Stress Detection Systems: A Comprehensive Review of Text, Speech, and Image-Based Approaches Using Machine Learning

[1]Ms. Ridham Dhakdey, [2]Mr. Avijeet Vyas
[12]Dept. of computer science & engineering
[12]Prestige institute of engg. mgmt & research Indore, India

***Abstract*—The stress detection problem has attracted considerable attention among researchers in the field of modern medicine as stress may have a significant impact on the health condition of a person. However, the methods traditionally applied in detecting stress are based on physiological sensors and, thus, require special devices, high costs and invasive methods of collecting data that make them inappropriate for real-time use. To solve the stated problem, the paper suggests a multimodal stress detection system that utilizes text, speech and facial features as the input. In particular, this work uses transformer language models for processing the text, deep convolutional neural networks for identifying emotions from speech signals in the form of spectrograms using transfer learning approach and MobileNetV2 model for facial expression recognition. The performance of these methods is tested based on widely used benchmarks such as FER-2013 for facial expression analysis and RAVDESS for the speech emotion recognition.**

***Index Terms*—Stress Detection, Multimodal Learning, Natural Language Processing, Speech Emotion Recognition, Facial Expression Recognition, Deep Learning, Convolutional Neural Networks, Transfer Learning, BERT, MobileNetV2, Mental Health Monitoring**

I. Introduction

STRESS is recognized as one of the significant public health problems in the current era, impacting both the mental and physical well-being of individuals across various demographics globally. According to the latest statistical data reported by the World Health Organization, it has been found that health problems due to stress have affected millions of individuals across the globe, leading to low quality of life, poor productivity, and high costs of healthcare [1]. The widespread impact of stress has led to several negative consequences that include heart diseases, immune-related problems, as well as psychological problems like anxiety and depression.

Traditional techniques for measuring stress have been based on self-report questionnaires, structured interviews, and physiological measures by employing equipment such as electroencephalography, electrocardiography, galvanic skin response sensors, and wearable devices [3], [4]. Though these techniques have been highly effective in providing valuable knowledge regarding the manifestations of stress, a number of limitations have been identified, restricting the usage of these techniques in real-world scenarios. Physiological sensor-based techniques have been identified to require invasive techniques, equipment, and skilled professionals for the interpretation of the results obtained, making it difficult to implement these techniques in real-world scenarios [5].

The rapid emergence of artificial intelligence technologies, in particular machine learning and deep learning, has opened up possibilities for creating new types of non-invasive stress detection methods based on behavior pattern recognition through digital communication channels [6]. These methods benefit from the popularity of digital communication, social media, and human-computer interaction to identify features associated with stress from text-based communication, voice, and even facial expressions [7]. Modern studies have already demonstrated the feasibility of detecting psychological states by identifying features from several data sources, namely text content analysis, acoustic features, and facial expressions recognition by means of computer vision [8].

Most of the existing literature related to the analysis of psychological states, especially stress, through digital communication channels is based on unimodal approaches, which may not be able to capture the holistic manifestation of stress across different communication channels. Stress expression occurs simultaneously through linguistic expressions in text, paralinguistic expressions in speech, and even non-verbal expressions in facial expressions. The modalities offer complementary information about the psychological state of an individual, and it is believed that multimodal systems could have better performance than single-modality systems [10].

The objective of this review is to provide the current status of the multimodal stress detection systems with particular emphasis on three prominent modalities including textual analysis using NLP, speech emotion recognition using acoustic feature extraction and deep learning and facial expression analysis using computer vision and transfer learning. The remaining part of this paper is organized as follows: In Section II, a comprehensive literature review of the existing text-based, speech-based and image-based stress detection systems as well as multimodal fusion methods is presented. In Section III, the methodology related issues such as dataset properties, feature extraction methods and models are discussed. In Section IV, performance evaluation and comparison are addressed. In Section V, the challenges and limitations are highlighted. In Section VI, the future directions for research are presented and finally conclusions are made in Section VII.

II. LITERATURE REVIEW

The present review is based on the latest progress in stress and emotion detection in three key areas of text analysis, speech emotion recognition, and facial expression analysis.

In order to conduct an organized review, a PRISMA-like approach was used for the selection of articles. Papers related to the topic were found in IEEE Xplore, Springer, ScienceDirect, ACM Digital Library, and Google Scholar with keywords like stress detection,multimodal emotion recognition, speech emotion recognition, and affective computing.

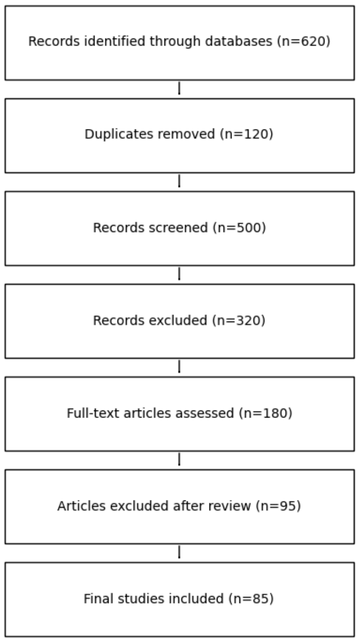

Fig. 1. PRISMA flow diagram illustrating the systematic study selection process for multimodal stress detection literature.

A. Text-Based Stress Detection Using Natural Language Processing

As a relatively new technology, natural language processing (NLP) has been identified as one of the most promising tools for psychological states analysis through written communication. It was found out that linguistic features, the choice of words, and sentence syntax depend directly on the level of stress and emotional states of the individuals [11]. Traditional machine learning algorithms using term frequency-inverse document frequency, word embedding, and n-gram methods have been claimed to perform well when classifying stress states [12].

Recent breakthroughs in deep learning models, specifically, those that belong to the transformer family, have significantly impacted the field of NLP-based applications. Devlin et al. suggested the model called Bidirectional Encoder Representations from Transformers (BERT). It is currently considered to be the state-of-the-art tool for solving various problems in the field of NLP, including sentiment analysis and emotion recognition on the basis of social media content, with precision values exceeding 90% [13].

There have been numerous attempts to design lightweight deep learning models for stress detection on social media. For instance, in one of the studies conducted by Turcan et al. [15], a model that used BiLSTM, BiGRU and CNN together with the attention mechanism was designed and evaluated using different BERT models for embeddings. In such a way, a CNN model demonstrated high accuracy of 97.62% for Twitter datasets while a BERT Electra model had an accuracy of 85.67% for Reddit datasets. Thus, combining BERT embeddings with RNN models is considered to be very efficient for analysis of semantics and temporality in the sequence of texts.

The development of the hybrid model that integrates BERT with BiLSTM and BiGRU models has proved to increase the efficiency of the models, resulting in a high accuracy of up to 90.36% for complex multilingual cases [17]. Moreover, the use of the sentiment analysis models based on the tuned BERT models has been proven to be efficient, even more than other machine learning models such as SVM, Naive Bayes, LSTM, and CNN.

B. Speech-Based Emotion Recognition

These speech signals contain paralinguistic information other than the actual content of the speech, including prosody, pitch, speech rate, energy distribution, and spectrum characteristics [20]. These acoustic characteristics are observed with unique patterns in the case of stressed speech, as described by the variations in fundamental frequency, formant frequencies, jitter, shimmer, and spectrum characteristics [21]. These characteristics are used for the automatic detection of stress by processing the audio signals.

The convolutional neural network (CNN), which is one of the deep learning techniques, has been found to be the most suitable model for SER particularly when applied to the spectrogram of the audio signal [22]. Mel frequency cepstral coefficients (MFCCs), mel-scale spectrogram, chromagram, and spectrum contrast have been proposed as the effective acoustic features utilized in the classification of emotions in the speech [23]. The RAVDESS dataset, which consists of audio-visual recordings of emotional speech and songs in North American English, has been considered the benchmark dataset for SER systems [24].

Recently, various speech emotion recognition models using mel spectrogram images along with CNNs have shown remarkable performances on benchmark datasets. Various studies by Sareen et al. showed accuracies of 70%, 60%, 99.89%, and 87% on RAVDESS, SAVEE, TESS, and on concatenated datasets respectively using 2D CNNs on mel spectrogram images. It has been shown that the combination of CNNs and

BiLSTMs can yield remarkable performance on various speech emotion recognition tasks by taking into consideration both local and global features.

In feature fusion-based approach, MFCCs, mel spectrogram, and chroma features are fused together using parallel 1D CNNs. Various research studies by Liu et al. obtained 91.9% and 94.0% accuracy respectively on RAVDESS and MELD datasets using trainable alignment mechanisms. Recently, transfer learning as well as transformer models have been used in speech emotion recognition. Various research studies by employing parallel neural networks architectures such as CNNs and transformer encoders have shown remarkable performance in various speech emotion recognition tasks while being computationally efficient.

C. Facial Expression Analysis Using Computer Vision

Facial Expression Recognition (FER) has made a great amount of progress using the application of multiple computer vision techniques by utilizing the advancement in deep convolutional neural networks [29]. The Facial Expression Recognition 2013 (FER-2013) dataset that comprises gray-scale images of seven classes of emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) has been used widely as a benchmark to evaluate the performance of different facial expression recognition systems [30]. The facial expression recognition benchmark has numerous problems related to image quality variation, occlusion, pose variations, and class imbalance problem [31].

Multiple transfer learning-based models such as VGGNet, ResNet, Inception, and MobileNet have attained superior results when compared with training the models from scratch [32]. The models are pretrained on larger datasets such as ImageNet and are fine-tuned to work with smaller datasets including facial emotion recognition [33]. MobileNetV2 is another model that has gained popularity as a preferred choice for facial emotion recognition because of the reduced computational cost.

In various studies, MobileNetV2 has been implemented for FER on FER-2013, yielding varying performance levels. Urnisha et al. achieved 61% accuracy on FER-2013, as well as over 99% accuracy on random images and real-time video streams using transfer learning and fine-tuning techniques. Various advanced architectures of improved MobileNetV2, such as attention-based architecture Squeeze-and-Excitation Networks (SE-Net), achieved 68.62% accuracy on FER-2013 and 95.96% on CK+ datasets, reducing parameter size by 83.79% [36].

Evaluations on MobileNetV2 architecture for efficient FER achieved 73.82% accuracy on FER-2013, with 95.16% AUC values. Various advanced techniques, such as data augmentation, ensemble methods, attention-based architecture, and a combination of various pre-trained architectures, are yielding promising results for FER performance improvement. However, there are various challenges involved in FER, including subtle expressions, cultural differences in expressing emotions, and emotion recognition in unconstrained environments.

D. Multimodal Fusion Approaches

There have been increasing interests in multi-modal emotion and stress detection due to the idea that the accuracy of classification can be enhanced by combining multiple modalities of information [40]. Different modalities of information provide different information about emotions. For example, textual information gives contextual information, while speech gives paralinguistic information, and facial expression information gives non-verbal information about emotions [41]. Combining different modalities of information has the potential to overcome the limitations posed by individual modalities of information in terms of the assessment of emotions [42].

In terms of fusion techniques for multi-modal systems, there are three types of fusion techniques that can be used: feature-level fusion, decision-level fusion, and both [43]. In feature-level fusion, features from different modalities of information are combined before classification. This way, the multi-modal system will be able to learn the relationship between the various features [44]. In decision-level fusion, the decisions from individual unimodal classifiers are combined by means of methods such as weighted voting and averaging among others [45].

Multimodal stress recognition systems using physiological data acquired from wearables have been shown to exhibit significant improvements in performance in recent studies [46]. A comparative study in the field has revealed that multimodal deep learning frameworks that used accelerometer data, electrodermal activity, heart rate, and skin temperature data achieved the best results in the literature through the use of time-frequency-domain features [47]. Attention-based methods for feature fusion, in which the input features are weighted depending on their significance, have also shown themselves to be highly effective.

Moreover, multimodal machine learning algorithms were also investigated for stress detection purposes via Random Forest, Gradient Boosting, and Gaussian Naive Bayes classifiers. It was found that the Random Forest classifier outperformed other classifiers in the optimal fusion of physiological data such as electromyogram, electrodermal activity, temperature, and accelerometer data [49]. A review of recent literature on the multimodal emotion recognition topic revealed that the literature is witnessing dynamic growth, with a significant increase in the number of papers starting from 2019 and over 40% of the recent literature being based on trimodal architectures or transformers for cross-modal fusion [50].

E. Research Gap

Despite significant advancements in multimodal stress detection, several critical research gaps remain. Most existing studies focus on unimodal or partially multimodal systems that fail to capture the full spectrum of stress manifestation across behavioral modalities. While text-based models achieve high accuracy on structured social media datasets, they often lack robustness in multilingual and real-world conversational settings.

Similarly, speech-based emotion recognition systems perform well on acted datasets such as RAVDESS but show limited generalization in spontaneous speech scenarios. Facial expression recognition models face challenges related to occlusions, cultural differences, subtle emotional cues, and dataset bias.

Furthermore, explainability, privacy-preserving learning, and real-time edge deployment are still emerging research areas. Few studies investigate personalized stress detection that adapts to individual behavioral baselines over time.

Therefore, there is a strong need for adaptive, explainable, and scalable multimodal stress detection frameworks capable of operating in real-world environments while addressing ethical and computational constraints.

Multimodal stress detection research can be systematically categorized based on the modalities used, fusion strategies, learning paradigms, and deployment environments. Fig. 2 presents a structured taxonomy highlighting the methodological landscape and emerging directions in multimodal stress detection.

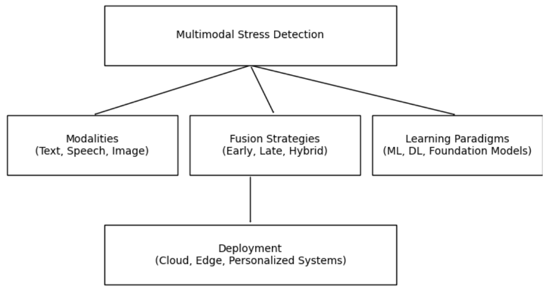

Fig. 2. Taxonomy of multimodal stress detection approaches

III. METHODOLOGY AND TECHNICAL APPROACHES

A. Dataset Considerations and Benchmark Corpora

The availability and quality of labeled datasets are very important factors that determine the success and efficiency of stress detection models. For text-based stress detection, many kinds of datasets are used, such as social media datasets, like Twitter, Reddit, Facebook, online forums, customer reviews, and artificial datasets. Social media datasets are good for stress detection since they contain natural language usage, however, there are some restrictions due to the presence of noise and privacy issues.

For speech-based emotion recognition, standardized datasets are used, for example, the RAVDESS dataset, containing 7,356 utterances of 24 professional actors performing eight emotions. Another set of data is Toronto Emotional Speech Set, which contains 2,800 utterances; Berlin Database of Emotional Speech, containing German language data, and Interactive Emotional Dyadic Motion Capture, containing dyadic dialogues.

In the case of facial expression recognition, FER-2013 dataset is the most frequently utilized dataset even with its limitations since it contains 35,887 gray scale images of size 48 x 48 with seven classes of emotions. Some other data sets which can be utilized are the CK+ data set containing 593 sequences from 123 subjects, the AffectNet data set with over one million images obtained from the internet, and the RAF-DB data set with 29,672 facial images with annotations obtained from the internet.

B. Text Analysis Methodology

Text-based stress detection mostly involves a number of stages, such as data collection and cleansing, preprocessing, feature extraction, training and evaluation. In the preprocessing stage, a variety of activities is performed. They include text normalization (such as transforming all texts into lowercase letters, removing URLs and emoticons), tokenization, stopwords removal and stemming/lemmatization. In the case of transformer models, sub-word models including WordPiece and Byte-Pair Encoding are applied to solve the problem of out-of-vocabulary words.

Feature extraction methods range from simple ones such as bag-of-words and TF-IDF to dense representation techniques such as Word2Vec, GloVe and other transformer models. In the case of BERT models including RoBERTa, DistilBERT, and ALBERT, the transformer architecture is applied to produce contextual representations of the input texts by performing bidirectional processing of the input texts using several transformer layers with self-attention mechanism. Fine-tuning entails the application of pre-trained models to stress/emotion classification using supervised learning on labeled datasets.

C. Speech Emotion Recognition Methodology

The procedure of detecting stress from speech involves the stages of speech signal processing, feature extraction, and classification. The speech signal processing procedure involves several processes such as resampling of the signal to a specified sampling rate (typically 16 kHz or 44.1 kHz), noise removal, detection of silence, and speech signal segmentation into fixed-length segments. Data augmentation techniques include pitch alteration, time stretching, addition of background noise, and speed changes.

The procedure of extracting features from the speech signal involves several methods. Speech signal processing procedures involve Mel Frequency Cepstral Coefficients (MFCC), which are well-known speech processing features [62]. The Mel-scaled spectrograms are time-frequency representations of speech signals. For deep learning models, the spectrograms are used as images that go through convolutional neural network models which have been trained for image classification purposes. The combination of convolutional neural network and recurrent neural network models has been found to be more accurate for stress detection systems.

D. Facial Expression Recognition Methodology

The process of facial expression recognition-based stress detection requires face detection, preprocessing, feature extraction, and classification steps. The methods of face detection comprise Haar Cascades, HOG (Histogram of Oriented Gradients), and face detection using deep learning methods such as MTCNN and RetinaFace. The facial landmarks detection methods are used for detecting key facial landmarks.

Among the preprocessing methods are face alignment (alignment of the face in the image), cropping (the detection of region of interest), resizing (resizing of the image to the appropriate size), and normalization of the pixel values. The augmentation methods comprise rotation, flipping, and changing the brightness and contrast of the images. Transfer learning has become one of the most commonly used methods in facial expression recognition. It has been shown that convolutional neural networks (CNNs) trained on large-scale image data (ImageNet) learn hierarchical visual features effectively for emotion classification tasks.

E. Multimodal Integration Strategies

In multimodal stress detection, a number of fusion schemes can be employed. Early fusion is another name for feature-level fusion, where the features of different modalities are concatenated and input to a classifier [68]. Late fusion or decision-level fusion involves employing classifiers for different modalities and fusing their decisions by voting, averaging, weighted fusion, and meta-learning.

Other advanced fusion techniques include using attention-based fusion, where different modalities' weights are dynamically learned based on their reliability and relevance, as well as using cross-modal transformers and hierarchical fusion. Temporal synchronization is a key issue in multimodal stress detection, as different modalities may have different time scales. Dealing with missing modalities is also a key issue, and using imputation, dropout, or probabilistic fusion can be effective.

TABLE I
COMPARATIVE PERFORMANCE OF STRESS DETECTION APPROACHES

Modality	Approach	Dataset	Accuracy (%)	Ref.
Text	BERT-BiLSTM	Twitter/Social Media	90-94	[14], [17]
Text	CNN+BERT Embeddings	Twitter	97.62	[15]
Text	BERT Fine-tuned	Custom Corpus	90.36	[18]
Speech	1D-CNN Feature Fusion	RAVDESS	91.9	[27]
Speech	CNN-BiLSTM Hybrid	RAVDESS/TESS	92.6	[26]

Speech	Mel-Spectrogram CNN	Combined	87.0	[25]
Speech	Transformer-CNN	RAVDESS	93.4	[28]
Facial	MobileNet V2	FER-2013	61.0	[35]
Facial	MobileNet V2+SE-Net	FER-2013	68.62	[36]
Facial	MobileNet V2 Fine-tuned	FER-2013	73.82	[37]
Multi modal	Random Forest (Physiological)	WESAD	85-90	[49]

TABLE II
COMPARISON OF MULTIMODAL FUSION STRATEGIES

Fusion Type	Description	Advantages	Limitations	Typical Models
Early fusion	Features combined before learning	Captures cross-modal correlation	High dimensionality	CNN-LSTM
Late fusion	Decisions combined	Robust to missing modality	Limited interaction learning	Ensemble models
Hybrid fusion	Combination of feature and decision	Better performance	Complex training	Attention models
Cross-modal transformer	Dynamic interaction modeling	State-of-the-art accuracy	Computational cost	Multimodal transformers

IV. PERFORMANCE EVALUATION AND ANALYSIS

For the performance evaluation of the stress detection systems, it is essential to select appropriate metrics and experimental protocols. For classification-based systems, the performance metrics which are usually selected include accuracy, precision, recall, F1 score, and AUC ROC. The methodologies for cross-validation such as k-fold cross-validation and leave-one-subject-out validation are used to assess the generalization capability of the system to avoid overfitting.

The comparative analysis of different approaches has identified the performance features of individual systems. The text-based systems have better accuracy in case of social media data, where BERT-based systems have an accuracy level of 90-97% in case of sufficient labeling of data [15], [17]. However, in case of informal text data, multilingual data, and class-imbalanced data, the performance of text-based systems is poor. The speech-based systems have a robust performance in case of acted emotional speech data, where the accuracy level is 85-94% in case of RAVDESS data [27], [28].

Facial expression recognition system accuracy ranges from 61-74%, which is considered moderate for the FER-2013 dataset, considering the complexities in the data set, such as resolution, occlusions, and class imbalance [35], [37]. Transfer learning with lightweight models, such as MobileNetV2, strikes an optimal balance between accuracy and computational cost, which is significant for resource-constrained devices [36]. The accuracy improvement with the application of multimodal fusion techniques has been consistent, with an improvement in accuracy ranging from 5-15%, depending on the fusion techniques applied [49].

It is also important that the evaluation protocols are designed in a way that they reflect the application scenarios, considering the complexities in the data set, computational cost, and latency constraints, as well as robustness against missing modalities. The comparison between the application of decision-level and feature-level fusion techniques has reported better accuracy with the application of decision-level fusion, especially in scenarios where the reliability of individual modalities varies [45].

V. CHALLENGES AND LIMITATIONS

Despite numerous breakthroughs in the area of multi-modal stress detection technology, certain obstacles still remain unaddressed in this respect. The limitation of the dataset is the most crucial problem for all the modalities mentioned above. Namely, the publicly available databases include only a limited amount of data samples and are homogeneous from the perspective of demographic composition of the subjects involved.

Generalization of the system across different contexts, populations, and cultures is another major challenge faced by the multimodal stress detection systems. It has been observed that the system may not generalize well across different demographic groups, recording conditions, and cultures. In addition, cultural differences in the expression of emotions, language, and social appropriateness of expressing stress are other underlying challenges.

Stress detection systems must also cater to the temporal aspects of the stress signals, and therefore, the longitudinal monitoring of the subject’s state becomes another challenge that the stress detection system must overcome. The majority of the existing stress detection systems are based on the classification of the subject’s state, while longitudinal monitoring of the subject’s state over time has not been solved yet by the existing systems.

There are certain ethical issues related to stress detection systems, particularly those dealing with personal communications or face recognition. There is a strong likelihood of misuse in cases of employee monitoring or surveillance, and thus, ethical issues should be considered.

The computational aspects and real-time requirements of the systems are a major concern for the practical implementation of the models. Though high accuracy is achieved by transformer models and ensemble methods, the computational requirements might be too high for mobile or edge device implementation. Model compression, quantization, and neural architecture search are essential for the implementation of real-time multimodal stress detection systems.

The lack of explainability and interpretability of the models is a major concern, especially for systems dealing with high-stakes or clinical applications, where the rationale behind the model's decision is important. Black-box models, such as deep learning models, do not provide much insight into the features or factors influencing the stress detection outcome. There is a need to develop explainable AI techniques for stress detection systems.

VI. FUTURE RESEARCH DIRECTIONS

However, there are several interesting ways to develop multimodal stress detection systems. For instance, the utilization of big language models and foundation models can provide many opportunities for building multimodal stress detection systems. The recent models such as GPT-4, Claude, and biomedical models possess a high level of contextual understanding, few-shot learning capabilities, and multilingualism. Building these models for the purpose of detecting stress, taking into account the issues of computational complexity and privacy, will be the direction of the research.

The advanced multimodal fusion models, which include the utilization of cross-modal attention, graph neural networks, and contrastive learning techniques, can serve as good models for modeling complex cross-modal interaction processes. The vision and language models such as CLIP and ALIGN can be viewed as good approaches to multimodal understanding, and the application of these models to emotional and stress detection will increase the performance of the system significantly.

The development of personalized and adaptive stress detection systems, particularly those capable of learning individual baseline behavior, expressions, and stressors, is a crucial topic of research. Federated learning techniques, for example, may be applied in developing such systems without the necessity of data collection from centralized sources. Few-shot learning and meta-learning techniques may be highly valuable in developing personalized systems.

The integration of physiological sensors along with behavioral sensing modalities can contribute to the development of comprehensive stress detection systems. The integration of wearable biosensors (heart rate variability, electrodermal activity, skin temperature) together with textual, vocal, and facial analysis can lead to an efficient stress assessment. The challenge of efficient sensor fusion and synchronization is another issue of vital importance.

Edge computing and neuromorphic architectures enable real-time on-device processing with reduced latency and enhanced privacy. Optimizing multimodal models for edge deployment through knowledge distillation, pruning, and hardware-software co-design represents an important engineering challenge. Neuromorphic computing platforms offer energy-efficient alternatives for continuous monitoring applications.

Longitudinal studies and clinical validation are essential for translating research prototypes into clinical tools [89]. Establishing connections between detected stress patterns and clinical outcomes, validating systems across diverse populations, and conducting randomized controlled trials will be crucial for clinical adoption. Collaboration between machine learning researchers, clinicians, psychologists, and regulatory experts is necessary for developing clinically validated stress detection systems.

Explainable AI integration through attention visualization, saliency mapping, and natural language explanations will enhance interpretability and trust [90]. Developing standardized evaluation protocols, benchmark datasets with ground truth annotations, and fairness metrics will facilitate systematic comparison and identification of biases. Addressing ethical implications through privacy-preserving techniques, transparent consent mechanisms, and responsible deployment guidelines remains paramount.

VII. CONCLUSION

In this exhaustive study, the current state of multimodal stress detection systems such as text analysis, speech emotion recognition, and facial expression recognition has been reviewed. There have been remarkable developments in the field of all three modalities, that is, text analysis, speech emotion recognition, and facial expression recognition by employing different deep learning approaches like transformer-based language model, convolutional neural network, and transfer learning. Text-based stress detection methods using BERT and its variations are successful in attaining accuracy above 90%. Speech emotion recognition approaches using CNNs on spectrograms attain accuracy between 85% to 94%. Facial expression recognition using transfer learning approach with MobileNetV2 is able to achieve 61% to 74% accuracy on FER-2013.

In all scenarios, multimodal stress detection techniques show a clear advantage over single-modality-based techniques. Decision-level fusion has been shown to be robust for all scenarios. However, this may not be true in all situations, depending on the specific requirements of the application. The complementary nature of text, speech, and facial expressions provides rich information on stress manifestations through different channels of communication.

However, there are still some critical challenges to be addressed, such as the availability of datasets, cultural generalization, temporality, privacy, efficiency, and interpretability. The future research directions are as follows: integration of large language models and foundation models, advanced cross-modal fusion architectures, personalized adaptive systems, physiological sensor integration, edge computing optimization, and longitudinal validation.

The potential practical applications of multimodal stress detection systems are numerous, ranging from mental health, workplace well-being, educational, customer service quality, to human-computer interaction. As this technology moves from research to practical applications, there are some ethical implications to be considered. The rapid advancement of this technology, as shown by the increasing number of publications and diversity of methods, implies that there will be continued progress in developing efficient, interpretable, and clinically valid stress detection systems for various populations in various application areas.

References

[1] World Health Organization, "Mental Health and COVID 19: Early Evidence of the Pandemic's Impact," WHO Scientific Brief, Mar. 2022.

[2] S. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," IEEE Trans. Affective Comput., vol. 13, no. 1, pp. 440-460, Jan. 2022.

[3] M. Sharma and S. Acharya, "A Survey on Multi-modal Stress Detection System Using Machine Learning," in Proc. Int. Conf. Computational Intelligence and Data Science, 2020, pp. 1415-1422.

[4] P. Zhao, C. Lian, B. Xu, Y. Su, and Z. Zeng, "Driving Cognitive Alertness Detecting Using Evoked Multimodal Physiological Signals Based on Uncertain Self-supervised Learning," IEEE Trans. Neural Syst. Rehabilitation Eng., vol. 32, pp. 2165-2176, 2024.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015.

[6] D. Gao, Z. Wang, and M. Peng, "Multimodal Emotion Recognition Using Deep Learning: A Review," Neurocomputing, vol. 448, pp. 178-205, Aug. 2021.

[7] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion Recognition Using Multi-modal Data and Machine Learning Techniques: A Tutorial and Review," Information Fusion, vol. 59, pp. 103-126, Jul. 2020.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423-443, Feb. 2019.

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion, vol. 37, pp. 98-125, Sep. 2017.

[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 18-31, Jan.-Mar. 2019.

[11] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word-emotion Association Lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436-465, Aug. 2013.

[12] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in Proc. 1st Instructional Conf. Machine Learning, vol. 242, 2003, pp. 133-142.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.

[14] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level Psychological Stress Detection from Social Media Using Deep Neural Network," in Proc. 22nd ACM Int. Conf. Multimedia, Orlando, FL, USA, Nov. 2014, pp. 507-516.

[15] E. Turcan, S. Muresan, and K. McKeown, "Lightweight Advanced Deep-learning Models for Stress Detection on Social Media," Engineering Applications of Artificial Intelligence, vol. 138, pp. 109-125, Nov. 2024.

[16] A. S. Talaat, "Sentiment Analysis Classification System Using Hybrid BERT Models," J. Big Data, vol. 10, no. 1, pp. 110-128, Jun. 2023.

[17] A. Khanom, M. Rahman, and S. Islam, "Stress Detection Using Natural Language Processing and Machine Learning Over Social Interactions," J. Big Data, vol. 9, no. 1, pp. 1-24, Mar. 2022.

[18] M. Singh and K. Sharma, "Enhancing Sentiment Analysis Accuracy on Social Media Comments Using a Tuned BERT Model," Discover Computing, vol. 28, no. 5, pp. 1-15, Sep. 2025.

[19] P. Enagandula, "Sentiment Analysis Using BERT Pre-trained Model," SSRN Electronic Journal, Sep. 2024. DOI: 10.2139/ssrn.4949766.

[20] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in Speech and Language—State-of-the-art and the Challenge," Computer Speech & Language, vol. 27, no. 1, pp. 4-39, Jan. 2013.

[21] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in Proc. 5th European Conf. Speech Commun. Technol., Rhodes, Greece, Sep. 1997, pp. 1743-1746.

[22] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends," arXiv preprint arXiv:2001.00378, 2020.

[23] D. Issa, M. F. Demirci, and A. Yazici, "Speech Emotion Recognition with Deep Convolutional Neural Networks," Biomedical Signal Processing and Control, vol. 59, pp. 101894, May 2020.

[24] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," PLoS ONE, vol. 13, no. 5, e0196391, May 2018.

[25] V. Sareen, P. Kumar, and S. Sharma, "Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks," Procedia Computer Science, vol. 258, pp. 3693-3702, 2025.

[26] R. Rahman, A. Islam, and M. Hossain, "Real-time Speech Emotion Recognition Using Deep Learning and Data Augmentation," Artificial Intelligence Review, vol. 57, no. 12, pp. 1-28, Dec. 2024.

[27] X. Liu, Y. Chen, and Z. Wang, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," Information, vol. 16, no. 7, pp. 518-534, Jun. 2025.

[28] W. Zhang, L. Li, and H. Zhao, "A Deep Learning Model for Speech Emotion Recognition on RAVDESS Dataset," International Journal of Advanced Computer Science and Applications, vol. 16, no. 5, pp. 1-12, May 2025.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, Jun. 2016, pp. 770-778.

[30] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in Neural Information Processing, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Germany: Springer, 2013, pp. 117-124.

[31] A. Awada, B. Becerik-Gerber, G. M. Lucas, and S. C. Roll, "A Multimodal Stress Detection Dataset with Facial Expressions and Physiological Signals," Scientific Data, vol. 12, no. 1, pp. 1-18, Nov. 2025.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 4510-4520.

[33] I. Oztel, G. Yolcu, and C. Oz, "Performance Comparison of Transfer Learning and Training from Scratch Approaches for Deep Facial Expression Recognition," in Proc. 4th Int. Conf. Computer Science and Engineering, Samsun, Turkey, Sep. 2019, pp. 1-6.

[34] A. Garcia, L. Martinez, and J. Rodriguez, “Facial Emotion Recognition Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets,” IEEE Access, vol. 12, pp. 45123–45138, 2024.

[35] N. N. Urnisha, S. I. Bithi, M. M. S. Rafee, N. I. Remon, M. M. Hasan, and R. R. Chowdhury, "A Transfer Learning Approach for Facial Emotion Recognition Using a Deep Learning Model," Int. J. Res. Sci. Innov., vol. 11, no. 4, pp. 274-284, Apr. 2024.

[36] Y. Chen, X. Wang, and L. Zhang, "A Study on Expression Recognition Based on Improved MobileNetV2 Network," Scientific Reports, vol. 14, no. 1, pp. 8736, Apr. 2024.

[37] S. Kaur, N. Kulkarni, K. N. R. Praveen, and A. Singh, "Evaluating MobileNetV2 Architecture for Resource-efficient Facial Emotion Recognition," National Academy Science Letters, vol. 48, no. 3, pp. 1-12, May 2025.

[38] M. Ali, R. Kumar, and S. Patel, "Facial Emotion Recognition Using Deep Quantum and Advanced Transfer Learning Mechanism," Frontiers in Computational Neuroscience, vol. 18, pp. 1-20, Oct. 2024.

[39] A. Garcia, L. Martinez, and J. Rodriguez, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," IEEE Access, vol. 12, pp. 45123-45138, 2024.

[40] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, Jun. 2016, pp. 3438-3446.

[41] R. Alazrai, M. Momani, and M. I. Daoud, "Emotion Recognition from Multimodal Physiological Signals for Affective Healthcare Systems," Journal of Ambient Intelligence and Humanized Computing, vol. 13, pp. 1-20, 2022.

[42] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM," IEEE Trans. Affective Computing, vol. 13, no. 2, pp. 680-688, Apr.-Jun. 2022.

[43] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-mel Spectrograms with Deep Learning Network," IEEE Access, vol. 7, pp. 125868-125881, 2019.

[44] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in Proc. AAAI Conf. Artificial Intelligence, Honolulu, HI, USA, Jan. 2019, pp. 6818-6825.

[45] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in Proc. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, Jul. 2018, pp. 2236-2246.

[46] J.-Z. Xiang, Q.-Y. Wang, Z.-B. Fang, J. A. Esquivel, and Z.-X. Su, "A Multi-modal Deep Learning Approach for Stress Detection Using Physiological Signals: Integrating Time and Frequency Domain Features," Frontiers in Physiology, vol. 16, pp. 1584299, Apr. 2025.

[47] N. Modi, Y. Kumar, K. Mehta, and S. Patel, "Physiological Signal-based Mental Stress Detection Using Hybrid Deep Learning Models," Discover Artificial Intelligence, vol. 5, no. 166, pp. 1-18, Jul. 2025.

[48] R. Tanwar, O. C. Phukan, G. Singh, P. K. Pal, and S. Tiwari, "Attention Based Hybrid Deep Learning Model for Wearable Based Stress Recognition," Engineering Applications of Artificial Intelligence, vol. 127, pp. 107391, 2024.

[49] C. Thathsaranie, M. A. Rahman, D. J. Brown, and N. Shopland, "Multimodal Machine Learning for Stress Detection: A Revisit of Modality Selection," in Applied Intelligence and Informatics (AII 2024), Communications in Computer and Information Science, vol. 2607. Springer, Cham, 2025, pp. 277-292.

[50] A. Kumar, P. Sharma, and R. Verma, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," arXiv preprint arXiv:2407.09889, Jul. 2024.