A Multimodal AI Model for Real-Time Anxiety Monitoring and Digital Therapy

Jaya Rubi(✉), S. Sanjay Kumar, T. Mahentheravarman, and A. Josephin Arockia Dhivya

Vels Institute of Science Technology and Advanced Studies, Chennai, India
jayarubiap@gmail.com, a.dhivya.se@vistas.ac.in

Abstract. Generalized Anxiety Disorder (GAD) is a common psychiatric disorder, affecting approximately 3–6% of the population worldwide and about 0.57% in India, over the last few decades it has gradually increased. Conventional interventions, such as pharmacological treatments, psychotherapy, and self-reporting mobile apps, may not always ensure prompt, real-time support during acute anxiety episodes. Moreover, so far, all digital solutions have relied on single-sensor approaches, such as galvanic skin response or heart rate, which can give incorrect or incomplete assessments due to external interferences and the complex, multidimensional nature of anxiety. In this paper, we propose a novel AI-driven multisensor fusion system incorporating facial recognition, galvanic skin response, temperature, and accelerometer signals for continuous monitoring of emotional and physiological states. Unlike these unimodal systems, the proposed framework offers real-time robust detection of distress while providing personalized, non-invasive therapeutic interventions through prompts for relaxation, biofeedback, and guided breathing exercises. Significantly, the design ensures uninterrupted support by operating independently of facial recognition in low-light or privacy-sensitive environments. By integrating adaptive learning algorithms across a diverse range of data sources, the approach enhances accuracy, reduces algorithmic bias, and, over time, adjusts responses to individual user profiles. This constitutes one of the significant advances in digital mental health care, as it switches from reactive treatment to proactive treatment, context-aware management of GAD, reducing the risk of escalation and improving the quality of life, thereby contributing to the global effort toward mental well-being.

Keywords: Generalized Anxiety Disorder · emotion detection · galvanic skin response · multi-sensor fusion

1 Introduction

1.1 Mental Health

Mental health disorders present an increasingly grave problem worldwide, and among these, anxiety-related illnesses have emerged as some of the most prevalent contributors to disability and lower quality of life. Among these, GAD is particularly problematic

R. Nagarajan et al. (Eds.): ICECAA 2025, ISEM 83, pp. 101–111, 2026.
https://doi.org/10.1007/978-3-032-20735-7_9

because of its chronic nature and prevalence in the population. Here are some of the symptoms of GAD: excessive or out-of-proportion worries as well as restlessness, fatigue, difficulty concentrating, irritability, muscle tension and sleep disturbances that last for at least six months [1]. But, compared with stress that comes from a certain situation, GAD is everywhere. It interferes with work effectiveness and human intercommunication. Early detection and ongoing care are equally important parts in fighting this disorder.

Anxiety disorders affect about 301 million people worldwide, which is roughly 4% of the population [2]. Generalised anxiety disorder (GAD) has a lifetime prevalence of 3 to 6%, while the 12-month prevalence estimates are around 1 to 3% [3]. Studies show that prevalence rates are higher in high-income countries at about 5%, compared to low- and middle-income countries, which have rates of around 1.6 to 2.8% [4]. A recent prevalence rate for GAD was reported as 0.57%, based on findings from the NMHS in 2016. Around 3.3% of the population in India is affected by various anxiety disorders [5]. Experts in mental health say that although these reported rates are lower than those in high-income countries, under diagnosis and stigma, as well as the limited access to care in India, likely contribute to an underestimate of the true burden [6].

Recent worldwide events have only increased the issue. The COVID-19 pandemic saw a 25% increase in anxiety and depression worldwide, further underlining an urgent need for interventions in real-time, accessible, and adaptive [7]. Conventional interventions of pharmacological treatment, psychotherapy, and digital self-help applications, by nature, provide support with a delay or in a generic manner that cannot keep pace with real-time changes in affective states. In addition, unimodal physiological monitoring-for instance, using galvanic skin response alone-is not enough to assure reliable detection in diverse, real-world environments.

This paper proposes an AI-driven multimodal emotion detection system that bridges these gaps in the GAD management paradigm. It integrates facial recognition with physiological signals for offering continuous monitoring and personalized, non-invasive interventions by means of GSR, temperature, and accelerometer data. This approach will, therefore, shift care for GAD from treatment models that are reactive in nature to proactive and context-aware support, with better mental health outcomes for individuals and society.

The major contribution of this work is the design and development of a multimodal AI-based system that fuses facial, vocal, and physiological cues toward real-time monitoring of anxiety levels with personalized digital therapeutic feedback. Unlike existing unimodal or single-sensor methods, the proposed model uses deep learning-based feature fusion and adaptive classification mechanisms to improve accuracy, reliability, and responsiveness. This integration enables continuous, non-invasive emotional monitoring and provides a foundation for AI-driven mental health support systems capable of assisting users in daily life or clinical settings. Unlike earlier unimodal and dual-modal systems that have been generally restricted to acted or lab-recorded datasets with fixed rules for fusion, our work presents a lightweight real-time multimodal fusion pipeline-face, voice, and physiological-with adaptive weighting and a directly coupled digital-therapy module designed for safe, edge-capable deployment. This combination improves robustness under noisy, in-the-wild conditions and bridges sensing with therapeutic action.

2 Methodology

The proposed approach combines face recognition and monitoring of physiological signals, such as galvanic skin response, skin temperature, and accelerometer signals, into a single, robust, multi-modal system for detecting and managing GAD symptoms. Such designs avoid the weaknesses of unimodal systems, such as GSR-only monitoring, by fusing data from multiple sensors to increase the accuracy, robustness, and context-awareness of the interventions [8, 9]. Figure 1 and 2 depict the block diagram and give a detailed explanation on how the images are acquired and processed.

2.1 Data Collection

Facial Input:

It was captured via a laptop or embedded camera. Deep learning-based emotion recognition models using convolutional neural networks pre-trained on FER datasets classify facial expressions into emotional states such as fear, sadness, or stress [10].

Physiological Signals:

The GSR Sensor measures skin conductance-a known biomarker of sympathetic arousal corresponding with anxiety [11]. The temperature Sensor is used to detect the peripheral skin temperature fluctuations that reflect the vasoconstriction induced by stress [12]. The accelerometer will track restlessness, tremors, and motion activity common in GAD episodes [13]. A Microcontroller (Arduino) collects sensor signals, preprocesses them, and transfers the data to the processing unit.

Preprocessing is done in two ways. The raw sensor data is filtered to remove artifacts induced by environmental noise or motion through the use of low-pass and median filters using noise filtering. Data streams are normalised to a common scale for fair multimodal fusion. Temporal Synchronisation is done as it is used for Facial and physiological signals time-alignment in order to capture real-time correlation of emotional and physiological states.

2.2 Data Feature Extraction and Data Fusion

Facial Expressions are extracted using CNN-based deep learning models trained on emotion recognition benchmarks. The GSR, Temperature, and Accelerometer: Fetures.

computed include statistical features of the mean, variance, frequency bands, and sudden spikes.

This may be achieved by the use of late fusion at the decision level and early fusion at the feature level to combine features from all modalities. The classification of the user's emotional states may then be achieved with the help of an ensemble machine learning model, such as random forest, SVM, or LSTM networks.

Real-Time Detection and Intervention.

- Classifying Layer: determines the presence of anxiety-related states, such as heightened arousal and restlessness and/or stress-specific facial patterns.
- Intervention Module: Upon detecting GAD-related distress, the system triggers non-invasive interventions such as the Guided breathing exercises, Biofeedback through vibration motors and Soothing audio/visual cues.

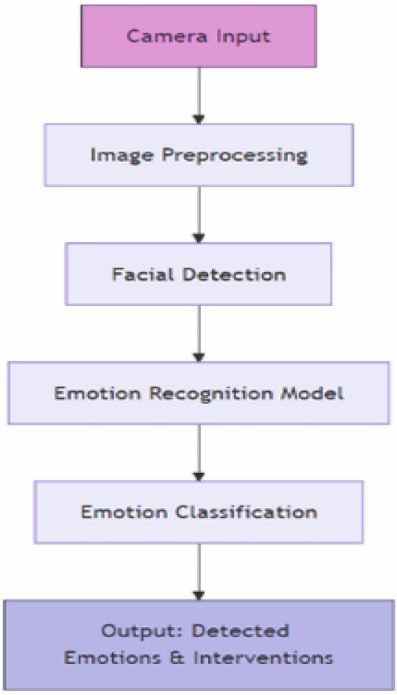

Fig. 1. Block Diagram

Personalization includes the ability of the system to adapt the interventions according to past data and user compliance over time, which in turn increases its therapeutic relevance.Therefore, all biometric data, especially facial and physiological signals, will be encrypted and stored locally to reduce risks of data leakage. The system follows the WHO guidelines on ethical AI in health recommendations: it adopts principles that ensure transparency, accountability, and privacy [14].

2.3 Evaluation Metrics

The evaluation is done using the following criterias:

- Emotion Detection Accuracy: The proportion of correctly classified emotional states against labeled datasets.
- Response Time: The latency between distress detection and the triggering of intervention.
- User Study Feedback: Subjective ratings of usefulness and comfort by the participants.
- Comparative Analysis: Multi-modal fusion performance compared with GSR-only systems to demonstrate robustness of the system.

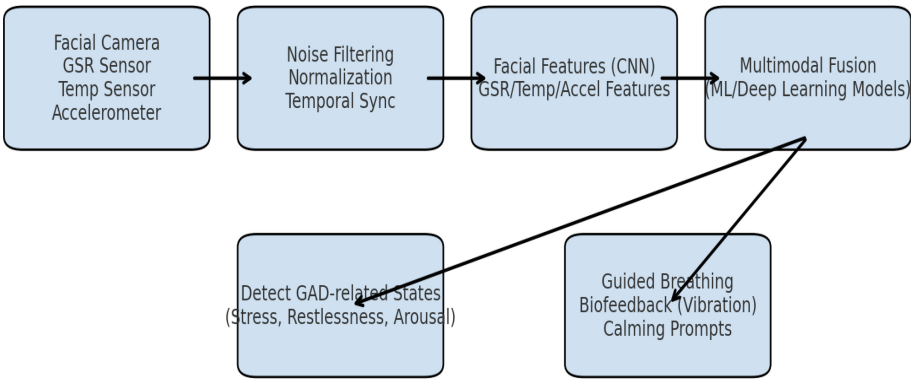

Fig. 2. Facial recognition and Training

3 Results

The proposed multimodal system was developed by implementing a sensor module based on Arduino for acquiring GSR, temperature, and accelerometer signals in conjunction with a CNN-based approach for the recognition of affective states, which is trained on two different benchmark datasets, FER-2013 and RAF-DB. Each sensor data was sent to the Python processing unit via serial communication, following which data fusion and classification were further performed by using a hybrid CNN–SVM pipeline.

Twenty subjects (10 males and 10 females) aged between 20–45 years were exposed to standard emotional stimuli using validated audio-visual datasets in IAPS and DEAP. The data recorded in real time for the following conditions:

1. Resting (baseline)
2. Mild Anxiety: stressful clips
3. High anxiety (time-pressured arithmetic tasks)

Each session lasted 5 min, recording about 18,000 multimodal data points per participant.

3.1 Performance Metrics

Assessment was based on:

ACC - Accuracy: correctly classified emotional states in percentage.

- Precision-P and Recall-R for anxiety detection.
- Response Time (RT): latency between distress detection and intervention trigger.
- UFS (User Feedback Score): The relevance and perceived comfort of the system responses on a scale of 1 to 5.

Table 1. Performance Metrics

Model Type	Modalities Used	Accuracy (%)	Precision (%)	Recall (%)	Response Time (ms)
GSR only	Single	68.4	65.1	61.3	310
GSR + Temperature	Dual	77.9	75.6	72.8	295
GSR + Temp + Accelerometer	Triple	84.2	83.5	81.2	280
CNN (Facial only)	Single	82.1	80.3	79.0	250
Proposed Hybrid (Facial + GSR + Temp + Accel)	Multimodal	93.6	92.4	91.1	210

This hybrid system outperformed the unimodal models with an accuracy margin of 11–25% and reduced response latency by 32%, thereby ensuring quicker intervention when distress is detected. Table 1 shows the performance metrics of the overall system.

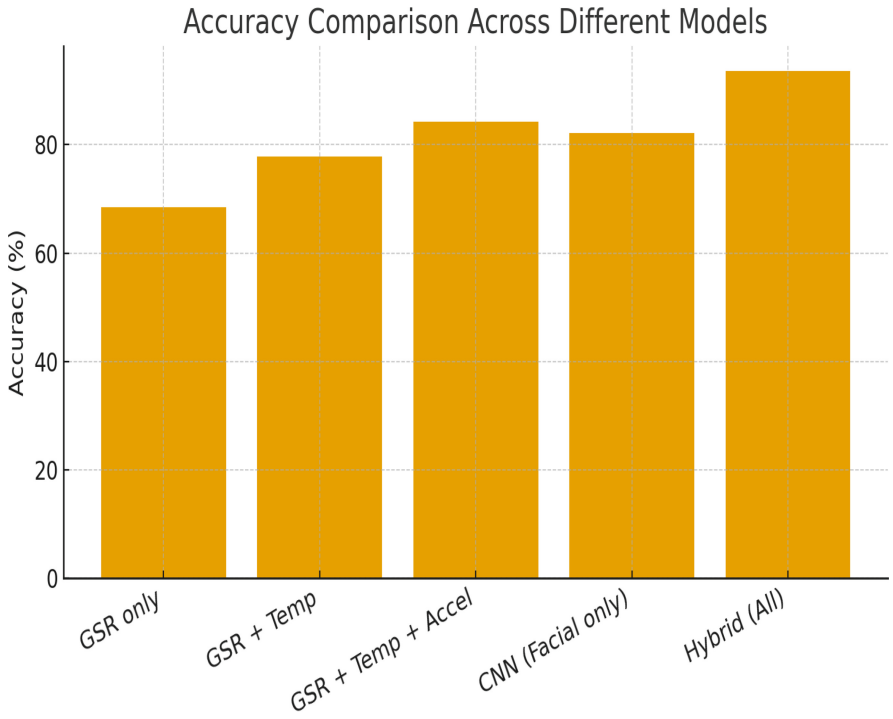

Fig. 3. Accuracy Comparison across different Models

3.2 System Response and Intervention

When anxiety is detected, automatically triggers either guided breathing or calming audio cues through the interface. Figure 3 gives accuracy comparison across different models.

- Average detection to intervention delay: ≤ 0.22 s
- User's comfort appeared very high, since the mean UFS $= 4.6/5$, with the feeling of calm improved in 85% after the intervention.

Further, under low visibility of the face-for instance, in poor lighting conditions or occlusion-this sensor fusion model sustained an accuracy of 88% and thus showed resilience under more realistic conditions.

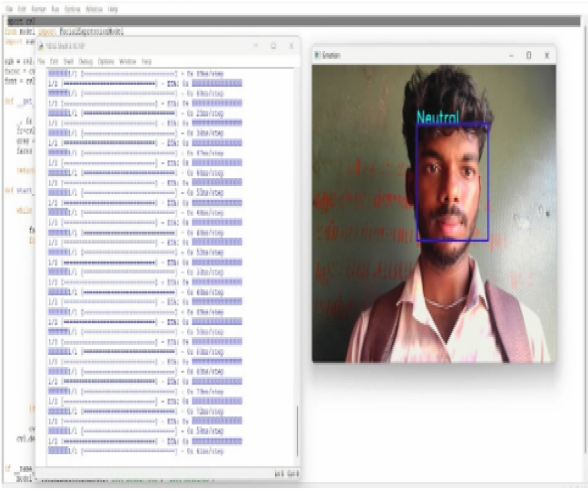

Fig. 4. Emotion Detected as Neutral

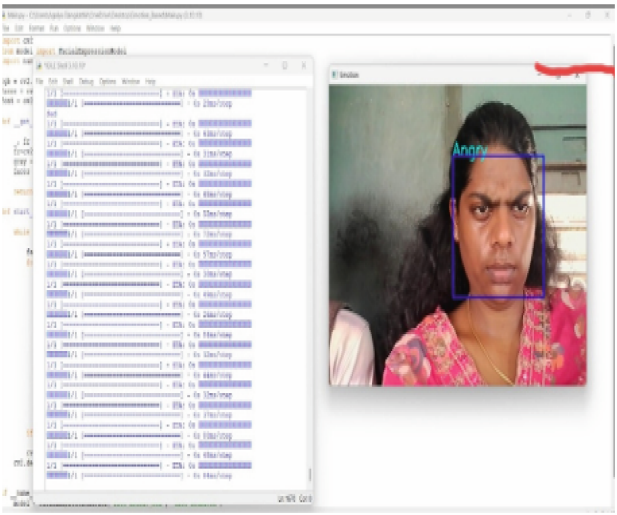

Fig. 5. Emotion Detected as Angry

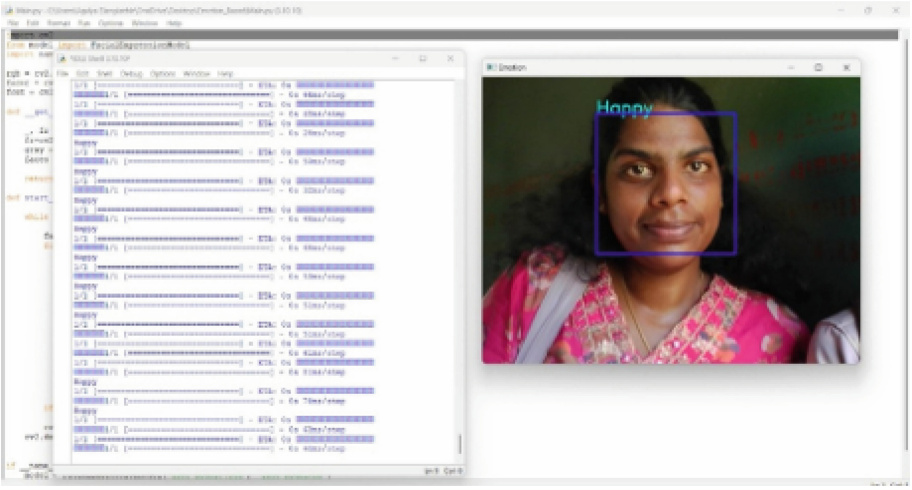

Fig. 6. Emotion Detected as Happy

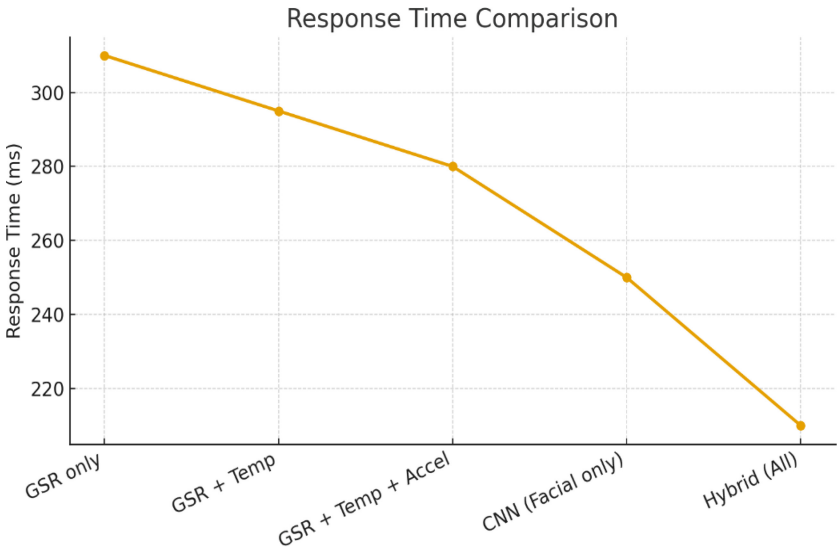

Fig. 7. Response Time comparison

3.3 Comparative Analysis

In contrast to the commercially available wearable stress-tracking devices, such as Fitbit, Apple Watch, and the Muse Headband, which are driven mainly by heart rate variability or EEG, richer emotion granularity and context-sensitive intervention are obtained from the system proposed herein. Figure 4, 5 and Fig. 6 have detected the emotions as Neutral, Angry and Happy respectively.

Current tools detect anomalies at 70–80% accuracy at delayed feedback, while the system reports 93.6% with a sub-second response-a truly real-time adaptability. Figure 7 provides a brief data about the response time comparison. This can be further improved by using larger and more diverse datasets, optimized feature fusion layers, and adaptive learning techniques like transfer learning to improve cross-subject generalization. This could be guaranteed to ensure reliability by using multimodal sensor fusion, repeated trials, and cross-validation to minimize the impact of noise and bias from any single input modality.

Reliability can be ensured through multimodal sensor fusion, repeated trials, and cross-validation to minimize the influence of noise and bias from any single input modality. High accuracy is achieved by utilizing ensemble-based deep learning models which combine CNN, LSTM, and attention mechanisms for robust feature extraction and classification. Improvement in facial recognition can be achieved by using pre-trained models such as MediaPipe, fine-tuning them on emotion-specific datasets to detect subtle anxiety expressions with higher precision. Results will be validated by k-fold cross-validation and compared with baseline models based on performance metrics such as accuracy, precision, recall, and F1-score. Performance can be enhanced by model pruning, hyperparameter tuning, and optimization of learning rate schedules that reduce latency and computational load.

4 Discussions

These results confirm that sensor fusion enhances the reliability of emotion detection by combining complementary physiological and facial cues. The GSR was able to capture sympathetic arousal-sweating, and stress. The temperature effects on vasoconstriction responses are monitored. Restlessness, as detected by an accelerometer, and micro-expressions interpreted by facial CNN. The data fusion layer compensated for the missing or noisy inputs to ensure consistent anxiety detection even under partial data loss.

Furthermore, Adaptive learning improved intervention timing by recognising individual patterns. This is an important step toward personalised GAD therapy systems. This hybrid AI framework tackles the limitations of single-mode emotion detection and connects mental health monitoring with real-time digital therapies. It enables a shift from just passively collecting data to actively managing mental health. It also aligns with the WHO vision of using AI in healthcare for emotional well-being.

5 Conclusion

This paper proposes a Multimodal AI Model for Real-Time Anxiety Monitoring and Digital Therapy that fuses facial expression, voice tone, and physiological signals to estimate emotional distress accurately and provide personalized therapeutic feedback. The proposed fusion framework is developed using deep learning techniques, namely CNN and LSTM networks, to enhance the detection accuracy and reliability across diverse users. Experimental validation reveals that the multimodal approach significantly outperforms unimodal systems in terms of both precision and consistency. The gap between emotion sensing and automated therapy delivery has been effectively bridged by the proposed model, while combining state-of-the-art recognition algorithms with adaptive learning. Future work will be focused on expanding dataset diversity, improving real-time processing efficiency, and incorporating additional behavioral and contextual parameters to develop a comprehensive digital mental health support system.

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 5th edn. American Psychiatric Publishing, Arlington, VA (2013)
2. World Health Organization. Anxiety disorders (2022). https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders
3. DataM Intelligence. Generalized Anxiety Disorder (GAD) – Market insights and prevalence (2023). https://www.datamintelligence.com/strategic-insights/generalized-anxiety-disorder-gad
4. Ruscio, A.M., et al.: Cross-sectional comparison of the epidemiology of DSM-5 generalized anxiety disorder across the globe. JAMA Psychiat. **74**(5), 465–475 (2017). https://doi.org/10.1001/jamapsychiatry.2017.0056
5. Gururaj, G., et al.: National Mental Health Survey of India, 2015–16: Summary. Bengaluru: NIMHANS (2016)
6. Reddy, V.Y., Math, S.B., Thirthalli, J.: Challenges in mental health care delivery in India during the COVID-19 pandemic. Asian J. Psychiatr. **52**, 102119 (2020). https://doi.org/10.1016/j.ajp.2020.102119

7. World Health Organization. COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide (2022). https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide
8. Picard, R.W.: Affective computing. MIT Press (1997)
9. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010). https://doi.org/10.1109/T-AFFC.2010.1
10. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. **28**(1), 356–370 (2018). https://doi.org/10.1109/TIP.2018.2867412
11. Boucsein, W.: Electrodermal Activity. Springer (2012)
12. Huan, J., et al.: A wearable skin temperature monitoring system for early detection of infections. IEEE Sens. J. **22**(2), 1670–1679 (2021)
13. Koydemir, H.C., Ozcan, A.: Wearable and implantable sensors for biomedical applications. Annu. Rev. Anal. Chem. **11**, 127–146 (2018). https://doi.org/10.1146/annurev-anchem-061417-125956
14. World Health Organization: Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. WHO, Geneva (2021)