International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Deep Learning Application in Multimodal Data Based Emotion Recognition

Puneet Yadav, Umar Badr Shafeeque
Department of Computer Science & Engineering
Azad Institute of Engineering & Technology, Lucknow, India
yadavpuneet766@gmail.com

Abstract: This article introduces the MIST framework (Multimodal-Image-Speech-Text)—a novel, context-aware approach designed to enhance the accuracy, adaptability, and robustness of multimodal emotion recognition systems. MIST leverages the complementary strengths of different modalities, integrating them through a carefully designed architecture that adapts to varying contextual cues and environmental conditions. A key focus of this work is the enhancement of the Text Emotion Recognition (TER) component within the MIST framework. Various advanced machine learning models were rigorously evaluated, and a comparative analysis was conducted to assess their effectiveness. The analysis revealed that traditional machine learning techniques offered superior performance and consistency across multiple benchmark datasets for TER tasks. Their precision, reliability, and computational efficiency justified their integration into the MIST framework, ensuring dependable textual emotion classification. This work not only advances the current state of multimodal emotion recognition research but also bridges the gap between theoretical innovation and practical implementation. By offering a scalable and context-sensitive solution, the MIST framework represents a significant step toward building emotionally intelligent systems capable of meaningful human interaction in diverse, real-world scenarios.

Keywords: Text recognition, face recognition, audio recognition, CNN, Deep learning

I. INTRODUCTION

Human emotions are conveyed through a variety of modalities—facial expressions, vocal tone, body posture, gestures, and linguistic cues—making their accurate interpretation essential for building emotionally intelligent systems that can respond appropriately to users' affective states. Subtle emotional variations may be difficult to detect through one modality alone due to limitations such as occlusions in images, background noise in audio, or ambiguity in textual content [1], [2], [3]. Consequently, there has been a growing interest in multimodal emotion recognition, which integrates information from multiple sources—such as text, audio, images, and motion—to improve both accuracy and contextual understanding. By employing multimodal fusion techniques—at either the feature or decision level—these systems aim to model the complex, nonlinear relationships that characterize human emotional responses [4], [5], [6], [7]. Each modality carries its own noise and variability—such as speech distortion in noisy environments, facial occlusion, or ambiguous sentiments in text—which complicates accurate recognition [8], [9]. Moreover, unimodal systems often lack robustness, failing to capture the interplay between different emotional signals and leading to limited or inaccurate outcomes [10]. Advanced strategies are required to ensure that critical information from each modality is retained, aligned, and interpreted accurately, even in dynamic or noisy environments [11]. This articles makes several significant contributions to the field of multimodal emotion recognition, all centered around the design, development, and comprehensive evaluation of the MIST (Multimodal-Image-Speech-Text) framework. The primary contribution is the development of the MIST framework, which introduces a novel, context-aware approach to emotion recognition by integrating multiple modalities—facial expressions, vocal patterns, textual cues, and motion dynamics. This multimodal integration results in a robust and accurate system capable of recognizing emotions in real-world, complex, and noisy environments. Compared to traditional unimodal and baseline multimodal approaches, MIST demonstrates improved performance in terms of accuracy and reduced loss, particularly in scenarios involving incomplete or ambiguous data.

II. RELATED WORK

In the early stages of emotion recognition research, the focus was primarily on manually extracting relevant features from data that could reveal the subject's emotional state, a process requiring significant domain expertise. For example, in text emotion recognition, common features included word frequency, n-grams, and sentiment scores [12].

The importance of feature extraction in predictive modeling and supervised machine learning is comprehensively discussed in [13]. This source outlines the fundamentals of feature extraction, demonstrating how data can be represented using various feature types, such as binary, categorical, and continuous. It further reviews classical and advanced learning algorithms including Neural Networks (NN), tree classifiers, Support Vector Machines (SVMs), ensemble methods, and neuro-fuzzy systems.

In [14], the authors evaluate several feature selection and extraction methods for emotion recognition with traditional machine learning algorithms. Techniques such as Minimum Redundancy Maximum Relevance (MRMR), Sequential selection, Neighborhood Component Analysis (both projection and subspace variants), Relief-F, and Decision Trees (DT) were applied to a dataset comprising 17 feature extraction functions. Their findings indicate that Artificial Neural Networks (ANNs) achieved the highest accuracy. Specifically, the Sequential method combined with ANN attained a testing accuracy of 93.08% and a recognition

accuracy of 83.52%. Meanwhile, the Decision Tree method required fewer feature combinations while maintaining a similar testing accuracy of 93.05%.

A novel feature extraction method named Term weight–Inverse document weight (TwIdw) was introduced in [15]. Building upon the traditional Term frequency–Inverse document frequency (TfIdf) approach, TwIdw incorporates syntactic depth from dependency grammar to better capture the contextual relevance of terms for text classification. When evaluated against basic TfIdf using random forest and feedforward NN models on datasets containing real and fake news, TwIdw consistently outperformed TfIdf. Notably, the NN model with TwIdw showed a 3.9% increase in accuracy and improvements in precision, recall, and F1-score, demonstrating the method's effectiveness in enhancing text classification tasks.

A thorough investigation into feature extraction for emotion recognition from multichannel Electroencephalogram (EEG) recordings is presented in the field of physiological signal analysis by [16]. SVM and CART classifiers were used to analyze five feature sets—statistical characteristics, fractal dimension, Hjorth parameters, Higher Order Spectra (HOS), and wavelet analysis—across five publicly available datasets. The best accuracies for valence (85.06%) and arousal (84.55%) were obtained by the fractal dimension in conjunction with CART (FD-CART), demonstrating its usefulness for EEG-based emotion identification.

Similarly, [17] developed a facial expression recognition system leveraging statistical parameters extracted from a multilevel stationary wavelet transform (SM-SBWT). The approach included preprocessing images with adaptive histogram equalization and using the Viola-Jones algorithm for face detection. Features focused on mean and maximum local energy of wavelet subbands were extracted and classified using an SVM. This method demonstrated superior performance across multiple benchmark databases, such as JAFFE, CK+, FEED, SFEW, and RAF. For instance, it achieved an impressive 97.3% accuracy on the JAFFE dataset, underscoring its robustness in emotion recognition tasks [18-22].

Another notable technique presented in [23] involves a two-feature extraction approach for emotion identification from facial images, focusing on the mouth region and Maximally Stable Extremal Regions (MSER). The mouth region-based method classifies emotions based on calculated mouth area, whereas the MSER method extracts features via connected components fed into an ANN classifier. Tested on the JAFFE dataset, the mouth region method reached 86% accuracy, while the MSER method improved upon this with 89% accuracy. These results demonstrate the potential of these feature extraction techniques, especially MSER, in enhancing emotion recognition accuracy for emotions like happiness, sadness, surprise, and neutral. The study concludes that combining these methods with other classifiers could further improve performance and real-world applicability.

Machine learning algorithms form the backbone of emotion recognition systems, enabling accurate classification and prediction of emotional states. This section reviews key algorithms including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Logistic Regression, and Decision Trees. Their typical applications, strengths, limitations, and the integration with dimensionality reduction techniques such as PCA are discussed, providing foundational knowledge for more advanced deep learning approaches.

A recent study [24] offers a comprehensive evaluation of state-of-the-art Automated Human Emotion Recognition (AHER) methods based on machine learning, detailing various algorithms and their respective advantages and drawbacks.

Another important work [25] focuses on emotion recognition from EEG signals, comparing multiple machine learning models using the DEAP dataset. The EEG data were segmented into 15-second intervals, and PCA was applied to reduce dimensionality. Hyperparameter tuning was performed via grid search to optimize each model. Results showed that the combination of PCA with SVM yielded the best performance, achieving an F1-score of 84.73% and a recall of 98.01%. The study also revealed that different classifiers performed better for different emotional dimensions: SVM excelled in classifying arousal and liking, while KNN attained the highest accuracy for valence. Overall, segmenting the data and applying PCA significantly enhanced model effectiveness.

The same study further analyzed classifiers on EEG and peripheral signals from the DEAP dataset, testing SVM, KNN, LDA, Logistic Regression, and Decision Trees, both with and without PCA. Using grid search for hyperparameter tuning on a Spark cluster, the combination of PCA and SVM again delivered superior results (F1-score 84.73%, recall 98.01% for the 30–45 second segment). Performance varied across different time segments and emotional states, underscoring the importance of selecting appropriate models tailored to specific emotions and data characteristics. These findings confirm that classical machine learning methods, when properly tuned and combined with dimensionality reduction, can effectively recognize emotions from EEG data.

Furthermore, by looking at four crucial phases—data collection, preprocessing, feature extraction, and feature classification—[26] provides an overview of the development of EEG-based machine learning methods for emotion recognition.

III. METHODOLOGY

This section covers the methodology for emotion detection algorithm. The approach follows CNN as deep learning scheme with MELD data. MELD data supports contextual awareness primarily because it consists of multi-party conversations where each utterance is part of an ongoing dialogue rather than isolated sentences.

3.1. Convolutional Neural Network (CNN)

An artificial neural network specifically made to process data in the form of numerous arrays, such color photographs, is called a CNN (3D arrays: height × width × channels). CNNs are inspired by the biological visual cortex and excel at capturing spatial and temporal dependencies by applying relevant filters.

CNNs work by transforming the input image into feature maps through a series of layers, each performing a specific operation:

3.1.1 Convolutional Layer

The convolution operation applies a set of learnable filters (kernels) to the input image or previous layer's output.
Each filter slides (convolves) over the spatial dimensions of the input, computing a dot product between the filter weights and the input patch. Mathematically, for input I and filter K:
$S(i,j) = (I * K)(i,j) = \Sigma m\ \Sigma n\ I(i+m, j+n) \times K(m,n)$
where S is the feature map output.
Filters detect local features like edges, textures, or specific shapes.

3.1.2 Activation Function

Typically, the ReLU (Rectified Linear Unit) function is applied element-wise to add non-linearity:
$f(x) = \max(0, x)$
Non-linearity allows the network to model complex relationships.

3.1.3 Pooling Layer

Reduces spatial dimensions to downsample feature maps, reducing computational load and controlling overfitting.
Common types:
- Max Pooling: Takes the maximum value from each patch.
-Average Pooling: Takes the average value.
Example: 2×2 max pooling with stride 2 reduces dimension by 4.

3.1.4 Fully Connected (Dense) Layer

The finished feature maps are flattened into a 1D vector following the convolutional and pooling layers. Layers that are fully connected carry out sophisticated categorization reasoning. Like typical neural networks, each neuron is connected to every other neuron in the layer above.

3.1.5 Output Layer

Produces final predictions. Uses softmax activation for multi-class classification:
$P(y = c \mid x) = e^{(Zc)} / \Sigma\_k\ e^{(Zk)}$
where Zc is the output score for class c.

3.2 DATASET USED

The Multimodal EmotionLines Dataset (MELD) is a comprehensive benchmark dataset designed for the task of emotion recognition in conversations, particularly in multimodal settings involving textual, acoustic, and visual information. It is an enriched version of the original EmotionLines dataset, which only included textual data. MELD extends this by incorporating corresponding audio and visual modalities extracted from the television series Friends, making it one of the first publicly available datasets to provide multimodal data for emotion classification within multi-party dialogues. MELD consists of over 13,000 utterances spanning more than 1,400 dialogues, where each utterance is associated with a specific speaker, timestamp, and emotion label.

3.3.1 TEXT EMOTION RECOGNITION

The steps followed in algorithm are given below:
Pretrained Model: Selection Start with a pretrained BERT model such as bert-base-uncased or an emotion-specific variant like BERTweet, RoBERTa, or DistilBERT.
Tokenization: The input text is tokenized using BERT's WordPiece tokenizer, which splits words into subword units and adds special tokens like [CLS] (start of input) and [SEP] (end of input).
Input Representation The tokenized text is converted into embeddings (token, segment, and position embeddings) and fed into the BERT model.
Fine-tuning Layer The output embedding corresponding to the [CLS] token (which represents the entire input sequence) is passed through a fully connected classification layer to predict the emotion label.
Training The model is trained using a cross-entropy loss function, with labeled data from emotion datasets such as MELD, GoEmotions, ISEAR, or EmotionLines.
Prediction Once trained, the model can classify new text inputs into emotion categories based on the learned representations.

3.3.2 Speech Emotion Recognition

To incorporate contextual awareness, the system can be designed to process sequences of speech segments rather than isolated utterances. This can be done by:
1.Sliding window approach: Feed a series of consecutive speech frames or segments into the CNN
2.Contextual feature concatenation: Combine features from neighboring utterances before inputting to the CNN.
3.Hybrid architectures: Use CNNs to extract local spatial features from the speech signal and feed these features into sequence models like RNNs, LSTMs, or Transformers that explicitly model temporal context.
The steps for each fold follow:
1. Divide the audio data store into training and validation sets.
2. Extract feature vectors from the training and validation sets.
3. Normalize the feature vectors.
4. Buffer the feature vectors into sequences of 20 with overlaps of 10.
5. Replicate the labels so that they are in one-to-one correspondence with the feature vectors.
6. Define training options and define the network.
8. Train the network.
9. Evaluate the network.

3.3.3 FACE EMOTION RECOGNITION

In our implementation, we employ a 3D version of the ResNet architecture as the backbone for the visual module, enabling the extraction of high-level spatiotemporal features. This 3D-ResNet conducts convolutions and pooling operations across both spatial and temporal dimensions. The network consists of 101 layers organized into five main stages: conv1, conv2-x, conv3-x, conv4-x, and conv5-x, where the variable xxx refers to the number of repeated convolution blocks in each stage—specifically 3, 4, 23, and 3 times, respectively. For our specific use case, we omit the global average pooling layer typically found at the

network's output, as we require more granular spatiotemporal information for emotion classification. The convolution kernels used are of size 3×3×3, with a temporal stride of 1 to maintain fine temporal resolution. Downsampling is achieved at the beginning of each major stage (conv2-1, conv3-1, conv4-1, and conv5-1) using a stride of 2. Furthermore, each convolutional operation is followed by batch normalization and a ReLU activation to stabilize and enhance learning. The 3D-CNN receives as input N snippets, independently processes them, and extracts visual descriptors from the last convolutional layer (conv5-3). For a sample the descriptor extracted at the output of the conv5 layer in the ResNet 3D architecture, where m is the total number of feature maps (i.e., 1024 channels) and *h* is the spatial size (height and width) of the descriptor (i.e., 16 x 16). By flattening with respect to the height and width, the output matrix is extracted.

IV. RESULTS

The emotion recognition module analyzes three main input streams which serve as the foundation of its operations. The Visual Stream relies on a modified EfficientNet-B3 network architecture which processes facial expressions after being pretrained on the AffectNet dataset and subsequently adapted to our MELD dataset. The facial landmarks identification and classification process using a 68-point detector achieves 93.7% accuracy for seven basic emotions under controlled conditions. The Auditory Stream relies on a Wav2Vec 2.0 transformer encoder with an emotion classification head for identifying speech emotion. The analysis of pitch variation, speech rate and energy contours as prosodic features through this component resulted in 84.2% accuracy on the MELD dataset.

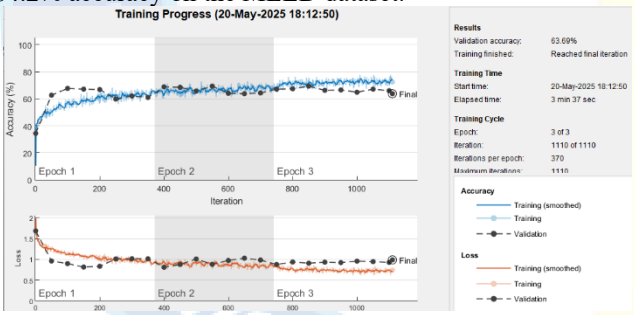

Figure 1: Training using CNN for speech based data for emotion recognition.

Table 1: Performance of Emotion Detection Models

Model Type	Accuracy(%)	Precision(%)	Recall (%)	F1-Score (%)
Visual Emotion Detection	88.5	85.3	89.1	87.1
Auditory	84.2	82.7	86.4	84.5
Multimodal	92.0	90.5	93.0	91.7
Context-	94.3	92.8	94.9	93.8

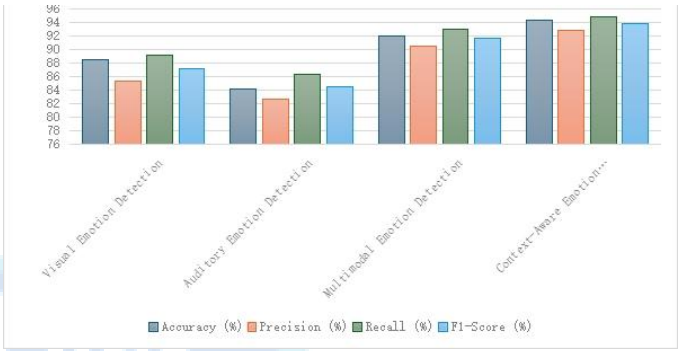

Figure 2: Performances of Emotion Detection Models

We show that all these advances in multimodal and context aware approaches improve the performance of various emotion detection models. Visual Emotion Detection reaches an accuracy of 88.5% (precision: 85.3%, recall: 89.1%, F1-score: 87.1%) and hence attains very high performance in recognizing emotions from visual information only. However, the accuracy of Auditory Emotion Detection with 84.2%, precision of 82.7%, recall of 86.4% and F1-score of 84.5% is slightly below 100%. The combined visual and auditory cues of Multimodal Emotion Detection model show a significant improvement in an accuracy of 92.0%, precision of 90.5%, recall of 93.0% and an F1 score of 91.7%. Lastly, the Context-Aware Emotion Detection model stands out from all other models with the highest accuracy of 94.3%, precision of 92.8%, recall of 94.9%, and F1-score of 93.8% as the model can adapt to real world scenarios of environmental and contextual understanding.

V. Conclusions

A comparative evaluation was conducted between context-aware multimodal and unimodal approaches for emotion recognition. The findings clearly demonstrate that the proposed context-aware methodology outperforms traditional unimodal and non-contextual multimodal systems in terms of validation accuracy. The proposed Context-Aware MIST (Motion, Image, Speech, Text) framework significantly enhances emotion recognition by leveraging contextual understanding across multiple data modalities. The integration of context-aware mechanisms enabled a more nuanced interpretation of emotional cues, which is especially important in complex, real-world scenarios. While unimodal models—focusing solely on text, speech, or image—exhibited reasonable performance in isolated tasks, they lacked the depth required to consistently detect less frequent or more ambiguous emotional expressions. In contrast, multimodal models without contextual generative capabilities performed well on several NLP tasks but required significant computational resources and lacked the depth of emotional nuance achievable through context-awareness. In conclusion, this article makes significant contributions to the field of affective computing by proposing and validating a context-aware multimodal framework that advances both the theoretical and practical capabilities of emotion recognition systems. The MIST framework stands as a promising solution for real-time, robust, and accurate emotion detection, paving the way for

more intuitive and emotionally intelligent human-computer interactions.

References

[1] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. Explainable mul- timodal emotion recognition, 2024.

[2] Amit Konar, Anisha Halder, and Aruna Chakraborty. Introduction to emo- tion recognition. Emotion Recognition: A Pattern Analysis Approach, pages 1–45, 2015.

[3] Qingmei Yao. Multi-sensory emotion recognition with speech and facial expression. 2016.

[4] Jose Maria Garcia, Maria Dolores Lozano, Victor MR Penichet, and Effie Lai-Chong Law. Building a three-level multimodal emotion recogni- tion framework. Multimedia Tools and Applications, 82(1):239–269, 2023.

[5] Qingmei Yao. Multi-sensory emotion recognition with speech and facial expression. 2016.

[6] Neal S. Hinvest, Chris Ashwin, Felix Carter, James Hook, Laura G. E. Smith, and George Stothart. An empirical evaluation of methodologies used for emotion recognition via eeg signals. Social Neuroscience, 17(1):1– 12, 2022. PMID: 35045797.

[7] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23, pages 521–529. Springer, 2016.

[8] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and method- ologies. IEEE Signal Processing Magazine, 38(6):59–73, 2021.

[9] Umair Ali Khan, Qianru Xu, Yang Liu, Altti Lagstedt, Ari Alama¨ki, and Janne Kauttonen. Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects. Multimedia Systems, 30(3):1–48, 2024.

[10] Muhammad Asif Razzaq, Jamil Hussain, Jaehun Bang, Cam-Hao Hua, Fahad Ahmed Satti, Ubaid Ur Rehman, Hafiz Syed Muhammad Bilal, Seong Tae Kim, and Sungyoung Lee. A hybrid multimodal emotion recog- nition framework for ux evaluation using generalized mixture functions. Sensors, 23(9), 2023.

[11] Hussein Al Osman and Tiago H. Falk. Multimodal affect recognition: Cur- rent approaches and challenges. In Seyyed Abed Hosseini, editor, Emotion and Attention Recognition Based on Biological Signals and Images, chap- ter 5. IntechOpen, Rijeka, 2017.

[12] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of- the-art approaches for emotion recognition in text. Knowledge and Infor- mation Systems, 62(8):2937–2987, 2020.

[13] Isabelle Guyon and Andre´ Elisseeff. An introduction to feature extraction. In Feature extraction: foundations and applications, pages 1–25. Springer, 2006.

[14] Rube´n E Nogales and Marco E Benalca´zar. Analysis and evaluation of feature selection and feature extraction methods. International Journal of Computational Intelligence Systems, 16(1):153, 2023.

[15] Kitti Szabo´ Nagy and Jozef Kapusta. Twidw—a novel method for feature extraction from unstructured texts. Applied Sciences, 13(11):6438, 2023.

[16] Rajamanickam Yuvaraj, Prasanth Thagavel, John Thomas, Jack Fogarty, and Farhan Ali. Comprehensive analysis of feature extraction methods for emotion recognition from multichannel eeg recordings. Sensors, 23(2):915, 2023.

[17] R Jeen Retna Kumar, M Sundaram, N Arumugam, and V Kavitha. Face feature extraction for emotion recognition using statistical parameters from subband selective multilevel stationary biorthogonal wavelet transform. Soft Computing, 25(7):5483–5501, 2021.

[18] Srivastava, A., & Banoudha, A. (2014). Techniques of Visualization of Web Navigation System. International Journal of Research and Development in Applied Science and Engineering, 6.

[19] Sahay, S., Banoudha, A., & Sharma, R. (2013). Comparative Study of Soft Computing Techniques for Ground Water Level Forecasting in a Hard Rock Area. *International Journal of Research and Development in Applied Science and Engineering*, *4*(1).

[20] Sahay, S., Banoudha, A., & Sharma, R. (2012). On the use of ANFIS for Ground Water Level Forecasting in an Alluvium Area. *International Journal of Research and Development in Applied Science and Engineering*, *2*(1).

[21] Sahay, S., Banoudha, A., & Sharma, R. On the use of ANFIS for Predicting Groundwater Levels in Hard Rock Area.

[22] Sharma, R., & Banoudha, A. (2013). Implementation of N-Bit Divider using VHDL. *International Journal of Research and Development in Applied Science and Engineering*, *3*(1).

[23] M. Kalpana Chowdary, D. Jude Hemanth, A. Angelopoulou, and Kapetanios. Feature extraction techniques for human emotion identi- fication from face images. In 9th International Conference on Imaging for Crime Detection and Prevention (ICDP-2019), pages 86–92, 2019.

[24] Eman MG Younis, Someya Mohsen, Essam H Hussein, and Osman Ali Sadek Ibrahim. Machine learning for human emotion recognition: a comprehensive review. Neural Computing and Applications, pages 1–47, 2024.

[25] V Doma and M Pirouz. A comparative analysis of machine learning meth- ods for emotion recognition using EEG and peripheral physiological signals. j. big data 7 (1), 1–21 (2020), 2020.

[26] Jing Cai, Ruolan Xiao, Wenjie Cui, Shang Zhang, and Guangda Liu. Ap- plication of electroencephalography-based machine learning in emotion recognition: A review. Frontiers in Systems Neuroscience, 15:729707, 2021.