frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 27 November 2024
DOI 10.3389/frai.2024.1467051

OPEN ACCESS

EDITED BY
Basabi Chakraborty,
Madanapalle Institute of Technology and Science, India

REVIEWED BY
Jorge Galvez,
University of Guadalajara, Mexico
Luigi Celona,
University of Milano-Bicocca, Italy

*CORRESPONDENCE
Huizilopoztli Luna-García
✉ hlugar@uaz.edu.mx
Klinge Orlando Villalba-Condori
✉ kvillalba@ucsm.edu.pe

RECEIVED 19 July 2024
ACCEPTED 04 November 2024
PUBLISHED 27 November 2024

CITATION
Espino-Salinas CH, Luna-García H, Celaya-Padilla JM, Barría-Huidobro C, Gamboa Rosales NK, Rondon D and Villalba-Condori KO (2024) Multimodal driver emotion recognition using motor activity and facial expressions.
Front. Artif. Intell. 7:1467051.
doi: 10.3389/frai.2024.1467051

Multimodal driver emotion recognition using motor activity and facial expressions

Carlos H. Espino-Salinas[1], Huizilopoztli Luna-García[1]*, José M. Celaya-Padilla[1], Cristian Barría-Huidobro[2], Nadia Karina Gamboa Rosales[1], David Rondon[3] and Klinge Orlando Villalba-Condori[4]*

[1]Laboratorio de Tecnologías Interactivas y Experiencia de Usuario, Universidad Autónoma de Zacatecas, Unidad Academica de Ingeniería Electrica, Zacatecas, Mexico, [2]Centro de Investigación en Ciberseguridad, Universidad Mayor de Chile, Providencia, Chile, [3]Departamento Estudios Generales, Universidad Continental, Arequipa, Peru, [4]Vicerrectorado de Investigación, Catholic University of Santa María, Arequipa, Peru

Driving performance can be significantly impacted when a person experiences intense emotions behind the wheel. Research shows that emotions such as anger, sadness, agitation, and joy can increase the risk of traffic accidents. This study introduces a methodology to recognize four specific emotions using an intelligent model that processes and analyzes signals from motor activity and driver behavior, which are generated by interactions with basic driving elements, along with facial geometry images captured during emotion induction. The research applies machine learning to identify the most relevant motor activity signals for emotion recognition. Furthermore, a pre-trained Convolutional Neural Network (CNN) model is employed to extract probability vectors from images corresponding to the four emotions under investigation. These data sources are integrated through a unidimensional network for emotion classification. The main proposal of this research was to develop a multimodal intelligent model that combines motor activity signals and facial geometry images to accurately recognize four specific emotions (anger, sadness, agitation, and joy) in drivers, achieving a 96.0% accuracy in a simulated environment. The study confirmed a significant relationship between drivers' motor activity, behavior, facial geometry, and the induced emotions.

KEYWORDS

facial emotion recognition, motor activity, driver emotions, transfer learning, convolutional neural network, ADAS

1 Introduction

Road accidents are among the leading causes of death worldwide, with approximately 1.3 million people losing their lives in traffic accidents each year. Additionally, 20 to 50 million individuals suffer non-fatal injuries, many of which lead to long-term disabilities. These injuries result in significant economic losses for individuals, their families, and nations as a whole (WHO, 2018). Various factors contribute to the high incidence of accidents, underscoring the need for effective interventions. Identifying these interventions requires a thorough analysis and classification of the factors that lead to accidents.

One of the most common causes of road accidents is the high cognitive load placed on drivers. They must continuously process a stream of visual information from the road, traffic signs, pedestrians, other vehicles, and the environment. Several situational factors can increase this cognitive load, including the presence of similarly aged passengers who may distract the driver, fatigue (particularly common among young drivers), and socioeconomic status, which can influence risk-taking behavior behind the wheel (Rezapour and Ksaibati, 2022).

The growing use of mobile phones has further increased the risk of accidents, especially among young drivers. The cognitive and behavioral demands of phone usage while driving divert attention from the road. Additionally, alcohol and drug consumption impairs cognitive processes and increases crash risks across all age groups (Celaya-Padilla et al., 2021). According to the National Highway Traffic Safety Administration (NHTSA), driver distraction occurs when attention shifts from driving to other activities, contributing to accidents. In 2021 alone, 3,522 people were killed, and approximately 36,241 were injured in traffic accidents caused by distracted driving. Research shows that both internal and external factors can lead to road accidents. One significant internal factor is emotional state, which can affect driving behavior and lead to erratic or inadequate driving. Emotions, which are often difficult to control, can unexpectedly influence a driver's behavior (Maldonado et al., 2020). They are transient mental states that can change rapidly in response to significant events, triggering behavioral responses that may be difficult to regulate (Zimasa et al., 2017).

Negative emotions, in particular, can significantly affect drivers. Studies have demonstrated a link between negative emotions and impaired driving performance. For example, sadness has been shown to increase location error rates, while anger slows drivers' ability to identify road elements (Dozio et al., 2024). Emotions such as anger, hostility, and nervousness are strongly associated with aggressive driving behaviors (Stephens et al., 2024). These negative emotions can impair cognitive processes and compromise road safety (Zhang Q. et al., 2020). A study by Dingus et al. (2016) found that drivers experiencing sadness, anger, or agitation were nearly ten times more likely to be involved in an accident.

Advanced Driver Assistance Systems (ADAS) must ensure safe transportation by taking into account drivers' vulnerability to accidents and recognizing that systems should be designed to accommodate human error (WHO, 2018). To assess a driver's readiness, these systems need to monitor the driver's physical, emotional, and physiological state and communicate relevant information effectively. Various real-time emotion recognition systems have been developed within the fields of affective computing and ADAS, with the goal of adapting to users' emotions for more natural and efficient interactions (Schuetz and Venkatesh, 2020). Emotion recognition allows interactive vehicle systems to interpret human emotions and use this data to make decisions. However, current ADAS largely implement basic mechanisms for emotional state recognition. If ADAS could account for a driver's emotional state, they could make more contextualized decisions based on the driver's potential reactions. Developing ADAS that continuously recognize both the driver's emotions and performance remains a significant challenge (Davoli et al., 2020). Emotion recognition has therefore become a central feature of vehicle systems, relying on various measurements–such as facial expressions, speech, gait patterns, physiology, and eye-tracking–analyzed using advanced techniques like artificial intelligence (Cai et al., 2023).

Despite the significant progress in developing less intrusive and more accurate methods for emotion recognition in automotive environments, numerous challenges remain. Traditional approaches often rely on camera-based systems, which can face issues such as occlusion, lighting variability, and differences in drivers' physical characteristics in uncontrolled environments. Additionally, control mechanisms based on biophysiological data, though potentially effective, tend to be intrusive and may cause emotional discomfort to drivers. Research into the motor activity and driving behavior of drivers, while promising, has been limited, with only a few studies utilizing basic artificial intelligence techniques to explore these characteristics.

To address the challenges of emotion recognition, a novel approach is proposed: the development of a multimodal artificial intelligence model for objective emotion recognition. This model will integrate data from motor activity and facial geometric changes to improve emotion recognition accuracy.

The remainder of this article is structured as follows: In Section 2, The most recent works in the area of emotion recognition related to the present study are mentioned. Section 3 explains the materials and methods used to generate an optimal and efficient multimodal emotion recognition model. Section 4 describes the results of the emotion induction phase and the developed emotion recognition model. In Section 5, a comprehensive discussion of the results obtained is made, emphasizing the contribution of the research to the existing body of knowledge. Finally, in Section 6, final conclusions and proposals for future research aimed at improving emotion recognition systems in drivers covering the analysis and processing of various information sources are presented.

2 Related studies

Related works in emotion recognition encompass a broad spectrum of research endeavors aimed at understanding and interpreting human emotions through various modalities. These works often explore the utilization of machine learning techniques, including deep learning algorithms, to detect, classify, and interpret emotional states. Some key areas of research and notable contributions.

2.1 Facial emotion recognition

Thanks to recent and continuous improvements in the application of artificial neural networks, many architectures have been proposed and employed for facial emotion recognition, each of which has surpassed its predecessors, thus improving the accuracy and performance of the latest generation (Ko, 2018).

Some studies propose the implementation of deep learning algorithms, such as Convolutional Neural Networks (CNNs), for facial emotion recognition. These studies often train CNNs with facial emotion data and test different architectures, including

VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161, using facial image datasets. For example, Mehendale (2020), proposes a CNN-based approach for facial emotion recognition. This network consists of two stages: the first stage removes the background of the image, while the second stage focuses on extracting facial features. The two-level CNN operates in series, with the final layer of the perceptron adjusting weights and exponent values after each iteration. This approach contrasts with the single-level CNN strategies typically used and results in improved accuracy. Similarly, Sarvakar et al. (2023), classified facial expressions into one of seven emotions using various models on an emotion dataset. The models tested include decision trees, feed-forward neural networks, and CNNs, achieving reasonably acceptable accuracy. Additionally, Modi and Bohara (2021) presents a CNN-based facial expression recognition framework in which the network classifies facial expressions as happy, sad, or neutral. In a related study, Khattak et al. (2022), addresses emotion recognition by applying a deep learning technique using CNNs to classify facial emotions and detect age and gender from facial expressions. The experimental results demonstrate that the proposed model can identify emotions, age, and gender with a high degree of accuracy.

Also, some researches has made significant advancements in facial emotion recognition using multimodal models. An example is the work of (Mocanu et al., 2023), which proposes an innovative methodology that integrates simultaneous video and audio analysis. For visual analysis, they employ a three-dimensional CNN, and for auditory analysis, they use a two-dimensional CNN. This study implements the ResNet-101 and ResNet-18 architectures, achieving impressive results with an average accuracy of 89.25% on the RAVDESS database and 84.57% on the CREMA-D database. These results represent a substantial improvement over previous approaches, with accuracy increases ranging from 1.72% to 11.25%.

Other research efforts have focused on analyzing different CNN architectures to compare their performance in facial emotion recognition. For instance, Chowdary et al. (2023), present a facial emotion recognition system utilizing transfer learning. The study employs pre-trained CNNs, including VGG-19, ResNet-50, Inception-v3, and MobileNet, with experiments conducted on the CK+ database. The results show accuracies of 96% for VGG-19, 97.7% for ResNet-50, 98.5% for Inception-v3, and 94.2% for MobileNet. Notably, MobileNet achieved the highest accuracy among the four networks, demonstrating its effectiveness in emotional facial recognition. Similarly, Sahoo et al. (2023) reports comparable results, emphasizing that the pre-trained VGG-19 model outperformed other models, such as AlexNet and SqueezeNet, on most benchmark databases. Compared to state-of-the-art technologies, the VGG-19 model achieved an accuracy of 99.7%. These results are considered a reference point for implementing transfer learning with the VGG-19 network to extract the probability vector in the present investigation.

Emotion recognition is currently used in various fields such as education, gaming, robotics, medical care and also in the automotive field, which is why new emotion models require more research to address the various challenges that exist around emotion recognition through facial expressions. The work of Bakariya et al. (2024), creates a real-time system that can analyze unstructured data capable of recognizing human faces, evaluate emotions and even make recommendations based on a deep learning approach. The accuracy of their proposal is 73.02%, objectively recognizing 6 emotions such as: anger, fear, joy, neutrality, sadness and surprise. In the same way Talaat et al. (2024), developed a real-time emotion identification system to detect emotions but in autistic children, using an autoencoder for feature extraction and selection, and applying transfer learning with different CNNs as a reference due to the reduced number of data. The Xception model achieved the highest performance with an accuracy of 95.23% demonstrating the ability of the procedure to recognize emotions. The study in question also establishes the feasibility of using transfer learning which is a critical point within the present research, well as Gursesli et al. (2024), which proposes a significant reduction of computational power and complexity for emotion recognition based on existing architectures such as MobileNetV2. Similarly, Ravikumar et al. (2024), used transfer learning and data augmentation procedures for model generalization using multiple reference data, concluding that a deep learning model based on transfer learning is recommended for recognizing emotions from facial expressions.

However, work such as Mehrotra et al. (2024), states that previous research focuses primarily on accuracy without taking into account prediction time which is also critical for an optimal emotion recognition system. Their suggested approach achieved an accuracy of 71.61% in a time of 58 minutes in the training process using the FER dataset.

2.2 Speech signals for emotion recognition

Some studies have proposed the development of Speech Emotion Recognition (SER) systems based on features extracted from spectrograms, implementing artificial neural network architectures. Mustaqeem et al. (2020) presents a significant method for selecting essential speech signal segments using a Radial Basis Function Network (RBFN). The selected segments are converted to spectrograms and passed to a CNN model to extract silent and discriminative features. These CNN features are normalized and fed into a deep bi-directional long short-term memory (BiLSTM) network for learning temporal features to recognize emotions. Similarly, Yao et al. (2020) developed an integrated framework combining Deep Neural Networks (DNN), CNN, and Recurrent Neural Networks (RNN). In their approach, the utterance-level outputs of high-level statistical functions (HSF), segment-level Mel-spectrograms (MS), and frame-level Low-Level Descriptors (LLDs) are inputted to DNN, CNN, and RNN, respectively. This yields three separate models–HSF-DNN, MS-CNN, and LLD-RNN. A multi-task learning strategy is employed across the models to extract generalized features by simultaneously performing emotional attribute regression and discrete emotion category classification.

Despite the rise of deep learning techniques, recent studies propose less computationally expensive methodologies in terms of time and performance. For example, Daqrouq et al. (2024) evaluates the performance of various machine learning algorithms,

such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and neural networks, by using discrete wavelet transform (DWT) with linear predictive coding (LPC). These findings can help guide the selection of appropriate classifiers and feature extraction methods for future research and real-world applications that use speech as a source of information.

The Mel Frequency Cepstral Coefficient (MFCC) method is widely employed for analyzing speech signals and has demonstrated superior performance in speech-based emotion recognition systems compared to other features. Alluhaidan et al. (2023) presents an emotion recognition model using hybrid features extracted from MFCC and the temporal domain, with a One-Dimensional Convolutional Neural Network (1D CNN). They use publicly available datasets such as EMO-DB, SAVEE, and RAVDESS to evaluate their method's performance, achieving precision rates of 96.6% for EMO-DB, 92.6% for SAVEE, and 91.4% for RAVDESS. The fusion of hybrid features with 1D CNN proved effective for speech emotion recognition, outperforming both conventional and deep learning approaches. Similarly, Bhangale and Kothandaraman (2023) explores emotion recognition using acoustic features and 1D CNN. Their study focuses on analyzing various acoustic features, such as MFCC, Linear Predictive Cepstral Coefficients (LPCC), Wavelet Packet Transform (WPT), Zero-Crossing Rate (ZCR), and Root Mean Square (RMS), to enhance the distinctiveness of speech signals. They develop a deep 1D CNN to reduce computational complexity in emotion recognition, testing its effectiveness on datasets like EMO-DB and RAVDESS. Their results demonstrate high accuracy for recognizing various emotions, including 94.83% accuracy for anger, 91.38% for calmness, 89.66% for disgust, 89.66% for fear, and 91.38% for happiness.

Recent studies continue to explore variations of MFCC. For instance, Mishra et al. (2024) extracted the MFCC coefficient matrix from various datasets, calculating features such as the statistical mean, MFCC-based approximate entropy, and MFCC-based spectral entropy. Their model achieved classification accuracies of 85.61%, 77.54%, and 76.26% across three different speech datasets.

There are also cases, such as in Khan et al. (2024), where it is suggested that reliable and robust multimodal speech emotion recognition systems are necessary to efficiently recognize emotions across multiple modalities, such as speech and text. In their proposal, a deep feature fusion technique for audio and text signals is applied to predict the emotion label. The proposed model processes raw speech and text signals using a CNN and employs encoders for semantic and discriminative feature extraction. The authors evaluate their model on various datasets and conduct extensive experiments, obtaining significant results that highlight the robustness and versatility of models trained on data from different sources.

However, the complexity of speech signal characteristics continues to present many challenges in emotion recognition. A study by Yang et al. (2024) introduces a multi-feature approach that aims to reduce the dimensionality of features to effectively address overfitting issues. Their experiment achieved remarkable accuracy on diverse datasets, with their model reaching accuracies of 98.47% and 98.87%, demonstrating the ability to accurately discern emotions from speakers. These findings underscore the importance of incorporating feature reduction in models that use speech cues as a primary source of information. This is an important consideration for the research presented in this manuscript, as reducing the dimensionality of motor cues will likely contribute to developing an optimal model for emotion recognition.

2.3 Biophysiological signals

Physiological signals are biochemical responses to stimuli that can be useful in identifying emotions. These data may include Electrocardiogram (ECG) signals, Electroencephalogram (EEG) signals, Electromyogram (EMG) signals, Galvanic Skin Response (GSR), and Heart Rate (HR). Methods based on biophysiological signals have shown promising results in emotion recognition. Recent studies, such as Yang et al. (2023), propose an LSTM system that combines smartphone sensors to capture images of the driver and a bracelet to record electrodermal activity, accurately determining the user's emotional state. The system was evaluated through a user study with 45 participants, using affective responses (facial expressions, speech, keystroke typing) and physiological responses (blood volume, electrodermal activity, and skin temperature) induced by visual stimuli.

Alternatively, other researchers have explored different methodologies for detecting emotions through ECG. For instance, Wu and Chang (2021) conducted experiments using ECG to investigate the impact of music on emotions. Their findings indicated that fast, intermediate, and slow music influenced the autonomic nervous system in different ways: fast music stimulated it, intermediate music inhibited it, and slow music had no significant effect. Additionally, they suggested that music could help alleviate psychological pressure. Hu and Li (2022) collected 140 ECG signal samples triggered by Self-Assessment Manikin (SAM) experiments using the International Affective Picture System. They employed a Wasserstein Generative Adversarial Network (WGAN) with a gradient penalty to augment different classes of samples. The results showed that increasing the quantity of data improved the accuracy and weighted F1 scores for all three classifiers.

Similarly, Fang et al. (2024) applied an emotion recognition method using random convolutional kernels for ECG signals. This approach reduces computational complexity and training time compared to methods that rely on multiple physiological signals or deep neural networks. It was validated on three publicly available datasets, achieving average recognition accuracies of 93.7%, 95.5%, and 91.5% in the valence, arousal, and dominance domains, following the three-dimensional approach to emotions proposed by Russell. Likewise, the study presented by Sweeney-Fanelli and Imtiaz (2024) implements deep learning techniques for emotion recognition using ECG signals, achieving accuracies of 98.68% for arousal and 97.30% for valence on two publicly available datasets. The results highlight the potential of temporal convolutional neural networks to enhance human-computer interactions and healthcare monitoring systems through improved emotion recognition. Additionally, Arslan et al. (2024) focuses on the processing and analysis of ECG and GSR signals to develop a predictive model

for emotion classification. Their methodology involves extracting key features such as heart rate variability, morphological features, and Hjorth parameters. A feature selection process based on statistical analysis is applied to optimize and adapt the data for machine learning techniques, resulting in a classification accuracy of 97.78%. This demonstrates the feasibility of real-time emotion recognition through statistical feature selection and machine learning algorithms.

Changes in human emotions involve a complex process that often triggers spatiotemporal brain activity, which can be detected by EEG. EEG signals contain valuable information, as they reflect the activity of countless neurons in the cerebral cortex, providing real-time insights into brain functioning. Moreover, EEG recording is relatively simple and cost-effective, making EEG-based methods highly attractive for emotion recognition and evaluation.

Studies have shown a correlation between EEG signals and different emotional states. For example, Andreu-Perez et al. (2021) conducted a study in which participants played the video game "League of Legends" while their brain activity was monitored using functional near-infrared spectroscopy (fNIRS), along with video recordings of their facial expressions. They decoded the players' expertise level within a multimodal framework, achieving a tri-class classification precision of 91.44%. Similarly, Zhang J. et al. (2020) measured EEG signals, extracted features, and processed the data using a modified radial basis function neural network algorithm. Their experimental results demonstrated the superiority of this modified algorithm over others. In a related study, Zangeneh Soroush et al. (2020) reconstructed EEG phase space, extracted Poincaré intersections as features, and integrated them into a classification model, introducing an effective method for emotion recognition and nonlinear signal processing. Likewise, Lu et al. (2020) analyzed people's physiological responses to different lighting conditions based on EEG signals. Their findings revealed that illumination levels and color temperatures significantly impact the visual center's response, which can help design optimal lighting environments.

In recent times, EEG-based emotion recognition models continue to evolve. Cai et al. (2024) introduced a new EEG input format called EEG spectral imaging, which integrates spatial domain features using Azimuthal Equidistant Projection (AEP) and frequency domain features through differential entropy. Experiments performed on the SEED and SEED IV datasets demonstrated superior performance compared to benchmark methods and state-of-the-art models. Their results showed relative improvements of 0.6% and 0.08% in subject-dependent experiments, achieving accuracies of 80.07% and 66.72%, respectively. The author also notes that existing approaches (Trujillo et al., 2024; Al-Asadi et al., 2024; Tokmak et al., 2024; Jha et al., 2024) mainly focus on the time and frequency domain characteristics of EEG signals.

Unlike behavioral data, physiological data are considered more objective because they typically reflect involuntary responses that are difficult to consciously conceal or alter (Lin and Li, 2023). However, much of the research in this field depends on medical-grade physiological sensing equipment, which tends to be invasive, expensive, and requires technical expertise, thus limiting its application in real-world settings (Dunn et al., 2018).

In the study by Siam et al. (2023), an approach is proposed to identify the mental stress of automotive drivers based on selected biosignals such as ECG, EMG, GSR, and respiration rate. Six different machine learning models were employed to classify stress and relaxation states. The proposed Stress Detection Technique (SDT) consists of three main phases: biosignal preprocessing, feature extraction, and classification. The results show that Random Forest outperformed other techniques, achieving a classification accuracy of 98.2%, sensitivity of 97%, and specificity of 100% using a public driving dataset. This research aims to integrate biosignals with the automotive industry to develop an applicable Advanced Driver Assistance System (ADAS). Additionally, Waheed Awan et al. (2024) suggests a method based on a one-dimensional convolutional neural network and Vision Transformer, where the process involves decomposing signals into segments, removing noise, and extracting features. These features are then integrated into a single vector for classification using a set of classifiers. The results are synthesized using Model Agnostic Meta Learning (MAML) to improve prediction accuracy. The model was validated on the AMIGOS and DEAP datasets, achieving up to 98.2% accuracy with 10-fold cross-validation, leveraging physiological signals for comprehensive emotion assessment.

2.4 Emotion recognition in automotive field

Although it is challenging to present a definitive or statistical figure for accidents caused by drivers' emotions, some researchers have worked on proposals to determine drivers' emotional states in real time, based on the analysis and processing of various sources of information, in order to prevent accidents in advance. The following are studies related to emotion recognition in an automotive environment.

Different methods for objectively determining drivers' emotions have been explored. For example, Wang et al. (2020) used feature fusion of multiple ECGs to detect driver emotions based on a backpropagation network and the Dempster-Shafer evidence method. In their approach, they selected ECG signals, time-frequency domain, waveform, and nonlinear features as parameters for emotion recognition, specifically identifying drivers' calmness and anxiety while driving. The results demonstrated that after fusing the ECG parameters, the proposed model could recognize drivers' emotions, with an accuracy of 91.34% for calmness and 92.89% for anxiety. The study concludes that this method holds theoretical and practical significance for improving road safety.

Other proposals present new multimodal frameworks for emotion recognition, integrating facial expressions and heart rate data. For instance, Du et al. (2021) established a deep learning model called the Bidirectional Convolutional Long-term Memory Neural Network (CBLNN). This model predicts drivers' emotions based on geometric features extracted from changes in RGB components. The facial features obtained using the CNN serve as intermediate variables for Bidirectional LSTM (BI-LSTM) heart rate analysis. The BI-LSTM output is then used as input to the

CNN module to extract features. CBLNN applies multimodal factorized bilinear clustering to fuse the extracted information and classify five common emotions. This emotion detection method achieved recognition rates of 91.6%, 90.50%, 91.51%, and 89.15% for happiness, anger, sadness, and neutrality, respectively.

The implementation of artificial intelligence algorithms has been a critical tool for objectively recognizing emotions in drivers, due to their capacity to extract features that identify the emotions we experience throughout the day. Naqvi et al. (2020) based their method on gaze changes and facial emotions using NIR camera sensors and an illuminator installed in the vehicle. They acquired time-series data from aggressive and normal drivers by simulating driving scenarios with racing and truck driving games. Software was used to capture the driver's image data, extracting images of the face and eyes to detect gaze changes with a CNN and classify facial emotions. Score-level fusion was applied to the scores obtained from gaze changes and facial emotions to classify aggressive and normal driving. The accuracy of their method, measured through a self-generated test database, achieved a classification accuracy of 98.93%. Similarly, Cui et al. (2020) implemented CNN for emotion recognition, where a multitask network processed facial expressions under varying lighting conditions. This was accomplished by simultaneously restoring corrupted images using a pre-trained CNN and predicting specific emotions. Their model, evaluated on the FerPlus and Cohn-Kanade (CK+) datasets (Lucey et al., 2010), achieved an accuracy of 83.1%.

Recent studies continue to advance emotion recognition techniques, with a focus on image analysis and processing. For example, Xiao et al. (2022) proposed a method that includes three modules: facial detection, a data resampling module, and an emotion recognition module based on a deep CNN pre-trained with FER (Barsoum et al., 2016) and CK+ datasets. This method, designed for real-time emotion recognition in drivers, collected on-road facial expression data in various driving scenarios, achieving an emotion recognition performance of 97.2%. Similarly, Zaman et al. (2023) used CNN, RNN, and multilayer perceptron classification models to develop a facial expression recognition system. Their model, built on the faster region-based enhanced CNN (R-CNN) for real-time face detection, fused CNN model features to train an emotion classification model. By incorporating InceptionV3 into the model, they improved accuracy, achieving recognition rates of 98.01%, 99.53%, 99.27%, 96.81%, and 99.90% across several datasets, including JAFFE (Lee and Kang, 2020), CK+, FER2013 (Zahara et al., 2020), AffectNet (Mollahosseini et al., 2017), and their own dataset. Additionally, Sukhavasi et al. (2022) proposed a hybrid methodology combining CNN and Support Vector Machine (SVM) to enhance classification predictions. By fusing Local Binary Patterns (LBP) and Gabor filters to extract robust features, their technique achieved accuracies of 84.41%, 95.05%, 98.57%, and 98.64% on FER-2013, CK+, KDEF (Goeleven et al., 2008), and KMU-FED (Kumar et al., 2022), respectively.

Another innovative study by Jain et al. (2023) proposed the development of an algorithm called Squirrel Search Optimization with Deep Learning Enabled Facial Emotion Recognition (SSO-DLFER) for detecting emotions in autonomous vehicle drivers. This algorithm employed the RetinaNet (Lin et al., 2017) for face detection and the NASNet-Large (Zoph et al., 2018) feature extractor with the Gated Recurrent Unit (GRU) classifier for emotion recognition. Hyperparameter tuning based on SSO enhanced the model's performance, achieving a maximum accuracy of 99.50% across multiple datasets, including KDEF and KMU-FED.

A notable study that deviates from image analysis is presented by Chen et al. (2024), who explored the relationship between EEG signals and emotions in a simulated driving environment. Their method used vehicle speed as a variable to simulate obstacle avoidance at different danger levels. For data processing, graphical neural networks with functional connectivity and attention mechanisms were employed to simulate the brain's physiological structure. Their binary classification result achieved an F1 score of 91.5%, demonstrating the effectiveness of capturing EEG signals and monitoring emotional states through deep learning models.

In another study, Mou et al. (2023) introduced a multimodal fusion framework for driver emotion recognition, employing a ConvLSTM network with a hybrid attention mechanism to integrate eye, vehicle, and environmental data. Their research revealed correlations between driver emotions and stress, with participants exhibiting higher levels of valence and emotional dominance under stressful conditions. The model achieved average precision values of 97.64% for valence, 97.27% for arousal, and 96.47% for dominance, further validated through ablation experiments.

Several studies have examined drivers' behavioral characteristics, such as motor activity signals influenced by emotional states. For instance, CAN bus signals are commonly used, though access to these signals is restricted to in-house developers (Zepf et al., 2020). Despite the growing interest in driver behavior, many studies lack sufficient detail about the characteristics and behaviors associated with different emotions. One of the least explored methodologies in a driving context is the use of multimodal artificial intelligence (AI) models. These models, which process data from various sources, have shown potential in emotion recognition. Oh et al. (2021) proposed an emotion recognition model that fused facial expression data with electrodermal activity, achieving an accuracy of 86.8%. Likewise, Zhou et al. (2023) proposed a multimodal model integrating driver voice, facial images, and video sequences using CNN, Bi-LSTM, and hybrid attention modules, recognizing six negative emotions (e.g., sadness, anger, fatigue) with an accuracy of 85.52%. Ying et al. (2024) similarly employed audio and video features to recognize driver emotions, enhancing the safety and humanization of advanced driver-assistance systems (ADAS).

The current state of the art highlights the feasibility of developing reliable emotion recognition systems suitable for real-world implementation. However, the exploration of multimodal AI models that integrate behavioral and facial data remains largely unexplored. Mou et al. (2023) have shown potential in emotion identification tasks within automotive environments. Such advancements can improve the driving experience and enhance road safety by mitigating aggressive or distracted behaviors, ultimately benefiting society.

Exploring alternatives for accurate, efficient emotion recognition remains complex, yet research on behavioral signals continues to provide insights. For instance, Paredes et al.

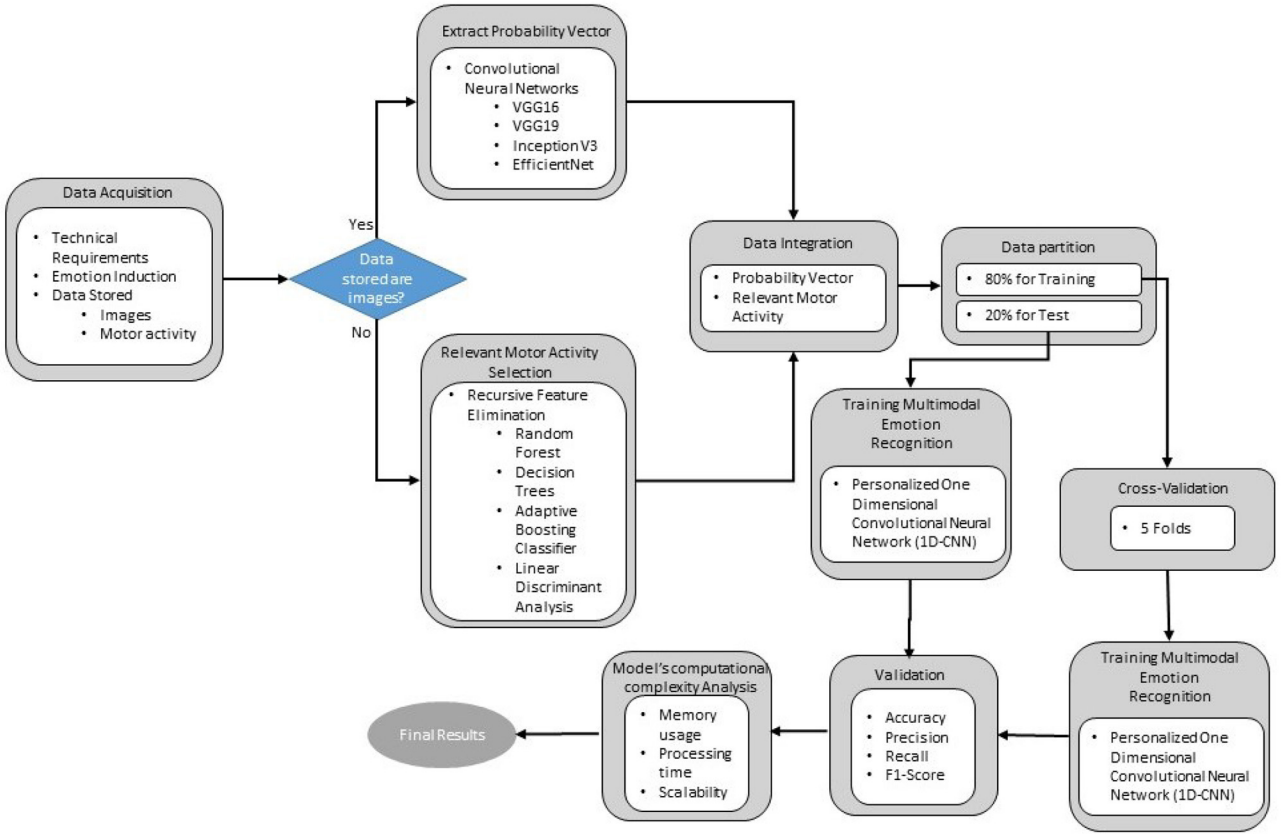

FIGURE 1
Proposed methodology for generating a dataset of motor activity and facial images of drivers under different induced emotional states, and a multimodal emotion recognition model using convolutional neural networks.

(2018) demonstrated the ability to measure driver strain using a steering angle and a mass-spring-damper model. Other studies have established correlations between emotional states and driving behaviors. Hu et al. (2024) proposed a multimodal emotion recognition model using facial videos and driving behavior (e.g., brake pedal force, Y-axis position, vehicle Z-axis position), achieving an accuracy of 63.83% in a first approach to intelligent emotion recognition based on facial and behavioral features.

In conclusion, while significant progress has been made, challenges remain in improving the accuracy and robustness of emotion recognition in dynamic environments. Integrating facial data with motor activity data, such as steering wheel and pedal interactions, could provide a richer context for recognizing drivers' emotions, enhancing both the accuracy and practicality of emotion recognition systems for real-world applications.

3 Materials and methods

Figure 1 presents the methodology followed in this research to develop a multimodal emotion recognition model. The model primarily utilizes the motor activity or behavior of drivers and geometric patterns of facial expressions as inputs. The first stage, data acquisition, focuses on generating a dataset of motor activity signals obtained from key vehicle elements, such as steering wheel angle, pedal movement, and braking, alongside visual data like facial images of the participants. These data are collected during the induction of four emotions in a simulated driving environment, using emotion neutralization and the augmented autobiographical recall technique.

In the second stage, the probability vector of images for each induced emotion is extracted using a pre-trained Convolutional Neural Network (CNN). The third stage involves selecting the most relevant driving signals for emotion recognition by applying intelligent feature selection methods with machine learning algorithms such as Random Forest Classifier (RFC), Decision Trees (DT), Adaptive Boosting Classifier (ABC), and Linear Discriminant Analysis (LDA). The fourth stage processes both motor signals and probability vectors through a One-Dimensional Convolutional Neural Network (1D-CNN) to generate a multimodal model capable of recognizing a limited set of emotions. This step marks a preliminary attempt at analyzing and processing these types of data using advanced artificial intelligence algorithms simultaneously.

In the fifth and final stage, the model is validated using key performance metrics from the field of artificial intelligence to evaluate its accuracy in identifying emotions in a driving environment. Additionally, a computational complexity analysis is performed to determine the model's viability for real-time inference in automotive systems.

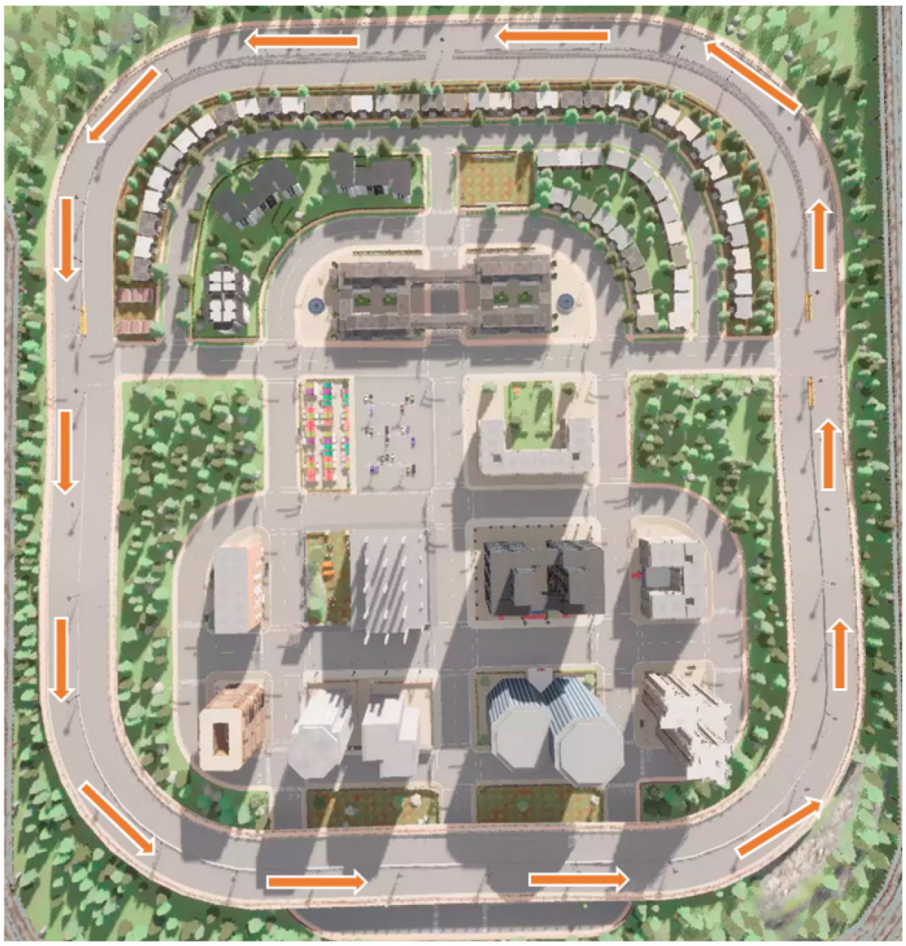

FIGURE 2
Map of the CARLA simulator used for driving simulation and the established route. The orange lines represent the route followed by each of the participants.

3.1 Data acquisition

The experimental tests were conducted using the open-source driving simulator CARLA 0.9.13 for safety reasons. CARLA was developed to support the creation, training, and validation of autonomous vehicles, and it is widely used for advanced driver assistance system research, including algorithm training for perception tasks. CARLA is freely available, and its sensor configuration settings allow for the collection of signals that can be used to train driving strategies (Dosovitskiy et al., 2017). For this research, the same scenario as shown in Figure 2 (whose specifications can be found here: https://carla.readthedocs.io/en/latest/map_town05/) was used for all participants and emotions tested. Each participant followed a pre-established route under uniform virtual conditions, adhering to real-world traffic rules. As the tests were conducted in a controlled driving environment, only active driving data were collected. This included recording the steering wheel angle (-180 to 180 degrees), brake pedal movement (-1 to 1), and throttle pedal movement (-1 to 1), as driver behavior can be influenced by emotional state, especially in interactions with the vehicle such as steering adjustments and pedal usage (Zepf et al., 2020).

To capture motor activity data, the Logitech Driving Force G29, which includes the steering wheel, throttle pedal, and brake pedal, was used. These peripherals are designed specifically for driving simulations, making them ideal for collecting essential data. The simulator's built-in properties were leveraged to record critical information, including the steering wheel angle and the amount of movement in both the brake and throttle pedals.

For capturing images of the region of interest (ROI)—in this case, the driver's facial geometry—a LOGITECH C270 camera with 720 megapixels was used.

The computer used for the experiment was equipped with an Intel Core i5-9400F processor running at 2.90 GHz, 32 GB of RAM, and an NVIDIA GeForce GTX 1070 Ti graphics card.

Each participant signed an informed consent form, following the ethical guidelines established in the Helsinki Declaration.

Various methods exist for inducing emotions, but augmented autobiographical recall has proven particularly effective in driving environments, as suggested by Braun et al. (2018). This method is advantageous because it allows the participant to generate the emotional stimulus themselves, reducing the risk of misinterpretation. Additionally, it can be smoothly integrated into driving tasks, offering

TABLE 1 Song titles, artist who performs it and the type of emotion it can evoke, obtained from the DEAP data set.

Emotion	Artist	Title
Happiness	M. Franti and spearhead	Say Hey (I Love You)
Sadness	James Blunt	Goodbye my lover
Anger	Dead To fall	Bastard set of dreams

a seamless transition from emotional provocation to the driving experience.

In this method, participants are asked to recall and write about a past event to evoke a specific emotion. They are encouraged to provide as many details as possible and vividly recount the events. Crucially, participants recall the story to themselves, without the experimenter's presence (López-Cano et al., 2020). To further enhance the emotional experience, scientifically validated songs known to evoke specific emotions were played through headphones. These auditory stimuli, sourced from the DEAP dataset (Koelstra et al., 2012), intensified the emotions participants experienced during the drive. Table 1 lists the songs used in the experiment.

Before beginning the augmented autobiographical recall, participants evaluated their emotional state using an affective annotation platform that featured the Self-Assessment Manikin (SAM). SAM is a widely used tool for assessing emotional states, featuring a graphic scale for valence and arousal from 1 to 9 (Veeranki et al., 2021). After this initial evaluation, participants were asked to write about a neutral event, such as their morning toothbrushing routine, to establish a neutral emotional state. This step follows the methodology of Sanghavi et al. (2020), which demonstrates that recalling a mundane event effectively induces neutrality. Once participants had completed this task, they put on headphones and began the driving simulation, during which they listened to a song designed to maintain a neutral emotional state. This initial phase not only familiarized participants with the simulator but also prepared them for the emotional recall process.

After a 5-minute neutral driving session, participants reassessed their emotional state on the platform. They were then asked to recall a moment that elicited one of the target emotions for the study (Happiness, Anger, or Sadness) and repeated the same tasks as in the neutral driving phase. Finally, participants evaluated their emotional state again, using the augmented autobiographical recall method during driving.

All collected data, including motor activity and facial images, were organized by the emotion associated with each participant, ensuring control over records and allowing for efficient storage. This organization adhered to the emotion categories of Happiness, Anger, and Sadness, selected based on Plutchik's emotional model. According to Plutchik, these basic emotions serve as building blocks for more complex emotions (Semeraro et al., 2021). Additionally, focusing on a smaller set of emotions allowed for a more representative and high-quality dataset, ensuring that the data were well-labeled and balanced (Khoo et al., 2024). Figure 3 summarizes the data acquisition process.

3.2 Probability vector extraction

In this study, pretrained Convolutional Neural Network (CNN) models, including VGG16, VGG19, Inception V3, and EfficientNet, were employed. These models are widely used and straightforward, and recent studies (Zaman et al., 2023; Gite et al., 2023; Tauqeer et al., 2022; Oh et al., 2021; Ahmad et al., 2024; Wawage and Deshpande, 2022) have demonstrated their remarkable performance across a variety of applications, including emotion recognition. The objective of the pretrained CNNs is to calculate the probability vector of facial geometry images for each induced emotion obtained from participants during simulated driving using transfer learning (Kusal et al., 2024).

The probability vector represents the distribution of probabilities across different emotions, which helps reduce the dimensionality of the data compared to raw image feature extraction. Emotion recognition often involves addressing variability in facial expressions, lighting conditions, and other environmental factors. The probability vector captures the uncertainty associated with these variations, making the model more robust to changes in the input data. By focusing on the probability distribution of emotions, rather than specific image features, the model achieves a more nuanced understanding of the underlying semantics of facial expressions. In some cases, using a probability vector enables end-to-end learning, allowing the model to directly map input images to probability distributions of emotions (Zhao et al., 2020).

The probability vector is generated before assigning a label to each input image. The number of elements in the probability vector corresponds to the number of induced emotions in the data acquisition process (Neutral, Happy, Angry, and Sad), with each element representing the likelihood of one specific emotion. Since the total probability is 1, the sum of all elements in the probability vector equals 1, and each element's value ranges between 0 and 1. The emotion associated with the highest value in the probability vector is selected as the detected emotion. Figure 4 illustrates the probability vector extraction process. These probability vectors, derived from the facial images, are a key complement to the tabular dataset (e.g., throttle pedal movement, brake pedal movement, and steering wheel angle) in the integration phase.

The VGG16 CNN was developed by the Visual Geometry Group (VGG) at the University of Oxford and became well-known after winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in the object identification category. This model has also shown very promising results in emotion recognition in other studies (Verma and Choudhary, 2018a,b). The goal of this network is to demonstrate that increasing the depth of the network can improve performance in certain tasks (Shahzad et al., 2023). The VGG19 network, proposed by Simonyan and Zisserman, consists of 19 layers, including 16 convolutional layers and 3 fully connected layers, and it is trained to classify 1000 different objects. VGG19 is trained on the ImageNet database, which contains over one million images (Bansal et al., 2023).

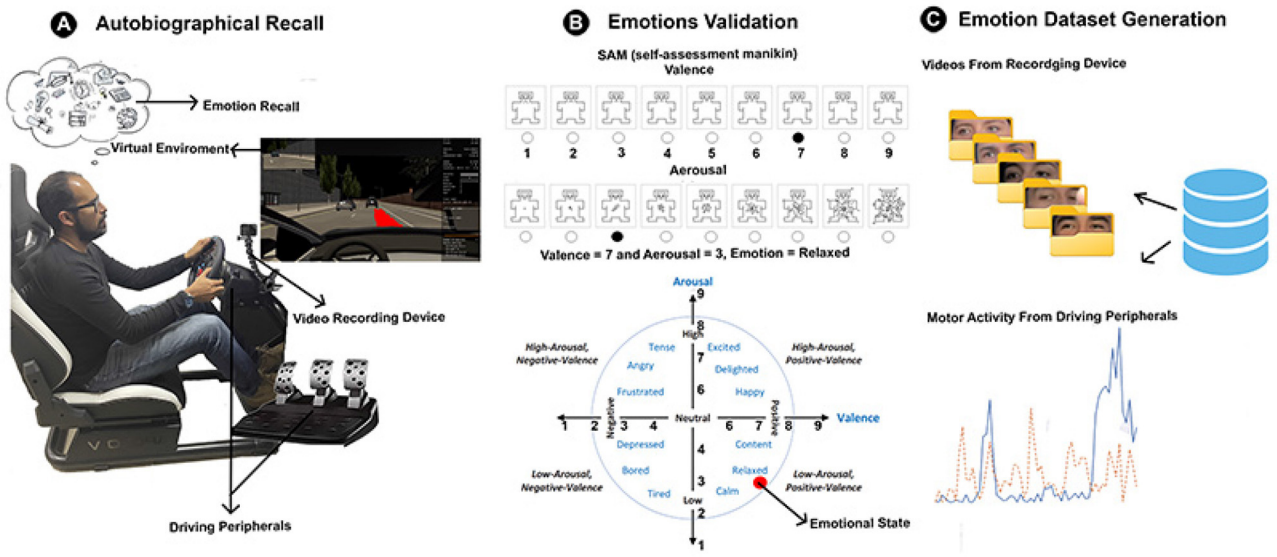

FIGURE 3
Proposed methodology for generating a dataset of motor activity and facial images of drivers under different induced emotional states, and a multimodal emotion recognition model using convolutional neural networks.

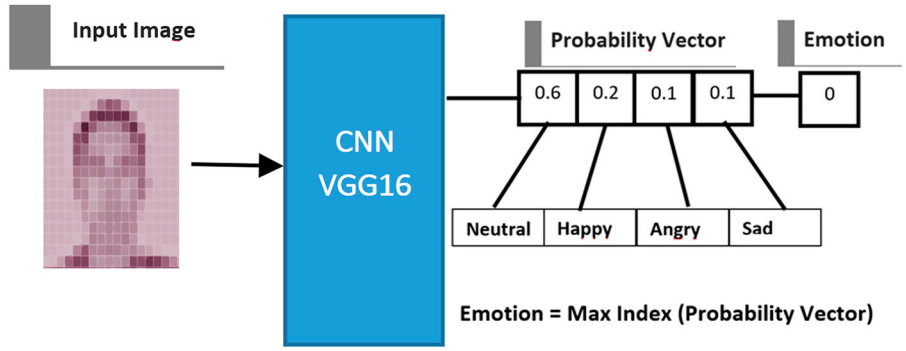

FIGURE 4
Process of extracting probability vector from an image using a CNN.

In contrast, the Inception V3 network is a deep architecture developed for the 2014 ImageNet visual recognition challenge. One of its main advantages over VGGNet is its faster execution speed (Cao et al., 2021). Finally, EfficientNet is another pretrained CNN, designed for transfer learning in image classification tasks. This architecture, developed by Google AI in May 2019, is available through TensorFlow and GitHub libraries (Marques et al., 2020).

3.3 Feature selection

Once the tabular data from the driving simulator peripherals, including throttle movement, braking, and steering wheel angle, are collected for the different induced emotional states of the participants, a feature selection process is performed using the Recursive Feature Elimination (RFE) technique. This is necessary due to the high dimensionality of the motor activity data and the probability vectors derived from the images.

RFE is a commonly used feature selection technique in machine learning. The main idea behind RFE is to iteratively train a model on subsets of features and eliminate the least important ones at each iteration until the desired number of features is reached. The general algorithm is as follows (Ba et al., 2023):

- Initialization:
  Let F represent the set of all features, initially $F = \{1, 2, .., p\}$ where p is the total number of features
- Iteration:
  For each iteration i (where $i = 1, 2, ..$):
  Train the model M using the features X_F and target variable y.
  Asses the importance of each feature based on some criterion, denoted as $I(f)$ for feature f in F.

identify the least important feature, $f_{min} = argmin_{f \epsilon F} I(f)$. Remove the least important feature from $F : F \leftarrow F \backslash \{f_{min}\}$.

- Stopping Criteria:
  Repeat the iteration until the desired number of features is reached or a predefined stopping criterion is satisfied.

The motor activity data vector consists of windows of 50 data points per record for steering wheel angle, throttle pedal movement, and braking movement. These data points are processed using RFE in conjunction with various machine learning techniques to identify the optimal data windows for emotion recognition based on variations in specific data segments collected under different emotional states. Ultimately, the most relevant data segments will be integrated with the probability vectors derived from the images. The machine learning techniques implemented for analysis and processing are described below.

3.3.1 Random forest classifier

RFC is a machine learning algorithm that builds multiple decision trees, where each tree is generated using a random subset of the data. Each tree casts a vote, and the most popular class is selected to classify the input vector (Amiri et al., 2024).

RFC uses the Gini index as a measure for selecting attributes, which quantifies the impurity of an attribute with respect to the target classes. The Gini index is shown in Equation 1.

$$\sum \sum_{j \neq i} (f(C_i, T)/ \mid T \mid)(f(C_j, T)/ \mid T \mid) \quad (1)$$

where $f(C_i, T)$ represents the frequency of class C_i in dataset T, and $|T|$ is the total number of instances in T.

3.3.2 Decision trees

Decision trees are supervised predictive models known for their interpretability and robustness, and they are widely applied in various domains. The fundamental idea behind a decision tree is to recursively divide the dataset into smaller subsets based on specific features until a strong prediction for the target variable is achieved. Each division is made in such a way that it maximizes the homogeneity of the resulting subsets in terms of the target variable (Costa and Pedreira, 2023).

3.3.3 AdaBoost classifier

The goal of the AdaBoost algorithm is to combine multiple weak learners to form a strong learner, thereby improving the classification or prediction model. The algorithm works by adjusting the weights of the misclassified points at each iteration, giving more weight to incorrectly classified samples. A weak learner is trained using these weighted data points. A coefficient is assigned to each weak learner based on its performance. For misclassified points, their weights are increased, and the weights of correctly classified points are decreased. The process is repeated until all data points are correctly classified or a stopping criterion is met.

The AdaBoost algorithm is commonly used for binary classification problems but can be extended to handle multiclass classification using methods such as One-vs-All (OvA) or One-vs-One (OvO). The equation for the combined classifier $H(\mathbf{x})$ is presented below:

$$H(\mathbf{x}) = \arg\max_k \sum_{t=1}^{T} \alpha_t \cdot \mathbb{I}(h_t(\mathbf{x}) = k)$$

where T is the number of iterations, α_t is the weight of the t-th weak classifier, $h_t(\mathbf{x})$ is the prediction of the weak classifier, and $\mathbb{I}$ is the indicator function.

3.3.4 Linear discriminant analysis

LDA is a supervised dimensionality reduction technique that aims to find a linear combination of features that maximizes the between-class variance while minimizing the within-class variance. In the transformed space, samples of the same class are separated as much as possible. For multiclass problems, LDA can be extended using Fisher's discriminant analysis to find a subspace that captures the maximum variability between classes (Zhu et al., 2022).

Suppose that each class C has a mean μ_i and a shared covariance matrix Σ. The between-class scatter matrix Σ_b can be defined as the covariance of the class means:

$$\Sigma_b = \frac{1}{C} \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

where μ is the mean of the class means.

3.4 Model generation

Once the data integration process was completed, and the most significant motor signals were identified using machine learning algorithms, dimensionality reduction was performed using the Recursive Feature Elimination (RFE) technique. The resulting dataset comprised 3,361 observations and 13 columns, with the first 9 columns representing motor activity data and the last 4 columns representing the extracted probability vectors. This adjustment ensured consistency between the significant motor activity data used for emotion recognition and the number of extracted probability vectors, given that the number of images was much smaller than the motor activity dataset.

With the final dataset established, a multimodal emotion recognition model was developed using a proposed one-dimensional convolutional neural network (1D-CNN). 1D-CNNs are commonly utilized to analyze one-dimensional signals, such as vectors, time series data, and other sequential data types, and have been applied in various fields, including bioengineering, physiological signal analysis, traffic analysis, marketing, and network analysis (Tang et al., 2020).

The proposed network consists of five convolutional layers with filter sizes of 64, 128, 256, 512, and 1024, each using the Softplus activation function and kernel sizes of 3, 3, 2, 2, and 1, respectively. Softplus, known for its smoothness and non-zero

gradient, was introduced by Dugas et al. (2000) in 2001 and, is defined as follows:

$$\text{Softplus}(x) = \ln(1 + e^x) \tag{3}$$

Additionally, max-pooling is applied at the end of the convolutional layers using a 3x3 kernel size. Four dense layers are then added with sizes 512, 256, 128, and 64, each with a Softplus activation function. The final layer, consisting of 4 units, uses a softmax activation function to predict the emotions.

The loss function used is Sparse Categorical Cross Entropy (SCCE), commonly employed in classification tasks where the target labels (y_{true}) are provided as integers (class indices) instead of one-hot encoded vectors. SCCE simplifies the process when handling integer labels, though its formula is similar to that of categorical crossentropy (Chaithanya et al., 2021).

Equation 5 presents the mathematical representation of sparse categorical crossentropy:

$$SCCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot \log(\hat{y}_{ij}) \tag{4}$$

Where:

- N is the number of samples.
- C is the number of classes.
- y_{ij} is a binary indicator of whether class j is the true class for sample i.
- $\hat{y}_{ij}$ is the predicted probability that sample i belongs to class j.

In summary, our approach integrates two types of CNNs: a two-dimensional network that extracts the probability vector from the visual dataset, and a one-dimensional network that processes motor activity signals in conjunction with the probability vector. This integration allows the model to process information from multiple sources and accurately and objectively identify emotions in drivers. The proposed architecture is summarized in Figure 5.

This data integration process has been scientifically validated in the study titled *Detection of Pedestrians in Reverse Camera Using Multimodal Convolutional Neural Networks* conducted by Reveles-Gómez et al. (2023).

3.5 Validation

To validate the performance of the model used in this study, a comprehensive set of evaluation metrics was employed. These metrics quantify the model's performance and assess its ability to effectively distinguish between different emotions. Each metric provides insights into various aspects of the model's quality, offering a holistic understanding of its behavior in classifying the emotions defined in this research.

The evaluation criteria applied include a range of metrics to ensure a thorough assessment of the model's efficacy. Among the metrics used are accuracy, recall, F1-score, and K-fold cross-validation. Together, these metrics provide valuable information about the model's classification performance, its ability to correctly identify each emotion category, and its robustness when evaluated through different validation techniques.

Accuracy is a fundamental metric that indicates the proportion of correctly classified instances among the total number of instances evaluated. Recall, on the other hand, measures the model's ability to correctly identify instances of a specific emotion class from all instances that truly belong to that class. The F1-score takes both precision and recall into account, providing a balanced assessment of the model's performance, particularly useful in cases where class distributions are imbalanced.

In addition to these metrics, K-fold cross-validation was used to evaluate the model's generalization capabilities and consistency across different subsets of the dataset. This technique involves dividing the dataset into K equally sized folds, training the model on $K - 1$ folds, and then evaluating its performance on the remaining fold. This process is repeated K times, with each fold serving as the validation set once, ensuring the model's performance is not overly reliant on any single subset of the data.

By employing this comprehensive suite of evaluation metrics, we gain a deeper understanding of the model's strengths and weaknesses, ensuring a rigorous assessment of its performance in emotion classification tasks.

4 Results

For data acquisition, 50 participants (comprising 10 females and 40 males) aged between 18 and 39 years (with an average age of 25.26, a standard deviation of 5.36, and a median of 25.5) were recruited from the Autonomous University of Zacatecas (UAZ). These participants underwent 55 simulated driving tests at the Interactive Technologies and User Experience Laboratory (L.I.T.U.X) and had a minimum of one year of driving experience (with an average of 6.13 years, a standard deviation of 5.20, and a median of 5 years).

Each of the 50 participants selected for the experiment underwent an initial process of emotion neutralization before inducing the emotions established in this study, using the method of augmented autobiographical recall. Inducing emotions through imagination and music is particularly suitable for measuring direct effects, such as anger caused by aggressive driving, as established in the work of Steinhauser et al. (2018).

Each participant recalled and wrote down an event that produced a certain emotional state, which they later recalled during the test. Below is an example of a happy autobiographical memory:

"When my son was born, holding him and feeling him in my arms was something incredibly special. Every day he hugs me and tells me he loves me, which makes me very happy."

The effectiveness of this methodology was assessed using the SAM (Self-Assessment Manikin) test, which measures emotional states across two dimensions: valence and arousal. During the driving tests, participants continuously characterized their emotional states using the SAM test. The Induced Emotion (IE) for each participant needed to closely align with the one characterized in the continuous model. As noted by Oh et al. (2021), if the induced emotion did not match the assigned values within the emotional range of the continuous model, the participant's data were discarded. Additionally, following the approach proposed

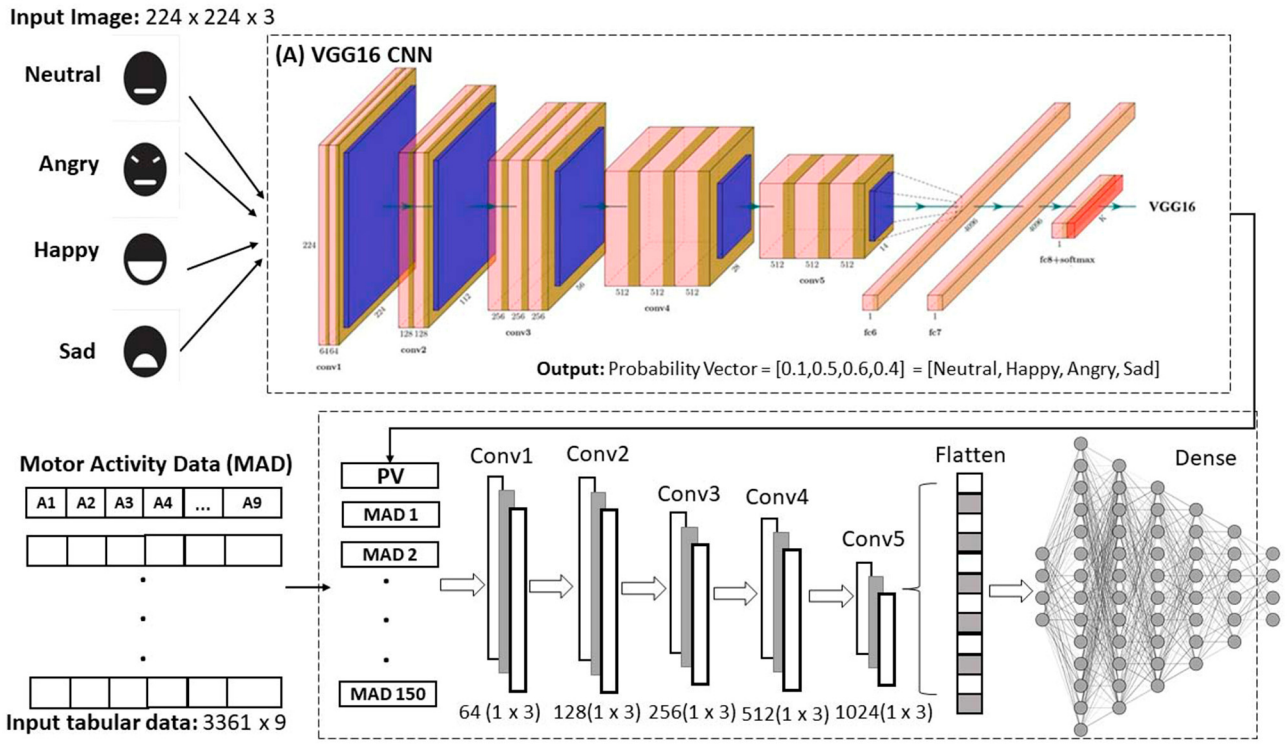

FIGURE 5
Proposed scheme for emotion recognition by fusing the image probability vector using the pre-trained CNN and tabular motor activity data using a one-dimensional convolutional neural network.

by Li et al. (2022), the data were normalized using the min-max normalization method, as shown in Equation 5, to ensure uniform treatment of each arousal-valence value.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

- X_{norm} represents the normalized values.
- X is the original value.
- X_{min} is the minimum value in the dataset.
- X_{max} is the maximum value in the dataset.

Figure 6 shows the distribution of the 50 participants after normalization of their SAM test results. Since the emotional states are diverse and each corresponds to a specific region in the two-dimensional plane (based on arousal and valence), this research categorized the regions where the three target emotions are found: happiness, anger, and sadness. If a participant's induced emotion, based on their arousal-valence values, fell within the expected region for the emotion, their data were considered valid.

Based on the results obtained, among the 50 participants subjected to emotion induction tests in a simulated driving environment, only 42% of the induced emotions matched the actual emotions experienced. This equated to 21 participants in total, with 9 indicating happiness, 9 indicating anger, and 3 indicating sadness. The remaining 58% of participants did not match the induced emotion with the actual one. The dataset contains information from 21 participants, including motor activity data (throttle pedal movement, brake pedal movement, and steering wheel angle) totaling 302,626 records (Neutral = 67,462, Happy = 99,148, Angry = 70,608, Sad = 65,408).

Additionally, visual data (images) collected from the participants' facial geometry totaled 3,361 (Neutral = 1,464, Happy = 1,037, Angry = 366, and Sad = 494), based on the discrete emotional model proposed by Paul Ekman. This research combined discrete and continuous emotional models to assess emotions, based on the premise that facial expressions do not always fully reflect the participant's emotional state, as suggested by Ekman. Therefore, the arousal-valence model proposed by James Russell was also used. The emotion prediction process characterized dimensional emotion labels using both continuous and discrete representations. Recent studies, such as Mihalache and Burileanu (2021), have shown performance gains when converting continuous labels into a discrete set, despite some label quantization error. AlBadawy and Kim (2018) demonstrated the effectiveness of using joint representations of discrete and continuous emotions in describing dynamically changing affective behavior.

Given this, the present study induced emotional states, characterized them using the continuous model, and verified that the induced emotions matched the actual ones via the SAM tool. Next, the participant's visual data were analyzed to ensure that their facial expressions aligned with the expected outcomes in the discrete model. Although the two models differ, the data collection procedure for each emotional state was as follows: if a specific emotion, such as happiness, was induced and matched the real

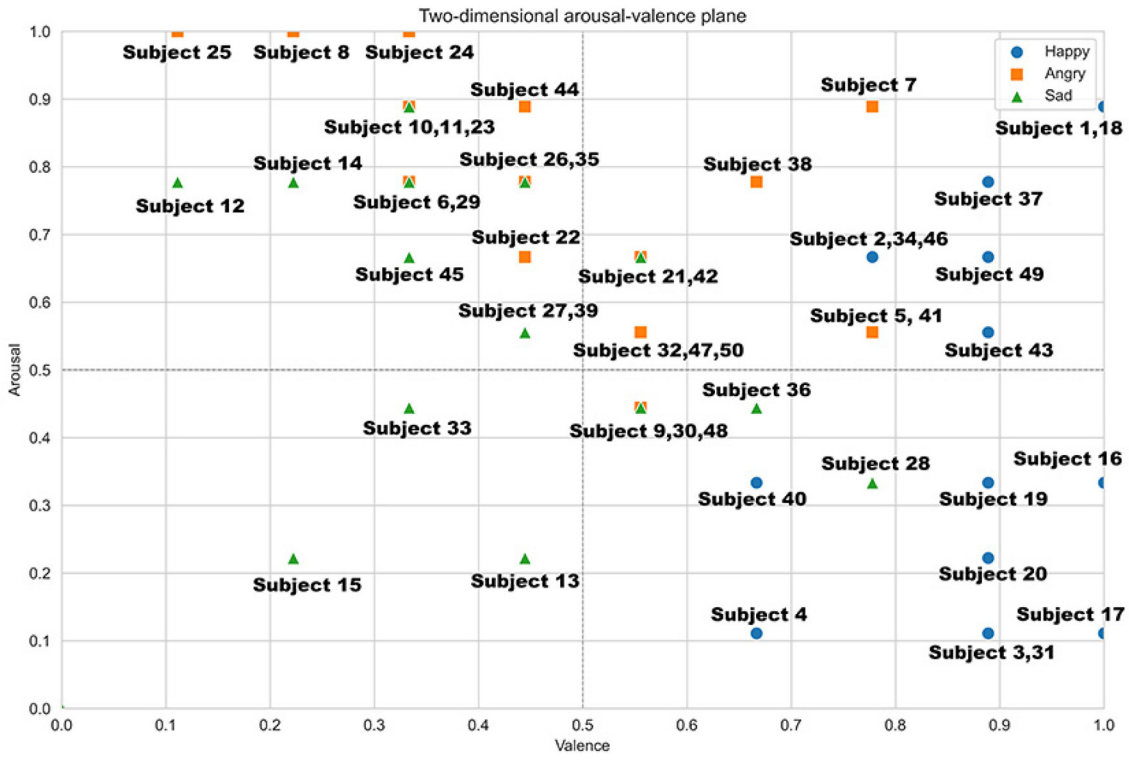

FIGURE 6
Emotional characterization of the 50 participants of the continuous model in the two-dimensional excitation-valence plane.

emotion, and if the participant's arousal-valence values fell within the corresponding range (e.g., 6–9 for both valence and arousal), the motor data and corresponding facial geometry data were considered valid for that emotion.

Paul Ekman states that there are six basic emotions universally expressed by humans in response to psychological triggers. For the neutral emotion, images that did not fall into any of the six basic emotions were collected. Figure 7 presents images showing the discrete emotional characterization of the neutral, happy, angry, and sad states during the driving tests.

After collecting and constructing the dataset, an in-depth analysis of motor activity was performed. Human behavior is complex and constantly changing, so it was essential to study drivers' behaviors across different emotional states while interacting with basic driving elements. Figure 8 illustrates the signals generated from steering wheel angle measurements in different emotional states, visually demonstrating variations in motor activity.

Windows of 50 records were analyzed and processed in steps of 5 from motor activity data to identify segments that provide relevant information for recognizing motor activity associated with the target emotional states. While various methods, techniques, algorithms, and transforms can perform these complex tasks, implementing feature selection through machine learning algorithms proved to be an effective tool for identifying key data related to certain emotions. Feature selection is essential for the successful application of machine learning and data mining algorithms in real-world scenarios. Numerous methods for relevant feature selection have been proposed in the literature, including RFE, which reduces irrelevant, redundant features, noisy data, and high dimensionality (Jeon and Oh, 2020).

The process involved applying the RFE technique with different machine learning algorithms using the motor activity dataset. Eighty percent of the data were used for training, and the remaining 20% for blind testing. Table 2 shows the performance of the models in terms of accuracy, which evaluates how often the model correctly predicts outcomes (Pacurari et al., 2023). Precision was also used, measuring how often the model correctly predicts true positives relative to the total number of positive predictions made (Imani and Arabnia, 2023). Additionally, recall (or sensitivity) was employed to measure how often the model correctly identifies true positives from all positive samples in the dataset. Lastly, the F1-Score, a metric derived from precision and recall, was used to assess the models' performance. F1-scores range from 0 to 1, with higher values indicating better model performance (Imani and Arabnia, 2023).

Based on the results, the model generated by the Random Forest Classifier (RFC) algorithm achieved the best performance in identifying motor activity segments that best characterize target emotional states, achieving 89% accuracy.

As a result of this procedure, 9 significant motor activity features were identified: 5 related to steering wheel angle, 0 related to brake movement, and 4 related to throttle movement. By identifying key signal segments that improve the final emotion recognition model, driving behavior could be distinguished across different emotional states. For example, the behavior related to

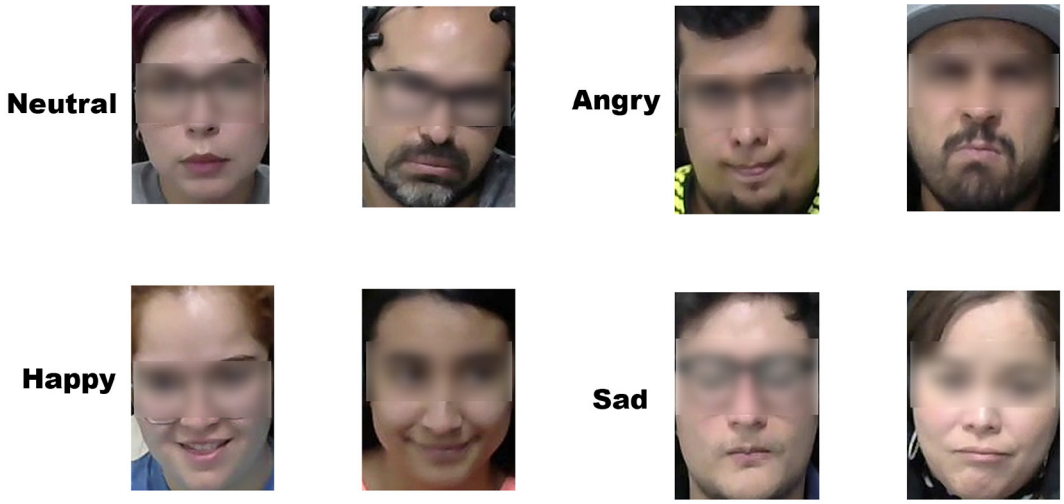

FIGURE 7
Sample images of the four emotion categories of the generated dataset.

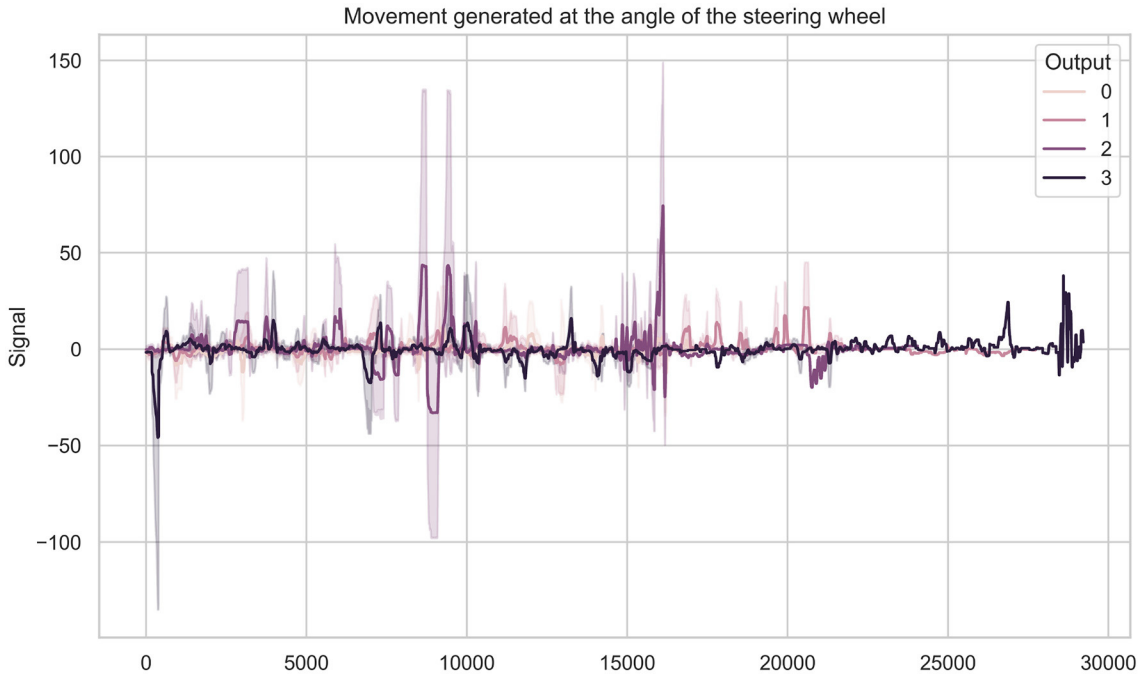

FIGURE 8
Steering wheel angle signal in different emotional states.Where each numerical value represents an emotion 0 for neutral, 1 for happy, 2 for angry, and 3 for sad.

steering wheel movement was significantly more intense during anger than for other emotions. Similarly, the neutral emotion also showed greater steering wheel movement, though not as intense as the anger emotion.

Although accelerator pedal movement signals seemed to fluctuate within similar ranges, there were notable differences between emotions. In particular, anger showed the highest levels of accelerator pedal engagement, indicating that participants pressed harder on the pedal when feeling angry.

The CNN architectures, VGG16 and VGG19, demonstrated the highest potential in recognizing emotions from the 3,361 images of drivers in various emotional states, achieving accuracy rates of 98% and 99%, respectively. These models generated probability vectors between 0 and 1 for each image, which were subsequently integrated with the engine dataset that had been correctly classified using the RFE algorithm. In the case of the Inception V3 network, promising results were also obtained, with an accuracy of 97.2%. However, the accuracy and loss plots for

TABLE 2 Results obtained from emotion classification models using different learning algorithms and motor activity as data source.

Algorithm	Emotion	Precision	Recall	F1-Score
RFC	Neutral	0.91	0.90	0.90
	Happy	0.88	0.90	0.89
	Angry	0.89	0.88	0.88
	Sad	0.89	0.88	0.88
Accuracy				0.89
DT	Neutral	0.80	0.82	0.81
	Happy	0.80	0.79	0.80
	Angry	0.79	0.79	0.79
	Sad	0.77	0.76	0.76
Accuracy				0.79
ABC	Neutral	0.42	0.37	0.40
	Happy	0.42	0.50	0.46
	Angry	0.57	0.52	0.54
	Sad	0.44	0.40	0.42
Accuracy				0.46
LDA	Neutral	0.20	0.08	0.11
	Happy	0.33	0.83	0.47
	Angry	0.44	0.14	0.21
	Sad	0.26	0.02	0.03
Accuracy				0.32

The bold values represents the overall accuracy of each of the implemented machine learning algorithms.

Inception V3 did not display as stable behavior as those for the VGG networks. According to Pathar et al. (2019), for a model to be validated satisfactorily–in this case, emotion recognition–the validation loss should be similar to or greater than the training loss. If the validation loss is lower than the training loss, the model may be underfitted and should be trained for more epochs. In the models generated using VGG16 and VGG19, this condition was satisfied, as the validation loss followed a similar pattern to the training loss. Finally, EfficientNet achieved an accuracy of only 43.5%, leading to its early dismissal from further consideration.

After identifying the most relevant motor activity features and reducing redundancy and dimensionality through feature selection, as well as extracting probability vectors for facial expressions using deep learning, a mid-fusion technique was implemented. This approach was necessary due to the distinct nature of the data: motor signals (such as steering wheel angle, pedal movement, and braking data) are inherently time series and require specialized feature extraction or vectorization methods, whereas facial data, typically extracted from images or video frames, require CNNs to identify relevant geometric patterns associated with emotions. These two data types differ significantly in their structures and processing requirements. Mid-fusion enables each modality to be processed independently, using feature extraction techniques specifically tailored to the characteristics of each data type, preserving the unique features of each modality prior to fusion (Hassani et al., 2024).

TABLE 3 1D-CNN architecture hyperparameters for training.

Hyperparameter	1D-CNN
Input	1 X 13
Activation functions	softplus
Epochs	500
Optimizer	Adam
Convolutional layers	5
Kernels	64,128,256,512,1024
Loss	sparse categorical crossentropy
Output function	softmax
Number of classes	4
Dense layers	4 of 512,256,128,64 neurons
Batch size	32

In contrast, early fusion involves combining raw data from different modalities at the initial stage. While this approach can integrate data quickly, it may result in information overload and make it difficult for the model to extract meaningful features from each modality. This is especially true when raw data from different formats (e.g., image data vs. time-series data) are combined. Early fusion often requires significant preprocessing to handle discrepancies between modalities, potentially diminishing the richness of features that can be learned (Gadzicki et al., 2020).

On the other hand, mid-fusion allows for independent feature extraction from each modality (e.g., probability vectors for facial emotions and key motor activity features). This preserves the unique characteristics of each modality and facilitates a more meaningful integration of information at a higher level of abstraction, where patterns between the modalities can be more effectively recognized.

Late fusion, in which decisions are made independently for each modality and then combined, may fail to capture cross-modal interactions, as each modality is treated in isolation before fusion (Boulahia et al., 2021). By fusing at an intermediate stage (mid-fusion), the model can capture relationships between motor activity and facial expressions, leading to a more nuanced and effective emotion recognition system.

In this approach, facial data were processed using a CNN to extract geometric facial features, while motor activity data were analyzed using feature selection algorithms (such as random forests, decision trees, etc.). This fusion methodology ensures that the most relevant features from each domain are incorporated into the final model for multimodal emotion recognition, which was built using a one-dimensional convolutional neural network (1D-CNN). The hyperparameters for this model are presented in Table 3.

With the architecture and hyperparameters of the one-dimensional convolutional neural network (1D-CNN) defined, we conducted a k-fold cross-validation process. This method involves dividing the dataset into k subsets, as described by Wong and Yeh

(2020). Each subset takes a turn serving as test data, while the remaining subsets are used for training. The validation process runs for k iterations, corresponding to the number of folds, so that each fold is used as test data exactly once. To determine the model's overall performance, we calculate the arithmetic mean of the results from all iterations. In this study, we selected k = 5, which is a common choice for k-fold validation as it balances computational efficiency and robust model evaluation. The k-fold validation method is widely regarded as a robust way to assess the effectiveness of classification models, as it partitions the dataset and treats each group as an independent validation set (Wong and Yeh, 2020). The results of the k-fold cross-validation are presented in Table 4.

A confusion matrix is a tabular representation of the actual labels versus the model's predictions, providing insight into the performance of the model. Each row of the confusion matrix represents the instances predicted to belong to a specific class, while each column represents the actual instances from the dataset. This matrix serves as the foundation for calculating other key performance metrics (Heydarian et al., 2022). Figure 9 shows the confusion matrices generated for classifying the target emotions identified in this research. Of the 361 integrated motor activity datasets combined with probability vectors from facial expression recognition, the 1D-CNN model correctly identified 363 instances of neutral emotion, 1,036 of happiness, 1,445 of anger, and 385 of sadness.

Figure 10 illustrates the network's performance in processing multimodal data over 500 epochs for each fold using the 1D-CNN. As seen in the training process, the accuracy and loss curves were consistently aligned, indicating stable model performance across epochs. This confirms that overfitting was not an issue, affirming the network's ability to generalize in recognizing emotions across the dataset.

Finally, an analysis of the computational complexity of the model was carried out. Since the model is intended for real-time inference in automotive systems, it is crucial to address memory usage, processing time and scalability. As a result, the following results were obtained 383.87 MB of memory usage, with an average processing time of 0.50 s. For real-time models, it is important to see how the model responds to multiple simultaneous requests. Simulations of 100 concurrent inference requests were performed to measure how the response time changes, with an average of 32.37 s.

The results obtained in this section are highly significant, demonstrating the model's ability to accurately identify four universal emotions (neutral, happiness, anger and sadness) in the discrete model, as well as three emotions in the continuous model, reaching a statistically significant accuracy of over 90%. This innovative methodology represents a valuable advance in the field of affective computing in driving environments.

However, it is important to note that the inference time of the model needs improvement. Accurately recognizing emotions in real time is crucial because emotions can be brief and subject to rapid change. Optimizing the inference time would improve the system's ability to respond effectively to these momentary emotional changes.

TABLE 4 Classification results per k-fold for multimodal emotion recognition.

K	Emotion	Precision	Recall	F1-score
1	Neutral	1.00	1.00	1.00
	Happy	1.00	1.00	1.00
	Angry	0.98	0.99	0.98
	Sad	0.96	0.95	0.95
Accuracy				0.99
2	Neutral	1.00	1.00	1.00
	Happy	1.00	1.00	1.00
	Angry	1.00	0.99	0.99
	Sad	0.97	1.00	0.98
Accuracy				1.00
3	Neutral	1.00	1.00	1.00
	Happy	1.00	1.00	1.00
	Angry	0.99	0.98	0.98
	Sad	0.94	0.96	0.95
Accuracy				0.99
4	Neutral	1.00	1.00	1.00
	Happy	1.00	1.00	1.00
	Angry	0.75	1.00	0.85
	Sad	0.00	0.00	0.00
Accuracy				0.85
5	Neutral	1.00	0.96	0.98
	Happy	1.00	1.00	1.00
	Angry	1.00	0.98	0.99
	Sad	0.92	0.99	0.95
Accuracy				0.99
Overall classification report	Neutral	1.00	0.99	1.00
	Happy	1.00	1.00	1.00
	Angry	0.93	0.99	0.96
	Sad	0.95	0.78	0.85
Overall accuracy				0.96

The bold values represents the accuracy obtained in each K-fold.

5 Discussion

There are several studies with promising approaches to emotion recognition in drivers that have achieved statistically significant results in terms of accuracy. However, studies focusing on multimodal artificial intelligence, which essentially involves processing and understanding information from different sources, have been less explored in driving environments. Table 5 presents a comparison between state-of-the-art methods and the proposed

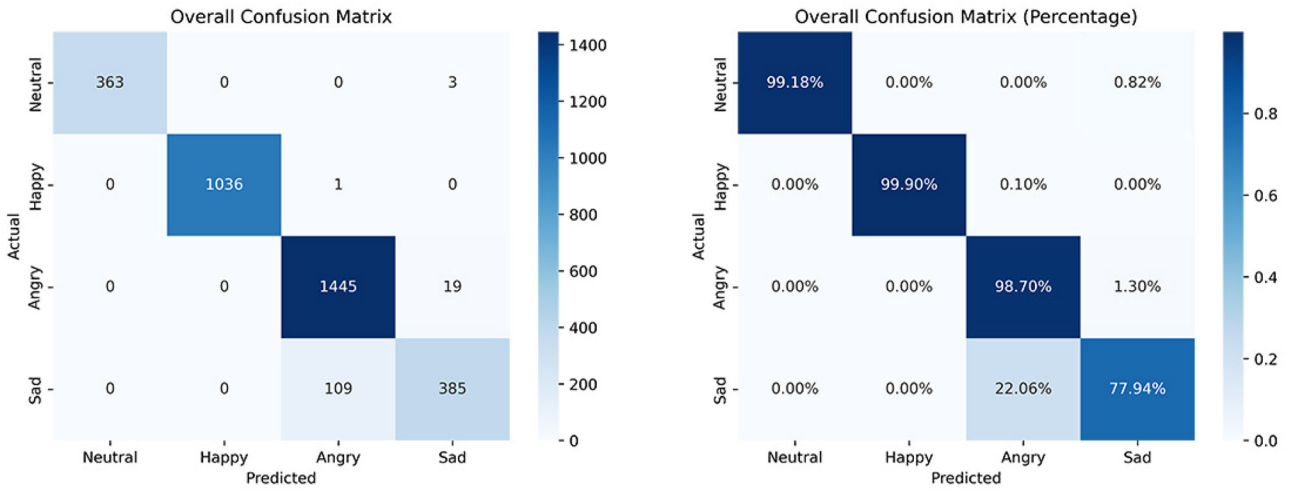

FIGURE 9
Confusion matrices resulting from the multimodal emotion recognition model. On the left side is shown the matrix with the number of correctly classified observations and on the right side the percentage of correctly classified observations.

methodology, examining the data sources, algorithms used, and the performance achieved.

Training CNNs with images has yielded significant results for driver emotion recognition, as shown in the table. However, as mentioned in the problem statement, accurately identifying emotions through constant monitoring of drivers presents limitations, primarily due to occlusion caused by various factors. Current research, such as that by Mou et al. (2023), combines large datasets related to both the driver and the surrounding environment. This research demonstrated that using data from the driver's eyes, vehicle data, and the environment across different scenarios to generate recognized emotions can be processed through a ConvLSTM network with an accuracy exceeding 80%. However, the model still relies heavily on accurate visual data from drivers in real time to achieve these levels of accuracy, which brings back the same issues found in methodologies based on video capture devices.

This limitation highlights the importance of multimodal modeling approaches, which address such problems by incorporating alternative data sources that either relieve or complement image analysis systems. These models have also proven to be highly relevant in the field of driver emotion recognition, demonstrating efficient performance.

The work presented by Shafaei et al. (2019) demonstrates the feasibility of classifying emotions using vehicle parameters generated by 16 participants, along with facial expressions from datasets such as CK+ and JAFFE, using SVM as a classifier. However, the limited potential of these traditional machine learning techniques may impede their real-world application. Despite this, their findings align with this study in a critical way: drivers tend to exhibit more active and abrupt behaviors when they are angry, happy, or excited. Conversely, their driving becomes more passive with fewer eye or body movements when they are tired or sad. These behavioral patterns correlate with emotional categories in the continuous model. Although the study is relevant for emotion identification, it becomes somewhat outdated due to the lack of more sophisticated AI algorithms like CNNs, which are superior tools for creating models that can be implemented in real-world settings.

Similarly, Du et al. (2021) proposes identifying emotions through heart rate and facial features from 16 volunteers in a simulated environment, yielding promising results. However, current devices for acquiring cardiac data can be invasive for drivers, leading to poor ergonomics and frequent emotional disturbances. Despite these limitations, their results demonstrate the ability of their methodology to recognize the emotions studied using convolutional neural networks (CNNs). Although the study lacks concrete details on the emotion induction process, it is inferred that participants assumed various emotional roles and behaviors. This suggests a subjective emotion labeling process, which limits the methodology's applicability to other scenarios. Moreover, the study did not base its work on an emotion theory, making its recognition approach less objective and precise than this research.

In contrast, Oh et al. (2021) developed a multimodal model combining facial images and dermal activity data from 13 volunteers (six men and seven women). This less invasive method could potentially be integrated into vehicle manufacturing. However, its performance is suboptimal compared to the results of this study. One key difference is the emotion induction techniques used. The authors applied methods such as movie watching and passage writing, which are effective but limited to generating emotions during the activity itself. In contrast, autobiographical memory techniques–used in this study–work well because participants recall emotional moments from their lives while driving. Additionally, the study's sampling frequency (10 Hz) was higher than the 5 Hz used in this research. The lower frequency, combined with AI-based selection of samples from the continuous motor activity signal, reduced noise and significantly improved recognition model performance.

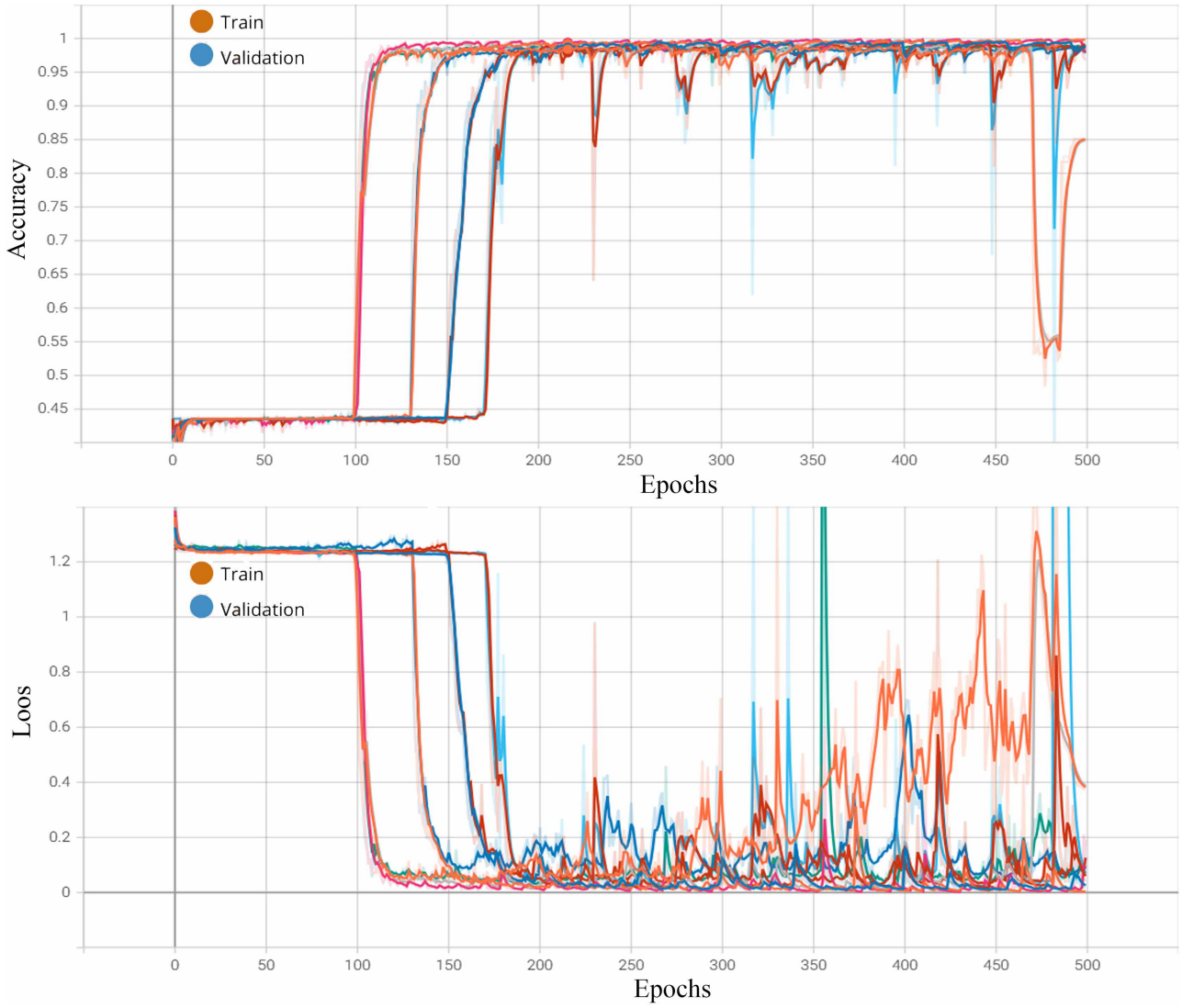

FIGURE 10
Loss and accuracy curves over 500 epochs during the training and validation of the multimodal emotion recognition model.

Among the studies closely related to this research that include data beyond facial geometric changes, Hieida et al. (2023) fused several physiological signals from drivers using Sparse Logistic Regression (SLR) on multimodal data to recognize negative emotions. They achieved a 74.0% area under the curve (AUC) for successful emotion classification, integrating the results into an ADAS system for greater driver comfort. However, their proposal's performance is still much lower compared to other studies, including this one. The use of SLR, while simpler and faster with lower computational requirements, is less effective for this type of data compared to more complex algorithms like CNNs. CNNs generally provide higher performance due to their ability to handle multimodal signals and automatically extract relevant features. Additionally, the ability of motor activity alone to recognize different emotional patterns has been demonstrated, further validating this study's findings.

The current proposal by Hu et al. (2024) demonstrated the potential of facial videos and driver behavior (brake pedal force, Y-axis position, and vehicle Z-axis position) as inputs in a multitask training approach. However, not considering other elements of this second data source significantly affects model performance, an important aspect that the present study does consider to improve accuracy in emotion identification. Also, the process that is carried out is more complex compared to the one presented, which could result in a longer inference time in real time.

Despite the various approaches discussed, most of the efforts have focused on image-based models that utilize deep learning algorithms like CNNs for emotion recognition. However, Vision Transformers (ViTs) are gaining recognition as an effective alternative to CNNs for various vision tasks, as they have been shown to be more robust against image distortions. ViTs take a different approach by exploring topological relationships between image patches, allowing them to capture more global and far-reaching connections, although they require more data-intensive training (Dai et al., 2021). ViT performance also relies heavily on factors like optimizer selection, dataset-specific hyperparameters,

TABLE 5 State-of-the-art, data source, algorithms and accuracy for emotion recognition.

Author	Data source	Algorithm	Accuracy
Lee et al. (2018)	Facial Geometry	CNN	99.95%
Verma and Choudhary (2018a)	Facial Geometry	CNN	98.80%
Patil and Veni (2019)	Facial geometry	SVM	86.70%
Shafaei et al. (2019)	Multimodal	SVM	94.00%
Wang et al. (2020)	ECG	Artificial neural network	91.11%
Cui et al. (2020)	Facial geometry	CNN	83.10%
Naqvi et al. (2020)	Facial geometry	CNN	98.93%
Du et al. (2021)	Multimodal	CNN	84.32%
Oh et al. (2021)	Multimodal	CNN	86.80%
Xiao et al. (2022)	Facial geometry	CNN	97.20%
Zaman et al. (2022)	Facial geometry	CNN	97.91%
Sukhavasi et al. (2022)	Facial geometry	CNN and SVM	94.09%
Mou et al. (2023)	Mutimodal	CNN	85.34%
Hieida et al. (2023)	Mutimodal	Sparse logistic regression	67.00%
Hu et al. (2024)	Mutimodal	Multitask learning network	67.92%
Our method (2023)	**Mutimodal**	**CNN**	**96.00%**

The bold values represents the overall accuracy obtained from the multimodal emotion recognition model in drivers proposed in this research.

and network depth, more so than CNNs. Preprocessing with overlapping convolutional filters of smaller size (stride < size) has been shown to contribute to performance and stability (Xiao et al., 2021).

CNNs, in contrast, deliver outstanding results even with relatively smaller datasets, compared to the larger datasets required by ViTs. This performance difference is largely attributed to the different inductive biases inherent to each architecture. CNNs' filter-based structure allows for quick identification of specific image features, but this same architecture limits their ability to capture more complex global relationships (Raghu et al., 2021).

This study leverages all the intrinsic properties of CNNs to address the challenges identified in the literature. By focusing on data related to drivers' interactions with vehicle elements, primarily the steering wheel, accelerator, and brake, this study demonstrates superior performance compared to state-of-the-art multimodal models that rely on motor activity data. The methodology presented here achieves higher performance due to the factors discussed throughout this section.

6 Conclusions

In conclusion, this study aimed to contribute with a new and innovative model for emotion recognition in drivers through motor activity data and facial expressions, implementing deep learning techniques such as CNNs, where it was possible to obtain an accuracy of 98.0% by giving equal weight to motor information and facial geometry data using feature selection algorithms to avoid outliers, redundancy and performance loss. These results demonstrate a high viability of the model for implementation in real environments, in addition to filling some gaps found in current studies based on cameras and driver behavior, where the performance of the presented proposal competes with any model developed to date.

A significant contribution was also made as a first approximation to the technique of augmented autobiographical memory as a method of inducing emotions in drivers, offering precise figures of success of the process in a specific experimental environment, guiding researchers to explore other options for inducing emotions in simulated driving environments.

It is essential to mention that this type of proposals could potentially help reduce the number of accidents related to negative emotions, since these emotions are among the main psychological factors that can influence driving behavior (Maldonado et al., 2020; Šeibokait et al., 2017). This study also concluded that different emotional states effectively cause a change in driving behavior and style. This allows us to identify emotions through the interaction we have with the vehicle. In addition, this type of proposals could improve the user experience, provided that vehicle systems know the emotions of drivers in real time, giving rise to more appropriate and personalized systems for each individual.

On the other hand, the study of motor activity is not limited to the automotive industry, but could also be extended to everyday life, where various emotions could be identified from data generated by mobile devices such as smartphones, smart watches and other smart devices for daily use. Although there are still different sources of information and emotions related to them, this study represents an important contribution as a first approximation to the recognition of emotions in drivers, with the aim of improving the quality and safety of transport systems.

6.1 Future work

As future work, it is proposed to deepen in different methodologies to neutralize and induce emotions in drivers, since one of the major limitations of the present study is the low level of participants who were successfully induced an emotion through the SAM tests, carried out in experimental simulated driving environments. Nevertheless, these results offer guidance, cautions and limitations of the technique in case we wish to implement it in future research. In addition, it is suggested to expand the group of participants to ensure a broader spectrum of emotions and related motor activity and facial geometry data. It is also proposed to explore other scenarios such as drowsiness, drunkenness and distraction, where this model could potentially be applied, in order to mitigate traffic accidents. Even the validation process could

be extended, particularly with testing under different conditions. This extension will improve the robustness and applicability of the emotion recognition model. Nevertheless, the present research establishes a first approach for real-time emotion recognition.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Universidad Autónoma de Zacatecas. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CE-S: Conceptualization, Data curation, Formal analysis, Investigation, Software, Writing – original draft. HL-G: Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing. JC-P: Conceptualization, Methodology, Resources, Writing – review & editing. CB-H: Project administration, Writing – original draft. NG: Project administration, Supervision, Validation, Writing – original draft. DR: Supervision, Writing – review & editing. KV-C: Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We want to express our sincere gratitude to the federal institution CONAHCYT for their invaluable support and funding in the development of our research project. Their backing has been essential for the success and accomplishment of our academic and scientific goals. We are deeply thankful for their commitment to excellence in research and innovation in our country.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, F., Hariharan, U., Muthukumaran, N., Ali, A., and Sharma, S. (2024). Emotion recognition of the driver based on KLT algorithm and ShuffleNet V2. *Signal, Image Video Proc.* 18, 3643–3660. doi: 10.1007/s11760-024-03029-z

Al-Asadi, A. W., Salehpour, P., and Aghdasi, H. S. (2024). A robust semi-supervised deep learning approach for emotion recognition using EEG signals. *Int. J. Mach. Learn. Cybernet.* 15, 4445–4458. doi: 10.1007/s13042-024-02158-8

AlBadawy, E. A., and Kim, Y. (2018). "Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches," in *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction* (Boulder, CO: ACM Digital Library).

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., and Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Appl. Sci.* 13:4750. doi: 10.3390/app13084750

Amiri, A. F., Oudira, H., Chouder, A., and Kichou, S. (2024). Faults detection and diagnosis of pv systems based on machine learning approach using random forest classifier. *Energy Convers. Managem.* 301:118076. doi: 10.1016/j.enconman.2024.118076

Andreu-Perez, A. R., Kiani, M., Andreu-Perez, J., Reddy, P., Andreu-Abela, J., Pinto, M., et al. (2021). Single-trial recognition of video gamer expertise from brain haemodynamic and facial emotion responses. *Brain Sci.* 11:1. doi: 10.3390/brainsci11010106

Arslan, E. E., Akahin, M. F., Yilmaz, M., and Ilgn, H. E. (2024). Towards emotionally intelligent virtual environments: classifying emotions through a biosignal-based approach. *Appl. Sci.* 14:8769. doi: 10.3390/app14198769

Ba, F., Peng, P., Zhang, Y., and Zhao, Y. (2023). Classification and identification of contaminants in recyclable containers based on a recursive feature elimination-light gradient boosting machine algorithm using an electronic nose. *Micromachines* 14:2047. doi: 10.3390/mi14112047

Bakariya, B., Singh, A., Singh, H., Raju, P., Rajpoot, R., and Mohbey, K. K. (2024). Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. *Evol. Syst.* 15, 641–658. doi: 10.1007/s12530-023-09506-z

Bansal, M., Kumar, M., Sachdeva, M., and Mittal, A. (2023). Transfer learning for image classification using VGG19: CALTECH-101 image data set. *J. Ambient Intell. Humaniz. Comput.* 14, 3609–3620. doi: 10.1007/s12652-021-03488-z

Barsoum, E., Zhang, C., Ferrer, C., and Zhang, Z. (2016). *Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution* (Tokyo: ACM Digital Library), 279–283.

Bhangale, K., and Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics* 12:839. doi: 10.3390/electronics12040839

Boulahia, S. Y., Amamra, A., Madi, M. R., and Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vis. Appl.* 32:6. doi: 10.1007/s00138-021-01249-8

Braun, M., Weiser, S., Pfleging, B., and Alt, F. (2018). "A comparison of emotion elicitation methods for affective driving studies," in *Adjunct Proceedings - 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2018* (Toronto, ON: ACM Digital Library), 77–81.

Cai, M., Chen, J., Hua, C., Wen, G., and Fu, R. (2024). Eeg emotion recognition using EEG-SWTNS neural network through eeg spectral image. *Inf. Sci*. 680:121198. doi: 10.1016/j.ins.2024.121198

Cai, Y., Li, X., and Li, J. (2023). Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. *Sensors* 23:2455. doi: 10.3390/s23052455

Cao, J., Yan, M., Jia, Y., Tian, X., and Zhang, Z. (2021). Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals. *EURASIP J. Adv. Signal Process.* 2021:1. doi: 10.1186/s13634-021-00740-8

Celaya-Padilla, J. M., Romero-Gonzlez, J. S., Galvan-Tejada, C. E., Galvan-Tejada, J. I., Luna-Garca, H., Arceo-Olague, J. G., et al. (2021). In-vehicle alcohol detection using low-cost sensors and genetic algorithms to aid in the drinking and driving detection. *Sensors* 21:7752. doi: 10.3390/s21227752

Chaithanya, B. N., Swasthika Jain, T. J., Usha Ruby, A., and Parveen, A. (2021). An approach to categorize chest X-ray images using sparse categorical cross entropy. *Indon. J. Elect. Eng. Comp. Sci.* 24:3. doi: 10.11591/ijeecs.v24.i3.pp1700-1710

Chen, J., Lin, X., Ma, W., Wang, Y., and Tang, W. (2024). EEG-based emotion recognition for road accidents in a simulated driving environment. *Biomed. Signal Process. Control* 87:105411. doi: 10.1016/j.bspc.2023.105411

Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human-computer interaction applications. *Neural Comp. Appl.* 35, 23311–23328. doi: 10.1007/s00521-021-06012-8

Costa, V. G., and Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. *Artif. Intellig. Rev*. 56:5. doi: 10.1007/s10462-022-10275-5

Cui, Y., Ma, Y., Li, W., Bian, N., Li, G., and Cao, D. (2020). Multi-EmoNet: a novel multi-task neural network for driver emotion recognition. *IFAC-PapersOnLine* 53, 650–655. doi: 10.1016/j.ifacol.2021.04.155

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). "CoAtNet: marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems*, 5.

Daqrouq, K., Balamesh, A., Alrusaini, O., Alkhateeb, A., and Balamash, A. (2024). Emotion modeling in speech signals: Discrete wavelet transform and machine learning tools for emotion recognition system. *Appl. Comp. Intellig. Soft Comp*. 2024, 1–12. doi: 10.1155/2024/7184018

Davoli, L., Martalò, M., Cilfone, A., Belli, L., Ferrari, G., Presta, R., et al. (2020). On driver behavior recognition for increased safety: a roadmap. *Safety* 6:55. doi: 10.3390/safety6040055

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., et al. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proc. Nat. Acad. Sci*. 113, 2636–2641. doi: 10.1073/pnas.1513271113

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning, volume 78 of Proceedings of Machine Learning Research*, eds. S. Levine, V. Vanhoucke, and K. Goldberg (New York: PMLR), 1–16.

Dozio, N., Bertoni, M., and Ferrise, F. (2024). Driving emotions: using virtual reality to explore the effect of low and high arousal on driver's attention. *Virtual Real*. 28:1. doi: 10.1007/s10055-024-00950-z

Du, G., Wang, Z., Gao, B., Mumtaz, S., Abualnaja, K. M., and Du, C. (2021). A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Trans. Intellig. Transp. Syst*. 22, 4570–4578. doi: 10.1109/TITS.2020.3007357

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. (2000). "Incorporating second-order functional knowledge for better option pricing," in *Advances in Neural Information Processing Systems*, eds. T. Leen, T. Dietterich, and V. Tresp (Cambridge, MA: MIT Press).

Dunn, J., Runge, R., and Snyder, M. (2018). *Wearables and the Medical Revolution, Vol. 15*. Oxfordshire: Taylor & Francis, 429–448. doi: 10.2217/pme-2018-0044

Fang, A., Pan, F., Yu, W., Yang, L., and He, P. (2024). Ecg-based emotion recognition using random convolutional kernel method. *Biomed. Signal Process. Control* 91:105907. doi: 10.1016/j.bspc.2023.105907

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. (2020). "Early vs late fusion in multimodal convolutional neural networks," in *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020* (Rustenburg: IEEE).

Gite, S., Mane, D., Doshi, I., Mohite, P., Upadhyay, B., and Joshi, S. (2023). Real-time driver sentiment analysis using hybrid deep learning algorithm. *Int. J. Intellig. Syst. Appl. Eng.* 12, 377–389. doi: 10.1016/j.aej.2022.12.009

Goeleven, E., De Raedt, R., Leyman, L., and Verschuere, B. (2008). The Karolinska directed emotional faces: a validation study. *Cogn. Emot*. 22:6. doi: 10.1080/02699930701626582

Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A., and Lanata, A. (2024). Facial emotion recognition (FER) through custom lightweight cnn model: Performance evaluation in public datasets. *IEEE Access* 12, 45543–45559. doi: 10.1109/ACCESS.2024.3380847

Hassani, S., Dackermann, U., Mousavi, M., and Li, J. (2024). A systematic review of data fusion techniques for optimized structural health monitoring. *Inform. Fusion* 103:102136. doi: 10.1016/j.inffus.2023.102136

Heydarian, M., Doyle, T. E., and Samavi, R. (2022). Mlcm: Multi-label confusion matrix. *IEEE Access* 10, 19083–19095. doi: 10.1109/ACCESS.2022.3151048

Hieida, C., Yamamoto, T., Kubo, T., Yoshimoto, J., and Ikeda, K. (2023). Negative emotion recognition using multimodal physiological signals for advanced driver assistance systems. *Artif. Life Robot*. 28:2. doi: 10.1007/s10015-023-00858-y

Hu, C., Gu, S., Yang, M., Han, G., Lai, C. S., Gao, M., et al. (2024). *MDEmoNet: A Multimodal Driver Emotion Recognition Network for Smart Cockpit*. Las Vegas, NV: IEEE, 1–6.

Hu, J., and Li, Y. (2022). "Electrocardiograph based emotion recognition via wgan-gp data enhancement and improved cnn," in *Intelligent Robotics and Applications*, eds. H. Liu, Z. Yin, L. Liu, L. Jiang, G. Gu, X. Wu, and W. Ren (Cham. Springer International Publishing), 155-164.

Imani, M., and Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis. *Technologies* 11:167. doi: 10.3390/technologies11060167

Jain, D. K., Dutta, A. K., Verdú, E., Alsubai, S., and Sait, A. R. W. (2023). An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image Vis. Comput*. 133:104659. doi: 10.1016/j.imavis.2023.104659

Jeon, H., and Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Appl. Sci*. 10:9. doi: 10.3390/app10093211

Jha, S. K., Suvvari, S., and Kumar, M. (2024). Emotion recognition from electroencephalogram (EEG) signals using a multiple column convolutional neural network model. *SN Comp. Sci*. 5:2. doi: 10.1007/s42979-023-02543-0

Khan, M., Gueaieb, W., El Saddik, A., and Kwon, S. (2024). MSER: multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Syst. Appl.* 245:122946. doi: 10.1016/j.eswa.2023.122946

Khattak, A., Asghar, M. Z., Ali, M., and Batool, U. (2022). An efficient deep learning technique for facial emotion recognition. *Multimed. Tools Appl*. 81:2. doi: 10.1007/s11042-021-11298-w

Khoo, L. S., Lim, M. K., Chong, C. Y., and McNaney, R. (2024). Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors* 24:2. doi: 10.3390/s24020348

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors* 18:2. doi: 10.3390/s18020401

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comp*. 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kumar, A., Sangwan, S. R., Arora, A., and Menon, V. G. (2022). Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT. *Appl. Soft Comput*. 122. doi: 10.1016/j.asoc.2022.108863

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., and Vora, D. (2024). Transfer learning for emotion detection in conversational text: a hybrid deep learning approach with pre-trained embeddings. *Int. J. Inform. Technol*. doi: 10.1007/s41870-024-02027-1

Lee, H. S., and Kang, B. Y. (2020). *Continuous Emotion Estimation of Facial Expressions on JAFFE and CK+ Datasets for Human Robot Interaction*.

Lee, K. W., Yoon, H. S., Song, J. M., and Park, K. R. (2018). Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. *Sensors* 18, 1–22. doi: 10.3390/s18040957

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., et al. (2022). *EEG Based Emotion Recognition: A Tutorial and Review*. Hong Kong: ACM Computing Surveys.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Lin, W., and Li, C. (2023). Review of studies on emotion recognition and judgment based on physiological signals. *Appl. Sci*. 13:4. doi: 10.3390/app13042573

López-Cano, M. A., Navarro, B., Nieto, M., Andrés-Pretel, F., and Latorre, J. M. (2020). Autobiographical emotional induction in older people through popular songs: Effect of reminiscence bump and enculturation. *PLoS ONE* 15:e238434. doi: 10.1371/journal.pone.0238434

Lu, M., Hu, S., Mao, Z., Liang, P., Xin, S., and Guan, H. (2020). Research on work efficiency and light comfort based on eeg evaluation method. *Build. Environ*. 183:107122. doi: 10.1016/j.buildenv.2020.107122

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, 94–101 (San Francisco, CA: IEEE).

Maldonado, A., Torres, M., Catena, A., Cndido, A., and Megas-Robles, A. (2020). From riskier to safer driving decisions: the role of feedback and the experiential automatic processing system. *Transp. Res. Part F: Traffic Psychol. Behav.* 73, 307–317. doi: 10.1016/j.trf.2020.06.020

Marques, G., Agarwal, D., and de la Torre Díez, I. (2020). Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl. Soft Comp. J.* 96:106691. doi: 10.1016/j.asoc.2020.106691

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* 2:3. doi: 10.1007/s42452-020-2234-1

Mehrotra, M., Singh, K. P., and Singh, Y. B. (2024). "Facial emotion recognition and detection using convolutional neural networks with low computation cost," in *2024 2nd International Conference on Disruptive Technologies (ICDT)* (Greater Noida: IEEE), 1349–1354.

Mihalache, S., and Burileanu, D. (2021). Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition. *UPB Sci. Bull. Series C: Elect. Eng. Comp. Sci.* 83:4. doi: 10.48550/arXiv.2002.10061

Mishra, S. P., Warule, P., and Deb, S. (2024). Speech emotion recognition using MFCC-based entropy feature. *Signal, Image Video Proc.* 18, 153–161. doi: 10.1007/s11760-023-02716-7

Mocanu, B., Tapu, R., and Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image Vis. Comput.* 133:104676. doi: 10.1016/j.imavis.2023.104676

Modi, S., and Bohara, M. H. (2021). "Facial emotion recognition using convolution neural network," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1339–1344.

Mollahosseini, A., Hasani, B., and Mahoor, M. (2017). Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comp.* 10, 18–31. doi: 10.1109/TAFFC.2017.2740923

Mou, L., Zhao, Y., Zhou, C., Nakisa, B., Rastgoo, M. N., Ma, L., et al. (2023). Driver emotion recognition with a hybrid attentional multimodal fusion framework. *IEEE Trans. Affect. Comp.* 14, 2970–2981. doi: 10.1109/TAFFC.2023.3250460

Mustaqeem, S. M., and Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access* 8, 79861–79875. doi: 10.1109/ACCESS.2020.2990405

Naqvi, R. A., Arsalan, M., Rehman, A., Rehman, A. U., Loh, W. K., and Paul, A. (2020). Deep learning-based drivers emotion classification system in time series data for remote applications. *Remote Sens.* 12, 1–32. doi: 10.3390/rs12030587

Oh, G., Ryu, J., Jeong, E., Yang, J. H., Hwang, S., Lee, S., et al. (2021). Drer: Deep learning–based driver's real emotion recognizer. *Sensors* 21, 1–29. doi: 10.3390/s21062166

Pacurari, A. C., Bhattarai, S., Muhammad, A., Avram, C., Mederle, A. O., Rosca, O., et al. (2023). Diagnostic accuracy of machine learning ai architectures in detection and classification of lung cancer: A systematic review. *Diagnostics* 13:13. doi: 10.3390/diagnostics13132145

Paredes, P. E., Ordo nez, F., Ju, W., and Landay, J. A. (2018). "Fast & furious: detecting stress with a car steering wheel," in *Conference on Human Factors in Computing Systems - Proceedings* (Montreal, QC: ACM Digital Library).

Pathar, R., Adivarekar, A., Mishra, A., and Deshmukh, A. (2019). "Human emotion recognition using convolutional neural network in real time," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (Chennai: IEEE), 1–7.

Patil, M., and Veni, S. (2019). "Driver emotion recognition for enhancement of human machine interface in vehicles," in *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019* (Chennai: IEEE), 0420–0424.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). "Do vision transformers see like convolutional neural networks?," in *Advances in Neural Information Processing Systems* (ACM Digital Library), 15.

Ravikumar, K., Panigrahi, B., Sahoo, S., Reddy, A., Manchala, Y., and Swain, N. (2024). Cnn based face emotion recognition system for healthcare application. *EAI Endors. Trans. Pervas. Health Technol.* 10:5458. doi: 10.4108/eetpht.10.5458

Reveles-Gómez, L. C., Luna-García, H., Celaya-Padilla, J. M., Barría-Huidobro, C., Gamboa-Rosales, H., Solís-Robles, R., et al. (2023). Detection of pedestrians in reverse camera using multimodal convolutional neural networks. *Sensors* 23:17. doi: 10.3390/s23177559

Rezapour, M., and Ksaibati, K. (2022). Identification of factors associated with various types of impaired driving. *Humanit. Soc. Sci. Commun.* 9:1. doi: 10.1057/s41599-022-01041-7

Sahoo, G. K., Das, S. K., and Singh, P. (2023). Performance comparison of facial emotion recognition: a transfer learning-based driver assistance framework for in-vehicle applications. *Circuits, Systems, and Signal Proc.* 42:7. doi: 10.1007/s00034-023-02320-7

Sanghavi, H., Zhang, Y., and Jeon, M. (2020). "Effects of anger and display urgency on takeover performance in semi-automated vehicles," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '20* (New York, NY: Association for Computing Machinery), 48–56.

Sarvakar, K., Senkamalavalli, R., Raghavendra, S., Santosh Kumar, J., Manjunath, R., and Jaiswal, S. (2023). Facial emotion recognition using convolutional neural networks. *Mater. Today: Proc.* 80, 3560–3564. doi: 10.1016/j.matpr.2021.07.297

Schuetz, S., and Venkatesh, V. (2020). Research perspectives: the rise of human machines: how cognitive computing systems challenge assumptions of user-system interaction. *J. Assoc. Inform. Syst.* 21:2. doi: 10.17705/1jais.00608

Šeibokait, L., Endriulaitien, A., Sullman, M. J., Markšaityt, R., and Žardeckait-Matulaitien, K. (2017). Difficulties in emotion regulation and risky driving among Lithuanian drivers. *Traffic Inj. Prev.* 18:7. doi: 10.1080/15389588.2017.1315109

Semeraro, A., Vilella, S., and Ruffo, G. (2021). PyPlutchik: Visualising and comparing emotion-annotated corpora. *PLoS ONE* 16:9. doi: 10.1371/journal.pone.0256503

Shafaei, S., Hacizade, T., and Knoll, A. (2019). Integration of driver behavior into emotion recognition systems: a preliminary study on steering wheel and vehicle acceleration. *Lect. Notes Comp. Sci.* 11367, 386–409. doi: 10.1007/978-3-030-21074-8_32

Shahzad, H. M., Bhatti, S. M., Jaffar, A., Akram, S., Alhajlah, M., and Mahmood, A. (2023). Hybrid facial emotion recognition using cnn-based features. *Appl. Sci.* 13:9. doi: 10.3390/app13095572

Siam, A. I., Gamel, S. A., and Talaat, F. M. (2023). Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. *Neural Comp. Appl.* 35, 12891–12904. doi: 10.1007/s00521-023-08428-w

Steinhauser, K., Leist, F., Maier, K., Michel, V., Pärsch, N., Rigley, P., et al. (2018). Effects of emotions on driving behavior. *Transp. Res. Part F: Traffic Psychol. Behav.* 59:12. doi: 10.1016/j.trf.2018.08.012

Stephens, A. N., Collard, J., and Koppel, S. (2024). Don't sweat the small stuff; anger rumination and lack of forgiveness are related to aggressive driving behaviours. *Curr. Psychol.* 43:7. doi: 10.1007/s12144-023-04744-5

Sukhavasi, S. B., Sukhavasi, S. B., Elleithy, K., El-Sayed, A., and Elleithy, A. (2022). A hybrid model for driver emotion detection using feature fusion approach. *Int. J. Environ. Res. Public Health* 19:5. doi: 10.3390/ijerph19053085

Sweeney-Fanelli, T. C., and Imtiaz, M. H. (2024). Ecg-based automated emotion recognition using temporal convolution neural networks. *IEEE Sens. J.* 24, 29039–29046. doi: 10.1109/JSEN.2024.3434479

Talaat, F. M., Ali, Z. H., Mostafa, R. R., and El-Rashidy, N. (2024). Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. *Soft Computing* 28, 6695–6708. doi: 10.1007/s00500-023-09477-y

Tang, W., Long, G., Liu, L., Zhou, T., Blumenstein, M., and Jiang, J. (2020). Rethinking 1DCNN for time series classification: A stronger Baseline. *arXiv* [preprint] *arXiv:2002.10061*.

Tauqeer, M., Rubab, S., Khan, M. A., Naqvi, R. A., Javed, K., Alqahtani, A., et al. (2022). Drivers emotion and behavior classification system based on internet of things and deep learning for advanced driver assistance system (adas). *Comput. Commun.* 194, 258–267. doi: 10.1016/j.comcom.2022.07.031

Tokmak, F., Subasi, A., and Qaisar, S. M. (2024). "Chapter 2 - artificial intelligence-based emotion recognition using ecg signals," in *Applications of Artificial Intelligence in Healthcare and Biomedicine, Artificial Intelligence Applications in Healthcare and Medicine*, eds. A. Subasi (Cambridge, MA: Academic Press), 37–67.

Trujillo, L., Hernndez, D., Rodriguez, A., Monroy, O., and Villanueva, O. (2024). Effects of feature reduction on emotion recognition using eeg signals and machine learning. *Expert Syst.* 41:13577. doi: 10.1111/exsy.13577

Veeranki, Y. R., Kumar, H., Ganapathy, N., Natarajan, B., and Swaminathan, R. (2021). A systematic review of sensing and differentiating dichotomous emotional states using audio-visual stimuli. *IEEE Access* 9, 124434–124451. doi: 10.1109/ACCESS.2021.3110773

Verma, B., and Choudhary, A. (2018a). "A framework for driver emotion recognition using deep learning and grassmann manifolds," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* (Maui, HI: IEEE), 1421–1426. doi: 10.1109/ITSC.2018.8569461

Verma, B., and Choudhary, A. (2018b). "Deep learning based real-time driver emotion monitoring," in *2018 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2018* (Madrid: IEEE), 1–6. doi: 10.1109/ICVES.2018.8519595

Waheed Awan, A., Taj, I., Khalid, S., Muhammad Usman, S., Imran, A. S., and Usman Akram, M. (2024). Advancing emotional health assessments: a hybrid deep learning approach using physiological signals for robust emotion recognition. *IEEE Access* 12, 141890–141904. doi: 10.1109/ACCESS.2024.3463746

Wang, X., Guo, Y., Ban, J., Xu, Q., Bai, C., and Liu, S. (2020). Driver emotion recognition of multiple-ECG feature fusion based on BP network and D-S evidence. *IET Intellig. Transp. Syst.* 14, 815–824. doi: 10.1049/iet-its.2019.0499

Wawage, P., and Deshpande, Y. (2022). Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. *Int. J. Perf. Eng.* 18:791. doi: 10.23940/ijpe.22.11.p4.791797

WHO (2018). *WHO: Road Traffic Injuries*. Geneva: World Health Organization.

Wong, T. T., and Yeh, P. Y. (2020). Reliable accuracy estimates from k-fold cross validation. *IEEE Trans. Knowl. Data Eng.* 32:8. doi: 10.1109/TKDE.2019.2912815

Wu, M.-H., and Chang, T.-C. (2021). Evaluation of effect of music on human nervous system by heart rate variability analysis using ecg sensor. *Sensors Mater*. 33:739. doi: 10.18494/SAM.2021.3040

Xiao, H., Li, W., Zeng, G., Wu, Y., Xue, J., Zhang, J., et al. (2022). On-road driver emotion recognition using facial expression. *Appl. Sci*. 12:2. doi: 10.3390/app12020807

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. (2021). "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems* (ACM Digital Library), 36.

Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., et al. (2023). Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Trans. Affect. Comp*. 14, 1082–1097. doi: 10.1109/TAFFC.2021.3100868

Yang, Z., Li, Z., Zhou, S., Zhang, L., and Serikawa, S. (2024). Speech emotion recognition based on multi-feature speed rate and lstm. *Neurocomputing* 601:128177. doi: 10.1016/j.neucom.2024.128177

Yao, Z., Wang, Z., Liu, W., Liu, Y., and Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Commun*. 120:5. doi: 10.1016/j.specom.2020.03.005

Ying, N., Jiang, Y., Guo, C., Zhou, D., and Zhao, J. (2024). A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform. *IEEE Intern. Things J*. 33:1. doi: 10.1109/JIOT.2024.3363176

Zahara, L., Musa, P., Prasetyo, E., Karim, I., and Musa, S. (2020). *The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm Based Raspberry PI*, 1–9.

Zaman, K., Sun, Z., Shah, S. M., Shoaib, M., Pei, L., and Hussain, A. (2022). Driver emotions recognition based on improved faster R-CNN and neural architectural search network. *Symmet*. 14:4. doi: 10.3390/sym14040687

Zaman, K., Zhaoyun, S., Shah, B., Hussain, T., Shah, S. M., Ali, F., et al. (2023). A novel driver emotion recognition system based on deep ensemble classification. *Complex Intellig. Syst*. 9, 6927–6952. doi: 10.1007/s40747-023-01100-9

Zangeneh Soroush, M., Maghooli, K., Setarehdan, S. K., and Nasrabadi, A. M. (2020). Emotion recognition using eeg phase space dynamics and poincare intersections. *Biomed. Signal Process. Control* 59:101918. doi: 10.1016/j.bspc.2020.101918

Zepf, S., Hernandez, J., Schmitt, A., Minker, W., and Picard, R. W. (2020). *Driver Emotion Recognition for Intelligent Vehicles: A Survey*.

Zhang, J., Zhou, Y., and Liu, Y. (2020). EEG-based emotion recognition using an improved radial basis function neural network. *J. Ambient Intellig. Humaniz. Comp*. doi: 10.1007/s12652-020-02049-0

Zhang, Q., Qu, W., Ge, Y., Sun, X., and Zhang, K. (2020). The effect of the emotional state on driving performance in a simulated car-following task. *Transp. Res. Part F: Traffic Psychol. Behav*. 69, 349–361. doi: 10.1016/j.trf.2020.02.004

Zhao, L., Wang, Z., Zhang, G., and Gao, H. (2020). Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector. *Multimed. Tools Appl*. 79, 35–36. doi: 10.1007/s11042-020-09259-w

Zhou, D., Cheng, Y., Wen, L., Luo, H., and Liu, Y. (2023). Drivers comprehensive emotion recognition based on ham. *Sensors* 23:19. doi: 10.3390/s23198293

Zhu, F., Gao, J., Yang, J., and Ye, N. (2022). Neighborhood linear discriminant analysis. *Pattern Recognit*. 123:108422. doi: 10.1016/j.patcog.2021.108422

Zimasa, T., Jamson, S., and Henson, B. (2017). Are happy drivers safer drivers? Evidence from hazard response times and eye tracking data. *Transp. Res. Part F: Traffic Psychol. Behav*. 46, 14–23. doi: 10.1016/j.trf.2016.12.005

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8697–8710.