Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 236 (2024) 428–435

Procedia Computer Science

www.elsevier.com/locate/procedia

International Symposium on Green Technologies and Applications (ISGTA'2023)

Modeling Speech Emotion Recognition via ImageBind representations

Adil CHAKHTOUNA[a,*], Sara SEKKATE[b], Abdellah ADIB[a]

[a]*Team Data Science & Artificial Intelligence, Laboratory of Mathematics, Computer Science and Applications (LMCSA), Faculty of Sciences and Technologies, Hassan II University, Mohammedia, Morocco*

[b]*Higher National School of Arts and Crafts, Hassan II University, Casablanca, Morocco*

Abstract

Speech Emotion Recognition (SER) refers to the ability of Machine Learning (ML) and Deep Learning (DL) techniques to accurately predict people's emotional states from speech signals. Significant progress has been achieved in the SER domain involving the incorporation of DL models to introduce novel features extraction processes. This paper introduces the use of deep representations learned from the multi-modal Large Language Model (LLM) called ImageBind. These representations were subsequently provided as input to the Nu-Support Vector Machine (Nu-SVM) with RBF kernel for the classification task. The experiments were executed using the IEMOCAP database within the context of a Speaker-Dependent (SD) scenario. The method achieved a noteworthy overall accuracy rate of 80.58% for the four emotions of IEMOCAP, representing a substantial improvement over well-established methods in the existing body of literature. Thus, affirming that the proposed methodology, founded upon ImageBind representations, introduces a novel perspective to the field of SER.

Keywords: ImageBind; Speech Emotion Recognition; Embedding representations; IEMOCAP; Nu-SVM.

1. Introduction

Speech Emotion Recognition (SER) is a multidisciplinary field that integrates diverse domains, starting with the processing of speech signals and extending to the comprehension of emotional states. It is widely regarded as a fundamental component of Human-Computer Interaction (HCI). SER aims to autonomously discern human emotions through the analysis of spoken statements. This process contributes significantly to the advancement of applications within speech-based HCI, particularly in fields such as intelligent service assistants [1], distance education [2], healthcare [3], and automated contact centers [4]. The underlying objective of this study is to explore and investigate the potential of SER systems into the domain of green technologies. The incorporation of SER into this domain is moti-

* Corresponding author.
E-mail address: adilchakhtouna10@gmail.com

10.1016/j.procs.2024.05.050

vated by the desire to create smarter, more adaptive, and user-centric green technologies, thereby advancing the goals of environmental sustainability and energy efficiency.

Characterizing and categorizing human emotions poses a formidable challenge, consequently, numerous research investigations seek to address this challenge by employing diverse modalities in both mono or multi-modal approaches using various data types, such as images, text, speech, physiological signals, among others [5, 6, 7]. The primary objective of SER is to use speech as a pathway that contains both para-linguistic and linguistic information, thereby enabling the extraction of a wide range of features. The process of extraction is categorized into two distinct groups of features: handcrafted and deep ones. Handcrafted features, often referred to as low-level features, demand specialized expertise for extraction encompassing domains such as prosody, spectral, and cepstral characteristics [8, 9]. Conversely, DL methods are introduced to automate the extraction of high-level features, reducing the manual effort associated with handcrafted feature extraction.

Recently, Large Language Models (LLMs) and transformers have significantly propelled advancements across numerous domains, with notable impacts observed, particularly in the Natural Language Processing (NLP) tasks. Concurrently, multi-modal LLMs, such as GPT-4[1], ImageBind [10], and PALM-E [11], have undertaken extensive investigations into the ability of LLMs to effectively understand multi-modal information. Motivated by this reality, the present research aims to harness the potential of the multi-modal ImageBind framework and leverage its efficiency in the context of the SER task. The method under consideration employs the audio modality of the ImageBind framework to acquire a unified representation space, utilizing diverse sets of image-paired data. Subsequently, ML algorithm, namely, Nu-Support Vector Machine (Nu-SVM) was applied to effectively classify distinct emotional states within the IEMOCAP dataset.

The structure of this research manuscript is arranged as follows: The subsequent Section 2 provides a concise overview of the most recent research in the field of SER. Section 3 outlines the methodology proposed in this study. Section 4 presents a comprehensive account of the conducted experiments, highlighting the results achieved in this paper. Lastly, in Section 5, we summarize the conclusions drawn from the study and discuss potential future directions for SER.

2. SER related works

Prior to the inclusion of DL methods into the SER, numerous studies relied on conventional ML approaches. After that and as a result, the modeling of the SER task can be accomplished through the utilization of ML, DL algorithms, or a combination of both. DL typically demonstrates enhanced performance, whereas ML can outperform in specific scenarios, depending on factors like dataset characteristics, feature extraction methodologies, and the specific case study in question. Identifying emotions using the speech modality is a challenging task, which motivates researchers to investigate a variety of techniques aimed at achieving more effective emotion classification. Below, we present the foremost SER approaches employed by diverse authors in the literature. Hao et al. [12] created an architecture referred to as ADRNN, which incorporates a dilated Convolutional Neural Network (CNN) along with residual blocks and a Bidirectional Long Short-Term Memory (BiLSTM) based on the attention mechanism. The input for the proposed algorithm consisted of 3-D spectrograms derived from raw signals, encompassing Log-Mel, Deltas, and Delta-deltas components. The research has undergone testing through mono-corpus and cross-corpus experiments, covering both SD and SI modes.

Sajjad et al. [13] presented a framework employing K-means clustering in conjunction with a Radial Basis Function Network (RBFN) to generate a sequential segment. This segment is subsequently converted into a spectrogram through the application of the Short-Time Fourier Transform (STFT) and then processed using the ResNet101 model to extract deep features. Finally, a BiLSTM architecture is employed to capture temporal information for emotion recognition. The proposed approach addressed the SER task in both SD and SI scenarios, achieving remarkable recognition rates. However, it is worth noting that the implementation of K-means, ResNet101, and BiLSTM increased the computational complexity of the method.

[1] https://openai.com/research/gpt-4

In [14], they proposed a fusion technique for combining two sets of representations, local and global features, which were extracted from distinct regions within the speech spectrogram. Specifically, the local representations were acquired through CNN, while the global ones were derived using a Dense Capsule Network (CapsNet). For the purpose of the classification task, the Extreme Learning Machine (ELM) was adopted. It is observed that the choice to combine the CNN and CapsNet models is seen as a wise one, effectively mitigating the individual limitations of each model while mutually enhancing their capabilities. As a result, the notable performance improvements achieved when these models are combined.

In [15], the authors used the unsupervised pre-trained model called Vector Quantization Variational Automatic Encoder (VQ-VAE) to learn latent representations from speech emotion signals. The sequence information obtained from VQ-VAE was concatenated with deep representations derived through the CNN architecture, and subsequently input into the Temporal Attention Convolutional Network (TACN) for the purpose of emotion recognition. this research explored the embedding features extraction avoiding to use the traditional handcrafted features. The study investigated the extraction of integrated features as an alternative to avoid the manually-crafted feature engineering.

From these various studies discussed, our selection process was guided by the choice of the specific dataset utilized, IEMOCAP, the designated data split ratio (80/20), and the inclusion of the DL approaches either during feature extraction or for the classification task. Our current research aims to contribute to the field of SER by incorporating cutting-edge techniques in Artificial Intelligence (AI), particularly focusing on the feature extraction stage. To the best of our knowledge, we employed the ImageBind architecture for the first time in SER, which was initially designed for multi-modal LLMs. We adapted ImageBind as a feature extractor within the context of our SER study. The embedding representations derived from ImageBind, along with the implementation of the selected Nu-SVM algorithm for classification, have demonstrated their efficiency compared to existing State-Of-The-Art (SOTA) approaches.

3. Proposed Methodology

This study presents an innovative approach to SER, which involves the integration of ImageBind representations acquired for each audio signal within the IEMOCAP dataset. The proposed approach is structured around two primary components: (1) feature extraction, and (2) model training, as illustrated in Fig. 1.

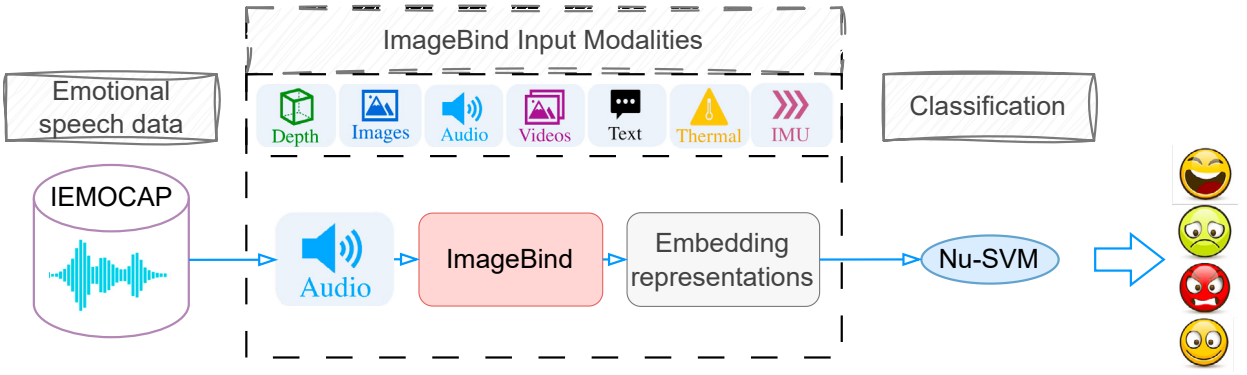

Fig. 1. The proposed SER flowchart.

3.1. ImageBind embedding representations

ImageBind [10] is an open source framework recently developed by Meta AI[2] for the purpose of multi-modal LLMs tasks. The model, by learning a shared representation space, serves to unify various data modalities, encompassing

[2] https://ai.meta.com/

images, audio, videos, text as well as sensory data types like Depth (3D), Thermal, and Inertial Measurement Unit (IMU).

ImageBind employs modality pairs, denoted as $(\mathcal{I}, \mathcal{M})$, wherein $\mathcal{I}$ denotes images, and $\mathcal{M}$ represents another modality, for the purpose of learning a single embedding. To illustrate this process further, when provided with an image I_i and its corresponding observation in the alternate modality M_i, these inputs are encoded into normalized embeddings q_i and k_i via two separate deep networks, labeled as f and g as follows:

$$\begin{cases} q_i = f(I_i) \\ k_i = g(M_i) \end{cases} \tag{1}$$

Here is the mathematical formulation for the loss function employed in the process of learning embeddings and the encoders:

$$\mathcal{L}_{(\mathcal{I},\mathcal{M})} = -\log \frac{\exp(q_i^T k_i/\tau)}{\exp(q_i^T k_i/\tau) + \sum_{j\neq i} \exp(q_i^T k_j/\tau)} \tag{2}$$

where, τ is a scalar parameter that regulates the degree of smoothness within the Softmax distribution, while the index j representing unrelated observations.

ImageBind was applied to a range of six modalities, structured in paired data format, which included combinations such as (image, text), (video, audio), (image, depth), (image, thermal), and (video, IMU). The datasets used for this purpose were drawn from large-scale web pairs datasets, including Audioset [16], SUN RGB-D [17], LLVIP [18], and Ego4D [19].

In the context of SER, we exclusively employed the audio modality of the ImageBind framework to embed the speech cues present in the IEMOCAP database. Firstly, each speech file is resampled from 48 KHz to 16 KHz. Subsequently, it is transformed into Mel spectrograms, comprising 128 Mel frequency bins. Following this, data normalization was applied to each Mel spectrogram. Finally, after the completion of these processing steps, the speech signal A_i was encoded into a normalized embedding through the deep networks within the ImageBind architecture. At the output, the generated vector of representations for each signal equal to 1024 hidden states.

3.2. Classification

Following the extraction of various embedded representations from the pre-trained ImageBind framework, the ensemble of features is then inputted into our selected downstream model Nu-SVM for the purpose of emotion classification.

Nu-Support Vector Machine

Support Vector Machine (SVM) [20] based classification represents a broad area of research activity, offering a powerful way of solving classification tasks across a wide range of domains [21, 22], including SER. SVM is highly suitable for diverse datasets types, proficiently handling both non-linear and linear inputs data using various kernel functions such as RBF, Linear, Polynomial, etc.

In 2000, Schölkopf et al. introduced the Nu-Support Vector Machine (Nu-SVM) [23], which distinguishes itself by employing a parameter denoted as ν instead of the C parameter found in standard SVM. The integration of ν serves to regulate the number of support vectors effectively. The ν parameter is defined as a fractional value in the interval [0, 1], representing an upper limit on the fraction of learning errors and a lower limit on the fraction of support vectors, which specifies the desired balance between the margin and the number of support vectors.

4. Experimentation, Results and Discussion

4.1. Experimental protocol

This section provides a comprehensive overview of all experiments conducted using the proposed methodology. For evaluation, we used the four widely adopted performance measures, namely precision, recall, F1-score and accuracy. Furthermore, t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization [24] is employed to furnish a graphical representation of the learned ImageBind representations. The evaluation considered the context of speaker dependency, where the dataset was randomly divided into two subsets: a training set and a testing set, with distribution proportions of 80%, and 20% respectively. For the Nu-SVM implementation, two key hyper-parameters were considered: the choice of the RBF kernel and the value of $\nu = 0.1$. The training process involved to derive the optimal learning parameters for the model, followed by evaluating the model's performance on the test set.

The well known IEMOCAP database was chosen for this study: **IEMOCAP**[3] [25] is an emotion recognition database including modalities from speech, text, video, and facial expressions. It comprises a total duration of approximately 12 hours, featuring dual conversations derived from both scripted and improvised dialogues. These conversations were expressed by 10 speakers including 5 men and 5 women and then partitioned into five sessions, each session contains 1 man and 1 woman. This study aligns with the existing SOTA works according to the speech modality and categorical emotions, in particular the four emotions: happiness, neutral, anger and sadness. The total number of samples included in this study is 5531.

4.2. Results & discussion

The outcomes obtained using the various metrics are presented in Table 1. An examination of these findings, employing ImageBind's integrated representations alongside with the Nu-SVM as downstream model, demonstrated favorable performance across the spectrum of emotional states. Specifically, the neutral emotional state displayed the highest recall rate at 83.63%, followed by happiness, sadness and anger with almost equal values of 79.82%, 78.80%, and 78.73% respectively. In the case of the accurately predicted positive instances, the obtained results achieved values of 87.69% for sad, 80.56% for happy, 77.68% for anger, and 78.57% for neutral. In a more extensive perspective, the comprehensive recognition rate of the proposed methodology reached an overall performance level of 80.58%.

Table 1. Method performance for SER on the IEMOCAP dataset.

Metrics	**Emotions**				**Average**
	Anger	**Happy**	**Neutral**	**Sad**	
Precision (%)	77.68	80.56	78.57	87.69	81.12
Recall (%)	78.73	79.82	83.63	78.80	80.24
F1-score (%)	78.20	80.18	81.02	83.01	80.60
Overall Accuracy (%)			**80.58**		

Another method for evaluating the proposed approach involves the utilization of the t-SNE algorithm. This technique was implemented to generate a three-dimensional (3D) visualization, allowing us to discover the transformations induced by the acquired ImageBind representations within the high-dimensional space of Nu-SVM. Fig. 2 explains this process, the distribution of data points demonstrated a strong performance, manifesting as well-defined clusters in the graph on the left for the training set and in the graph on the right for the test set. In addition, the illustration shows that the proposed method achieves a distinct separation between all emotional states, where each emotional class is

[3] https://sail.usc.edu/iemocap/index.html

represented by a specific color group. It is crucial to highlight that the proposed method does not achieve a perfect separation for all emotional categories within the test set, as is evident from the occurrence of misclassifications among specific emotional instances.

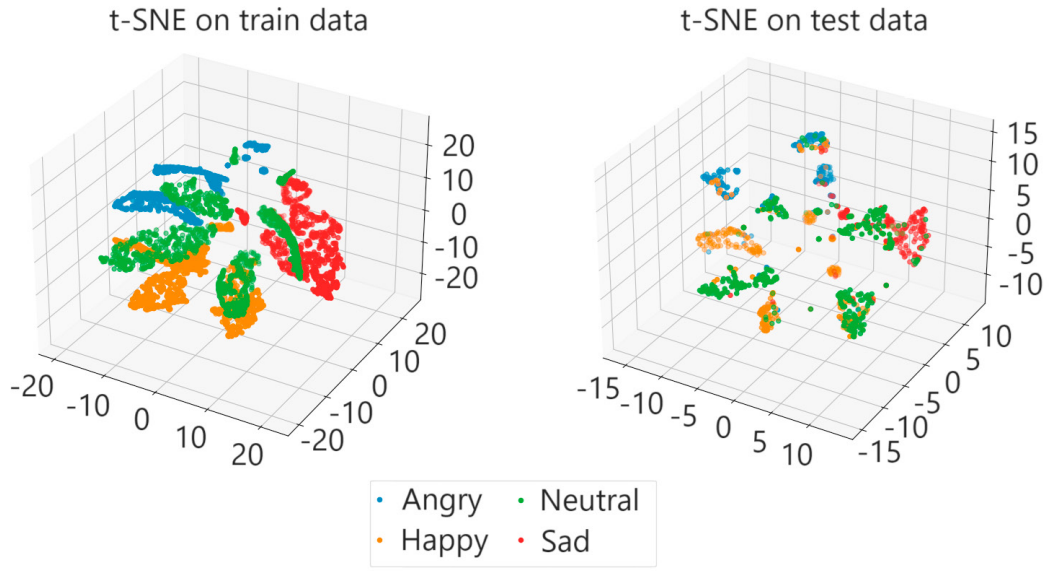

Fig. 2. 3D t-SNE visualization of the learned representations on train and test data.

4.3. Comparative study

The principal objective of this subsection is to conduct a comprehensive evaluation of the proposed method based on the embedded representations of ImageBind framework alongside with the Nu-SVM algorithm in comparison to the prior SOTA approaches for the SER task. Making a reasonable comparison between the outcomes of various prior studies is often exceptionally challenging, due to the extensive diversity of configurations employed within each investigation. In order to address this matter and establish an equitable comparison, we respected the following considerations: same dataset (IEMOCAP), same sample distributions for each emotion (in total 5531), SD scenario, 80/20 data split-ratio, and naturally, recognizing that the distinguishing factor among these studies lies in the distinct methodologies (features & model) proposed by each one.

Table 2 displays the performance results of our developed approach in comparison to several other methods documented in the existing literature [12, 14, 15]. The examination is based on the accuracy achieved by each of the assessed methods. The insights derived from Table 2 underscore the effectiveness of the proposed approach, which relies on ImageBind representations in conjunction with the Nu-SVM. It should be mentioned that, to the best of the authors' knowledge, the integration of ImageBind representations in the SER domain was explored for the first time in this study and was not previously investigated in the literature.

5. Conclusion

The central focus of this study lies around the extraction of embedded features denoted as ImageBind representations, which are derived from the newly developed Meta AI framework, especially the Audio modality of ImageBind. The SER system under consideration was trained and tested employing the Nu-SVM model, incorporating a RBF kernel, and evaluated on the widely recognized emotional benchmark database IEMOCAP. In addition, the study explored the SD experimental setup, and the obtained findings proved robust performance within this context, resulting in a noteworthy recognition rate.

Table 2. Comparison study between the SER proposed approach and the previous SOTA studies.

Reference	Database	Method (Features & Model)	Split-ratio	Accuracy (%)
[15]	IEMOCAP (5531)	VQ-VAE + deep CNN features & TACN	80/20	70.16
[14]	IEMOCAP (5531)	Spectrogram & TFCNN + DenseCap + ELM	80/20	70.78
[12]	IEMOCAP (5531)	3D Mel spectrogram & Dilated CNN + BiLSTM	80/20	74.96
Our	**IEMOCAP (5531)**	**ImageBind representations & Nu-SVM**	**80/20**	**80.58**

The comparative analysis carried out in relation to SOTA methodologies reveals that the developed system outperforms the existing approaches in SER. As prospective avenues for future research, we intend to evaluate the efficacy of this proposed method in the SI scenario, as well as we plan to investigate the utilization of the updated version of ImageBind called "ImageBind-LLM", which is coupled with LLMs. Furthermore, we are interested to get into the multi-modal emotion recognition by combining text and facial modalities, with the assistance of ImageBind technology.

Acknowledgments

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/01).

References

[1] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, D. Giri, Real-time speech emotion analysis for smart home assistants, IEEE Transactions on Consumer Electronics 67 (1) (2021) 68–76.
[2] D. Tanko, S. Dogan, F. B. Demir, M. Baygin, S. E. Sahin, T. Tuncer, Shoelace pattern-based speech emotion recognition of the lecturers in distance education: Shoepat23, Applied Acoustics 190 (2022) 108637.
[3] Z. Tariq, S. K. Shah, Y. Lee, Speech emotion detection using iot based deep learning for health care, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 4191–4196.
[4] M. Płaza, S. Trusz, J. Kęczkowska, E. Boksa, S. Sadowski, Z. Koruba, Machine learning algorithms for detection and classifications of emotions in contact center applications, Sensors 22 (14) (2022) 5311.
[5] C. Gautam, K. Seeja, Facial emotion recognition using handcrafted features and cnn, Procedia Computer Science 218 (2023) 1295–1303.
[6] V. Revathy, A. S. Pillai, F. Daneshfar, Lyemobert: Classification of lyrics' emotion and recommendation using a pre-trained model, Procedia Computer Science 218 (2023) 1196–1208.
[7] A. Pradhan, S. Srivastava, Hierarchical extreme puzzle learning machine-based emotion recognition using multimodal physiological signals, Biomedical Signal Processing and Control 83 (2023) 104624.
[8] A. Chakhtouna, S. Sekkate, A. Adib, Improving speech emotion recognition system using spectral and prosodic features, in: International Conference on Intelligent Systems Design and Applications, Springer, 2021, pp. 399–409.
[9] A. Chakhtouna, S. Sekkate, A. Adib, Improving speaker-dependency/independency of wavelet-based speech emotion recognition, in: International Conference on Networking, Intelligent Systems and Security, Springer, 2022, pp. 281–291.
[10] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, I. Misra, Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.
[11] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., Palm-e: An embodied multimodal language model, arXiv preprint arXiv:2303.03378 (2023).
[12] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE access 7 (2019) 125868–125881.

[13] M. Sajjad, S. Kwon, et al., Clustering-based speech emotion recognition by incorporating learned features and deep bilstm, IEEE access 8 (2020) 79861–79875.
[14] J. Liu, Z. Liu, L. Wang, L. Guo, J. Dang, Speech emotion recognition with local-global aware deep representation learning, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7174–7178.
[15] J. Liu, Z. Liu, L. Wang, Y. Gao, L. Guo, J. Dang, Temporal attention convolutional network for speech emotion recognition with latent representation., in: INTERSPEECH, 2020, pp. 2337–2341.
[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 776–780.
[17] S. Song, S. P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 567–576.
[18] X. Jia, C. Zhu, M. Li, W. Tang, W. Zhou, Llvip: A visible-infrared paired dataset for low-light vision, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3496–3504.
[19] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al., Ego4d: Around the world in 3,000 hours of egocentric video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18995–19012.
[20] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.
[21] A. Chakhtouna, S. Sekkate, A. Adib, Speaker and gender dependencies in within/cross linguistic speech emotion recognition, International Journal of Speech Technology (2023) 1–17.
[22] S. Akil, S. Sekkate, A. Adib, Classification of credit applicants using svm variants coupled with filter-based feature selection, in: International Conference on Networking, Intelligent Systems and Security, Springer, 2022, pp. 136–145.
[23] B. Schölkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett, New support vector algorithms, Neural computation 12 (5) (2000) 1207–1245.
[24] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11) (2008).
[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.