Journal of Information Systems Engineering and Management
2025, 10(16s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Stress Detection using Multimodal Representation Learning, Fusion Techniques, and Applications

[1,2*]Yogesh J. Gaikwad, [3]Kalyani Kadama, [4]Rutuja Rajendra Patil, [5]Dr. Gagandeep Kaur
[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune, Maharashtra, India· phdgrad.yogesh.gaikwad@siu.edu.in*
[2]Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, 411038, Maharashtra, India, yogesh.gaikwad@mitwpu.edu.in
[3]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune, Maharashtra, India, kalyanik@sitpune.edu.in
[4]Computer Science and Engineering-Artificial Intelligence & Machine Learning Department, Vishwakarma Institute of Information Technology, Pune, 411037, Maharashtra, India, rutujapat@gmail.com
[5]Assistant Professor, Symbiosis Institute of Technology, Nagpur, gagandeep.kaur@sitnagpur.siu.edu.in

ARTICLE INFO

Received: 10 Dec 2024

Revised: 28 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

The fields of speech recognition, image identification, and natural language processing have undergone a paradigm shift with the advent of machine learning and deep learning approaches. Although these tasks rely primarily on a single modality for input signals, the artificial intelligence field has various applications that necessitate the use of several modalities. In recent years, academics have placed a growing emphasis on the intricate topic of modelling and learning across various modalities. This has attracted the interest of the scientific community. This technical article provides a comprehensive analysis of the models and learning methods available for multimodal intelligence. Specifically, this work concentrates on the fusion of video and language processing modalities, which has become a crucial area in both computer vision and natural language research. In this article, we explore recent research on multimodal deep learning from three different perspectives: learning multimodal representations, combining multimodal inputs at different levels, and multimodal applications. Regarding the learning of multimodal representations, the article delves into the concept of embedding, which involves the combination of different types of signals into a unified vector space. This enables cross-modal signal processing, which has significant implications for various applications. Moreover, several forms of embedding created and trained for common downstream tasks are examined. Regarding multimodal fusion, the research focuses on specific designs that merge representations of unimodal inputs for a specific purpose.

Keywords: Multimodality, representation, multimodal fusion, deep learning, speech, vision, text-to-image generation, visual question answering, visual rezoning, and visual reasoning.

1. Introduction

The field of machine learning has advanced significantly in the last few years, primarily due to the rapid progress of deep learning algorithms [1]- [6]. A remarkable achievement in 2010 was the substantial improvement in the precision of large-scale automatic speech recognition using fully connected deep neural networks (DNNs) and deep auto-encoders [7]-[17]. The field of computer vision (CV) has witnessed significant advancements in recent years, especially with the use of deep convolutional neural network (CNN) models [18]. These models have achieved major breakthroughs in large-scale picture categorization [19]-[22] and large-scale object recognition [23]-[25], while relying solely on a single input modality. In natural language processing (NLP), recurrent neural network (RNN)-based semantic slot filling approaches [26] have set a new standard for spoken language comprehension. Furthermore, the integration of attention mechanisms in RNN-encoder-decoder models [27], also known as sequence-to-sequence models [28], has resulted in remarkable end-to-end machine translation outcomes [29], [30]. In NLP, particularly in question answering (QA) for machine reading ability, generative pre-training has been shown to set the standard for cutting-edge performance when dealing with limited training data [31-33]. This approach involves unsupervised training or self-training, where parameters are transferred from a pre-trained language model (LM) on a large out-of-domain dataset, followed by fine-tuning on smaller in-domain datasets. Despite significant

progress in vision, speech, and language processing, many artificial intelligence problems require multiple input modalities, such as smart personal assistant mechanisms that need to recognize interpersonal interactions based on spoken words, body posture, and pictorial languages [34]. Therefore, there is a growing interest in researching modeling and training techniques that can handle a wide range of modalities [35].

Due to the progress in image processing and language comprehension [36], tasks that involve combining images and text have gained significant attention. These tasks include interpreting references made in videos and localizing words [37]-[39], as well as generating captions for images and videos [40]-[45], visual question-answering (VQA) [46]-[48], generating images from text [49]-[51], and navigating through visual and linguistic inputs [52]. The understanding of natural language plays a critical role in these tasks by helping machines "comprehend" the content of images. This involves capturing the underlying semantic connections between the language and visual data derived from the images. In addition to text, sound speech recognition [53]-[55], speaker recognition [56]-[58], speaker diarization [59], [60], speech separation [61], [62], and augmentation [63] can also be performed using visual and voice inputs.

This article offers an academic summary of the models and training methodologies used for multimodal intelligence. To give a clear viewpoint, the review is organized around three key elements: representations, fusion, and applications. The purpose is to provide readers with an orderly and thorough grasp of the subject matter. The paper's content is unique and does not contain any plagiarized material.

Deep learning has emerged as a critical method for extracting representations from raw data. Obtaining concurrent data across many senses might be difficult in the case of multimodal activities. Pre-trained representations with certain qualities, such as those ideal for zero-shot or few-shot learning, may help to solve this difficulty. Using these representations to solve this problem has shown to be an effective method. This article examines all supervised and unsupervised training-based multimodal representation learning approaches. The objective is to find the most effective ways for learning and applying those representations in a variety of contexts. To maintain academic integrity, it is critical to guarantee that the representations employed are both correct and free of plagiarism.

The integration of representations from several modalities is a basic difficulty in any multimodal job. We separate relevant research based on the procedures employed during fusion instead of the fusion stage in a method to classify them. This technique is required since recent complicated approaches make stage-based classification challenging.

Our examination centers around three applications: picture captioning, text-to-image creation, and video quality assurance (VQA). These examples show how representation learning and fusion may be applied to specific tasks and provide insight into the current state of multimodal applications, particularly those using natural language and vision. We also look at visual reasoning techniques.

2. Research Strategy

2.1 Source and Methods

Bibliometrics refers to a widely-used set of quantitative tools for evaluating academic literature and scholarly communication in research evaluation metrics [38,39]. In our study, we have conducted extensive research and analyzed a vast amount of scientific data using bibliometric analysis. This approach has enabled us to generate high-impact research, obtain a one-stop summary, identify knowledge gaps, and generate unique ideas for further research [40]. Using fundamental measures in bibliometric analysis, we analyzed published research and retrieved data on the most active or noteworthy scholars and their organizations, collaboration patterns, widely used keywords, and a variety of publications on the subject. This necessary material is available from reputable archives such as "Scopus" and "Web of Science." Scopus [41] is a major peer-reviewed abstract and citation database launched by Elsevier in 2004.

2.2 Data Selection and Extraction

The search strategy employed to obtain results from the Scopus database were shown in table 2. It is crucial to include appropriate and relevant terms when conducting a literature search. In this study, the primary keyword used was " Stress Detection" while two additional keywords, "Multimodal" and "Fusion" were identified from abstracts. "Stress Detection" was the sole phrase given to the keyword section of research publications. The table below contains the detailed query. Among the entire outcomes obtained, all findings up to 2023 were evaluated, and relevant research

papers were discovered from databases, as shown in Figure 4. For the analysis, only journals, conferences, and review publications were considered. The Scopus search yielded a total of 36 results.

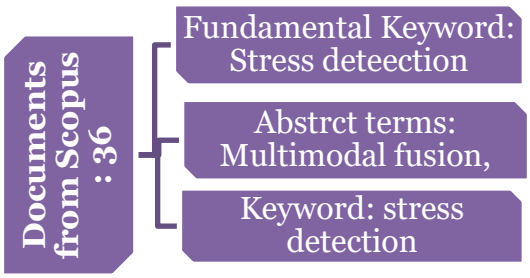

Figure 1. Search Strategy.

In order to gather the required information for the analysis, different metadata pertaining to the research papers were gathered, which included the title of the article, year of publication, source, citation count, author's name, author's keywords, cited references, organization, and country. It should be noted that the data utilized in the analysis was obtained on April 23, 2023. A summary of the search approach utilized in this study can be found in Table 1, which outlines the important keywords employed to identify pertinent research papers.

Table 1. List of Keywords used

Keyword	Occurrences	Total link strength
Multi-modal	13	52
Multi-modal fusion	9	44
Sentiment analysis	8	44
Multimodal fusion	9	43
Multimodal sentiment analyse	6	40
Multimodal sentiment analysis	6	40
Affective computing	7	32
Emotion recognition	6	31
Stress detection	12	27
Long short term memory	7	25

Query in Scopus:

TITLE-ABS-KEY (multimodal AND fusion AND stress AND detection) AND (LIMIT-TO (PUBYEAR, 2023) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019))

Data Analysis Procedure

Information presented in graphical form is often more comprehensible and facilitates analysis, conclusion-making, and prediction, among other benefits. This article uses popular software tools, such as VosViewer [45], Gephi [46], and BibExel[47] for performing bibliometric analysis of deep neural network approaches utilized for ISR. These tools are commonly utilized to represent multi-dimensional data through graphical visualization. VoSViewer is an extensively used bibliometric analysis visualization tool that enables the creation of multiple networks based on keywords, citations, publication sources, authors, co-authors and other factors. All entities are represented as circles and are linked to one another by linkages. The distance between the items represents the relationship between them, with the closer entities having shorter distances. Additionally, Gephi, a popular graphical clustering tool, is also utilized in this study.

The application is cross-platform and employs the OpenGL 3D engine to allow arranging of data based on various parameters such as scale, qualities, classification etc. BibExel is a free software created by Olle Persson, an

information scientist, exclusively for non-commercial educational purposes. It is a helpful tool for scholars looking to conduct bibliometric analysis. The analysis in this study is split into two main categories: quantitative and qualitative. The data utilized in the analysis is sourced from the reliable Scopus database, ensuring its credibility.

The following quantitative analysis is conducted using several parameters:

- Examining documents by year of publication
- Conducting a citation-based analysis
- Identifying top keywords using Scopus and Web of Science
- Analyzing document types
- Examining publication data by geographical location
- Analyzing publication sources
- Conducting co-occurrence analysis of author keywords.

This study focuses on Stress detection using multimodal fusion and their qualitative analysis, which will cover their historical background, available datasets for research, proposed deep learning-based techniques, and performance evaluation metrics.

3. Quantitative Analysis

3.1 Analysis of documents by year

In the 1990s, stress detection methods were primarily based on subjective measures such as self-report questionnaires and interviews. These methods relied on individuals reporting their own experiences of stress, including symptoms such as increased heart rate, sweating, and muscle tension. Today, stress detection has evolved to incorporate more objective measures, such as physiological monitoring and behavioural tracking. Physiological measures include heart rate variability, cortisol levels, and skin conductance, which can be used to measure stress responses in the body. Behavioural tracking involves monitoring changes in behavior, such as sleep patterns, exercise habits, and social interactions, which may be indicative of stress. Advancements in technology have also made it possible to monitor stress levels in real time, using wearable devices such as smartwatches or fitness trackers, which can track physiological and behavioural changes throughout the day. Additionally, machine learning algorithms can be used to analyze these data streams and provide insights into an individual's stress patterns and triggers. As shown in Figure 4, there has been a steady growth in the number of publications in this field since 2019.

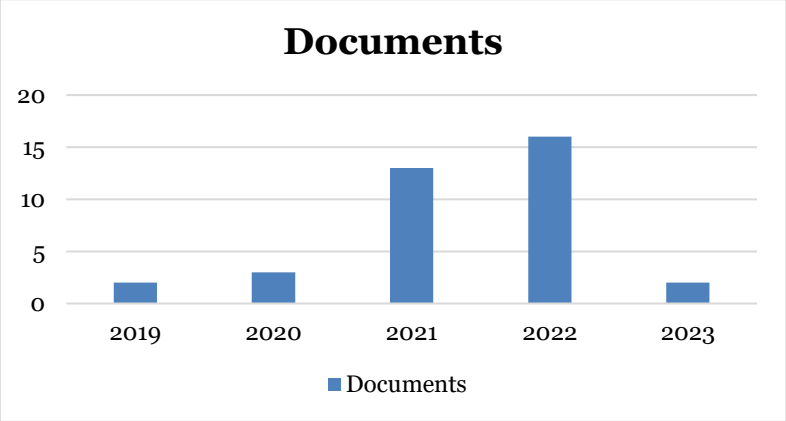

Figure 4. Comparative analysis of publications per year.

3.2 Citation Based Analysis

The published document includes several citations that highlight the significance and relevance of the text's solution to the problem at hand. Table 2 provides a detailed analysis of the number of citations received by publications listed in the Scopus databases, organized by year. The data shows a consistent increase in the number of citations since 2019, indicating that substantial work is being carried out globally.

Table 3. Year wise citation analysis.

Year	2019	2020	2021	2022	2023
Scopus Citation	0	12	42	140	44

4. Research Virtue

4.1 Top 10 keywords extracted from Scopus database

As shown in Figure 7. The most frequently occurring keyword is Deep Learning, with a count of 39, followed by other keywords such as Artificial Intelligence and Machine Learning. Interestingly, Unmanned Aerial Vehicle is also among the top 10 keywords, ranked at number five with a count of 15. To gain further insights, we will conduct co-occurrence analysis on all available documents in this domain, focusing on the author keywords.

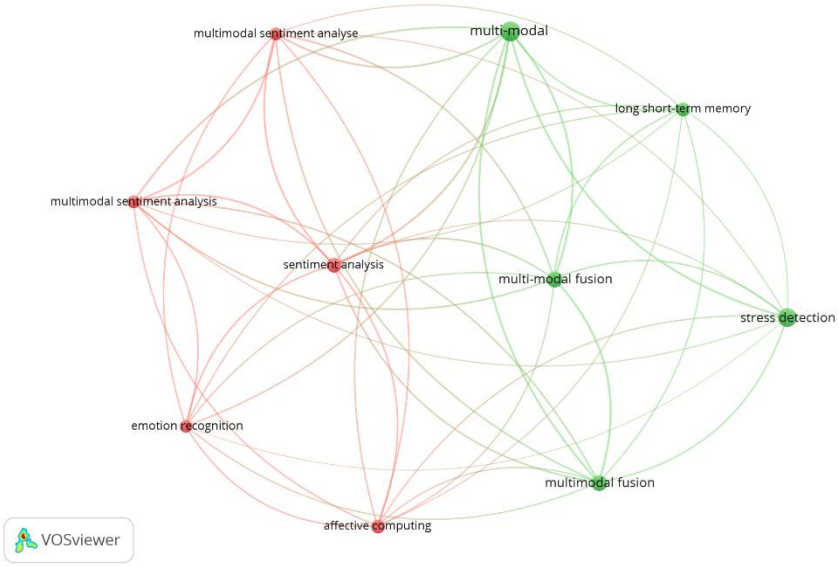

Figure 5. Top keywords used in Scopus.

4.2 Analysis of document type

Table 5 contains information on documents published in the subject of Stress detection with multimodal fusion. In all, 121 articles have been published in Scopus indexed events. Figure 8 depicts the full distribution of publication categories.

Based on Figure 8, it can be observed that over 50% of all documents are released in journals, highlighting their significant contribution to the publication industry. Another crucial category for publications is journal/articles. However, it is worth noting that there are limited survey articles available on the topic. However, there is no bibliometric study in the topic of Stress detection with multimodal fusion. One of the latter parts discusses a comprehensive review of the publication's sources and their citation count.

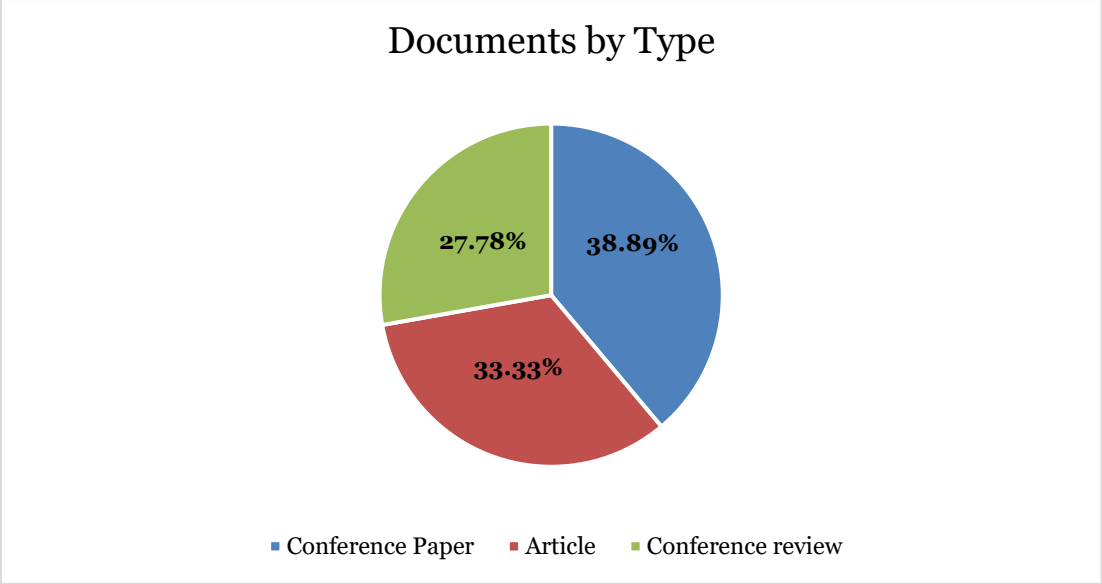

Figure 6. Category of publication.

Table 4. Publication count by type of document

Type of Publication	Scopus
Conference Paper	38.89%
Article	33.33%
Conference Review	27.78%
Total	**100**

4.3 Analysis of Geographical area

Research documents released by various countries and regions can shed light on the ongoing research efforts. Figure 9 presents a breakdown of the number of documents published in Scopus-indexed publications worldwide, organized by country. China leads the list with 54 Scopus-indexed articles, followed by the United States. India is ranked on the top of the list.

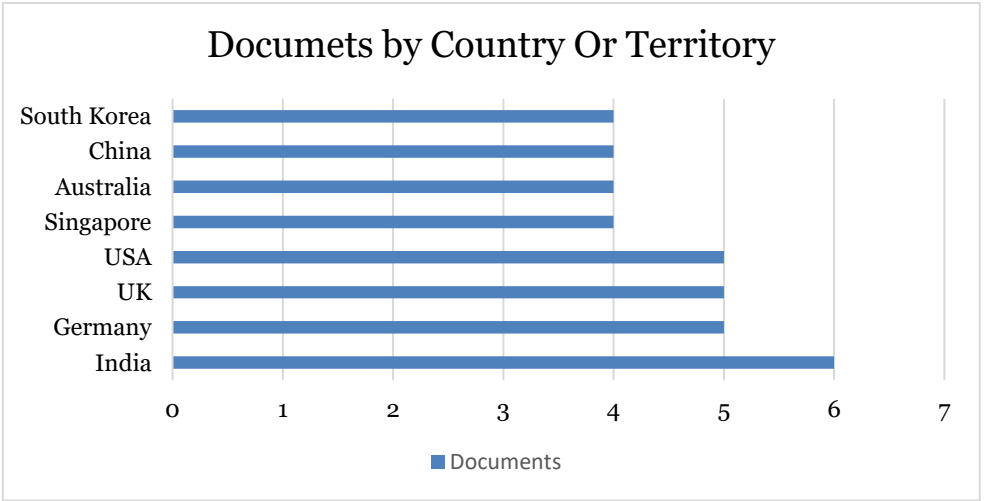

Figure 7. Analysis of Geographical Area (Country wise)

Table 5 displays a categorization of the leading eight countries with the greatest volume of publications in the field. This breakdown is the result of a thorough assessment of all accessible papers.

Country	Territory Documents
India	06
United States	05
Germany	05
United Kingdom	05
Singapore	04
Australia	04
China	04
South Korea	04

4.4 Analysis of Publication's by source

Figures 11 provide a visual representation of the primary sources of Scopus-indexed articles. LNCS is the predominant field of publication, with the majority of researchers choosing to publish their work through Springer Nature. Additionally, the Computer Vision Foundation (CVF), a non-profit organization, is a significant publisher of research focused on picture super resolution. The analysis encompasses all available papers within the domain.

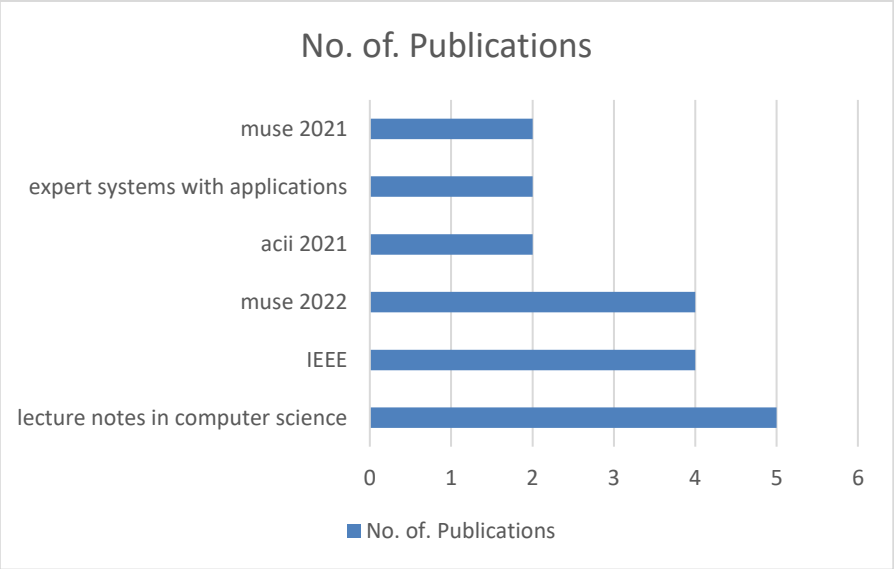

Figure 8. Publishers in Scopus.

4.5 Co-occurrence analysis (Author keywords)

The results of a co-occurrence analysis of author keywords extracted from various documents retrieved from Scopus databases are presented in Figure 13. The analysis reveals that "Multimodal fusion" and "Stress detection" are the most frequently occurring keywords. Additionally, commonly used terms include "Emotion recognition," " Sentiment analysis," and other related phrases.

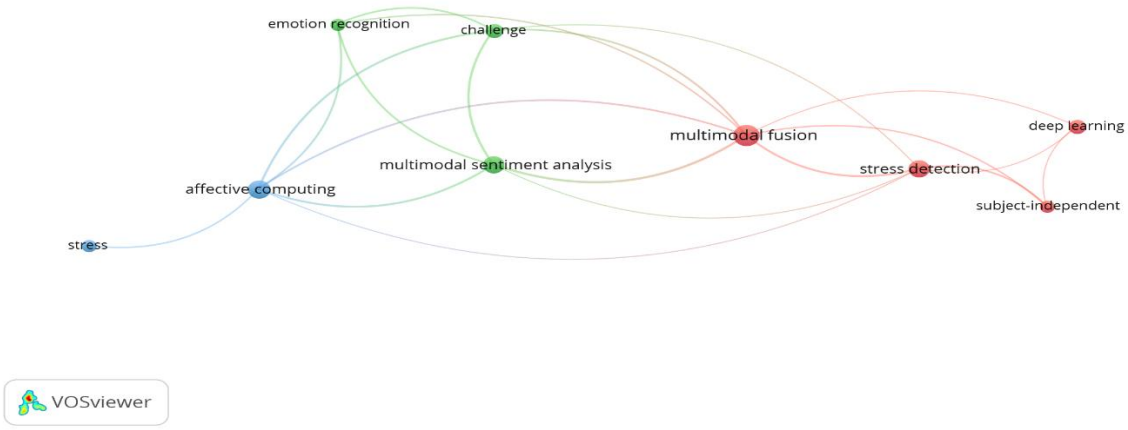

Figure 09. Co-occurrence analysis (Author keywords)

Table 6 presents comprehensive information regarding the keywords and their respective Total Link Strength (TLS) values. TLS and link strength are two parameters that carry different weights, utilized to evaluate the potency of co-authorship links connecting the author and their associates. The TLS value is a crucial metric that provides an overall assessment of the researcher's collaborative strength within their peer network.

Table 6. Keywords and their respective Total Link Strength (TLS)

Keyword	Occurrences	Total Link Strength (TLS)
Multi-Modal	13	71
Stress Detection	12	50
Multi-Modal Fusion	9	58
Multimodal Fusion	9	54
Sentiment Analysis	8	51
Stresses	8	30
Affective Computing	7	36
Deep Learning	7	27
Long Short-Term Memory	7	37
Electrocardiography	6	25
Emotion Recognition	6	37
Multimodal Sentiment Analyse	6	46
Multimodal Sentiment Analysis	6	46
Physiology	6	27
Convolutional Neural Network	5	22
Forecasting	5	27

4.6 Citation analysis of documents

The number of citations in a publication is a good indicator of its impact in the field. To identify the most significant publications, co-citation analysis can be used. A detailed analysis of document citations is presented in Figure 14 and Table 9. These findings can help to understand the most influential works in the domain.

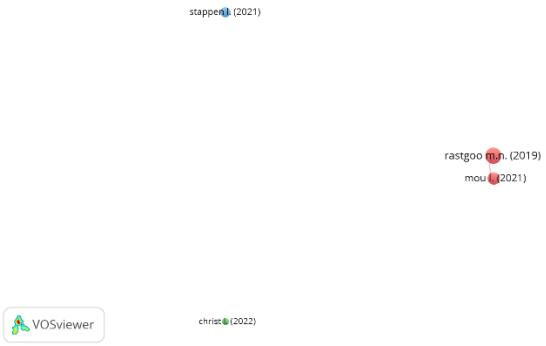

Figure 10. Citation analysis of documents.

Table 7. Top 14 documents with highest citations.

Author	Citations	Total Link Strength
rastgoo m.n. (2019)	73	1
mou l. (2021)	43	1
stappen l. (2021)	27	0
christ l. (2022)	14	0

This analysis takes into account all research publications from both databases. The author who has received the most citations is rastgoo m.n. (2019), with a total of 73 citations. A less number of documents and significant number of citations highlight the need for further research in the field of stress detection using Multimodal fusion. Therefore, this analysis is conducted on all available papers in this domain.

4.7 Citation analysis of source

Detailed information on the sources of publication for papers in the field of study is provided in Figure 15 and Table 10. An analysis of document types reveals that approximately 50% of articles are published in conferences, with most papers appearing in conference proceedings. Notably, IEEE transactions have published the largest number of documents in this domain, with seven articles to their credit. The next prominent source is "Lecture Notes in Computer Science," which has published six papers. Conferences associated with IEEE, such as IEEE Potentials and IEEE Access, are popular sources for publishing research in this field. It's worth mentioning that this analysis is based on all available documents in this domain.

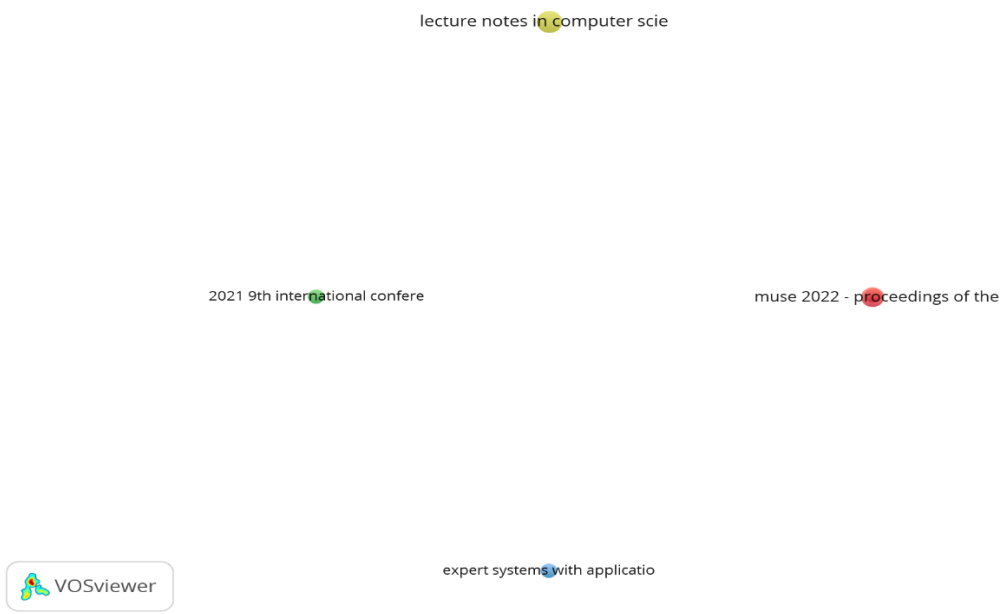

Figure 11. Citation analysis of Source.

Table 08. Citation analysis by source.

Source	Documents	Citations	Total Link Strength
lecture notes in computer science	5	0	0
muse 2022	4	14	1
2021 9th international conference on affective computing and intelligent interaction, acii 2021	2	2	0
expert systems with applications	2	116	0
muse 2021 - proceedings of the 2nd multimodal sentiment analysis challenge, co-located with acm mm 2021	2	30	1
lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)	5	0	0

4.8. Citation Analysis of Author

According to Table 9, Chen H., He Z., Shi B., and Zhong T. are the authors with the highest number of publications and the most cumulative citations, totaling 41. This detailed analysis was conducted on all available documents in the domain to identify the most prolific and impactful authors.

Table 09. Citation analysis by author.

Author	Citations	Total Link Strength
rastgoo m.n. (2019)	73	1
mou l. (2021)	43	1
stappen l. (2021)	27	0
christ l. (2022)	14	0

4.9. Bibliographic Coupling of Documents

The term "bibliographic" indicates that when two texts share references, they must also share technical substance. A detailed study of bibliographic coupling for all publications is presented in Figure 12 and Table 10. Among the authors with the most links, christ l. (2022), stappen l. (2021a), Shi B., have a total of 51 links.

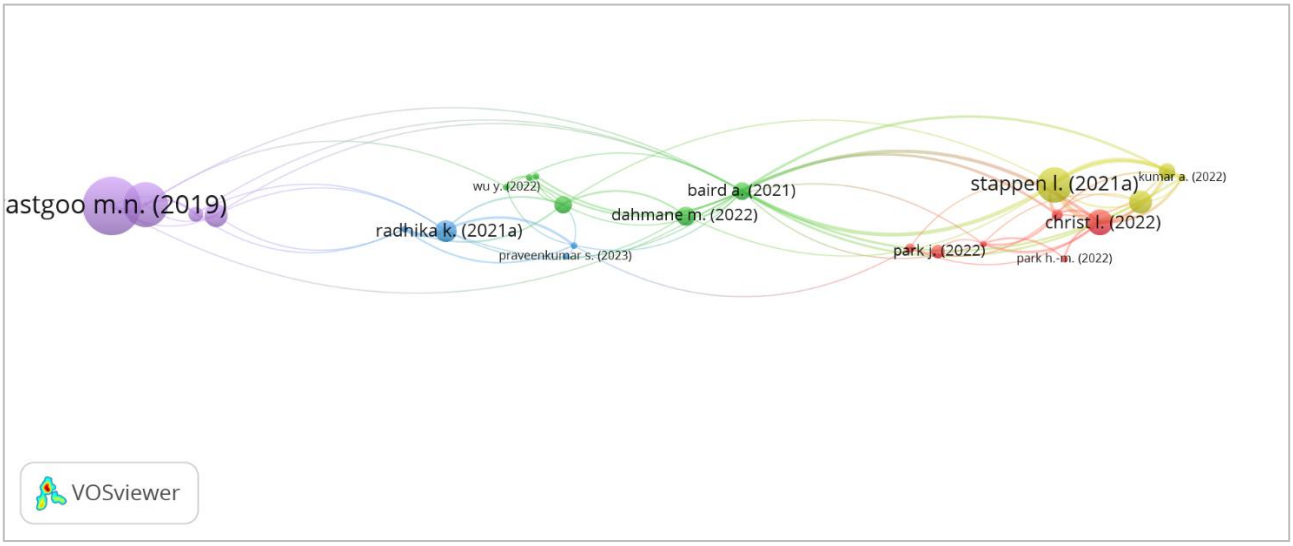

Figure 12. Bibliographic analysis of documents.

Table 10. Bibliographic analysis of documents.

Document	Citations	Total Link Strength
christ l. (2022)	14	51
stappen l. (2021a)	27	51
baird a. (2021)	7	44
hamieh s. (2021)	3	24
stappen l. (2021b)	6	24
kumar a. (2022)	0	21
mou l. (2021a)	43	21
radhika k. (2021a)	10	20
amiriparian s. (2022)	11	17
rastgoo m.n. (2019)	73	17
radhika k. (2021b)	1	16

li j. (2022)	0	13
yan m. (2022)	7	12
kuttala r. (2023)	0	11

I. EXISTING MACHINE LEARNING TECHNIQUES

Table 2. Several Existing Machine Learning Techniques.

ML Techniques	Advantages	Disadvantages	Applications
Fuzzy logic [22], [25]	This approach doesn't require precise inputs and uses a small amount of memory since the code is compact. It's capable of solving complex problems and has a simple structure that makes it easy to construct. Additionally, it can mimic human thought processes, which is especially beneficial for uncertain or approximate reasoning. Overall, it's a highly useful method for a variety of tasks.	The accuracy of a fuzzy logic control system can be compromised by inaccurate data and inputs. To maintain accuracy, it is essential to update the system's rules regularly. Validation and verification testing with hardware is also necessary to ensure the system's performance. Fuzzy logic systems are highly dependent on human knowledge and skill, which can limit their acceptance due to inaccuracies in results.	The areas of focus include medicine, defense, transportation systems, industry, naval control, auto transmission, fitness management, and washing machines.
K-Nearest Neighbour (KNN) [20], [23]-[24], [26],[34], [38], [45]-[50].	The algorithm can be used for both regression and classification tasks, and is adept at identifying outliers. Additionally, it is not limited by the requirement that classes be linearly separable, and has been shown to provide high levels of accuracy. Furthermore, the algorithm is straightforward to interpret and implement, and does not make any assumptions about the underlying data.	High-dimensional data is not a suitable fit for certain methods due to their wide time and space complexities, as well as the costly testing of each instance. Additionally, these methods may produce less meaningful distance numbers due to their sensitivity to noisy or irrelevant attributes.	Recommendation systems Semantic document search Credit card fraud detection Banking systems Political science Economic forecasting
Support Vector Machine (SVM) [20], [26],[35]-[40], [42],[45]-[50], [53]-[57], [60]-[62], [64], [65]	This method effectively handles unstructured and semi-structured data and has the ability to solve complex problems using an appropriate kernel function. It is capable of scaling high-dimensional data and has a lower risk of over-fitting. Additionally, it is more	Training large datasets takes a considerable amount of time, and the results may not always be transparent. Additionally, the performance of the model may suffer in noisy environments.	Handwriting and text recognition, the inverse geosounding problem, facial expression classification, speech recognition, steganography detection in digital images, cancer diagnosis and prognosis, and intrusion detection.

	memory efficient in comparison to other methods.		
Logistic Regression [18],[21], [40], [47],[49], [61]-[62]	The method is easy to implement and understand. It does not make any assumptions about the distributions of classes in feature space, and can be easily extended to handle multiple classes. It can quickly classify unknown records and performs well on linearly separable datasets. Furthermore, it is less likely to overfit, and regularization can be used to prevent this issue. Additionally, the method can provide great training efficiency in some cases with low computation power.	Linear boundary construction may result in overfitting when the number of observations is lower than the number of features. It can only be used to predict discrete functions, and it requires a lack of multicollinearity or low average multicollinearity between independent variables. Linear decision surfaces cannot solve non-linear problems. It's challenging to capture complex relationships with linear boundaries, and they are sensitive to outliers.	online credit card transactions, email spam detection, credit scoring, medicine, text editing, hotel booking, and gaming.
Naïve bayes [23]-[24], [35], 47], [54], [59], [61]	The method is quicker, capable of handling multi-class prediction problems, and more appropriate for categorical input variables than numerical ones. Additionally, it requires significantly less training data, provided that its assumptions regarding the independence of features remain valid.	To address the zero-frequency issue, it is important to note that estimations may be inaccurate in certain instances. As a result, the probability outputs may be less dependable, particularly since all predictors are seldom independent in real-world scenarios.	Text classification, recommendation system, sentiment analysis.
Principal Component Analysis (PCA) [20], [26], [49], [65]	Efficiently removing correlated features can improve algorithm performance, reduce overfitting by decreasing the number of features, and enhance data visualization by transforming high-dimensional data into low-dimensional data.	To avoid losing information and making independent variables less interpretable, it is necessary to standardize the data before applying PCA.	facial recognition, spike-triggered covariance analysis in neuroscience, image compression, detection and visualization of computer network attacks, quantitative finance, anomaly detection, and medical data correlation.

Ensemble methods [23], [56]	Overfitting can be avoided by creating a robust and stable model that can accurately predict outcomes. Such a model should be able to capture both linear and non-linear relationships within the data, thereby enabling it to handle the various requirements of complex problems. This requires developing hypotheses that are tailored to the specific needs of the problem.	The complexity of the model has increased, which has resulted in a reduction in its interpretability. Additionally, the design and computation time required for this model are high, making it unsuitable for real-time applications.	emotion recognition, medicine, financial decision-making, computer security, remote sensing, fraud detection, and face recognition.
Decision Trees [24],[35],[37]-[38],[59]	The output of this method is user-friendly and can be applied for classification and regression tasks, accommodating both continuous and categorical variables. It employs a rule-based approach that eliminates the need for feature scaling and can handle missing values and outliers automatically. Additionally, it has a short training period.	Overfitting occurs when a model is not suitable for large datasets, as even small amounts of noise can cause instability and result in incorrect predictions. This can lead to high variance in outputs, causing numerous errors in final estimations.	Demographic data is utilized in various fields such as client prospecting, business and customer relationship management, engineering, fraud detection, energy consumption analysis, healthcare management, and fault diagnosis.
Random forest [23],[49],[54], [61]-[62],[66]	This method ensures accuracy via cross-validation, mitigates overfitting in decision trees, handles categorical and continuous values, and is applicable for classification and regression. It also automatically deals with missing data and uses a rule-based approach, eliminating the need for data normalization or feature scaling.	Building multiple decision trees to combine their outputs requires significant computational power and resources, which in turn increases training time. Additionally, ensembles of decision trees can suffer from interpretability issues and fail to accurately determine the impact of individual variables.	Banking, healthcare, stock market, and e-commerce are distinct sectors.
Artificial Neural Network (ANN) [37], [53], [71]	The network stores information throughout, not just in a database. It can produce output even with incomplete data after training, and is fault-tolerant and has distributed memory. Additionally, it can handle multiple tasks	The network's functioning depends on the hardware and is unexplained during probing. There is no specific rule for determining the structure of problems, which must first be translated into numerical values before being introduced to ANN. Additionally, the	The applications include solving the traveling salesman problem, predicting stock exchange trends, compressing images, recognizing handwriting, speech, characters, signatures, and human faces.

	simultaneously, learns through examples, and deals with attribute-value pair problems.	network's duration is unknown.	

II. REPRESENTATION

Deep learning is a type of representational learning that uses artificial neural networks to automatically find appropriate features from raw data. Improved representations simplify subsequent learning problems. With abundant data and improved deep learning, effective and robust representations can be trained for text and images. Developing multimodal representations is a difficult task as it involves intricate cross-modal interactions and inconsistencies between test and training data for each mode. This section offers a summary of individual modal representations such as text and images, which are utilized for acquiring multi-modal representations. Moreover, it examines supervised and unsupervised techniques for acquiring a joint representation space.

A. Unimodal Embedding's

1) Visual representations
2) Language representations
3) Vector arithmetic for word and image embedding's
4) Speaker representations

B. Multimodal Embedding's

1) Unsupervised training methods
2) Supervised training methods
3) Methods for Zero Shot Learning
4) Transformer based methods

While there have been significant improvements in representing language and vision, relying solely on unimodal input to model human concepts is inadequate. Non-visual means like natural language struggle to convey visual concepts such as a "beautiful image". Therefore, it is essential to develop joint embedding's that utilize multimodal data to create more accurate representations.

1) ***Unsupervised training methods:*** To achieve joint embedding for multimodal data, different approaches can be used. One way is by using multiple deep Boltzmann machines or auto-encoders to reconstruct inputs and share layers for a common representation space[107]-[109]. To avoid plagiarism, one approach is to transform pre-trained representation spaces of various modalities into a shared space, using techniques that apply to a single modality. For example, Fang et al. extended the Deep Structured Semantic Model (DSSM)[110] from text to images, creating the Deep Multimodal Similarity Model (DMSM), to generate embedding's that exist in the same vector space. The combination of word and image embedding was achieved by employing addition or concatenation [109], [111]. To enhance the similarity between textual and visual embedding's, they can be trained [112]. A recent study aimed to maximize the correlation and mutual information between embedding's of diverse modalities [113], [114]. Additionally, the similarity between word embedding's can be altered based on their visual representations [115]. The method utilizes unsupervised clustering of abstract scenes to identify visual representations. It adjusts the distances between word embedding's based on the similarity of their visual expressions.

Researchers have utilized picture fragments to correlate with sentence fragments and attribute words to create fine-grained multimodal embeddings[116]. They achieved this by automatically aligning images with phrase fragments. Another study harmonized idea embedding's at various levels, such as objects, characteristics,

relationships, and whole scenes[117]. Two models were proposed for image-text matching. The stacked cross-attention network was suggested to learn aligned embeddings, while the deep attention multimodal similarity model (DAMSM) incorporated an additional loss function to assess similarity between sub-regions of images and words. [118, 51]

2) ***Supervised Training Method:***

Multimodal representation learning can be improved using supervised training. These representations can be divided into two types of components: multimodal discriminative factors, which are used for supervised training, and intra-modality generative factors, which are used for unsupervised training [119]. The discriminative factors are shared across all modalities and can be utilized for performing discriminative tasks. On the other hand, generative factors can be used to regenerate missing modalities. According to certain studies, a potential method for learning word embeddings from natural scene photos or image areas with comprehensive text annotations is through visual co-occurrences (ViCo) [120]. By more precisely capturing similarities and contrasts between visual concepts, ViCo can enhance GloVe text embeddings which are unable to extract such information from text corpora alone. Additionally, different supervised training tasks have been utilized on different layers of vision-language encoders [121]. The curricular learning concept establishes the arrangement of training tasks to gradually enhance the complexity of training goals.

3) ***Zero-Shot Learning Methods****:*

Zero-shot learning is a frequently utilized method for vision-based tasks, which is necessary as gathering an adequate number of labeled images for every possible object category is challenging. However, not all multimodal representations are appropriate for zero-shot learning, as some require paired data from various modalities simultaneously. To tackle this problem, researchers have investigated methods that rely on additional language sources. A deep learning approach to zero-shot learning involves the creation of a linear mapping layer between multiple pre-trained embeddings [122], [138]. Researchers developed a sophisticated model that combines Skip-gram text embeddings with AlexNet visual features to create a deep visual-semantic embedding system. This approach allows for simultaneous training of both types of pre-existing models through a linear mapping layer [123]. The researchers put the model to the test by applying it to a set of 1000 well-known classes and 2000 previously unknown classes on a large scale.

In a study [124], it was found that the use of correlated auto-encoders for rebuilding representations for each modality can lead to better representations for one-shot and few-shot picture retrieval. Another study [125] used word labels that were not related to the target class to create positive and negative visual priors from a pre-trained VGG network [139]. The aforementioned priors were utilized as inputs for a subsequent model, which enabled semantic image segmentation of novel object categories that were not included in the original training dataset. Multiple modalities can benefit from rich sources of information such as words extracted from Wikipedia articles and features generated from various CNN layers [126]. To improve the results of zero-shot learning, instead of using direct text attribute inputs, recurrent model-generated phrase embedding's can be used as a text interface [127].

4) ***Transformer-based Methods:***

Transformers are a type of encoder-decoder model that uses a sequence-based approach. They are built by stacking multiple blocks of feedforward layers and multi-head self-attention models, with shared parameters [128]. Unlike RNN-based models [27], transformers can perform better on longer sequences because they do not rely on the first-order Markovian assumption of RNNs. BERT, which is the encoder part of a transformer model that is pre-trained on a large text corpus using masked language modelling, is often used for text embedding tasks. It is possible to expand BERT's capabilities beyond text-only tasks by incorporating images and generating pre-trained bimodal embedding's. Expanding the scope of unimodal BERT to bimodal applications can be done by introducing new tokens that represent visual feature inputs, as described in sources [129]-[133]. Additionally, to enhance the transformer model, visual aspects can be incorporated by adding extra encoder or attention structures, as explained in sources [134]-[136]. For more information on changed buildings, please refer to Section III-B. Furthermore, recent research in NLP [137] has indicated that multitask learning can enhance the generalization capacity of BERT representations. As a result, most bimodal BERT-based models leverage multitask training to improve their performance on downstream tasks like VQA, picture and video captioning, and others.

III. FUSION

Fusion, or the process of combining two or more things into a single entity, is an essential technique that has been used in various fields. From nuclear physics to cooking, the concept of fusion is widely employed to create something new and valuable. However, in this article, we will focus on fusion techniques in the context of technology and engineering.

In technology, fusion techniques are used to combine two or more technologies to create a new and more advanced system. For instance, in the field of artificial intelligence, machine learning algorithms are fused with natural language processing techniques to develop advanced chatbots that can interact with humans in a more human-like manner. Similarly, in the field of robotics, researchers are fusing different types of sensors and actuators to create robots that can perform complex tasks autonomously.

One of the most widely used fusion techniques is data fusion, where data from different sources is combined to obtain a more complete and accurate picture of a system. Data fusion techniques are used in various applications such as surveillance, medical diagnosis, and weather forecasting. For instance, in surveillance, data from multiple cameras is fused to track an object or a person across different cameras. In medical diagnosis, data from multiple medical sensors is fused to obtain a more accurate diagnosis of a patient's health condition.

Another important fusion technique is sensor fusion, which involves combining data from multiple sensors to obtain a more accurate representation of the environment. Sensor fusion techniques are widely used in autonomous vehicles, where data from cameras, radar, lidar, and other sensors are fused to obtain a more accurate perception of the surrounding environment. Sensor fusion techniques are also used in aerospace, where data from multiple sensors is fused to monitor the health of a spacecraft or an aircraft.

In summary, fusion techniques are essential in various fields of technology and engineering. From data fusion to sensor fusion to material fusion, these techniques are used to create new and advanced systems with enhanced capabilities. As technology continues to advance, fusion techniques will play an increasingly important role in creating new and innovative solutions to the world's most pressing problems.

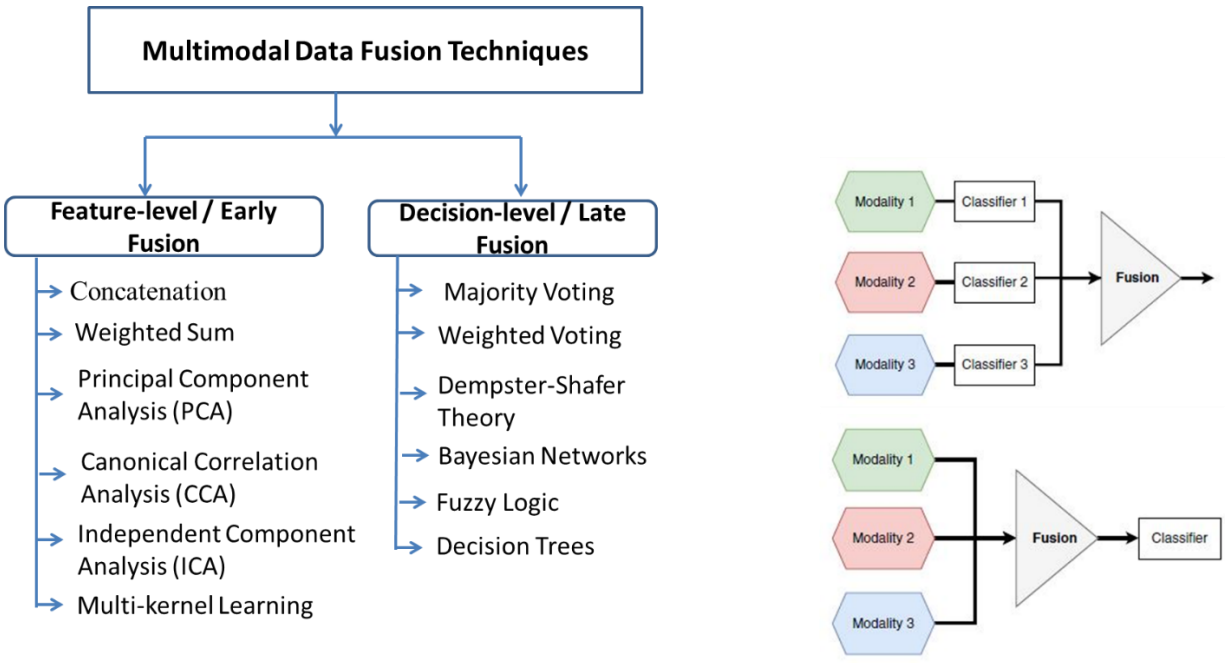

Figure: Categorization of Multimodal Data Fusion Techniques

4. Fusion Methods

A significant obstacle to multimodal sentiment analysis is the use of effective techniques to integrate feature information from several modalities. As seen in Fig. 2, we categorize 42 approaches in this section into 8 groups based on their fusion techniques. We go into great depth about each model's structure and enumerate its benefits and drawbacks so that readers might be motivated to complete their own work. This part concludes with a

thorough comparison of the fusion techniques of each classification, along with an explanation of the development motivated by the models' benefits and drawbacks.

4.1. Early fusion

Another name for early fusion is feature-level fusion. A joint representation is created by taking the characteristics of each modality and combining them at the input level. Sentiment classification is then carried out using this joint representation. This technique can have a simple framework since it does not require a specialized model design, instead learning view-specific and cross-view dynamics using general models such as deep neural networks or Support Vector Machines (SVM [56]). Nevertheless, view-specific dynamics are not well modeled in early stages of fusion, which results in the loss of contextual and temporal dependencies within each modality. This, in turn, impacts the modeling of cross-view dynamics and causes an overfitting of the data. Table 3 enumerates each model's benefits and drawbacks.

4.2. Late fusion

Decision-level fusion, or late fusion, starts with sentiment analysis based on each modality and suggests various mechanisms (e.g., averaging [59], majority voting [60], weighted sum [61], or learnable models) to integrate unimodal sentiment decisions into the final decision. Because its components are integrated, the fusion technique is often effective at simulating view-specific dynamics and produces lightweight, adaptable models that can easily adjust to variations in the number of modalities. The dynamics between various modalities are frequently more complicated than decision voting, hence inter-modal interactions are rarely well described as separate models are constructed for each modality. Table 4 enumerates the benefits and drawbacks of every approach.

4.3. Tensor-based fusion

To produce multimodal sentence representations, tensor-based approaches primarily compute the tensor product of unimodal sentence representations. This is a common non-concatenated feature fusion technique that calls for first turning the input representation into a high-dimensional tensor and then mapping it back to a low-dimensional output vector space. Because they may capture significant higher-order interactions over several modalities, feature dimensions, and time, tensors are an effective tool [63]. The computing complexity of this method increases exponentially, which is a disadvantage. Moreover, there is no fine-grained word-level interaction between cross-modalities. In order to investigate the dynamics within the three modalities—text, vision, and audio—the approach first embeds them. It then merges multimodal embedding representations to investigate the dynamic interactions between modalities. Table 5 enumerates each model's benefits and drawbacks.

4.4. Word-level fusion

By modeling interactions at each time step, the word-level fusion technique effectively examines time-dependent interactions while accounting for both view-specific and cross-view interactions. Two modules typically make up the model framework of this fusion method: one for temporal modeling and the other for attention. A temporal modeling network (LSTM, LSTHM, 1D temporal CNN, etc.) is used in the temporal modeling module to represent dynamics that are peculiar to a certain modality. After receiving the temporal modeling module's output, the attention module models crucial information in dynamic cross-modal interactions using the attention mechanism and its variations. Table 6 enumerates each model's benefits and drawbacks.

4.5. Translation-based fusion

This category is a translation-based approach to representing the interplay between modalities. Researchers suggest converting one modality to another in order to capture more significant interactions across modalities, motivated by the success of sequence to-sequence (Seq2Seq) models in machine translation. Another choice is to modify the transformer encoder's structure and apply a pretrained language model to record word interactions. Table 7 enumerates each model's benefits and drawbacks.

4.6. Feature space manipulation-based fusion

This kind of fusion technique is centered on understanding the relationship between features through a

sequence of mathematical operations or analyses, and mapping features into feature space following feature extraction. Table 8 enumerates the benefits and drawbacks of every approach.

4.7. Contextual-based fusion

The relationships between the utterances in the video are ignored by previous approaches, which consider each syllable as a separate entity. By taking into account the relationships between the target speech and other utterances in the context, contextual-based fusion produces superior outcomes. Recurrent neural network-based models are typically employed in these models to integrate contextual data. Table 9 enumerates each model's benefits and drawbacks.

4.8. Quantum-based fusion

The majority of currently used techniques rely on neural networks, which implicitly and incomprehensibly simulate multimodal interactions. Models may learn multimodal interactions from large-scale data in an end-to-end fashion thanks to neural networks, which frequently provide results with acceptable accuracy. However, these models implicitly incorporate multimodal interactions, functioning as a kind of black box with little numerical limitations, making it more challenging to comprehend multimodal interactions in human language. Because these models provide considerable performance advantages, researchers are trying to figure out how to comprehend the model and determine whether or not we can trust it enough to use it in practical applications [77], or whether or not it has privacy or security flaws [78]. They so started researching multimodal fusion techniques based on quantum mechanics. Table 10 enumerates each model's benefits and drawbacks.

4.9. Summary of different fusion methods

The two primary categories of early multimodal sentiment analysis techniques are early fusion and late fusion. These two forms of fusion don't require an extremely intricate fusion structure, making them very straightforward. Another name for early fusion is feature-level fusion. The input features of the whole model are created by splicing together the feature vectors of the three modalities at the input end. For sentiment classification, the feature is given to a later classifier, which may be an SVM or another type of deep learning network. This type of fusion has the advantage of not requiring any particular model architecture; instead, it just needs to think about how to build classifiers more effectively.

There is, however, a clear disadvantage: early feature fusion from several modalities results in incomplete modeling of particular view dynamics, which impacts cross-view dynamics modeling and causes overfitting. One may say that late fusion is the exact opposite process. It is also known as decision-level fusion because sentiment predictions are first established for each modality, and the outcomes based on these forecasts are then incorporated into the final result using various decision-making techniques. These decision-making techniques may include weighted, majority, average, or other statistical approaches. It follows that this type of fusion is effective at simulating view-specific dynamics. It is able to adjust well to variations in the quantity of modalities because of the integration of its modules. Low-level interactions across various modalities are disregarded as a result, and dynamic interactions between viewpoints cannot be properly investigated.

Tensor representation and interaction are utilized by tensor-based techniques. The tensor product is computed after each modality's feature representation. During the mapping process, tensors may record significant higher-order interactions spanning time, feature dimensions, and many modalities. They are also highly capable of probing crossmodal dynamics. Nevertheless, the outer dot product calculation of the tensor-based fusion approach uses a lot of processing power, which leads to an exponential increase in computational complexity. Additionally, during fusion, there is no fine-grained word-level interaction. In order to improve generalization, approaches under this paradigm primarily concentrate on lowering the computational complexity and resource use of fusion.

We categorize the three fusion approaches mentioned above as utterance-level fusion methods since they do not include finer-grained interactions. The three kinds of approaches—word-level fusion, translation-based fusion, and FSM-based fusion—that are discussed next focus primarily on the fine-grained interactions of modalities and are referred to as fine-grained fusion methods. Using an attention mechanism, the word-level fusion technique collects relevant information by modeling the interaction connection at each time step. As a result, the attention module and the temporal modeling module often make up the framework. Temporal networks like 1D temporal CNN and LSTM are included in the temporal modeling module. This module aims to investigate dynamics particular to a certain

modality.

In order to mine cross-modal interactions, the attention module models crucial information between modalities using the attention mechanism and its variations. Though merging unimodal characteristics with timestamps would omit an explicit and independent component to manage intra-modal and inter-modal interactions, word-level fusion approaches effectively examine time-dependent interactions.

The Seq2Seq model in machine translation is the source of inspiration for the translation-based fusion technique. More significant connections between modalities are discovered via translating one modality to another. The process of transformation can fill up the gaps in the modality's lacking information, strengthening its significance. Nevertheless, the word order information will be mostly ignored in the translation that is based on the connection of individual word representations. Another method is to record word interactions by including extra elements or changing the pre-trained language model's structure.

Feature Space Manipulation-based fusion is the full name of the FSM-based fusion. It focuses on examining the link between characteristics in the feature space using a number of mathematical studies or learning models. This fusion method's strength lies in its capacity to investigate the interactions between characteristics; nevertheless, the model does not take an efficient fusion technique into account.

Because it incorporates additional utterances in the context in addition to the target utterance, contextual-based fusion is also known as multi-utterance fusion. Recurrent neural network-based models are typically employed to concentrate on contextual data. A context sequence that includes the target speech and additional utterances in the context can aid in more accurately determining the polarity of the target utterance. Nevertheless, utterance-level sentiment analysis receives little attention, and overfitting can occur readily when contextual associations are extracted. Compared to existing neural networks, quantum-based fusion techniques describe multimodal interactions implicitly and incomprehensibly.

Human cognition may be better modeled by resolving the paradoxes of classical probability theory using quantum-inspired methods like superposition, entanglement, and interference, all of which have improved interpretability. But sentiment analysis's parallels don't quite add up, and quantum theory is riddled with contradictions.

IV. APPLICATIONS

1. **Healthcare:** Stress is a major risk factor for many chronic diseases such as cardiovascular diseases, depression, and anxiety. Multimodal data from sensors such as heart rate monitors, electroencephalography (EEG) sensors, and skin conductance sensors can be used to detect stress in patients. This can help healthcare providers monitor and manage stress levels in patients with chronic conditions [140].
2. **Workplace Safety:** Stress can negatively impact employee productivity, safety, and well-being. Multimodal data from sensors such as smart watches and activity trackers can be used to monitor employees' stress levels in real-time. Employers can use this information to identify high-stress situations and implement interventions to reduce stress levels and promote employee well-being.
3. **Education:** Stress can affect students' academic performance and mental health. Multimodal data from sensors such as EEG sensors and eye-tracking devices can be used to detect stress in students. This information can be used to implement interventions such as mindfulness-based stress reduction programs to help students manage stress levels.
4. **Automotive Safety:** Stress can affect a driver's ability to operate a vehicle safely. Multimodal data from sensors such as heart rate monitors and steering wheel sensors can be used to detect stress in drivers. This information can be used to alert drivers or trigger safety features such as automatic emergency braking systems to prevent accidents.
5. **Sports:** Stress can affect athletes' performance and recovery. Multimodal data from sensors such as heart rate monitors, GPS trackers, and sweat sensors can be used to monitor athletes' stress levels during training and competition. This information can be used to adjust training regimes and recovery protocols to optimize performance and prevent injuries.

V. MODALITIES

Multimodal sentiment analysis utilizes various modalities to detect affective states in a conversation. The most commonly used modalities include text, audio, and visual cues. Each modality contributes to better sentiment prediction, and research suggests that bimodal and tri-modal systems yield better results than unimodal systems. The accuracy of the analysis is improved by the significant contributions made by each modality.

Text is the most prevalent modality used for sentiment analysis, as it enables the identification of underlying emotions. Although textual sentiment analysis can yield effective results, it is worth noting that nowadays most opinionated data is shared in video format rather than text.

The use of **visual features** facilitates a better identification of sentiments or opinions. For instance, if the text reads "this is a pretty good mouse," it can be challenging to discern whether it pertains to the animal or computer device. In this context, visual cues prove helpful, and a bimodal system combining both text and visuals generates better outcomes than unimodal systems.

Acoustic features in **audio modality** are utilized to generate textual data from videos, and to identify the tone of the speakers. Combining all three modalities creates a more robust analysis model. While visuals may not always detect humor, sarcasm, or common sense accurately, the combination of modalities can correctly identify the sentiments.

Stress detection can be approached through various modalities that can provide different types of information about a person's emotional state. Here are some of the modalities commonly used in stress detection:

Physiological signals: This modality involves measuring biological signals, such as heart rate, skin conductance, and muscle tension, to detect stress. These signals can provide insight into a person's autonomic nervous system activity, which can be indicative of stress levels.

Facial expressions: Facial expressions can provide valuable information about a person's emotional state, including stress. Facial expression analysis can be performed using computer vision techniques, which involve detecting and analyzing changes in facial muscle activity.

Speech: Changes in speech patterns, such as speaking rate, pitch, and tone, can be indicative of stress. Speech analysis can be performed using natural language processing techniques, which involve analyzing the semantic and syntactic structure of spoken language.

Behavior: Changes in behavior, such as fidgeting or avoidance behaviors, can be indicative of stress. Behavioral analysis can be performed using motion capture or wearable sensors.

Eye-tracking: Eye-tracking can provide information about a person's attention and cognitive processing, which can be indicative of stress. Eye-tracking analysis can be performed using specialized hardware or software that tracks eye movements.

Brain activity: Brain activity can be measured using techniques such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). These techniques can provide insight into a person's cognitive processing and emotional state.

In summary, stress detection can be approached through various modalities that can provide different types of information about a person's emotional state. These modalities include physiological signals, facial expressions, speech, behavior, eye-tracking, and brain activity.

VI. DATASETS

Table 3.

Dataset	Age Group / Gender	Features	Sensors
WESAD	15 - 9 Males and 6 females aged between 19 and 31 years.	Electrocardiogram (ECG), Electrodermal Activity (EDA), Accelerometer, Respiration, Temperature, Blood Volume Pulse (BVP)	Empatica E4 wristband, Microsoft Band 2, Chest strap.

Affectiva-MIT Stress[21]	37 - 19 males and 18 females, aged between 18 and 35 years.	Heart rate, (EDA), Motion	Empatica E4 wristband, Smartphone (motion)
DEAP[22]	32 - 16 males and 16 females, aged between 19 and 37 years.	EEG (14 channels), Peripheral physiological signals	Biosemi ActiveTwo system (EEG), Empatica E3 wristband
AMIGOS	40 - 20 males and 20 females) aged between 19 and 25 years.	EEG (8 channels), Peripheral physiological signals	Emotiv EPOC+ headset (EEG), Empatica E4 wristband
DriveDB database	24 male drivers in Boston	ECG, EDA, EMG (Electromyogram), RESP	Empatica E4 Wristband
Multimodal Affect Recognition Challenge	88 - 58 males and 30 females) aged between 18 and 35 years.	Speech, Facial expressions, Body gestures	Audio and video recordings, Microsoft Kinect camera
Multi-modal Stress Challenge	21 - 11 males and 10 females) aged between 22 and 55 years.	EEG (4 channels), Peripheral physiological signals	B-Alert X10 EEG headset (EEG), Empatica E4 wristband

The WESAD (Wearable Stress and Affect Detection) dataset is a publicly available dataset that was created to aid the development of algorithms for detecting stress and affect using physiological signals obtained from wearable sensors.

The dataset was created by collecting physiological data from 15 participants using a variety of sensors, including electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), and temperature (TEMP) sensors.

While the WESAD dataset has been widely used and has contributed to advances in the field of affect and stress detection, **it also has some limitations. These include:**

1. **Small sample size:** The dataset contains data from only 15 participants, which limits its generalizability to larger populations.
2. **Limited demographic diversity:** The participants in the dataset are all young, healthy adults, which may limit the generalizability of the dataset to other age groups or individuals with health conditions.
3. **Limited range of stressors:** The dataset was collected in a laboratory setting, which may not accurately represent the range of stressors that individuals experience in their daily lives.
4. **Limited sensor modalities:** While the dataset includes several sensor modalities, it does not include other physiological signals that may be relevant for stress and affect detection, such as EEG (electroencephalogram) or eye-tracking data.
5. **Limited annotation:** The dataset includes only self-reported stress and affect ratings, which may be subject to biases and inaccuracies.

It is important to keep these limitations in mind when using the WESAD dataset for research or algorithm development, and to consider combining it with other datasets or sources of information to improve the generalizability and accuracy of stress and affect detection algorithms.

REFERENCES

[1] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, pp. 504–507, 2006.

[2] Y. Bengio, “Learning deep architectures for AI,” Foundations and Trends in Machine Learning, vol. 2, pp. 1–127, 2009.

[3] L. Deng and Y. Dong, “Deep Learning: Methods and Applications,” Foundations and Trends in Signal Processing, vol. 7, pp. 197–387, 2014.
[4] J. Schmidhuber, “Deep learning in neural networks: An overview,”
NeuralNetworks, vol. 61, pp. 85–117, 2015.
[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.
[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsuperv Press, 2016.
[7] D. Yu, L. Deng, and G. Dahl, “Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition,” Proc. NIPS Workshop, 2010.
[8] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep autoencoder,” Proc. Interspeech, 2010.
[9] L. Deng, “An overview of deep-structured learning for information processing,” in Proc. APSIPA ASC, 2011.
[10] D. Yu, L. Deng, F. Seide, and G. Li, “Discriminative pre-training of deep nerual networks,” in U.S. Patent No. 9,235,799, 2011.
[11] G. Dahl, D. Yu, and L. Deng, “Large-vocabulry continuous speech recognition with context-dependent DBN-HMMs,” in Proc. ICASSP, 2011.
[12] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer,
G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, “Recent advances in deep learning for speech research at Microsoft,” in Proc. ICASSP, 2013.
[13] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Transactions on Audio, Speech and Language Processing, vol. 20, pp. 30–42, 2012.
[14] F. Seide, L. Gang, and Y. Dong, “Conversational speech transcription using context-dependent deep neural networks,” in Proc. Interspeech, 2011.
[15] G. Hinton, L. Deng, Y. Dong, G. Dahl, A.-R. Mohamed, N. Jaitly,
A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, vol. 29, pp. 82–97, 2012.
[16] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” Proc. ICASSP, 2013.
[17] D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach. Springer, 2015.
[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, pp. 2278–2324, 1998.
[19] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012.
[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.
[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, 2016.
[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. CVPR, 2015.
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. CVPR, 2014.
[24] R. Girshick, “Fast R-CNN,” in Proc. ICCV, 2015.
[25] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real- time object detection with region proposal networks,” in Proc. NIPS, 2015.
[26] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani- Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, “Using recurrent neural networks for slot filling in spoken language understanding,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 23, pp. 530–539, 2015.
[27] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. ICLR, 2015.
[28] I. Sutskever, O. Vinyals, and Q. Le, “Sequence to sequence learning with neural networks,” in Proc. NIPS, 2014.
[29] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah,
M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C.

Young, J. Smith,
J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," in arXiv:1609.08144, 2016.
[30] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," in Proc. EMNLP, 2015.
[31] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in Proc. NAACL, 2018.
[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," in 2018.
[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.
[34] H.-Y. Shum, X. He, and D. Li, "From Eliza to XiaoIce: Challenges and opportunities with social chatbots," Frontiers of Information Tech- nology & Electronic Engineering, vol. 19, pp. 10–19, 2018.
[35] S. Bengio, L. Deng, L. Morency, and B. Schuller, Perspectives on Predictive Power of Multimodal Deep Learning: Surprises and Fu- ture Directions. Chapter 14 in Book: The Handbook of Multimodal- Multisensor Interfaces. ACM and Morgan & Claypool Publishers, 2019.
[36] L. Deng and Y. Liu, Deep Learning in Natural Language Processing. Springer, 2018.
[37] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in Proc. EMNLP, 2014.
[38] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg, "Modeling context in referring expressions," in Proc. ECCV, 2016.
[39] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and L. S., "Flickr30k entities: Collecting region-to phrase correspondences for richer image-to-sentence models," in Proc. ICCV, 2015.
[40] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proc. CVPR, 2015.
[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. CVPR, 2015.
[42] J. Johnson, A. Karpathy, and F.-F. Li, "Densecap: Fully convolutional localization networks for dense captioning," in Proc. CVPR, 2016.
[43] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in Proc. CVPR, 2016.
[44] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in Proc. CVPR, 2016.
[45] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proc. CVPR, 2016.
[46] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," in Proc. NAS, 2015.
[47] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batral, C. Zitnick, and D. Parikh, "VQA: Visual question answering," in Proc. ICCV, 2015.
[48] L. Yu, E. Park, A. Berg, and T. Berg, "Visual Madlibs: Fill in the blank description generation and question answering," in Proc. ICCV, 2015.
[49] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in Proc. ECCV, 2016.
[50] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in Proc. ICML, 2016.
[51] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in Proc. CVPR, 2018.
[52] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Su¨nderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in Proc. CVPR, 2018.
[53] S. Dupont and J. Luettin, "Audio-visual speech modeling for contin- uous speech recognition," IEEE Transactions on Multimedia, vol. 2, pp. 141–151, 2000.
[54] M. Cookea, J. Barker, S. Cunningham, and X. Shao, "An audio- visual corpus for speech perception and automatic speech recognition," Journal of Acoustic Society of America, vol. 120, pp. 2421–2424, 2006.

[55] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. Early Access, pp. 1–13, 2018.
[56] B. Maison, C. Neti, and A. Senior, “Audio-visual speaker recognition for video broadcast news: Some fusion techniques,” in Proc. MMSP, 1999.
[57] Z. Wu, L. Cai, and H. Meng, “Multi-level fusion of audio and visual features for speaker identification,” in Advances in Biometrics (D. Zhang and A. Jain, eds.), pp. 493–499, Springer Berlin Heidelberg, 2005.
[58] J. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in Proc. Interspeech, 2018.
[59] I. Gebru, S. Ba, X. Li, and R. Horaud, “Audio-visual speaker diarization based on spatiotemporal Bayesian fusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 1086–1099, 2018.
[60] J. Chung, B.-J. Lee, and I. Han, “Who said that?: Audio-visual speaker diarisation of real-world meetings,” in Proc. Interspeech, 2019.
[61] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, “Time domain audio visual speech separation,” in Proc. ASRU, 2019.
[62] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim,
W. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separa- tion,” ACM Transactions on Graphics, vol. 37, pp. 112:1–11, 2018.
[63] T. Afouras, J. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” in Proc. Interspeech, 2018.
[64] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 1798–1828, 2013.
[65] P.-S. Huang, X. He, G. J., L. Deng, A. Acero, and L. Heck, “Learning deep structured semantic models for web search using clickthrough data,” in Proc. CIKM, 2013.
[66] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, “Learning semantic representations using convolutional neural networks for web search,” in Proc. WWW, 2014.
[67] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, “Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, pp. 694–707, 2016.
[68] D. Rumelhart, G. Hinton, and R. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, pp. 533–536, 1986.
[69] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural prob- abilistic language model,” Journal of Machine Learning Research, vol. 3, pp. 1137–1155, 2003.
[70] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. ICLR, 2013.
[71] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Dis- tributed representations of words and phrases and their compositional- ity,” in Proc. NIPS, 2013.
[72] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.
[107] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, “Multimodaldeep learning,” in *Proc. ICML*, 2011.
[109] C. Silberer and M. Lapata, “Learning grounded meaning representa- tions with autoencoders,” in Proc. ACL, 2014.
[110] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dolla´r,
J. Gao, X. He, M. Mitchell, J. C. Platt, et al., “From captions to visual concepts and back,” in Proc. CVPR, 2015.
[111] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, “Distributional semantics in technicolor,” in Proc. ACL, 2012.
[112] S. Kottur, R. Vedantam, J. Moura, and D. Parikh, “Visual Word2Vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes,” in Proc. CVPR, 2016.
[113] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. Bernal, and J. Luo, “Deep multimodal representation learning from temporal data,” in Proc. CVPR, 2017.
[114] P. Bachman, R. Hjelm, and W. Buchwalter, “Learning representations by maximizing mutual information across views,” in Proc. NeurIPS, 2019.
[115] A. Lazaridou, N. Pham, and M. Baroni, “Combining language and vision with a multimodal skip-gram model,” in Proc. NAACL, 2015.
[116] A. Karpathy, A. Joulin, and F.-F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” in Proc. NIPS, 2014.

[117] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma, “Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations,” in Proc. CVPR, 2019.
[118] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in Proc. ECCV, 2018.
[119] Y.-H. Tsai, P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” in Proc. ICLR, 2018.
[120] T. Gupta, A. Schwing, and D. Hoiem, “ViCo: Word embeddings from visual co-occurrences,” in Proc. ICCV, 2019.
[121] D.-K. Nguyen and T. Okatani, “Multi-task learning of hierarchical vision-language representation,” in Proc. CVPR, 2019.
[122] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in Proc. NIPS, 2013.
[123] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “DeViSE: A deep visual-semantic embedding model,” in Proc. NIPS, 2013.
[124] Y.-H. Tsai, L.-K. Huang, and R. Salakhutdinov, “Learning robust visual-semantic embeddings,” in Proc. ICCV, 2017.
[125] D. Golub, R. Mart´ın-Mart´ın, A. El-Kishky, and S. Savarese, “Leverag- ing pretrained image classifiers for language-based segmentation,” in Proc. WACV, 2020.
[126] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, “Predicting deep zero-shot convolutional neural networks using textual descriptions,” in Proc. ICCV, 2015.
[127] S. Reed, Z. Akata, B. Schiele, and H. Lee, “Learning deep represen- tations of fine-grained visual descriptions,” in Proc. CVPR, 2016.
[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez,
L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. NIPS, 2017.
[129] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, “Unicoder- VL: A universal encoder for vision and language by cross-modal pre- training,” in arXiv:1908.06066, 2019.
[130] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “VL- BERT: Pre-training of generic visuallinguistic representations,” in arXiv:1908.08530, 2019.
[131] L. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visual- BERT: A simple and performant baseline for vision and language,” in arXiv:1908.03557, 2019.
[132] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “VideoBERT: A joint model for video and language representation learning,” in Proc. ICCV, 2019.
[133] C. Alberti, J. Ling, M. Collins, and D. Reitter, “Fusion of detected objects in text for visual question answering,” in Proc. ICMLC, 2019.
[134] H. Tan and B. Mohit, “LXMERT: Learning cross-modality encoder representations from transformers,” in Proc. EMNLP, 2019.
[135] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining task- agnostic visiolinguistic representations for vision-and-language tasks,” in Proc. NeurIPS, 2019.
[136] S. Pramanik, P. Agrawal, and A. Hussain, “OmniNet: A unified architecture for multi-modal multi-task learning,” in arXiv:1907.07804, 2019.
[137] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” in Proc. ACL, 2019.
[138] Mane, D., Ashtagi, R., Suryawanshi, R., Kaulage, A.N., Hedaoo, A.N., Kulkarni, P.V., Gandhi, Y. (2024). Diabetic retinopathy recognition and classification using transfer learning deep neural networks. Traitement du Signal, Vol. 41, No. 5, pp. 2683-2691. https://doi.org/10.18280/ts.410541
[139] Malwade, Sulakshana. (2024). Predicting Heart Diseases in IoT-Based Electronic Health Records: A Federated Learning Approach. Journal of Electrical Systems. 20. 472-484. 10.52783/jes.1465.
[140] Shanthi Kunchi, Vijaya N. Aher, Sharayu Ikhar, Kishor Pathak, Yatin Gandhi, and Kirti Wanjale. 2024. Risk Factor Prediction for Heart Disease Using Decision Trees. In Proceedings of the 5th International Conference on Information Management & Machine Intelligence (ICIMMI '23). Association for Computing Machinery, New York, NY, USA, Article 110, 1–6. https://doi.org/10.1145/3647444.3647937