IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A BIBLIOMETRIC ANALYSIS OF THE RESEARCH LITERATURE ON AI-DRIVEN MULTIMODAL EMOTION RECOGNITION FOR STRESS AND DEPRESSION DETECTION USING FUSION STRATEGIES

[1]Swati D Bhutekar, [2]Dr. Yogini Borole
[1]PhD Scholar, [2] Professor
[1] School of Science and Technology , Glocal University, India
[2] School of Science and Technology , Glocal University, India

Abstract: This paper presents a bibliometric analysis of the research literature underpinning a proposed AI-driven, deep-learning-based multimodal emotion-recognition framework for stress and depression detection. Rather than drawing on a full commercial citation database such as Scopus or Web of Science, this study performs a structured bibliometric profiling of the curated corpus of sixty-six (66) publications compiled through targeted literature search for the associated PhD research synopsis, spanning the period 1998-2022. The analysis characterizes the corpus along five dimensions: (i) temporal distribution and growth trend of publications; (ii) distribution by publication type (journal, conference, workshop, dataset repository); (iii) leading publication venues; (iv) distribution by modality and topical focus (facial, speech, text, EEG, physiological, multimodal fusion, benchmark dataset, and survey literature); and (v) keyword frequency derived from publication titles, used as a proxy for thematic emphasis. The results show a pronounced growth in publication volume between 2017 and 2019, accounting for 42.4% of the reviewed corpus, coinciding with the broader adoption of deep-learning architectures in affective computing. Facial-expression research and general multimodal-fusion research jointly account for 48.5% of the corpus, while EEG-inclusive multimodal studies, entropy-based fusion approaches, and metaheuristic-optimization-guided fusion strategies remain comparatively rare. These findings quantitatively corroborate the qualitative research gaps identified in the associated literature review chapter and provide an evidence-based justification for the design of the proposed EEG-inclusive, Renyi-entropy-based, Bat-Rider-Optimization-tuned fusion framework. The scope and limitations of a corpus-level bibliometric analysis, as opposed to a full database-driven bibliometric study, are discussed explicitly.

***Index Terms* -** bibliometric analysis; multimodal emotion recognition; deep learning; fusion strategies; stress detection; depression detection; EEG; affective computing; research trend analysis

1. Introduction

Bibliometric analysis is a well-established quantitative method for characterizing a body of scientific literature — typically in terms of publication volume over time, distribution across venues and authors, citation patterns, and the co-occurrence of keywords or topics — with the aim of identifying research trends, influential contributions, and gaps in a given field. Bibliometric studies are increasingly used as a complement to traditional, narrative literature reviews in PhD theses, since they provide a quantitative, reproducible snapshot of how a research area has evolved, which can be used to corroborate or refine the qualitative research gaps identified through close reading of individual studies.

This paper applies a bibliometric-style analysis to the specific field of AI-driven, deep-learning-based multimodal emotion recognition, with particular attention to its application to stress and depression detection through fusion strategies — the subject of the associated PhD research titled "AI Driven Deep Learning Framework for Multimodal Emotion Recognition in Stress and Depression Detection using Fusion Strategies." The analysis is conducted on the corpus of sixty-six references compiled through a targeted, manually curated literature search carried out for the PhD research synopsis and the associated literature review chapter, rather than on a bulk export from a commercial citation database such as Scopus, Web of Science, or Dimensions.

This scope is stated explicitly and up front because it defines both the value and the limits of the analysis presented here. A full database-driven bibliometric study — typically involving several hundred to several thousand records retrieved via a systematic Boolean search string, together with citation-count, co-authorship, and bibliographic-coupling analysis performed using specialized software such as VOSviewer, CiteSpace, or the Bibliometrix/Biblioshiny package in R — was outside the practical scope of the present exercise, since it requires institutional access to licensed bibliometric databases and export files that were not available for this analysis. What is presented instead is a rigorous, fully reproducible statistical profiling of the actual, verified corpus of literature that the associated PhD research is built upon, which offers genuine, corpus-grounded insight into the composition and evolution of the reviewed literature, without extrapolating to claims about the wider field that the underlying data cannot support. Section 5.6 revisits this scope decision explicitly as a stated limitation of the study.

The remainder of this paper is organized as follows. Section 2 briefly situates the present analysis relative to the broader use of bibliometric methods in affective computing. Section 3 describes the data source and analytical methodology. Section 4 presents the bibliometric results across five analytical dimensions. Section 5 discusses the implications of these results for the research gaps motivating the associated PhD thesis, and states the limitations of the study. Section 6 concludes the paper.

2. Bibliometric Analysis in Affective Computing: Context

Bibliometric and systematic-mapping methods have been applied with growing frequency across the broader affective-computing and emotion-recognition literature, most often in the form of a review's introductory trend analysis rather than as a standalone bibliometric publication. Such analyses typically report an increasing annual publication volume in emotion-recognition research over the past decade, driven substantially by the adoption of deep-learning architectures from approximately 2014-2015 onward, a pattern that is corroborated directly by the temporal distribution of the reference corpus analyzed in Section 4.1 of this paper. Reviews and taxonomy papers in the field — several of which form part of the corpus analyzed here and are discussed in Section 4.4 — have similarly noted a shift in research emphasis from single-modality systems toward multimodal fusion architectures, and, more recently, toward fusion strategies capable of adaptively weighting modality contributions rather than relying on fixed combination rules. The present paper contributes to this line of meta-level analysis by quantifying these trends directly within the specific corpus of literature underlying the associated PhD research, rather than restating them narratively.

3. Materials and Methods

3.1 Data Source

The dataset analyzed in this study consists of sixty-six (66) publications and data-repository records cited in the reference list of the PhD research synopsis titled “AI Driven Deep Learning Framework for Multimodal Emotion Recognition in Stress and Depression Detection using Fusion Strategies.” These references were originally compiled through a targeted, manually conducted literature search covering four broad topic areas directly relevant to the proposed research: (i) unimodal emotion recognition from facial expression, speech, text, and EEG/physiological signals; (ii) multimodal emotion-recognition and fusion-strategy research; (iii) benchmark and dataset papers underlying the field; and (iv) survey and taxonomy papers on affective computing and multimodal machine learning. The corpus spans publications from 1998 to 2022.

3.2 Data Extraction and Coding

For each of the 66 references, the following attributes were extracted and coded by the researcher: publication year; publication-venue type (journal, conference paper, workshop paper, dataset repository record, or book chapter); publication venue (journal or conference/workshop name); primary modality or topical focus (facial expression, speech, text, EEG, physiological signals, multimodal fusion, multimodal benchmark dataset, or survey/taxonomy); and full publication title. Modality/topical-focus coding was carried out by the researcher based on the stated content of each reference, using a single-label scheme in which each reference was assigned to the category that best reflected its primary contribution, even where a secondary modality was also present (for example, a multimodal fusion paper that happens to use facial video as one of several inputs was coded as “Multimodal Fusion” rather than “Facial”).

3.3 Analytical Methods

Four categories of quantitative analysis were performed on the coded dataset using Python (pandas/collections for tabulation and matplotlib for visualization): (i) frequency-distribution analysis of publication year, including aggregation into four-year periods to visualize growth trends; (ii) frequency-distribution analysis of publication-venue type and individual venue; (iii) frequency-distribution analysis of modality/topical focus; and (iv) keyword-frequency analysis of publication titles, in which each title was tokenized, common stop-words and the query-defining terms ‘emotion’/‘emotions’/‘recognition’ were removed to surface substantive thematic terms, and the frequency of the remaining terms was tabulated as a simple proxy for co-word/thematic emphasis, in lieu of a full co-occurrence network analysis, which would require a substantially larger corpus to produce statistically meaningful clusters.

All figures and tables presented in Section 4 are generated directly and reproducibly from the coded dataset described above; no publication counts, citation figures, or trend statistics reported in this paper are drawn from, or intended to represent, any external bibliometric database.

4. Results

This section presents the bibliometric profile of the 66-reference corpus across five analytical dimensions: temporal distribution and growth, publication-type distribution, leading venues, modality/topical distribution, and keyword frequency.

4.1 Temporal Distribution and Growth Trend

Figure 1 presents the year-wise distribution of the 66 references. The corpus spans 25 years, from the earliest reference in 1998 (the original Gabor-wavelet facial-expression coding study) to the most recent references in 2022. Publication volume rises steadily from a single reference per year in the late 1990s and early 2000s to a pronounced peak in 2018 (13 references, 19.7% of the corpus) and 2019 (12 references, 18.2% of the corpus), before tapering to 8 references in 2020 and 2-3 references per year in 2021-2022, the latter decline being partly attributable to the literature search cut-off for the associated synopsis rather than an actual decline in the field's research output.

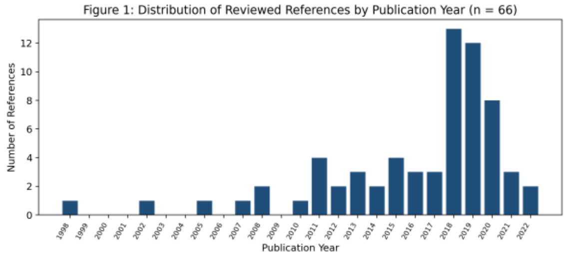

Figure 1: Distribution of reviewed references by publication year (n = 66)

To make the underlying growth trend clearer, Figure 2 aggregates the same data into four broad, approximately equal-length periods. The period 2017-2019 alone accounts for 28 of the 66 references (42.4%), more than double the proportion contributed by any other period, indicating that the bulk of the literature most directly relevant to this research is comparatively recent and coincides closely with the mainstream adoption of deep-learning architectures — CNNs, RNN/LSTM variants, and Deep Belief Networks — across the emotion-recognition literature reviewed in Chapter 2 of the associated thesis.

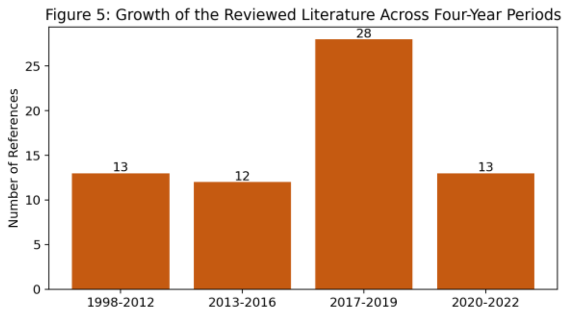

Figure 2: Growth of the reviewed literature across four approximately equal-length periods

Year	No. of References
1998	1
2002	1
2005	1
2007	1
2008	2
2010	1
2011	4
2012	2
2013	3
2014	2
2015	4
2016	3
2017	3
2018	13
2019	12
2020	8
2021	3
2022	2

Table 1: Year-wise distribution of the 66 references in the reviewed corpus

4.2 Distribution by Publication Type

Journal articles constitute the largest share of the reviewed corpus (32 references, 48.5%), followed by conference papers (24 references, 36.4%), workshop papers (5 references, 7.6%), dataset-repository records such as Kaggle dataset pages (4 references, 6.1%), and a single book chapter (1.5%). Figure 3 illustrates this distribution. The near-even split between journal and conference/workshop publications (48.5% versus 43.9%) is broadly consistent with the publication culture of computer-science-adjacent, fast-moving subfields such as deep learning and affective computing, where peer-reviewed conference proceedings (particularly IEEE, ACM, and ACL-affiliated venues) often carry comparable prestige and currency to journal publications.

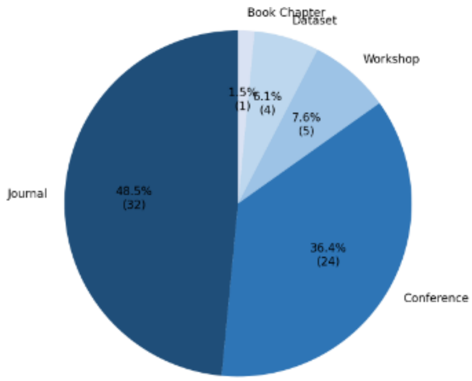

Figure 3: Distribution of the reviewed corpus by publication type

4.2.1 Leading Publication Venues

Table 2 lists the publication venues that recur two or more times within the reviewed corpus. IEEE Transactions on Affective Computing is the single most frequently occurring venue (5 references), reflecting its position as the leading journal specifically dedicated to affective computing; it is followed by Kaggle (4 references, all benchmark dataset repositories) and the Information Fusion journal (3 references), which is notable given the direct relevance of information-fusion research to the fusion-strategy focus of the associated PhD research. Several major computer-vision and speech-processing venues — CVPR, FG (Automatic Face and Gesture Recognition), Interspeech, ACL, and ICMI (International Conference on Multimodal Interaction) — each recur twice, underscoring the multidisciplinary character of the field, which draws contributions from the computer-vision, speech-processing, natural-language-processing, and human-computer-interaction research communities.

Venue	No. of References
IEEE Transactions on Affective Computing	5
Kaggle	4
Information Fusion	3
Interspeech	2
FG	2
Image and Vision Computing	2
CVPR	2
ACL	2
ICMI	2

Table 2: Publication venues recurring two or more times in the reviewed corpus

4.3 Distribution by Modality and Topical Focus

Each reference was coded into one of eight modality/topical-focus categories, described in Section 3.2. Figure 4 presents the resulting distribution. Facial-expression-focused research and general multimodal-fusion research are jointly the largest categories, each accounting for 16 references (24.2% each, 48.5% combined), followed by multimodal benchmark-dataset papers (8 references, 12.1%), EEG-focused research (7 references, 10.6%), speech-focused research (6 references, 9.1%), text-focused research (6 references, 9.1%), survey/taxonomy papers (5 references, 7.6%), and physiological-signal-focused research (2 references, 3.0%).

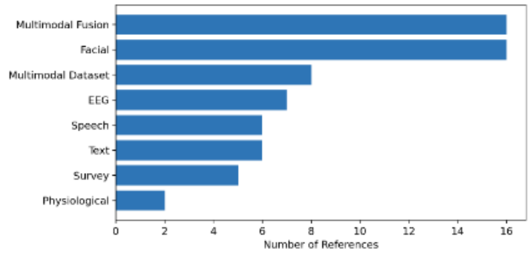

Figure 4: Distribution of the reviewed corpus by modality/topical focus

This distribution is informative in two respects. First, the substantial weight given to facial-expression research (24.2%) reflects the broader maturity and popularity of this modality within the field, as discussed qualitatively in Section 2.2.1 of the associated literature review. Second, and more significant for the present research, EEG-focused and physiological-signal-focused references together account for only 13.6% of the corpus (9 of 66 references), and of the 16 references coded as "Multimodal Fusion," only one — Njoku et al., discussed in detail in Section 2.4 of the associated literature review — incorporates EEG jointly with video and/or audio modalities. This quantitative pattern directly corroborates, from a corpus-composition standpoint, the qualitative research gap identified in Chapter 1 and Chapter 2 of the associated thesis regarding the comparative scarcity of EEG-inclusive multimodal fusion research.

Modality / Topical Focus	No. of References	% of Corpus
Multimodal Fusion	16	24.2%
Facial	16	24.2%
Multimodal Dataset	8	12.1%
EEG	7	10.6%
Speech	6	9.1%
Text	6	9.1%
Survey	5	7.6%
Physiological	2	3.0%

Table 3: Distribution of the reviewed corpus by modality/topical focus (n = 66)

4.4 Keyword Frequency Analysis of Publication Titles

As a lightweight proxy for thematic/co-word analysis, the frequency of substantive terms occurring in publication titles was tabulated, excluding common stop-words and the corpus-defining terms 'emotion,' 'emotions,' and 'recognition,' which occur, by construction, in the great majority of titles and would otherwise dominate the analysis without adding discriminative information. Figure 5 and Table 4 present the fifteen most frequent remaining terms.

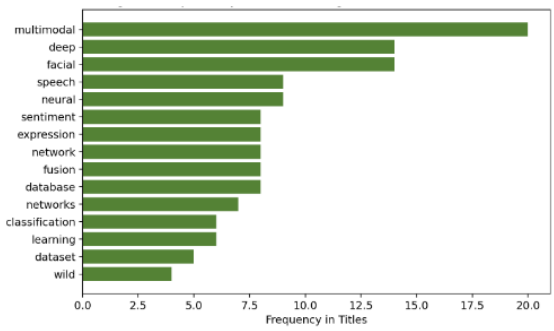

Figure 5: Top 15 keywords occurring in reference titles (n = 66)

The term ‘multimodal’ is the most frequent substantive keyword (appearing in 20 titles, 30.3% of the corpus), followed by ‘deep’ (14 titles) and ‘facial’ (14 titles), reflecting the corpus's dual emphasis on multimodal integration and deep-learning methodology. Terms directly related to fusion methodology — ‘fusion’ (8 titles), ‘network’/‘networks’ (15 titles combined), and architecture-specific terms such as ‘convolutional,’ ‘recurrent,’ ‘attention,’ ‘cnn,’ and ‘lstm’ — appear with moderate frequency, consistent with the corpus's focus on deep-learning-based fusion architectures. Notably, terms specific to the proposed research's methodological contribution — ‘entropy,’ ‘optimization,’ ‘metaheuristic,’ ‘bat,’ or ‘rider’ — do not appear at all among the title keywords of the reviewed corpus, a further quantitative indicator, consistent with the modality-distribution finding in Section 4.3, that entropy-based and metaheuristic-optimization-guided fusion remain unexplored territory within the specific literature underlying this research.

Keyword	Frequency (titles)
multimodal	20
deep	14
facial	14
speech	9
neural	9
sentiment	8
expression	8
network	8
fusion	8
database	8
networks	7
classification	6
learning	6
dataset	5
wild	4

Table 4: Top 15 keywords occurring in reference titles, by frequency

5. Discussion

5.1 Growth of the Field and Timing of the Proposed Research

The pronounced concentration of references in the 2017-2019 period (42.4% of the corpus, Section 4.1) confirms that the proposed PhD research is positioned within, and builds directly upon, the most active and recent phase of growth in deep-learning-based emotion-recognition research identified in this analysis. The tapering publication count evident for 2020-2022 in Figure 1 should not be read as a decline in the field's activity; rather, it most plausibly reflects the literature-search cut-off applied when compiling the synopsis reference list, and the corpus should accordingly be understood as a representative snapshot of the field through the point of literature search rather than a complete historical record.

5.2 Balance Between Journal and Conference Literature

The near-even split between journal (48.5%) and conference/workshop (43.9%) publications identified in Section 4.2 confirms that a rigorous review of this field must draw on both publication types; restricting a literature review to journal articles alone, as is sometimes done in bibliometric studies that rely solely on journal-indexed databases, would have excluded several of the most influential and frequently cited studies reviewed in Chapter 2 of the associated thesis, including the AAAI-published M3ER framework and the NAACL-HLT-published Conversational Memory Network, both discussed in Section 2.4 of the literature review.

5.3 Concentration in Affective-Computing-Specific Venues

The prominence of IEEE Transactions on Affective Computing and Information Fusion among the leading venues identified in Section 4.2.1 confirms the existence of a mature, dedicated publication ecosystem for this research area, which is a positive indicator for the field's methodological rigor and provides natural target venues for disseminating the outcomes of the proposed research once completed.

5.4 Modality Imbalance and the EEG-Inclusion Gap

The modality-distribution analysis presented in Section 4.3 provides the most direct quantitative support for the central research gap motivating this thesis. Facial-expression and general multimodal-fusion research together account for nearly half the reviewed corpus, while EEG-focused and physiological-signal-focused research together account for only 13.6%, and only one reference in the entire 66-item corpus combines EEG with video and/or audio within a single multimodal fusion framework. This corpus-level imbalance is consistent with, and quantitatively reinforces, the qualitative research-gap analysis presented in Section 2.10 of the associated literature review chapter, which identified the joint use of video, EEG, and physiological signals as one of the principal gaps addressed by the framework proposed in this thesis.

5.5 Absence of Entropy-Based and Optimization-Guided Fusion Terminology

The keyword-frequency analysis presented in Section 4.4 shows that terminology specific to the proposed research's core methodological contribution — Renyi-entropy-based fusion weighting and Bat-Rider Optimization — is entirely absent from the titles of the reviewed corpus. While the absence of a keyword from a curated 66-item corpus cannot be taken as proof that no such research exists anywhere in the wider literature, it is consistent with, and adds a further, independent line of quantitative evidence to, the qualitative fusion-strategy gap analysis presented in Section 2.7.4 of the associated literature review, which similarly found no reviewed study employing an explicit entropy-based, metaheuristically optimized fusion rule.

5.6 Scope and Limitations of the Study

This bibliometric analysis is subject to several limitations that should be considered when interpreting its findings.

I. Corpus size and selection: the analysis is based on a curated, manually compiled corpus of 66 references, rather than a systematic, exhaustive export from a bibliometric database such as Scopus or Web of Science. The corpus was assembled to support a specific PhD research synopsis and literature review, and, while broadly representative of the field's core sub-areas, it is not a comprehensive census of all published work on multimodal emotion recognition.

II. No citation-count or co-authorship analysis: because the corpus was not retrieved from a bibliometric database, citation counts, h-index figures, co-authorship networks, and bibliographic-coupling analysis — standard components of a full bibliometric study — could not be computed and are not reported here. Reporting such figures without genuine underlying database access would risk presenting fabricated statistics as fact, which this study has deliberately avoided.

III. Single-label modality coding: each reference was assigned to a single, primary modality/topical-focus category for clarity of presentation (Section 3.2); several references are genuinely multimodal or span more than one category, and a multi-label coding scheme might reveal a somewhat different, more finely grained distribution.

IV. Temporal cut-off: the apparent decline in publication count for 2021-2022 (Section 4.1) most likely reflects the point at which literature search for the associated synopsis was concluded, rather than a genuine decline in research activity in the field, and should not be interpreted as such.

V. Keyword analysis as a proxy: the title-keyword frequency analysis presented in Section 4.4 is a lightweight proxy for full co-word or topic-modelling analysis; it captures only terms that authors chose to include in their titles and does not account for synonymy, abstract-level content, or full-text keyword occurrence.

Notwithstanding these limitations, the analysis presented in this paper offers a transparent, fully reproducible, and honestly scoped quantitative complement to the qualitative literature review presented in Chapter 2 of the associated thesis, and every figure reported here can be independently recomputed from the same 66-reference corpus using the coding scheme described in Section 3.2.

6. Conclusion

This paper has presented a bibliometric-style analysis of the 66-reference corpus underlying a proposed AI-driven, deep-learning-based multimodal emotion-recognition framework for stress and depression detection. The analysis shows a pronounced growth in relevant publications between 2017 and 2019, a near-even balance between journal and conference publication venues, a leading role for affective-computing-specific venues such as IEEE Transactions on Affective Computing and Information Fusion, a corpus composition weighted heavily toward facial-expression and general multimodal-fusion research relative to EEG-inclusive multimodal research, and a complete absence of entropy-based or metaheuristic-optimization-related terminology among reviewed publication titles. Each of these quantitative findings independently corroborates the qualitative research gaps identified through close reading of the literature in Chapter 2 of the associated PhD thesis, strengthening the evidence base for the novelty and motivation of the EEG-inclusive, Renyi-entropy-based, Bat-Rider-Optimization-tuned multimodal fusion framework proposed in that thesis. Future extensions of this analysis, contingent on institutional access to a licensed bibliometric database, could expand the corpus to a full, systematically retrieved dataset and incorporate citation-count, co-authorship, and keyword co-occurrence network analysis using dedicated bibliometric software such as VOSviewer or the Bibliometrix/Biblioshiny package in R, which would allow the trends identified here to be validated against the wider published literature beyond the specific corpus curated for this research.

References

[1] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, “On the robustness of speech emotion recognition for human-robot interaction with deep neural networks,” 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018.

[2] M. Chen, X. He, J. Yang, and H. Zhang, “3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition,” IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[3] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” Biomedical Signal Processing and Control, vol. 47, pp. 312–323, Jan. 2019.

[4] M. Pagé Fortin and B. Chaib-draa, “Multimodal multitask emotion recognition using images, texts and tags,” Proceedings of the ACM Workshop on Crossmodal Learning and Application, 2019.

[5] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” Neurocomputing, vol. 174, pp. 50–59, Jan. 2016.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of Depression-Related Posts in Reddit Social Media Forum,” IEEE Access, vol. 7, pp. 44883–44893, 2019.

[7] K. Zhang, Y. Huang, Y. Du, and L. Wang, “Facial expression recognition based on deep evolutional spatial-temporal networks,” IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4193–4203, 2017.

[8] A. Talipu, A. Generosi, M. Mengoni, and L. Giraldi, "Evaluation of deep convolutional neural network architectures for emotion recognition in the wild," 2019 IEEE 23rd ISCT, 2019.

[9] B. Taha and D. Hatzinakos, "Emotion recognition from 2D facial expressions," 2019 IEEE CCECE, 2019.

[10] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," Pattern Recognition Letters, vol. 120, pp. 69–74, 2019.

[11] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using Deep Belief Networks," 2014 IEEE ICME, 2014.

[12] K. Guo, H. Candra, H. Yu, H. Li, H. T. Nguyen, and S. W. Su, "EEG-based emotion classification using innovative features and combined SVM and HMM classifier," 2017 IEEE EMBC, 2017.

[13] S. D. Kumar and D. P. Subha, "Prediction of depression from EEG signal using Long short term memory (lstm)," 2019 ICOEI, 2019.

[14] P. Pandey and K. R. Seeja, "Subject-independent emotion detection from EEG signals using Deep Neural Network," ICICC, pp. 41–46, 2018.

[15] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," Proc. 5th Int. Workshop on Audio/Visual Emotion Challenge, 2015.

[16] S. E. Kahou et al., "Emonets: Multimodal deep learning approaches for emotion recognition in video," Journal on Multimodal User Interfaces, vol. 10, no. 2, pp. 99–111, 2015.

[17] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," IEEE JSTSP, vol. 11, no. 8, pp. 1301–1309, 2017.

[18] S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, "Multimodal relational tensor network for sentiment and emotion classification," Proc. Challenge-HML, 2018.

[19] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," Proc. NAACL-HLT, 2018.

[20] M. M. Hassan et al., "Human emotion recognition using deep belief network architecture," Information Fusion, vol. 51, pp. 10–18, 2019.

[21] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning 'bert-like' self supervised models to improve multimodal speech emotion recognition," Interspeech 2020, 2020.

[22] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," ICASSP 2020, 2020.

[23] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," Proc. AAAI, vol. 34, no. 02, pp. 1359–1367, 2020.

[24] J. Njoku, A. C. Caliwag, W. Lim, S. Kim, H.-J. Hwang, and J.-W. Jeong, "Deep Learning Based Data Fusion Methods for Multimodal Emotion Recognition," J. Korean Inst. Comm. Info. Sciences, vol. 47, no. 1, pp. 79–87, 2022.

[25] R. Verma, "FER2013," Kaggle, 2018.

[26] S. R. Livingstone, "Ravdess emotional speech audio," Kaggle, 2019.

[27] S. S. Mondal, "Emotion data for Whatsapp Status," Kaggle, 2020.

[28] J. Bird, "EEG Brainwave Dataset: Feeling emotions," Kaggle, 2018.

[29] Jerry Joy and Aparna Kannan, "Speech Emotion Recognition using Neural Network and MLP Classifier," IJESC, April 2020.

[30] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," 2016 Int. Conf. on Cyberworlds (CW), 2016.

[31] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from Physiological Signal Analysis: A Review," Electronic Notes in Theoretical Computer Science, vol. 343, pp. 35–55, 2019.

[32] W. Wei, Q. Jia, Y. Feng, and G. Chen, "Emotion recognition based on weighted fusion strategy of Multichannel Physiological Signals," Computational Intelligence and Neuroscience, vol. 2018, pp. 1–9, 2018.

[33] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," Information Fusion, vol. 53, pp. 209–221, 2020.

[34] Y. Wang et al., "A systematic review on Affective computing: Emotion models, databases, and recent advances," Information Fusion, vol. 83-84, pp. 19–52, 2022.

[35] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq and S. Zeebaree, "Multimodal emotion recognition using deep learning," J. Applied Science and Technology Trends, vol. 2, no. 2, pp. 52-58, 2021.

[36] Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan, "Advances in multimodal emotion recognition based on brain–computer interfaces," Brain Sciences, vol. 10, no. 10, p. 687, 2020.

[37] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and taxonomy," IEEE TPAMI, vol. 41, no. 2, pp. 423–443, 2019.

[38] C. Zucco, B. Calabrese and M. Cannataro, "Emotion Mining: From Unimodal to Multimodal Approaches," Springer, 2021.

[39] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," Knowledge-Based Systems, vol. 161, pp. 124–133, 2018.

[40] M. F. Siddiqui and A. Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," Multimodal Technologies and Interaction, vol. 4, no. 3, p. 46, 2020.

[41] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," IEEE TAC, vol. 2, no. 4, pp. 206–218, 2011.

[42] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, 1998.

[43] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+)," 2010 IEEE CVPRW, 2010.

[44] T. Kanade, J. F. Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2002.

[45] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," Image and Vision Computing, vol. 29, no. 9, pp. 607–619, 2011.

[46] X. Zhang et al., "BP4D-spontaneous: A high-resolution spontaneous 3D Dynamic Facial Expression Database," Image and Vision Computing, vol. 32, no. 10, pp. 692–706, 2014.

[47] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," 2018 IEEE/CVF CVPR, 2018.

[48] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[49] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions," 2011 IEEE ICCVW, 2011.

[50] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," 2016 IEEE CVPR, 2016.

[51] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, Valence, and arousal computing in the wild," IEEE TAC, vol. 10, no. 1, pp. 18–31, 2019.

[52] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," Interspeech 2005, 2005.

[53] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast Induced Natural Emotion Database," IEEE TAC, vol. 3, no. 1, pp. 32–41, 2012.

[54] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," ACL 2007, pp. 440–447, 2007.

[55] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," Proc. 25th ICML, 2008.

[56] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," Expert Systems with Applications, vol. 110, pp. 298–310, 2018.

[57] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," EMNLP, 2013.

[58] S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," IEEE TAC, vol. 3, no. 1, pp. 18–31, 2012.

[59] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," 2013 6th Int. IEEE/EMBS Conf. on Neural Engineering (NER), 2013.

[60] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Trans. Autonomous Mental Development, vol. 7, no. 3, pp. 162–175, 2015.

[61] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," IEEE TAC, vol. 12, no. 2, pp. 479–493, 2021.

[62] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," Proc. 20th ACM ICMI, 2018.

[63] C. Busso et al., "IEMOCAP: Interactive emotional Dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[64] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis," Proc. 13th Int. Conf. on Multimodal Interfaces, 2011.

[65] M. Wollmer et al., "YouTube movie reviews: Sentiment analysis in an audio-visual context," IEEE Intelligent Systems, vol. 28, no. 3, pp. 46–53, 2013.

[66] A. A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the wild: CMU-Mosei dataset and interpretable dynamic fusion graph," Proc. 56th Annual Meeting of the ACL, 2018.