams in cyber physical environments. In: 2020 25th IEEE international conference on emerging technologies and factory automation (ETFA), volume 1, pp 929–936. IEEE
24. Wood Adrienne, Rychlowska Magdalena, Korb Sebastian, Niedenthal Paula (2016) Fashioning the face: sensorimotor simulation contributes to facial expression recognition. Trends Cognit Sci 20(3):227–240
25. Mitra S, Acharya T (2007) Gesture recognition: a survey. IEEE Trans Syst Man Cybern Part C (Appl Rev) 37(3):311–324
26. Happy SL, Routray A (2017) Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans Affect Comput 10(3):394–406
27. Verma GK, Tiwary US (2014) Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage 102:162–172
28. Gunes H, Piccardi M (2005) Affect recognition from face and body: early fusion versus late fusion. In: 2005 IEEE international conference on systems, man and cybernetics, vol 4, pp 3437–3443. IEEE
29. Zhou X, Jin Y, Zhang H, Li S, Huang X (2016).A map of threats to validity of systematic literature reviews in software engineering. In: 2016 23rd Asia-Pacific software engineering conference (APSEC), pp 153–160. IEEE

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.