IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 January 2023, accepted 31 January 2023, date of publication 9 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243854

SURVEY

A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions

ARNAB BARUA, MOBYEN UDDIN AHMED, AND SHAHINA BEGUM
School of Innovation, Design and Engineering, Mälardalen University, 72220 Västerås, Sweden

Corresponding authors: Arnab Barua (arnab.barua@mdu.se) and Mobyen Uddin Ahmed (mobyen.uddin.ahmed@mdu.se)

This work was supported in part by the Project FitDrive through the European Union's Horizon 2020 Research and Innovation Programme under Grant 953432; in part by the DIGICOGS Project through Vinnova funded by Vinnovas Diarienr under Grant 2019-0532; and in part by the Innovation Program Process Industrial IT and Automation (PiiA), Mälardalen University.

ABSTRACT Multimodal machine learning (MML) is a tempting multidisciplinary research area where heterogeneous data from multiple modalities and machine learning (ML) are combined to solve critical problems. Usually, research works use data from a single modality, such as images, audio, text, and signals. However, real-world issues have become critical now, and handling them using multiple modalities of data instead of a single modality can significantly impact finding solutions. ML algorithms play an essential role in tuning parameters in developing MML models. This paper reviews recent advancements in the challenges of MML, namely: representation, translation, alignment, fusion and co-learning, and presents the gaps and challenges. A systematic literature review (SLR) was applied to define the progress and trends on those challenges in the MML domain. In total, 1032 articles were examined in this review to extract features like source, domain, application, modality, etc. This research article will help researchers understand the constant state of MML and navigate the selection of future research directions.

INDEX TERMS Multimodal machine learning, systematic literature review, representation, translation, alignment, fusion, co-learning.

I. INTRODUCTION

Artificial intelligence (AI) has progressed rapidly in the last few decades. It impacts human livelihood, health care, science and technology. With the flow of progress, AI needs improvement in its techniques to tackle more critical real-world problems. ML is an application of AI that gives a system the capability to learn and improve from its experience automatically. The current trend of ML is high and involves several issues to provide solutions. It is rich in algorithms and uses them to build models that can process different kinds of data. Data are ubiquitous and hold information such as official reports, medical or financial records etc. The importance of data is increasing with the progress of AI. It contains information in several forms, such as numeric, text, signals etc. Data can come from a distinct range of modalities where modality means how something is experienced or happens [1]. Visual, auditory, haptic, physiological signals, etc., are examples of modality. Data can be defined as multimodal when multiple modalities are involved together [1]. Speech recognition is a multimodal example in which audio and visual data are combined to recognize what a person is saying [2]. The blend of multimodal data and ML frames the notion of MML, which focuses on building models to process multimodal data from multiple modalities. Data from various modalities are always heterogeneous. Heterogeneous data are always ambiguous, and learning from these multimodal data provides an opportunity to understand the relationships between modalities [3]. Five core technical challenges covered MML: representation, translation, alignment, fusion, and co-learning [1]. Figure 1 depicts the classification diagram of MML.

Representation is the first challenge, which means presenting data using information from modalities. Representing multiple modalities is crucial because data come from heterogeneous sources, contain noise, and may have missing

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

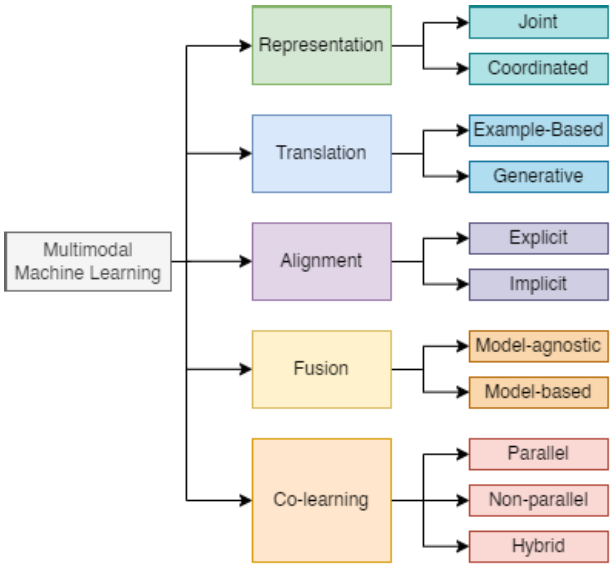


FIGURE 1. Classification diagram of multimodal machine learning.

information [4]. It has two types: joint representation, which merges unimodal data into the exact representation, and coordinated representation, which means coordinating each modality through a constraint [1]. The second challenge is translation, which means translating or mapping an entity of one modality to a different modality [5]. Multimodal translation approaches are usually modality specific because they share several unifying factors. It is categorised into two types: example-based and generative [1]. Example-based models translate modalities using a dictionary; on the other hand, generative models build a model that can translate. Alignment is the third challenge, where it tries to identify relationships and consistency between subelements of two or more separate modalities. Multimodal alignment has two types: explicit and implicit, where explicit aligns subcomponents of modalities, and implicit alignment is the intermediate step for another task [1]. The fourth challenge is fusion, which has a broad range of applications. Fusion joins information from multiple modalities for prediction [6]. Multimodal fusion is classified into two categories: model-agnostic and model-based [1]. In the model-agnostic, fusion is performed before applying the ML method. In contrast, in the model-based method, fusion takes place during the construction of the ML method. The final challenge is co-learning, where one model transfers knowledge to another [7]. Multimodal co-learning is categorised into parallel, nonparallel and hybrid [1]. In parallel, modalities share a set of instances, but in nonparallel, concepts or categories are shared instead of instances. In the hybrid, two nonparallel modalities are linked up by a shared modality. This SLR explores the recent adoption of these five core challenges to seek answers to research questions.

The remainder of this survey is organised as follows. Motivation and contribution of this study are presented in section II. Section III summarizes reviews of MML and its challenges. Details of the employed research methodology are described in Section IV. Section V presents the results with figures. Section VI provides a discussion of the performed analysis. Finally, Section VII concludes this article with a summary. The structure of this survey article, with all sections and subsections, is presented in Figure 2.

II. MOTIVATION AND CONTRIBUTION

There are plentiful surveys available related to MML and its challenges. Most of the surveys have a limitation in covering modalities and challenges. For example, a significant number of studies covered only specific domains. There is a lack of studies which cover not only all challenges but also different modalities. This constraint motivated the conduct of this study. In the section on related studies, several publications are listed and contrasted with this study. The primary focus is on seeking literature on MML and modalities and finding insights thought to give future directions. To summarise, the contributions of this article are the following:

- identified 374 articles on MML and its challenges.
- compared with 39 related studies.
- describes the definition of MML.
- outline research challenges and gaps.
- depicts the results to understand the current trend of research.
- points out the domains and applications used in MML
- highlight available modalities and their combination.
- clearly shows the algorithms used to build a model for MML.

III. RELATED STUDIES

This section presents existing survey or review works on MML and its challenges. We highlighted the differences and compared them with our proposed study. MML is an advanced research area and is growing very fast. A research team from Carnegie Mellon University conducted excellent survey research on MML [1], where they classified all the challenges of MML into several sections. The survey focuses on audio, video, and text modalities. Another article [8] reviewed applied methods and applications in multimodal deep learning, where the authors concentrated on a few common deep learning (DL) methods and applications. Apart from this article, we discovered a few surveys that discussed how MML could solve different problems related to modalities. In [9], [10], and [11], they surveyed meme classification, sentimental analysis, and content understanding by utilizing MML accordingly. Visual and language analysis is an attractive research area, and with MML, new possibilities are growing rapidly. Survey papers [12], [13], and [14] show advances and trends in computer vision, language and image analysis using MML techniques.

Researchers not only focus on MML but also on each core challenge to improve the quality of the research. All challenges have a specific role in MML. We encounter several surveys on each challenge, where most focus on applying techniques of challenges to handle data from modalities. The

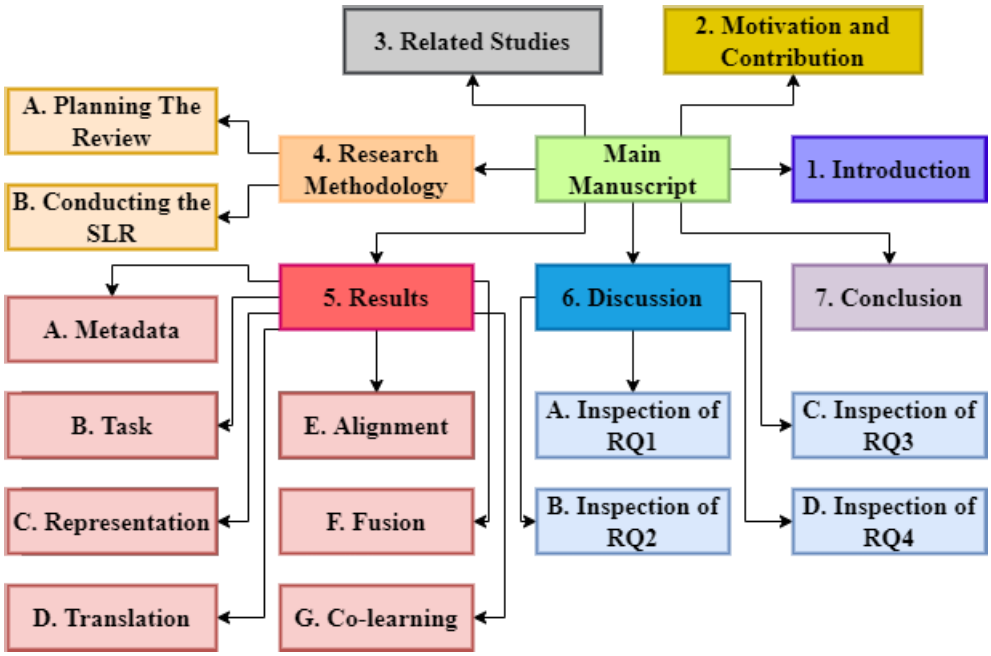


FIGURE 2. Structure of the review article.

first challenge of MML is the representation, which means representing and summarizing data to point out the complementarity and synchrony within modalities [1]. Two recent survey works [4], and [15] presented an overview of representation learning with proposed approaches. Article [16] reviewed representation learning development in unsupervised and deep learning. In [17], they introduced two categories for multi-view representation learning and presented an investigation of various essential applications. [18] performed a survey on representation to quantify its techniques in human affect recognition.

Translation is the second challenge of MML, which means translating or mapping data from one modality to another. Translation approaches are broad and often specific to the modality [1]. Until now, no survey article has entirely focused on multimodal translation techniques and advances. A recent survey [5] presented different aspects of multimodal translation on visual and speech datasets.

The third challenge is alignment, which identifies linear relationships in elements of two or more separate modalities [1]. Like translation, there is a lack of sufficient survey articles focused on multimodal alignment approaches. The paper [19] introduced a kernel method for manifold alignment (KEMA) that can match a random number of data sources without having similar pairs. In [17], they discussed multi-view representation alignment and its applications as one category of multi-view representation.

Fusion is the fourth and most studied challenge in MML. Joining information from two or more modalities is the primary purpose of multimodal fusion [1]. References [20], [21], and [22] are three recent survey works of multimodal fusion, and they classify approaches including different modalities. Articles [6], [23], [24], and [25] present an overview of the methods of fusion with challenges and prospects. Instead of focusing on fusion processes, there are surveys related to the use of fusion in several domains. Health care is an eye-catching area and, together with fusion, can create a significant impact. In [26], they conducted a comprehensive survey on the fusion of medical signals to facilitate intelligent healthcare systems. Fusion is famous for combining images, especially medical images such as in [27], [28], [29], [30], and [31], which survey different techniques on medical image fusion. Activity detection and monitoring systems have data from multiple modalities, and fusion techniques help to merge all modalities. In [32], [33], [34], and [35], they presented a survey on activity recognition leveraging fusion techniques. References [36] and [37] presented a survey on biometric systems where fusion techniques were applied to merge multiple data. Fusion is commonly used in audio-visual information fusing. References [38], [39], and [40] are good examples of audio-visual information fusion-related surveys.

Co-learning is the last challenge of MML and transfers knowledge between models to boost prediction [1]. Co-learning is not as familiar as fusion and has always received insufficient attention in research. However, its importance has increased in recent years. References [7] and [41] are the two recent survey works on co-learning where they discussed several approaches of co-learning with challenges, applications, recent advances and directions.

Except [1], above discussed related works mainly focused on one specific challenge and its application. Instead of focusing on a particular challenge, this survey included all

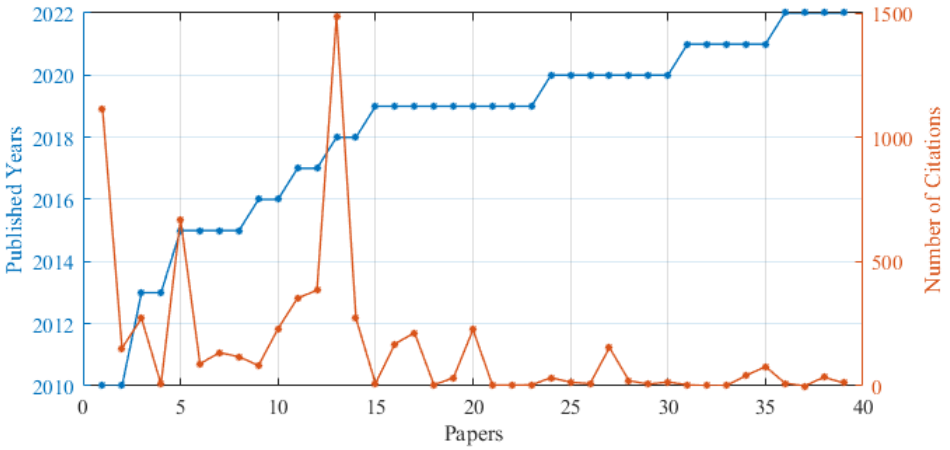


FIGURE 3. Article vs citation and years.

TABLE 1. Analysis of related surveys, Where T=Text, A=Audio, I=Image, V=Video, S=Sensor, N=Numerical, Si=Signal, Re=Representation, Tr=Translation, Al=Alignment, Fu=Fusion, Cl=Co-learning.

Refs	T	A	I	V	S	N	Si	Re	Tr	Al	Fu	Cl
[1]	✓	✓	✓	✓				✓	✓	✓	✓	✓
[8]	✓	✓	✓	✓			✓					
[9], [11], [20]	✓		✓								✓	
[10]	✓	✓	✓	✓				✓				
[12]			✓	✓				✓		✓	✓	
[13]	✓		✓					✓				
[14]			✓									
[4]		✓		✓				✓				
[15]	✓	✓	✓	✓				✓				
[16]								✓				
[17]	✓		✓					✓		✓		
[18]		✓	✓				✓	✓				
[5]		✓	✓	✓					✓			
[19]			✓		✓					✓		
[21], [25]		✓		✓							✓	
[22]			✓								✓	
[6]	✓	✓	✓		✓						✓	
[23]	✓	✓	✓	✓							✓	
[24]						✓					✓	
[26]							✓				✓	
[27]			✓				✓				✓	
[28]–[31], [36]			✓								✓	
[32], [33], [35]					✓						✓	
[34]			✓		✓						✓	
[37]			✓		✓		✓				✓	
[38]–[40]		✓	✓	✓							✓	
[7]	✓	✓	✓									✓
[41]	✓											✓
This Survey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

and discussed their current advances, gaps and challenges. Although article [1] focused on three modalities, this paper contained all the possible modalities. Article [8] discussed the current use of ML methods and applications in MML, but they limited their review by selecting typical ML methods and applications. On the contrary, this study presented all ML algorithms, domains, and applications available in the search range. To compare included related surveys and this work, an analysis of the associated surveys are presented in Table 1. The table clearly distinguishes between the inclusion of modalities and the challenges of MML in the study. Figure 3 shows the trend between the published year and the number of citations of included survey papers. From the figure, it is visible that after the year 2018, researchers are interested in working on MML and its challenges. However, the number of citations is lower than in previous years, but it will increase over time.

IV. RESEARCH METHODOLOGY

A systematic literature review (SLR) is conducted according to the guidelines of the article [42] to accomplish the objective of this research. SLR, also known as systematic review, aims to determine, review, and interpret every available study related to a specific research area or question. Three steps

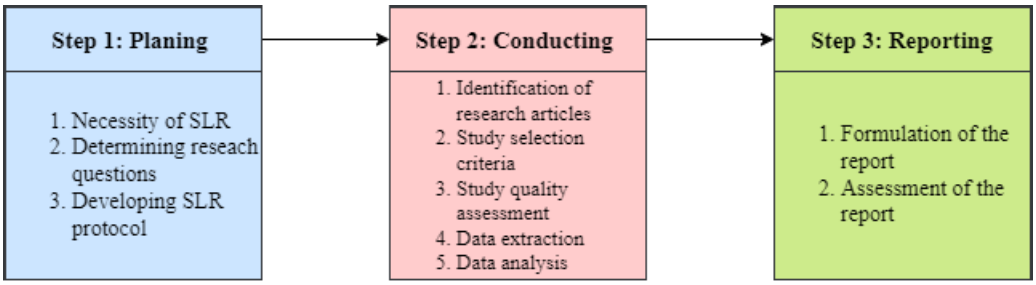


FIGURE 4. Steps of SLR methodology.

are involved in SLR: planning, conducting, and reporting the review. Figure 4 depicts all steps with substeps. This section discusses in detail all the steps.

A. PLANNING THE REVIEW

The planning phase is the first stage related to the set of tasks for designing and formulating the protocol. It includes identifying the importance of SLR in a specific area, defining research questions that SLR will address, and generating a review protocol for stating review procedures.

1) NECESSITY OF SLR

It is necessary to verify the importance of such a review before initiating. Researchers have recently focused on using MML techniques to solve multimodality problems. However, there is also a lack of articles discussing the techniques of challenges, and conducting procedures does not always provide a solution. It is also necessary to focus on the use of modality. The understanding of MML lies in the relationship between its challenges and used modalities. Therefore, SLR is essential to depict the importance and understanding of MML.

2) RESEARCH QUESTIONS (RQs)

In SLR, specifying RQs is the most crucial part. Analysing prior works on the challenges and understanding MML is the main objective of this SLR. It includes the explanation of MML, approaches to challenges, considered modalities, applied ML models, and gaps for future research. The four following questions, including one subquestion, are proposed to facilitate this research.

RQ1: What is the definition of multimodal machine learning?
RQ2: What are the challenges adopted when framing multimodal machine learning?
RQ2.1: What are the feasible gaps in challenges?
RQ3: What are the modalities considered in multimodal machine learning?
RQ4: Which machine learning models were applied?

3) DEVELOPING SLR PROTOCOL

The SLR protocol determines the methods to begin a specific review on a particular area. In this review study, a protocol was constructed to obtain the objective. Initially, we searched for primary studies from prominent bibliographic databases. In the second, we make a margin for the selection criteria. Data extraction and study quality assessment took place in the third and fourth steps accordingly. The final section will involve data analysis.

TABLE 2. List of databases used for article search.

Source Name	Type	URL
Science Direct - Elsevier	Digital Library	https://www.sciencedirect.com/
Scopus	Search Engine	https://www.scopus.com/
Web of science	Search Engine	https://www.webofscience.com/
PubMed	Search Engine	https://pubmed.ncbi.nlm.nih.gov/

B. CONDUCTING THE SLR

Conducting the SLR is vital and starts once researchers have agreed upon the protocol [42]. All the specified steps in the protocol are executed in this section to obtain the research goal. It is divided into five parts and discussed below.

1) IDENTIFICATION OF RESEARCH

To produce answers to all research questions, we used a few keywords to search underlying studies in renowned online databases in this SLR. Four major online databases were explored to maintain unbiased results during article searches. A rich library of journals and conferences is the main reason behind exploring those four significant databases. The list of four databases is depicted in Table 2.

The next step is the formation of procedures for seeking the scientific and technical articles that these searches provided. The process is divided into two parts: determining search keywords that cover the area of MML and its five challenges and determining queries by placing keywords between Boolean operators AND or OR. All the queries used are presented in Table 3.

We started searching on the 10th of January 2022 and considered published articles from 2009 until the search date 31st of December 2022. The search process included the title, abstract, keywords, and introduction. We also added papers from the references found in the primary studies during the search.

2) STUDY SELECTION CRITERIA

This study used a group of inclusion criteria (IC) and exclusion criteria (EC) to purify the search results and ignored

TABLE 3. List of search queries.

S/L	Queries
SQ1	((“Multimodal” AND “Machine Learning ” AND “Survey”))
SQ2	((“Multimodal” AND “Machine Learning ” AND “Representation”))
SQ3	((“Multimodal” AND “Machine Learning ” AND “Translation”))
SQ4	((“Multimodal” AND “Machine Learning ” AND “Alignment”))
SQ5	((“Multimodal” AND “Machine Learning ” AND “Fusion”))
SQ6	((“Multimodal” AND “Machine Learning ” AND “Co-learning”))

studies that failed to answer the EC. All IC and EC to determine relevant articles are included in Table 4.

This study followed the steps of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) from [43] to identify research articles. The stages of PRISMA include identification, screening, eligibility, and sorting, and Figure 5 illustrates it using a flow diagram. Initially, by applying keywords, a total of 1009 articles were collected from bibliographic databases, and 23 articles were collected from other sources. After eliminating duplicates, the total number of articles was 800. Of 800 articles, 650 were screened, and 150 were excluded due to unavailability. Initially, 593 articles were considered in the eligibility section, and 57 were excluded because they were not published in any journal or conference. Then, 545 papers were selected because they were scientific or technical, and 48 survey or review-related articles were excluded. From the 545 articles, 338 articles met the requirements of MML and its challenges, as opposed to 207 articles excluded. After that, 319 articles were thoroughly scanned, and 55 papers were included from the recursive reference search. Finally, 374 articles were considered for the review of this study.

3) STUDY QUALITY ASSESSMENT

This section evaluates each selected article depending on the established questions. All the questions were prepared by following the guidelines of articles [42], and [44] and are presented in Table 5. Each question was answered using Yes, Partly, and No, where Yes $= 1$, Partly $= 0.5$, and No $= 0$. The total value stays between 0 (very poor) and 6 (excellent) according to the number of questions in Table 5. For the selection, the scores of each article must be four or above. Table 6 provides an example of the assessment process using five articles.

4) DATA EXTRACTION

This step introduces extracted features, where all extracted features come from a different perspective: metadata, task, representation, translation, alignment, fusion, and co-learning. All extracted features from metadata involve articles underlying information. From the task outlook, extracted features will give the idea of applied ML algorithms, selected datasets, and their modalities. The last five outlooks are the challenges of MML and features related to their types. Table 7 shows a list of outlooks and corresponding features with definitions.

5) DATA ANALYSIS

A comprehensive analysis was performed on extracted data from multiple outlooks, as shown in Table 7 by the authors of this article. The source feature of metadata gives the notion of the venues of the published articles. Subsequently, articles were clustered depending on domains and applications. Papers were grouped into algorithm, data, and modality in the task prospect. The algorithm feature shows distinct ML algorithms used to tackle the challenges of MML. The modality feature gives a clear view of sources from which data appear. Additionally, the collected articles were clustered into individual elements of the last five perspectives. All the clustering and investigations performed in this review to understand and summarise current trends on MML.

V. RESULTS

A. METADATA

This section of the article presents all the metadata extracted from the collected paper. According to the inclusion and exclusion criteria, all selected articles were related to MML and its challenges. From 374 selected papers, 268 reports were published in journals, and 106 were in conferences. A total of 28 publication venues were identified from the selected documents. Figure 6 shows the highest five venues with the total number of journals and conferences. Most of the relevant articles were published in the IEEE Xplorer. ELSEVIER is another popular venue after IEEE Xplorer for publishing similar journals. The number of articles from different countries is analysed in this paper. In total, 51 countries were encountered; among them, China, the USA, Germany, and the UK dominated the most. Figure 7 presents the top ten countries with the most published articles in journals and conferences.

The total number of articles and citations according to the years 2009 to 2022 is displayed in figure 8. From the figure, it is visible that every year, research interest in MML is growing. The rapid growth is evident from 2018. Increment in citations each year is also apparent in the figure.

From the metadata, 34 domains and 35 associated applications are encountered. Figure 8 depicts the distribution of articles depending on domain and application. In the figure, all the domains are placed on the right side, and their arrowheads link to all associated applications. In the domain section, most of the papers are relevant to the medical domain, then domain agnostic, human activity, and emotion recognition accordingly. Conversely, in the application part, the maximum number of articles are related to recognition, classification, detection, and prediction. In this paper, selected articles are clustered based on the domains, wherein in each domain, articles are clustered again depending on applications. In the medical domain, from 122 articles, 19 were in classification [49], [50], [51], [52], [53], [54], [55], [56],

TABLE 4. List of inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
1. Studies must fall in the area of MML.	1. Studies not related to MML or its challenges.
2. Studies should answer one or two research questions.	2. Studies not written in English.
3. Published between the year 2009 to 2022.	3. Studies present a survey or review.
4. Must publish in a journal and conference.	4. Duplicate articles with a similar result.
5. Focus on MML or its challenges.	5. Unavailable in electronic format.

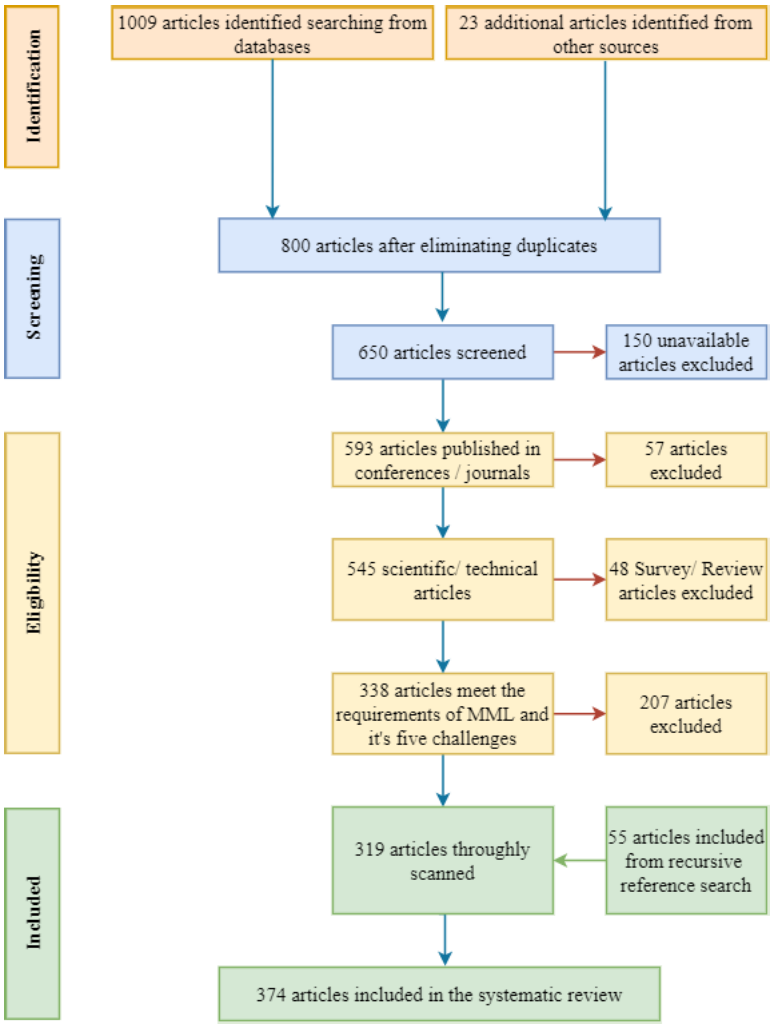


FIGURE 5. Flow diagram of article selection steps using PRISMA.

TABLE 5. Study quality assessment questions.

S/L	Questions
Q1	Does the topic of the study cover the answers to the research questions?
Q2	Are the objectives defined precisely?
Q3	Are the applied approaches described in the study?
Q4	Does the study include links to prior studies?
Q5	Are the results of the study adequately evaluated?
Q6	Is the study referenced accurately?

TABLE 6. Example of evaluating five articles applying quality assessment questions.

Refs	Q1	Q2	Q3	Q4	Q5	Q6	Total Score
[45]	1	1	0.5	1	0.5	1	5
[46]	0.5	1	1	0	0	1	3.5
[47]	1	0.5	1	1	1	1	5.5
[48]	1	1	1	1	1	1	6
[49]	1	1	1	0.5	0.5	1	5

[57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], 26 were in prediction [48], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], 19 were in detection [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], 16 were in diagnosis [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125],

TABLE 7. List of extracted features.

Outlook	Features	Definition
Metadata	Source	Name of journal or conference where the paper was published.
	Domain	The domain of the paper.
	Application	Particular application developed in the paper.
Task	Algorithm	Employed ML algorithms to perform classification.
	Data	Data used to train developed ML models.
	Modality	Form of the source from where distinct types of data appear.
Representation	Joint	Concatenating each modality for representation.
	Coordinated	Coordinate modality through constraint.
Translation	Example-Based	Translate data based on training data.
	Generative	Construct models instead of using unimodal sources.
Alignment	Explicit	Align instances of subcomponents from two or more modalities.
	Implicit	Latently align data during model training.
Fusion	Model-agnostic	Fusing data before knowing the model.
	Model-based	Data fuse while applying the model.
Co-learning	Parallel	Sharing a set of instances between modalities.
	No-parallel	Share categories instead of instances of modalities.
	Hybrid	Use of shared modality as a bridge between modalities.

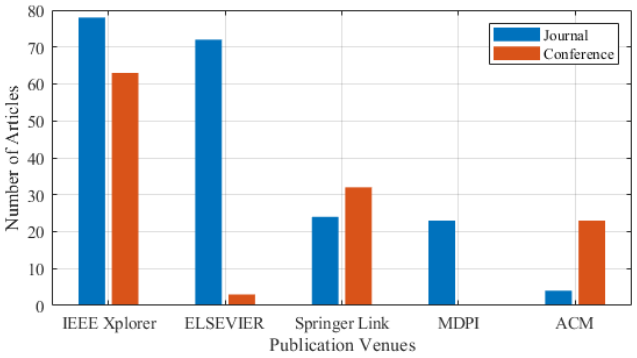


FIGURE 6. Articles published in journals and conferences from 2009 to 2022.

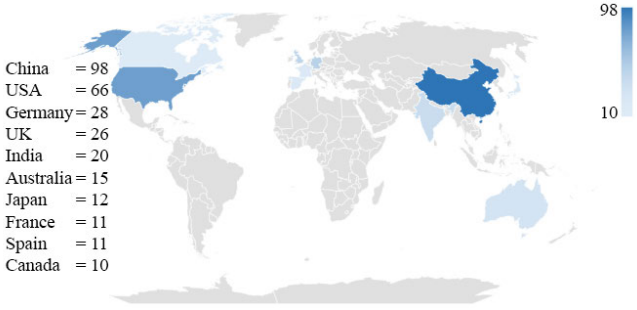


FIGURE 7. Aarticles published in different countries from 2009 to 2022.

[126], [127], 10 were in recognition [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], 9 were in segmentation [138], [139], [140], [141], [142], [143], [144], [145], [146], 10 were in analysis [19], [147], [148], [149], [150], [151], [152], [153], [154], [155], 3 were in image processing [156], [157], [158] and 2 were in assessment [159], [160]. Identification, augmentation, screening, evaluation, measurement, and differentiation each had 1 article [161], [162], [163], [164], [165], [69]. Language processing [166], [167] has 2 articles in the medical domain. In total, 60 articles were found related to the domain agnostic domain, where, 16 were in classification [168], [169], [170], [171], [172], [173], [174], [175], [176], [177], [178], [179], [180], [181], [182], [183], 9 were in detection [184], [185], [186], [187], [188], [189], [190], [191], [192], 11 were in analysis [193], [194], [195], [196], [197], [198], [199], [200], [201], [3], [202], 8 were in recognition [203], [204], [205], [206], [207], [208], [209], [210], 9 were in prediction [211], [212], [213], [214], [215], [216], [217], [218], [219], 1 was in language processing [220], 2 were in image processing [221], [222], 1 was in image retrieval [47], 2 were in integration [223], [224] and 1 was in segmentation [225]. Of 43 articles in the human activity domain, 20 were in recognition [226], [227], [228], [229], [230], [231], [232], [233], [226], [234], [235], [236], [237], [238], [239], [175], [240], [241], [242], [243], 8 were in detection [244], [245], [246], [93], [247], [248], [249], [250], 5 were in classification [251], [252], [253], [254], [255], 4 were in analysis [256], [257], [258], [259], 3 were in identification [260], [261], [262] and 1 was in comparison [263], monitoring [264] and assessment [265]. In the emotion recognition domain, 28 articles are encountered; from them, 21 were in recognition [266], [267], [268], [269], [270], [271], [272], [273], [274], [275], [276], [277], [278], [279], [280], [281], [282], [283], [284], [285], [286], 2 were in analysis [287], [288], and prediction [289], [290], and 1 was in detection [291], assessment [292], and estimating [293]. In total, 17 papers were found in the security domain, where 6 were in detection [186], [294], [295], [296], [297], [298], 4 were in identification [299], [300], [301], [302], 2 were in authentication [303], [304] and recognition [305], [306], and 1 was in determination [307], verification [308] and filtering [309]. Of the 13 papers in the biometric domain, 6 were in recognition [310], [311], [312], [313], [314], [315], 2 were in detection [316], [317], 3 were in authentication [318], [319], [320] and 1 was in transforming [321] and classification [180]. A total of 9 articles are related to the robotics domain; from them, 4 were in recognition [322], [323], [324], [325], 2 were in prediction [326], [327], 1 was in language processing [328] and 2 were in detection [329], [330]. From 8 articles in the image domain,

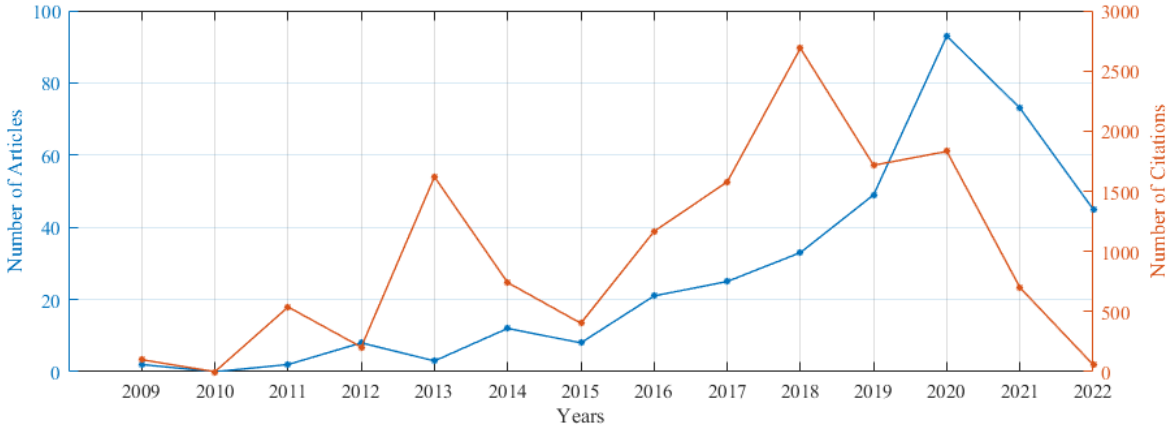


FIGURE 8. Figure of published articles and citations between 2009 to 2022.

2 were in analysis [331], [332], translation [333], [334] and caption generating [335], [336], and 1 was in retrieval [337] and description generating [338]. Of 8 papers in the geoscience domain, 5 were in classification [339], [340], [341], [342], [343], 1 was in recognition [344], detection [345], and precision [346]. In total 10 articles were identified in the social media domain, where 4 were in detection [347], [348], [349], [350], 2 were in analysis [351], [352] and 1 was in classification [353], prediction [354], recommendation [355] and verification [356]. In the vehicle domain, 6 articles were found; from them, 2 were in recognition [357], [358], 1 was in identification [359] and 3 were in detection [360], [361], [362]. There were 5 articles related to the text domain, with 4 were in processing [363], [364], [365], [366], and 1 was in classification [367]. Of the 5 papers in the visual domain, classification [368], detection [369], prediction [370], identification [213] and extraction [371] have 1. The entertainment domain has 6 relevant articles, where 2 were in segmentation [372], [373], 1 was in prediction [374], 2 were in detection [375], [376] and 1 was in recognition [377]. The sports domain has 4 papers: 2 in classification [378], [379] and 1 in recognition [380] and segmentation [381]. The audio domain has 3 articles, where recognition [382], detection [383] and translation [384] each have 1. Detection [385], prediction [386] and translation [387] each have one paper from 3 papers in the game domain. Each of the three articles in the biology domain pertained to recognition [388], prediction [389] and identification [390] accordingly. In total, 4 bird species domain relevant papers were identified, where 3 were in classification [45], [391], [392] and 1 was in integration [393]. There were 2 articles each in the signal processing and gender domains. In the signal processing domain, recognition [394] and analysis [395] each have 1; in the gender domain, classification [396] and recognition [397] each also have 1 article. The industry, fashion, e-commerce, energy, mapping, academia, product, food, cooking, plant, language, chemistry and mobile application domains have 1 relevant article in the following applications accordingly: prediction [398], retrieval [399], classification [400], classification [401], prediction [402], recognition [403], recognition [404], inspection [405], retrieval [406], segmentation [407], translation [408], analysis [409], and detection [410].

B. TASK

This section represents extracted information related to algorithms and used data from selected articles. Each article applied distinct types of ML algorithms to perform tasks. In this survey totally of 79 kinds of algorithms were extracted from the collected papers. All gathered algorithms clustered based on their types and methods. Supervised, semi-supervised, unsupervised, and reinforcement are four types of ML algorithms. First, the extracted algorithms were grouped into supervised, semi-supervised and unsupervised. The collected papers show that the number of supervised, semi-supervised and unsupervised algorithms is 63, 10 and 6 respectively. Then, the extracted algorithms were clustered into 11 types of standard models: neural networks (NNs), support vector machines (SVMs), ensemble models (EMs), nearest neighbour models (NNMs), tree-based models (TB), Bayesian models (BM), linear models (LM), genetic algorithms (GA), graph-based models (GBMs), encoder-decoder (ED), and k-means. Each of the models is associated with supervised, semi-supervised and unsupervised types. For example, algorithms related to NNs can be supervised, semi-supervised and unsupervised that's why there is a link between them visible in Figure 9. The total number of encountered NNs, SVMs, EMs, NNMs, TB, BM, LM, GA, GBMs, ED and k-means algorithms is 49, 10, 4, 2, 2, 4, 3, 1, 1, 2, and 1 respectively. Finally, the collected algorithms were clustered into 79 specific kinds, and each connected with 11 previously categorised types of standered models. The names of 79 algorithms are Artificial Neural Network (ANN), AlexNet, Bidirectional Recurrent Neural Network (BRNN), Convolutional Neural Network (CNN), CNN-Waterfall, CNN Long-Short Term Memory

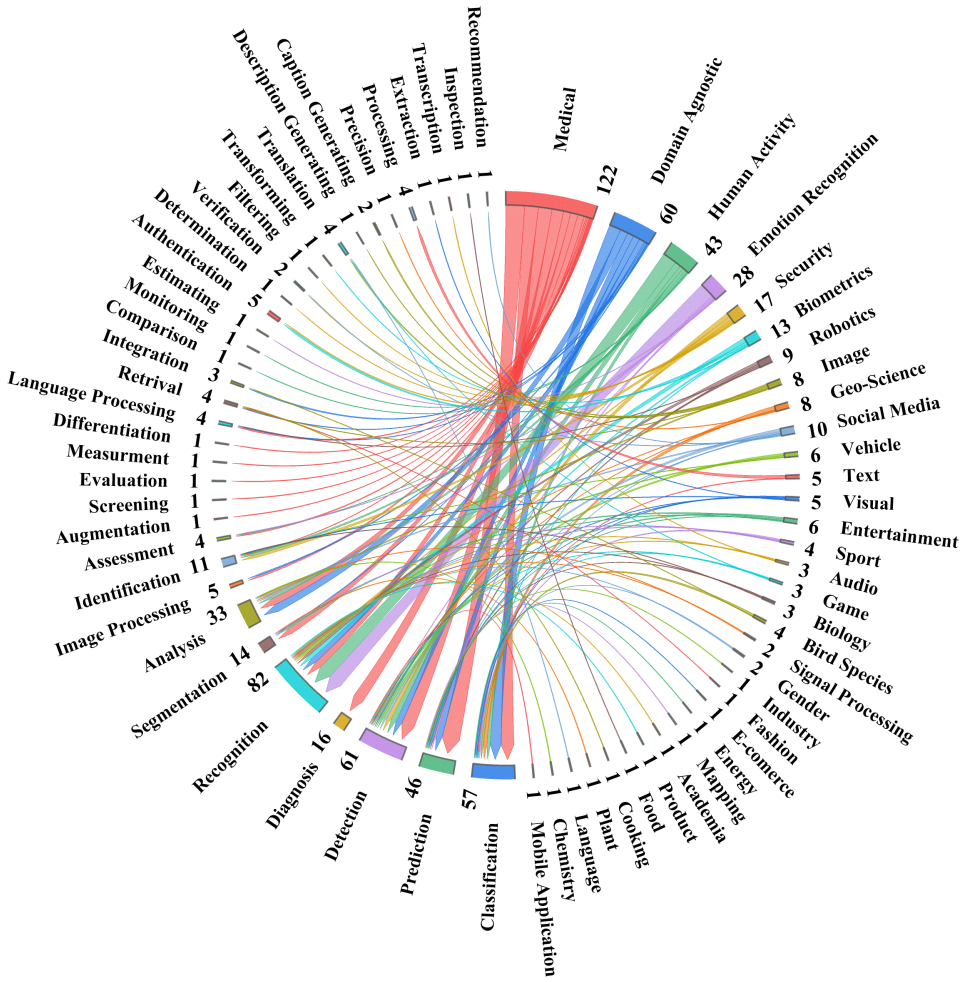


FIGURE 9. Distribution of articles by application and domain.

(CNN-LSTM), Convolutional Long-Short Term Memory Fusion Network (CNN-L), Chaotic Neural Network, Conditional Random Fields – RNN (CRF-RNN), Deep CNN (DCNN), Deep Neural Network (DNN), Deep Polynomial Network (DPN), Dilated and Transposed CNN (DT-CNN), Dynamic Context-Guided Capsule Network (DCCN), Extreme Learning Machine (ELM), Fast-Shaped-Based Network (FS-Net), Feed Forward Neural Network (FNN), Fully Connected Neural Network (FCNN), FCNN-CRF, iMageNet, Inception, LSTM, Liquid State Machine (LSM), MobileNet, Multi-Group Norm Constraint CNN (MGNC-CNN), Multiple Image CNN Long-Short Term Memory (MICNN-L), Recurrent Neural Network (RNN), Residual Neural Network (ResNet), Region-Based Convolutional Neural Network (RCNN), RNN-LSTM, Squeeze-and-Excitation Based ResNet (SE-ResNet), SE-CNN, TextNet, Text-CNN, Temporal CNN (T-CNN), Traffic Sign Recognition CNN (TSR-CNN), U-Net, Visual Geometry Group (VGG), VolNet, Bidirectional Long-Short Term Memory (Bi-LSTM), Bidirectional Convolutional Long-Short Term Memory (BC-LSTM), Gated Graph Convolutional Network (GGCN), Multimodal Semantic Model (MSM), Bidirectional GRU Attention (Bi-GRU-Attention), Cycle Generative Adversarial Network (CycleGAN), Deep Belief Network (DBN), Generative Adversarial Network (GAN), Point-Wise GBM (PGBM), Restricted Boltzmann Machine (RBM), LogDet Support Vector Machine (L-SVM), Multi-Kernel SVM (MK-SVM), Particle Swarm Optimization SVM (PSO-SVM), PCA-SVM, SVM with Radial Bias Function (SVM-RBF), Support Vector Regression (SVR), String Kernels SVM (SK-SVM), Support Vector Clustering (SVC), Singular Value Decomposition (SVD) Adaptive Boosting (AdaBoost), Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Reduced Error Pruning Tree (REPTree) k-Nearest Neighbors (k-NN), Distance-Weighted kNN (WKNN) Decision Tree (DT), J48 Bayesian, Bayesian Logistic Regression, Bayesian Rule List, Naïve Bayes Linear Regression (LnR), Logistic Regression (LgR), Linear Dicriminat Analysis (LDA), K-Means Clustering Correlational Multimodal Variational Autoencoder Via Triplet Network (CMVAE-TN), Multimodal Deep Autoencoder (MADE), Genetic Algorithm with aggressive mutation

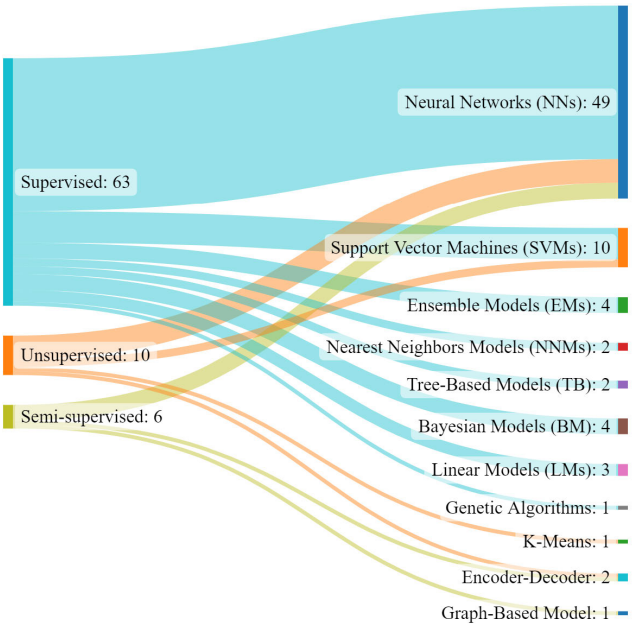


FIGURE 10. Sankey diagram of intensified algorithms.

(GAAM), Graph-based Transductive Learning (GTL). The total number of presences of each algorithm and links with previous categories is presented in supplementary file at table 1. Table 1 in the supplementary helps to understand what types of algorithms are used to solve problems related to MML.

This paper divided the extracted data from the collected articles into six modalities: video, image, audio, text, sensor, signal, and numerical. Data from each modality are applied individually or combined with others. In 374 articles, image modality was present in 247, which were image data individually used in 129 articles [19], [49], [51], [52], [54], [56], [57], [59], [60], [61], [63], [64], [66], [73], [74], [76], [79], [83], [84], [85], [87], [88], [89], [94], [95], [96], [99], [103], [105], [106], [108], [110], [112], [113], [114], [115], [117], [120], [121], [122], [123], [128], [130], [134], [135], [137], [138], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [150], [151], [153], [154], [156], [157], [159], [161], [170], [171], [172], [174], [176], [189], [192], [193], [194], [196], [200], [201], [206], [207], [209], [211], [212], [214], [218], [219], [221], [222], [225], [242], [250], [261], [262], [266], [285], [300], [301], [303], [304], [305], [310], [312], [313], [314], [315], [316], [320], [321], [180], [325], [329], [331], [333], [335], [336], [337], [338], [339], [340], [342], [351], [361], [364], [365], [388], [391], [393], [405], [407] and 118 articles combined with other types. Video modality was found in a total of 45 articles, with 11 papers [164], [215], [263], [322], [370], [371], [372], [373], [378], [379], [381] separately and 34 combined with data of different modalities. The use of audio modality was found in 72 articles; 5 were used individually [129], [239], [270], [278], [324], and 67 were used with others. The occupancy of text modality was found in 93 papers, and from them, it was individually used in 11 articles [68], [48], [82], [158], [166], [167], [198], [240], [309], [346], [410] and combinedly used in the rest of 82 articles. Sensor modality was identified in 40 papers, 16 of which were individually identified [65], [80], [136], [162], [165], [227], [228], [229], [241], [253], [264], [295], [298], [344], [385], [394] and 24 of which were grouped with others. In total, 41 studies utilised signal data, with 27 articles combined with others and 14 articles [97], [118], [131], [132], [133], [223], [243], [272], [273], [274], [283], [293], [317], [359] using it alone. Numerical data were utilised in 38 papers, 21 of which were used alone [70], [71], [72], [77], [78], [90], [92], [188], [217], [224], [291], [294], [296], [318], [343], [360], [374], [389], [390], [395], [409], and 17 of which were used with others. In total, 26 distinct combinations were identified using data of six different modalities. They are Video & Image & Audio & Text [398], Video & Image & Audio [377], Video & Image [233], Video & Audio & Text [186], [256], [259], [267], [268], [271], [288], [186], [368], [213], [384], Video & Audio & Sensor [93], [290], [400], Video & Audio [152], [169], [226], [247], [249], [260], [269], [275], [281], [286], [289], [292], [308], Video & Text [197], Video & Sensor [265], Video & Signal [245], [277], Image & Audio & Text [111], [173], [3], [210], [213], [257], [258], [282], [332], [387], [397], Image & Audio & Sensor & Signal [358], Image & Audio & Sensor [244], [326], [327], Image & Audio [175] [180], [182], [184], [187], [190], [205], [230], [231], [226], [234], [235], [175], [279], [330], [45], [403], Image & Text [47], [53], [69], [81], [86], [91], [102], [107], [124], [155], [160], [163], [168], [177], [179], [183], [191], [195], [199], [208], [216], [220], [246], [280], [284], [287], [299], [302], [307], [328], [334], [347], [348], [349], [350], [352], [353], [354], [355], [356], [362], [363], [366], [367], [376], [399], [404], [406], [408], Image & Sensor & Signal [93], Image & Sensor [50], [55], [58], [101], [149], [204], [232], [248], [255], Image & Signal [67], [98], [100], [104], [116], [125], [127], [185], [203], [237], [319], [357], [386], [402], Image & Numerical [62], [75], [119], [126], [69], [306], [323], [345], [396], [401], Audio & Text & Sensor [375], Audio & Text [178], [276], [369], [382], [383], Text & Signal [109], Text & Numerical [297], [341], Sensor & Signal [236], [238], [252], [380], Sensor & Numerical [181], Signal & Numerical [202], [251], [254], [311]. Figure 10 displays the extracted information related to each modality and data type with the links between them. The presented result shows that image data and the combination of image and text data are used more than others.

C. REPRESENTATION

This section briefly discusses the extracted paper on multimodal representation learning. According to article [1], the types of representation learning are divided into several subparts. Neural networks (NNs), graphical models, and sequential are subtypes of joint representation, and similarity and structured are two sub-types of coordinated representation. From 374 MML-related papers, 82 articles were associated with representation learning. Joint and coordinated

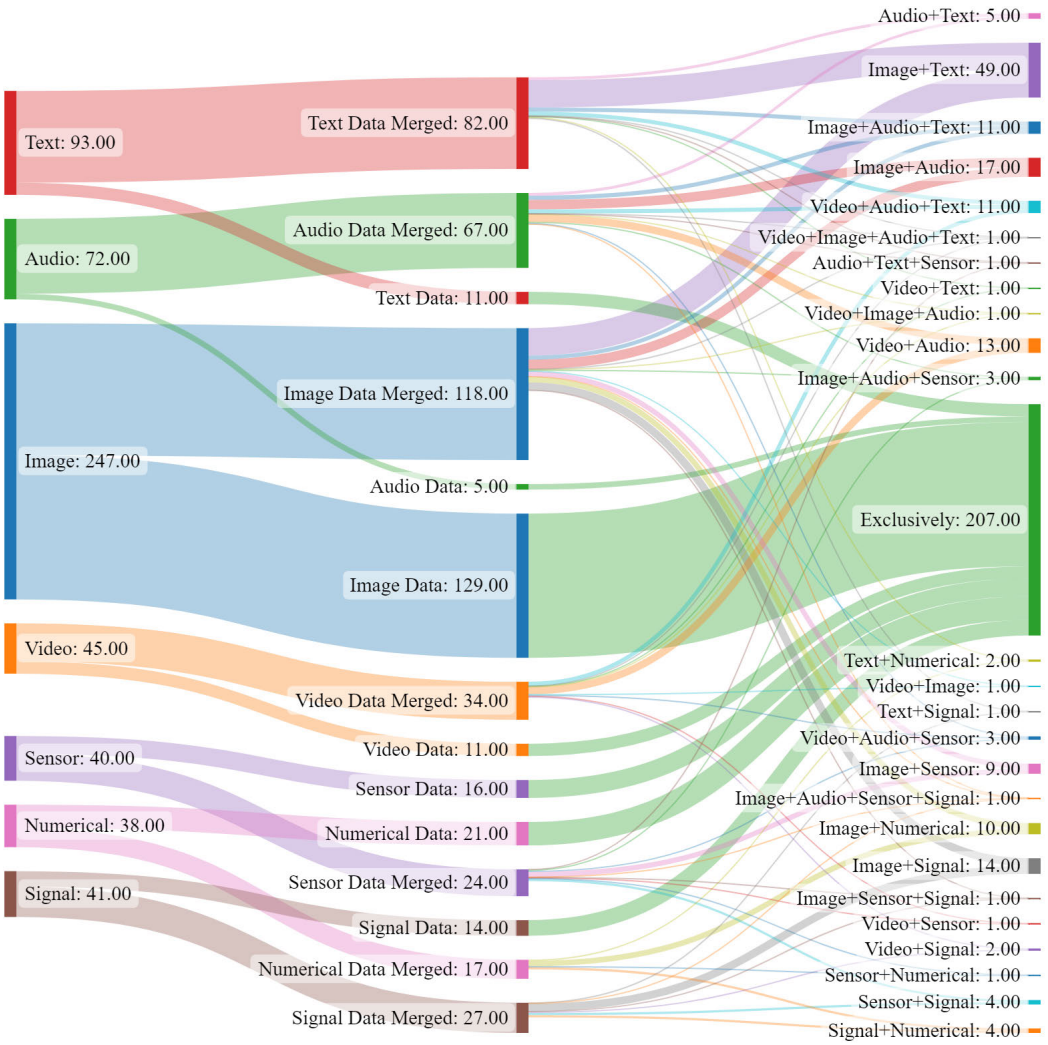


FIGURE 11. Modality wise data distribution.

representation each have 70 and 12 articles. In joint representation, for the NNs, graphical and sequential model subtypes each has 46 [47], [50], [68], [69], [70], [71], [72], [77], [88], [89], [90], [92], [93], [138], [139], [159], [174], [176], [177], [183], [184], [185], [196], [200], [213], [214], [217], [218], [227], [233], [251], [260], [261], [266], [269], [294], [310], [337], [340], [344], [348], [370], [398], [399], [409], [410], 10 [49], [112], [113], [128], [191], [201], [224], [244], [287], [374], and 14 [73], [94], [95], [116], [166], [226], [175], [267], [268], [283], [186], [328], [347], [354] articles, respectively. In total, 7 and 5 articles related to similarity [48], [51], [125], [168], [169], [262], [330] and structured [64], [197], [208], [209], [256] coordinated representation were found. Figure 11 shows the flow of representation learning in a Sankey diagram with article numbers.

D. TRANSLATION

In this section, collected publications on multimodal translation learning are discussed. This paper found 27 articles on translation learning. It has two types, example-based and generative, where example-based has two subtypes: retrieval and combination. A total of nine papers related to example-based were identified, six of which were related to retrieval [156], [167], [182], [222], [335], [387] and two of which related to combination [170], [198]. Grammar-based, encoder-decoder, and continuous are three subtypes of generative. Of 19 articles, one was grammar-based [364], twelve were encoder-decoder based [110], [154], [192], [219], [285], [331], [333], [334], [363], [365], [366], [408], and six were continuous-based [129], [175], [239], [288], [336], [338]. The flow of translation learning with its types is depicted in figure 11.

E. ALIGNMENT

Multimodal alignment learning is split into two kinds: explicit and implicit. Each category is subdivided into two types: supervised and unsupervised related to explicit and graphical models, and neural networks, related to implicit models. The flow of alignment learning in the Sankey diagram is presented in Figure 11. Eighteen alignment learning-related articles

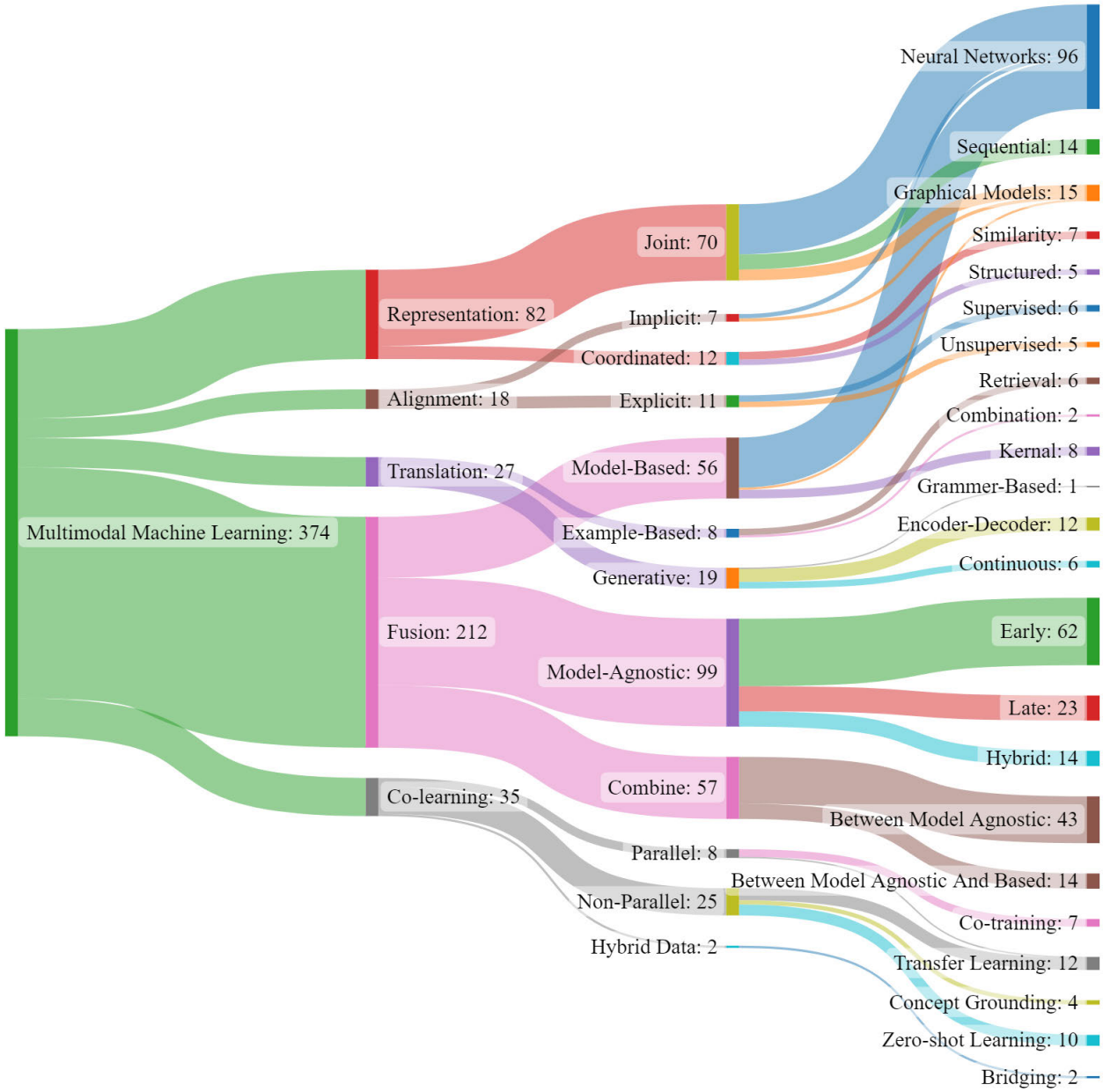


FIGURE 12. Challenge wise extracted result.

were found from the total number of articles. Of 18 articles, 11 were related to explicit alignment, and 7 were related to implicit alignment. In explicit alignment learning, six articles were related to supervised [52], [96], [19], [157], [379], [381] and five articles were related to unsupervised [53], [140], [259], [395], [406]. There are three articles on graphical models [54], [147], [407] and four about neural networks [130], [158], [240], [332] in implicit learning.

F. FUSION

Fusion is the most prevalent challenge to others. According to [1], it has two types: model-agnostic and model-based. Early, late and hybrid are three subtypes of model agnostic fusion. Model-based fusion is divided into Kernal, graphic models, and neural networks. A total of 212 articles related to fusion learning were encountered. Of 155 articles, 99 were model-agnostic, where 62 pertained to early [55], [56], [58], [59], [62], [63], [75], [76], [79], [98], [102], [103], [104], [105], [111], [115], [119], [120], [133], [141], [142], [165], [171], [204], [210], [236], [238], [245], [93], [248], [252], [253], [264], [265], [274], [276], [293], [296], [298], [299], [300], [306], [313], [317], [319], [322], [326], [329], [339], [341], [349], [351], [356], [359], [372], [373], [375], [382], [384], [388], [396], [397], 23 pertained to late [114], [127], [136], [152], [160], [69], [172], [178], [179], [215], [237], [250], [258], [270], [273], [301], [315], [352], [357], [358], [378], [380], [383] and 14 pertained to hybrid [180], [187], [247], [290], [303], [304], [305], [308], [309], [311], [312],

[316], [318], [371]. In all, 56 model-based studies were discovered, with 46 relating to NNs [91], [97], [99], [121], [124], [126], [132], [143], [144], [145], [148], [151], [153], [155], [161], [162], [163], [195], [3], [202], [211], [225], [228], [230], [226], [243], [255], [271], [284], [286], [289], [292], [320], [321], [343], [345], [360], [361], [376], [377], [389], [390], [393], [401], [402], [404], 8 to kernels [45], [57], [106], [122], [134], [186], [314], [386], and 2 to graphical models [61], [385]. The remaining 57 articles were related to combined models of model-based and agnostic-based models. Seven types of the combination were found: early & late [65], [78], [80], [86], [87], [109], [118], [123], [131], [135], [150], [164], [188], [203], [231], [232], [235], [246], [249], [263], [272], [275], [291], [295], [297], [307], [342], [350], [353], [355], [394], [403], [405], early & late & hybrid [74], [117], [229], [254], [257], [278], [323], [346], [362], early & late & kernel [149], [234], early & late & neural networks [60], [67], [181], [282], [400], early & neural networks [81], [223], [367], late & hybrid [302], and late & neural networks [100], [101], [137], [277]. The combination of early and late fusion was found most than others. The flow of fusion learning, along with the number of articles, is presented in Figure 12.

G. CO-LEARNING

Co-learning is the final challenge in MML. Parallel, non-parallel and hybrid is the co-learning types, where the parallel is subdivided into co-training and transfer learning. Transfer learning, concept grounding, and zero-shot learning are the subtypes of non-parallel learning. Hybrid has only one type, which is bridging. This paper extracted 35 articles related to co-learning, where eight related to parallel, 27 related to non-parallel and two related to hybrid. Of eight articles in parallel, seven pertained to co-training [66], [190], [206], [207], [180], [368], [213] and one related to transfer learning [199]. On the other side, from 25 non-parallel articles, 11 pertained to transfer learning [83], [107], [108], [146], [205], [212], [220], [241], [279], [280], [325], four related to concept grounding [82], [216], [324], [327] and 10 related to zero-shot learning [84], [85], [173], [189], [221], [242], [281], [369], [391], [392]. In hybrid co-learning, two articles were related to bridging [193], [194]. Figure 12 depicts the flow of co-learning with the number of articles.

In Figure 13, the number of articles related to five MML challenges between the years 2009 to 2022 is depicted. The figure shows that the number of publications increases each year.

VI. DISCUSSION

Initially, the primary search using keywords shows that interest in research on MML is gaining. However, one or two challenges of MML were known to researchers more than ten years ago, but it came into the spotlight after 2016 as a core research topic. All the defined research questions mentioned in Section IV are discussed in this section and introduce perceptions. Figure 14 shows various contributing parts of this review study.

A. INSPECTION OF RQ1

The first research question is made up to investigate the definition of MML. MML stands for building models that can process and describe data from multiple modalities. It is necessary to understand the relation between multimodal and ML to determine the definition of MML more precisely. Multimodal or multimodality is variously used in academic literature. According to [1], multimodality means including multiple modalities such as image, text, audio, etc. Based on the definition of multimodality from [1], MML handles data from multiple modalities using ML. A recent article [411] defines MML as a machine learning system that receives and processes data from multiple modalities. They also characterize multimodality in three types: human, machine and task centred. Human-centred means the way data communicate through humans. Machine-centred means data are encoded using the ML system before being processed. Task-centred is related to the job an ML system needs to perform, and based on that, data input and output are represented differently. Human and machine-centred definitions aim to catch the summary of multimodality in a task-agnostic manner. On the other hand, the task-centred approach attempts to discover the involvement of each input depending on the relations to the task. Therefore, from the above definitions, article [411] argued that the purpose of MML is not only to build ML models to process multiple modalities but also to focus on the relation between modalities and the given task. Depending on the overhead discussion, two statements can be possible to establish for MML. First, in MML, ML models build to process data; the second is those data come from multiple modalities and is related to tasks.

B. INSPECTION OF RQ2

RQ2 is designed to inspect the use of five challenges of MML in research. Each challenge of MML solves specific tasks; for example, fusion techniques are used to fuse information from two or more modalities, and translation maps one modality to another to translate information. The first challenge of MML is representation, which is crucial because of the difficulties in representing heterogeneous data in a meaningful way. A good representation of data is necessary to support the performance of ML models, which is visible in recent articles such as [49] and [50]. Paper [1] classified representation learning and extracted articles depending on it, which are presented in Figure 12. Figure 12 shows that neural networks (NNs), a subtype of joint representation, have more consideration than others to represent data. The use of NNs to represent visual, audio and text data is increasing [177], [183], [184], [213]. Graphical models frame representation using implicit random variables where probabilistic graphical models such as deep Boltzmann machines (DBM) [113] are used. The sequential representation uses sequential models where hidden states are considered to represent the data, such as recurrent neural networks (RNNs) [283], [347], [354]. From three subtypes of joint representation, NNs were used

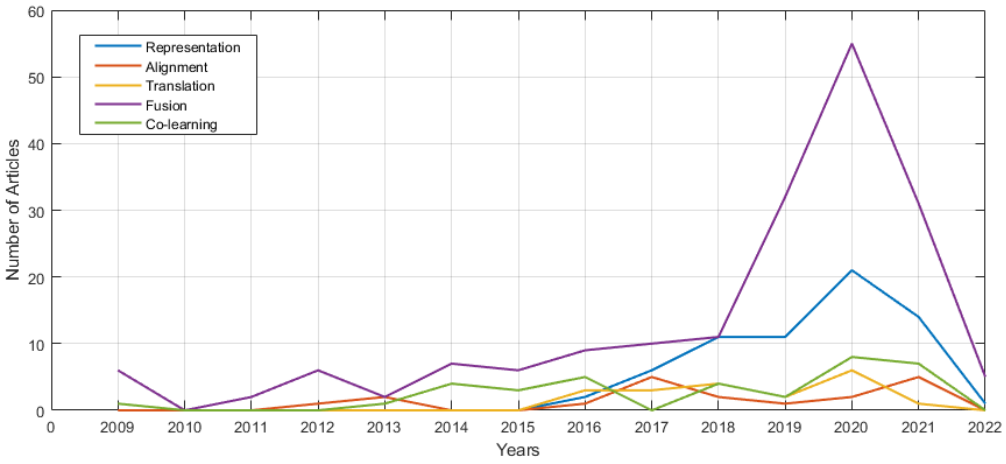


FIGURE 13. Number of published articles related to five challenges between 2009 to 2022.

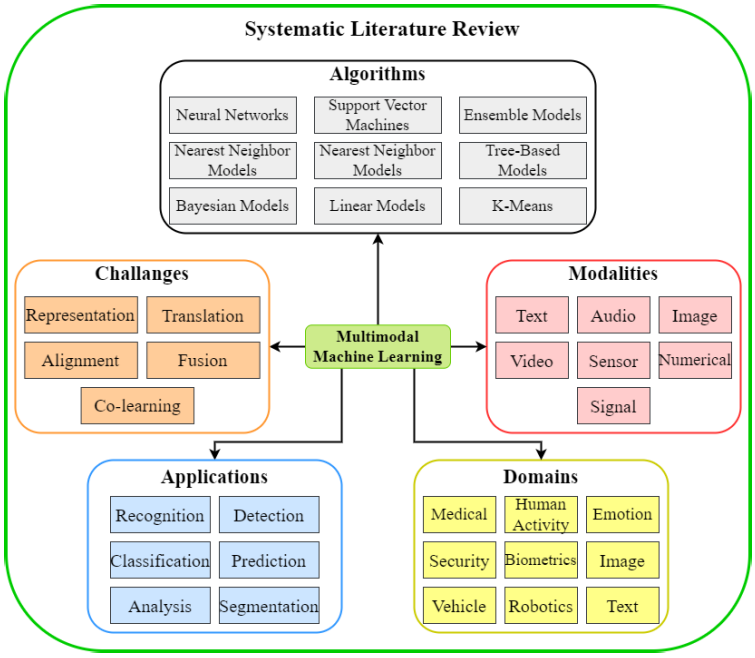


FIGURE 14. Various contributing part of the review.

primarily because of their superior performance, but they could not handle missing data. However, graphical models can handle missing data and the whole modality. On the reverse side, the sequential model is used to represent the sequences of data. Similarity and structure are two subtypes of coordinated representation. The similarity models work on the distances of two modalities [169], [330]. Structured models are typically used in crossed hashing [64], such as canonical correlation analysis (CCA) [197], [209]. In comparison, joint representation is best suited for all modalities as opposed to a coordinated representation suitable for application when only one modality is present at test time. More than two modalities are used in joint representation, but coordinate representation is limited to two.

The extracted results on multimodal translation learning and its types are also displayed in Figure 12, according to article [1]. Interest in research on this topic is currently growing. Retrieval is a simple example-based multimodal translation learning where it tries to find the nearest sample in the dictionary to produce a translated result [335], [387]. CNN and kernel canonical correlation analysis (KCCA) are famous for retrieval-based models such as [156], [182]. Conversely, combination-based models combine the piece from the dictionary meaningfully to generate better translation [170], [198]. Grammar-based models produce translation on a specific domain by applying grammatical conditions, as in [364]. The encoder and decoder model first encodes the origin modality to a latent representation and then decodes it to produce the target modality, where CNNs use most of the cases [331], [333], [334], [365]. Continuous generation models have outputs at each timestamp, such as sequence-to-sequence translation [239], [288]. RNNs with long short-term

memory (LSTM) usually use continuous generation models like [239], [288], and [338]. In contrast, example-based is simple, but it makes the model heavy because the model itself acts as a dictionary and sometimes generates unrealistic translations. Generative models are crucial to constructing since models need the capability to produce sequences of symbols. This reason has led many researchers to choose example-based models to provide the solution.

Figure 12 depicts the extracted results on multimodal alignment learning. Explicit and implicit are two kinds of alignment, where each is split into two types [1]. Unsupervised explicit alignment aligns a modality without having any direct labels, as in [53], [157], [259], [395], and [406], where CCA and dynamic time wrapping (DTW) with CCA models are usually applied. Supervised explicit alignment methods, on the other hand, count on the labelled instance, for example, [19], [52], [96], [157], [379], and [381]. Graphical implicit models align modalities based on mapping them, such as aligning images and text [54], [407], or images and signals [147]. Neural network implicit models align modalities using encoder-decoder [130], [158], [240] or cross-modal retrieval [332] techniques. In implicit alignment, latent information is used for alignment to perform tasks that provide better performance. In contrast, explicit alignment focuses on subcomponents of modalities.

Fusion is the most common challenge in MML. Extracted information regarding different types of fusion according to [1] is presented in Figure 12. Model-agnostic fusion has three kinds: early, late, and hybrid fusion, where early and late are used most often. Early fusion unites features directly when they are extracted as in [56], [151], [263], and [305]. On the other hand, late fusion is performed after making decisions such as voting [55], [114] and weighting [272] schemes. Late fusion is also known as decision fusion. Hybrid fusion integrates outputs of early fusion and individual unimodal predictors, for example, [74], [117], [257], and [323]. Multiple kernel learning, graphical models, and neural networks are subtypes of model-based fusion. Fusion using the kernel approach means finding similarities between data points; for instance, kernel support vector machines (SVMs) are used to find the views in the data [45], [57], [106], [186]. Graphical models utilise the local and temporal structure of the modality, such as DBN [61] and dynamic Bayesian networks [62]. Neural network models are largely used in multimodal fusion [412]. In the NN, latent information from layers is fused to achieve better performance, as in [97], [162], [228], [271], and [289]. There are few specific comparisons between model-agnostic and model-based approaches. Depending on modalities and tasks, all the approaches are performed accordingly.

Co-learning is the final challenge of MML, and papers on co-learning are extracted according to article [1] presented in Figure 12. Parallel data co-learning is divided into co-training and transfer learning, where co-training generates more data, as in [207], [213], and [368], and transfer learning transmits information from one model to another to perform well [369]. In non-parallel data co-learning, models share concepts with each other. It is divided into transfer learning, conceptual grounding, and zero-shot learning. Similar to parallel learning, transfer learning is also possible in non-parallel learning [189], [279], [280]. Conceptual grounding shares the semantic concept with modalities where most of them are related to linguistics, such as [216], [324], and [327]. Zero-shot learning classifies data without having any labels, and approaches such as cross-models [189], [221] and auto-encoders [391], [392] are used for the solution. Bridging is hybrid co-learning where two non-parallel modalities share information. For example, the article [193], and [194] used bridged non-parallel modalities using neural networks.

1) INSPECTION OF RQ2.1

This research question establishes gaps and future directions in the challenges of MML. From the above discussion, it is visible that it is not easy to handle all the challenges simultaneously. That is why adopting five challenges to solve one problem is not necessary. Addressing one or two challenges relying on the primary task is sufficient to frame MML.

Representation and fusion are more studied challenges than others in MML. The relationship between representation and fusion is very close. When fusion is performed on any data, fused data will change its form to another, which is ultimately a new representation of the source data. Model-based fusion is similar to the joint representation. For example, NNs fuse two or more modalities and obtain a unique data representation as output. However, representation not always depends on fusion. Data can be represented in new forms while data are preprocessing, converting characteristics, etc. Apart relationship with fusion, representation also relates to translation and alignment. Whenever data is translated or aligned, it always gets a new form of representation. Of the two types of representation, the use of joint representation techniques will increase more because researchers can get facilities to apply multiple modalities.

Fusion is applied more in MML research than others. There are gaps in understanding the types and subtypes of fusion. Based on the definition of MML discussed in RQ1, data must be processed by the ML system. However, model-agnostic fusion means fused data before performing ML methods [1]. Therefore, based on the article [1], it will be wrong if any article claims multimodal fusion by performing model-agnostic fusion approaches. Conversely, suppose a paper argues that data is fused using a late or hybrid model-agnostic fusion approach. In that case, it is possible to consider multimodal fusion because models are usually used during late and hybrid fusion. So, to claim multimodal fusion, authors must prove that they have performed fusion by using models and avoid early fusion. The naming of the fusion approach is also confusing for the model-agnostic process. Early fusion is also known as knowledge fusion and feature fusion though some articles used NNs for fusion and named it to feature fusion. There is a lot of work done using fusion techniques, and more is coming. Combining fusion techniques with other

challenges to facilitate research work can be considered a future research direction.

Multimodal alignment finds cross-modal connections and interactions within elements of multiple modalities. It helps align separate modalities, such as video caption generation and image-text classification. Alignment is difficult because of the implicit dataset, and it is difficult to design similar metrics. NNs-based models such as autoencoders are popularly used for visual data alignment. However, researchers can primarily focus on three aspects to solve multimodal alignment problems. They are the identification of the connection between modalities, concept-wise modality representation and tackling the ambiguity in high-dimensional data.

Apart from representation, alignment and fusion, other challenges need attention. A considerable problem in multimodal translation is that methods are very problematic to evaluate. Developing generative approach models, such as encode-decoder and continuous generation models, is critical and always complex. It is easy to build example-based models, but the possibility of unrealistic results is high. Multimodal co-learning is new to most researchers. Subtypes of each type of co-learning are individually known, but their concept related to co-learning is new. Co-learning is task-independent, and it can help to address other challenges. One common problem of co-learning is that biased training samples often lead to overfitting. There is a lack of enough research works in translation, alignment, and co-learning. Making more robust models for those three challenges can open new research directions.

Aside from focusing on the challenges of MML, researchers can also focus on domains and applications. Figure 9 presents the results of domains and applications extracted from collected data. The figure shows that most problems stayed in medical, human activity and emotion recognition because of heterogeneous data availability. Instead, researchers can focus on those domains where data is less available and more complex. Recognition, classification detection, prediction and analysis are the most prominent application used, according to Figure 9. But after analysing the data, it can be said that the application is not much more impactful than the domain. Because anyhow, the problem must follow an application while solving it. Most of the MML-related surveys focus on how to handle multimodal data or the challenges of MML. Alternatively, researchers can apply MML to emerging concepts like explainable artificial intelligence [413], [414] and digital twin [415] for industry etc. Those arising can act as a new research direction for researchers.

C. INSPECTION OF RQ3

Modalities represent the manner in which something occurs or is perceived [1]. Image, audio, and text are examples of modalities. Extracted information related to modalities is presented in Figure 11. From the figure, it is visible that image and text modalities are used the most. The audio and video modalities come next accordingly. The main reason behind using visual modalities is that most of the MML-related problems are connected to the image and video, which can have multiple modalities. Audio and text modalities are associated with the visual modality. Applications such as caption generation and speech-to-text conversion are examples where video, audio, and text modalities come together to solve specific problems. Sensors, signals, and numeric data are easy to handle, and much research has already been performed.

D. INSPECTION OF RQ4

Different kinds of ML algorithms are used to solve MML problems. A summary of the applied algorithms is presented in Figure 10. It shows that most of the research applied NN-related models to deal with MML-related issues. Two reasons work behind it. One is the modality, and the other is the type of task. Visual, audio, and text modalities are easy to process using NN models such as CNN, RNN, DBM, etc. Types of tasks also play an essential role in using NNs because most problems relate to image-to-text, speech-to-text, speech-to-speech generation, etc., where NN models perform better. Apart from NNs, SVMs, ensemble models (EMs), nearest neighbour models (NNMs), tree-based (TB), Bayesian models (BMs), linear models (LMs), K-means, encoder-decoder, genetic algorithms, and graph-based models have been applied by researchers. From figure 10, it is also evident that most applied algorithms are supervised learning. Only for some specific tasks were semi-supervised and unsupervised methods used, such as graphical related models, which use unsupervised methods.

E. INSPECTION OF RQ4

VII. CONCLUSION

This article performed a systematic literature review on MML and its challenges to provide an overview of recent trends. The study was conducted utilizing the PRISMA approach, and its selection procedure was reported in detail. In total, 374 articles were selected from an initial 1032 collected articles depending on their relevance to the four created research questions. The findings of this review reveal that MML depends on not only modalities but also ML algorithms and tasks. The investigation of algorithms and data shows that NNs-based algorithms and image data are the most used. This review also exposes different aspects of multimodal representation, translation, alignment, fusion, and co-learning and their possible gaps. However, this SLR displayed a summary of work done in MML, which needs to be expanded further, providing timely future work opportunities for researchers interested in this interdisciplinary field.

ABBREVIATIONS

In this section, we introduce a list of abbreviations that are used throughout the article. Table 8 contains abbreviations of key methods and algorithms.

TABLE 8. A list of abbreviations used in the article.

Abbreviation	Definition
MML	Multimodal Machine Learning
ML	Machine Learning
SLR	Systematic Literature Review
AI	Artificial Intelligence
KEMA	Kernel Method For Manifold Alignment
IC	Inclusion Criteria
EC	Exclusion Criteria
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
NNs	Neural Networks
SVMs	Support Vector Machines
EMs	Ensemble Models
NNMs	Nearest Neighbor Models
TB	Tree-based Models
BM	Bayesian Models
LM	Linear Models
GA	Genetic Algorithms
GBMs	Graph Based Models
ED	Encoder-Decoder
ANN	Artificial Neural Network
BRNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network - Long-Short Term Memory
CRF-RNN	Conditional Random Field - Recurrent Neural Network
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
DT-CNN	Dilated and Transposed Convolutional Neural Network
ELM	Extreme Learning Machine
FS-Net	Fast-Shaped-Based Network
FNN	Feed Forward Neural Network
FCNN	Fully Connected Neural Network
LSM	Liquid State Machine
MGNC-CNN	Multi-group Norm Constraint CNN
MICNNL	Multiple Image CNN Long-Short Term Memory
RNN	Recurrent Neural Network
SE-ResNet	Squeeze-and-Excitation Based ResNet
T-CNN	Temporal Convolutional Neural Network
VGG	Visual Geometry Group
Bi-LSTM	Bidirectional Long-Short Term Memory
GGCN	Gated Graph Convolutional Network
MSM	Multimodal Semantic Model
DBN	Deep Belief Network
GAN	Generative Adversarial Network
RBM	Restricted Boltzmann Machine
L-SVM	LogDet Support Vector Machine
MK-SVM	Multi-Kernel Support Vector Machine
PSO-SVM	Particle Swarm Optimization Support Vector Machine
SVM-RBF	Support Vector Machine with Radial Bias Function
SVR	Support Vector Regressor
SK-SVM	String Kernels Support Vector Machine
SVC	Support Vector Clustering
SVD	Singular Value Decomposition
RF	Random Forest
GBDT	Gradient Boosting Decision Tree
kNN	K-Nearest Neighbor
WKNN	Distance-Weighted k-Nearest Neighbor
LnR	Naïve Bayes Linear Regression
LDA	Linear Discriminant Analysis
MADE	Multimodal Deep Autoencoder
GAAM	Genetic Algorithm with Aggressive Mutation
GTL	Graph-based Transudative Learning
CCA	Canonical Correlation Analysis
KCCA	Kernel Canonical Correlation Analysis
DTW	Dynamic Time Wrapping

DATA

All the supported data used for this review study is available at the following link: https://doi.org/10.5281/zenodo.7615714

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5774–5778.

[3] Y. Ren, N. Xu, M. Ling, and X. Geng, "Label distribution for multimodal machine learning," *Frontiers Comput. Sci.*, vol. 16, no. 1, pp. 1–11, Feb. 2022.

[4] S.-F. Zhang, J.-H. Zhai, B.-J. Xie, Y. Zhan, and X. Wang, "Multimodal representation learning: Advances, trends and challenges," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2019, pp. 1–6.

[5] U. Sulubacak, O. Caglayan, S.-A. Grönroos, A. Rouhe, D. Elliott, L. Specia, and J. Tiedemann, "Multimodal machine translation through visuals and speech," *Mach. Transl.*, vol. 34, nos. 2–3, pp. 97–147, Sep. 2020.

[6] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[7] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, May 2022.

[8] J. Summaira, X. Li, A. M. Shoib, and J. Abdul, "A review on methods and applications in multimodal deep learning," 2022, *arXiv:2202.09195*.

[9] T. H. Afridi, A. Alam, M. N. Khan, J. Khan, and Y.-K. Lee, "A multimodal memes classification: A survey and open research issues," in *Proc. 3rd Int. Conf. Smart City Appl.* Cham, Switzerland: Springer, 2020, pp. 1451–1466.

[10] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[11] W. Chen, W. Wang, L. Liu, and M. S. Lew, "New ideas and trends in deep multimodal content understanding: A review," *Neurocomputing*, vol. 426, pp. 195–215, Feb. 2021.

[12] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, Aug. 2022.

[13] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, and A. Zadeh, "Multimodal research in vision and language: A review of current and emerging trends," *Inf. Fusion*, vol. 77, pp. 149–171, Jan. 2022.

[14] H. Moujahid, B. Cherradi, and L. Bahatti, "Convolutional neural networks for multimodal brain MRI images segmentation: A comparative study," in *Proc. Int. Conf. Smart Appl. Data Anal.* Cham, Switzerland: Springer, 2020, pp. 329–338.

[15] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[17] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.

[18] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524–543, Apr. 2021.

[19] D. Tuia and G. Camps-Valls, "Kernel manifold alignment for domain adaptation," *PLoS ONE*, vol. 11, no. 2, Feb. 2016, Art. no. e0148655.

[20] M. A. Wajid and A. Zafar, "Multimodal fusion: A review, taxonomy, open challenges, research roadmap and future directions," *Neutrosophic Sets Syst.*, vol. 45, no. 1, p. 8, 2021.

[21] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020.

[22] W. C. Sleeman IV, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," 2021, *arXiv:2109.09020*.

[23] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[24] S. A. Chhabria, R. V. Dharaskar, and V. M. Thakare, "Survey of fusion techniques for design of efficient multimodal systems," in *Proc. Int. Conf. Mach. Intell. Res. Advancement*, Dec. 2013, pp. 486–492.

[25] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[26] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, and T. H. Falk, "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," *Inf. Fusion*, vol. 76, pp. 355–375, Dec. 2021.

[27] L. Lazli, M. Boukadoum, and O. A. Mohamed, "A survey on computer-aided diagnosis of brain disorders through MRI based on machine learning and data mining methodologies with an emphasis on Alzheimer disease diagnosis and the contribution of the multimodal fusion," *Appl. Sci.*, vol. 10, no. 5, p. 1894, Mar. 2020.

[28] J. M. Dolly, "A survey on different multimodal medical image fusion techniques and methods," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICIICT)*, Apr. 2019, pp. 1–5.

[29] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, pp. 3–20, Nov. 2016.

[30] E. E. Tulay, B. Metin, N. Tarhan, and M. K. Arıkan, "Multimodal neuroimaging: Basic concepts and classification of neuropsychiatric diseases," *Clin. EEG Neurosci.*, vol. 50, no. 1, pp. 20–33, Jan. 2019.

[31] S. Dähne, F. Biessmann, W. Samek, S. Haufe, D. Goltz, C. Gundlach, A. Villringer, S. Fazli, and K.-R. Müller, "Multivariate machine learning methods for fusing multimodal functional neuroimaging data," *Proc. IEEE*, vol. 103, no. 9, pp. 1507–1530, Sep. 2015.

[32] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, pp. 147–170, Mar. 2019.

[33] H. Purohit and P. K. Ajmera, "Fusions of palm print with palm-phalanges print and palm geometry," in *Proc. Int. Conf. Adv. Comput. Netw. Informat.* Cham, Switzerland: Springer, 2019, pp. 553–560.

[34] J. Yashas and G. Shivakumar, "Hand gesture recognition: A survey," in *Proc. Int. Conf. Appl. Mach. Learn. (ICAML)*, May 2019, pp. 3–8.

[35] F. Fereidoonian, F. Firouzi, and B. Farahani, "Human activity recognition: From sensors to applications," in *Proc. Int. Conf. Omni-Layer Intell. Syst. (COINS)*, Aug. 2020, pp. 1–8.

[36] I. Nigam, M. Vatsa, and R. Singh, "Ocular biometrics: A survey of modalities and fusion approaches," *Inf. Fusion*, vol. 26, pp. 1–35, Nov. 2015.

[37] P. Akulwar and N. A. Vijapur, "Secured multi modal biometric system: A review," in *Proc. 3rd Int. Conf. I-SMAC*, Dec. 2019, pp. 396–403.

[38] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1368–1396, 2021.

[39] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[40] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[41] A. Zadeh, P. P. Liang, and L.-P. Morency, "Foundations of multimodal co-learning," *Inf. Fusion*, vol. 64, pp. 188–193, Dec. 2020.

[42] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Keele, U.K., Durham Univ., Durham, U.K., Joint Rep. EBSE 2007-001, Jul. 2007. [Online]. Available: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[43] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and G. Prisma, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, pp. 264–269, 2009.

[44] K. S. Khan, G. T. Riet, J. Glanville, A. J. Sowden, and J. Kleijnen, *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for Carrying Out or Commissioning Reviews*. York, U.K.: NHS Centre for Reviews and Dissemination, 2001.

[45] N. Bold, C. Zhang, and T. Akashi, "Bird species classification with audio-visual data using CNN and multiple kernel learning," in *Proc. Int. Conf. Cyberworlds (CW)*, Oct. 2019, pp. 85–88.

[46] M. Rane, T. Latne, and U. Bhadade, "Biometric recognition using fusion," in *Proc. ICDSMLA*. Cham, Switzerland: Springer, 2020, pp. 1320–1329.

[47] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 159–167.

[48] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A multimodal deep learning framework for predicting drug–drug interaction events," *Bioinformatics*, vol. 36, no. 15, pp. 4316–4322, Aug. 2020.

[49] Y. Gu, K. Vyas, M. Shen, J. Yang, and G.-Z. Yang, "Deep graph-based multimodal feature embedding for endomicroscopy image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 481–492, Feb. 2021.

[50] F. Heidarivincheh, R. McConville, C. Morgan, R. McNaney, A. Masullo, M. Mirmehdi, A. L. Whone, and I. Craddock, "Multimodal classification of Parkinson's disease in home environments with resiliency to missing modalities," *Sensors*, vol. 21, no. 12, p. 4133, Jun. 2021.

[51] J. Liu, Y. Pan, F.-X. Wu, and J. Wang, "Enhancing the feature representation of multi-modal MRI data by combining multi-view information for MCI classification," *Neurocomputing*, vol. 400, pp. 322–332, Aug. 2020.

[52] C. Zu, B. Jie, M. Liu, S. Chen, D. Shen, and D. Zhang, "Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment," *Brain Imag. Behav.*, vol. 10, no. 4, pp. 1148–1159, Dec. 2016.

[53] J. S. Lara, V. H. Contreras, S. Otálora, H. Müller, and F. A. González, "Multimodal latent semantic alignment for automated prostate tissue classification and retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 572–581.

[54] Z. Wang, X. Zhu, E. Adeli, Y. Zhu, F. Nie, B. Munsell, and G. Wu, "Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning," *Med. Image Anal.*, vol. 39, pp. 218–230, Jul. 2017.

[55] H. Monkaresi, M. S. Hussain, and R. A. Calvo, "Classification of affects using head movement, skin color features and physiological signals," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2012, pp. 2664–2669.

[56] Z. Ahmad, A. Tabassum, L. Guan, and N. Khan, "ECG heart-beat classification using multimodal image fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1330–1334.

[57] P. S. Pillai and T.-Y. Leong, "Fusing heterogeneous data for Alzheimer's disease classification," in *MEDINFO 2015: eHealth-Enabled Health*. Amsterdam, The Netherlands: IOS Press, 2015, pp. 731–735.

[58] R. F. Ahmad, A. S. Malik, N. Kamel, F. Reza, H. U. Amin, and M. Hussain, "Visual brain activity patterns classification with simultaneous EEG-fMRI: A multimodal approach," *Technol. Health Care*, vol. 25, no. 3, pp. 471–485, Jun. 2017.

[59] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "Diagnosis of Alzheimer's disease using view-aligned hypergraph learning with incomplete multi-modality data," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 308–316.

[60] D. Mitrea, R. Badea, P. Mitrea, S. Brad, and S. Nedevschi, "Hepatocellular carcinoma automatic diagnosis within CEUS and B-mode ultrasound images using advanced machine learning methods," *Sensors*, vol. 21, no. 6, p. 2202, Mar. 2021.

[61] S. Grbić, R. Ionasec, Y. Wang, T. Mansi, B. Georgescu, M. John, J. Boese, Y. Zheng, N. Navab, and D. Comaniciu, "Model-based fusion of multi-modal volumetric images: Application to transcatheter valve procedures," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2011, pp. 219–226.

[62] X.-A. Bi, X. Hu, H. Wu, and Y. Wang, "Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2973–2983, Oct. 2020.

[63] Q. Zhang, J. Xiong, Y. Cai, J. Shi, S. Xu, and B. Zhang, "Multimodal feature learning and fusion on B-mode ultrasonography and sonoelastography using point-wise gated deep networks for prostate cancer diagnosis," *Biomed. Eng.*, vol. 65, no. 1, pp. 87–98, Jan. 2020.

[64] O. D. Kose and M. Saraclar, "Multimodal representations for synchronized speech and real-time MRI video processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1912–1924, 2021.

[65] J. L. Kernec, F. Fioranelli, S. Yang, J. Lorandel, and O. Romain, "Radar for assisted living in the context of Internet of Things for health and beyond," in *Proc. IFIP/IEEE Int. Conf. Very Large Scale Integr.*, Oct. 2018, pp. 163–167.

[66] J. Zhu, J. Shi, X. Liu, and X. Chen, "Co-training based semi-supervised classification of Alzheimer's disease," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 729–732.

[67] L. Qiu, Y. Zhong, Z. He, and J. Pan, "Improved classification performance of EEG-fNIRS multimodal brain-computer interface based on multi-domain features and multi-level progressive learning," *Frontiers Hum. Neurosci.*, vol. 16, Aug. 2022, Art. no. 973959.

[68] H. Faris, M. Habib, M. Faris, H. Elayan, and A. Alomari, "An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine," *Informat. Med. Unlocked*, vol. 23, Jan. 2021, Art. no. 100513.

[69] M. Xu, L. Ouyang, L. Han, K. Sun, T. Yu, Q. Li, H. Tian, L. Safarnejad, H. Zhang, Y. Gao, F. S. Bao, Y. Chen, P. Robinson, Y. Ge, B. Zhu, J. Liu, and S. Chen, "Accurately differentiating between patients with COVID-19, patients with other viral infections, and healthy individuals: Multimodal late fusion learning approach," *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e25535.

[70] J. Yu, X. Wu, M. Lv, Y. Zhang, X. Zhang, J. Li, M. Zhu, J. Huang, and Q. Zhang, "A model for predicting prognosis in patients with esophageal squamous cell carcinoma based on joint representation learning," *Oncol. Lett.*, vol. 20, no. 6, p. 1, Oct. 2020.

[71] X.-N. Fan, S.-W. Zhang, S.-Y. Zhang, and J.-J. Ni, "LncRNA_Mdeep: An alignment-free predictor for distinguishing long non-coding RNAs from protein-coding transcripts by multimodal deep learning," *Int. J. Mol. Sci.*, vol. 21, no. 15, p. 5222, Jul. 2020.

[72] H. Lei, Y. Wen, Z. You, A. Elazab, E.-L. Tan, Y. Zhao, and B. Lei, "Protein–protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1290–1303, May 2019.

[73] Y. Wang, G. Ma, Le An, F. Shi, P. Zhang, D. S. Lalush, X. Wu, Y. Pu, J. Zhou, and D. Shen, "Semisupervised tripled dictionary learning for standard-dose PET image prediction using low-dose PET and multimodal MRI," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 3, pp. 569–579, Mar. 2016.

[74] D. Hu, H. Zhang, Z. Wu, F. Wang, L. Wang, J. K. Smith, W. Lin, G. Li, and D. Shen, "Disentangled-multimodal adversarial autoencoder: Application to infant age prediction with incomplete multimodal neuroimages," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4137–4149, Dec. 2020.

[75] L. Pei, L. Vidyaratne, M. M. Rahman, Z. A. Shboul, and K. M. Iftekharuddin, "Multimodal brain tumor segmentation and survival prediction using hybrid machine learning," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2019, pp. 73–81.

[76] A. Abrol, Z. Fu, Y. Du, and V. D. Calhoun, "Multimodal data fusion of deep learning and dynamic functional connectivity features to predict Alzheimer's disease progression," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 4409–4413.

[77] C.-Y. Hung, C.-H. Lin, C.-S. Chang, J.-L. Li, and C.-C. Lee, "Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2447–2450.

[78] K. C. Fraser, K. L. Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, "Predicting MCI status from multimodal language data using cascaded classifiers," *Frontiers Aging Neurosci.*, vol. 11, p. 205, Aug. 2019.

[79] Y. Gupta, R. K. Lama, and G.-R. Kwon, "Prediction and classification of Alzheimer's disease based on combined features from Apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers," *Frontiers Comput. Neurosci.*, vol. 13, p. 72, Oct. 2019.

[80] A. K. Chowdhury, D. Tjondronegoro, V. Chandran, J. Zhang, and S. G. Trost, "Prediction of relative physical activity intensity using multimodal sensing of physiological data," *Sensors*, vol. 19, no. 20, p. 4509, Oct. 2019.

[81] Y. Hao, M. Usama, J. Yang, M. S. Hossain, and A. Ghoneim, "Recurrent convolutional neural network based multimodal disease risk prediction," *Future Gener. Comput. Syst.*, vol. 92, pp. 76–83, Mar. 2019.

[82] M. Garagnani and F. Pulvermüller, "Conceptual grounding of language in action and perception: A neurocomputational model of the emergence of category specificity and semantic hubs," *Eur. J. Neurosci.*, vol. 43, no. 6, pp. 721–737, Mar. 2016.

[83] B. Cheng, M. Liu, H.-I. Suk, D. Shen, and D. Zhang, "Multimodal manifold-regularized transfer learning for MCI conversion prediction," *Brain Imag. Behav.*, vol. 9, no. 4, pp. 913–926, 2015.

[84] D. Mahapatra, B. Bozorgtabar, S. Kuanar, and Z. Ge, "Self-supervised multimodal generalized zero shot learning for Gleason grading," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Strasbourg, France: Springer, 2021, pp. 46–56.

[85] D. Mahapatra, "Multimodal generalized zero shot learning for Gleason grading using self-supervised learning," 2021, *arXiv:2111.07646*.

[86] S. Kanwal, F. Khan, and S. AlaMRI, "A multimodal deep learning infused with artificial algae algorithm—An architecture of advanced E-health system for cancer prognosis prediction," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2707–2719, Jun. 2022.

[87] L. K. Singh and M. Khanna, "A novel multimodality based dual fusion integrated approach for efficient and early prediction of glaucoma," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103468.

[88] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3833–3849, Nov. 2021.

[89] F. Grazioli, R. Siarheyeu, I. Alqassem, A. Henschel, G. Pileggi, and A. Meiser, "Microbiome-based disease prediction with multimodal variational information bottlenecks," *PLOS Comput. Biol.*, vol. 18, no. 4, Apr. 2022, Art. no. e1010050.

[90] J. Park, M. G. Artin, K. E. Lee, Y. S. Pumpalova, M. A. Ingram, B. L. May, M. Park, C. Hur, and N. P. Tatonetti, "Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer," *J. Biomed. Informat.*, vol. 131, Jul. 2022, Art. no. 104095.

[91] S. Tang, O. Razeghi, R. Kapoor, M. I. Alhusseini, M. Fazal, A. J. Rogers, and M. R. Bort, "Machine learning–enabled multimodal fusion of intra-atrial and body surface signals in prediction of atrial fibrillation ablation outcomes," *Circulat., Arrhythmia Electrophysiol.*, vol. 15, no. 8, 2022, Art. no. e010850.

[92] M. R. Karim, T. Islam, C. Lange, D. Rebholz-Schuhmann, and S. Decker, "Adversary-aware multimodal neural networks for cancer susceptibility prediction from multiomics data," *IEEE Access*, vol. 10, pp. 54386–54409, 2022.

[93] Y. Zhang, Y. Chen, and C. Gao, "Deep unsupervised multi-modal fusion network for detecting driver distraction," *Neurocomputing*, vol. 421, pp. 26–38, Jan. 2021.

[94] Y. Yoo, L. Y. W. Tang, T. Brosch, D. K. B. Li, S. Kolind, I. Vavasour, A. Rauscher, A. L. MacKay, A. Traboulsee, and R. C. Tam, "Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls," *NeuroImage, Clin.*, vol. 17, pp. 169–178, Oct. 2018.

[95] C. Zu, L. Zhu, and D. Zhang, "Iterative sparsity score for feature selection and its extension for multimodal data," *Neurocomputing*, vol. 259, pp. 146–153, Oct. 2017.

[96] G. Ometto, G. Montesano, S. S. Afgeh, G. Lazaridis, X. Liu, P. A. Keane, D. P. Crabb, and A. K. Denniston, "Merging information from infrared and autofluorescence fundus images for monitoring of chorioretinal atrophic lesions," *Transl. Vis. Sci. Technol.*, vol. 9, no. 9, p. 38, Aug. 2020.

[97] N. Fouladgar, M. Alirezaie, and K. Främling, "CN-waterfall: A deep convolutional neural network for multimodal physiological affect detection," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 2157–2176, Feb. 2022.

[98] T. Doherty, S. McKeever, N. Al-Attar, T. Murphy, C. Aura, A. Rahman, A. O'Neill, S. P. Finn, E. Kay, W. M. Gallagher, R. W. G. Watson, A. Gowen, and P. Jackman, "Feature fusion of Raman chemical imaging and digital histopathology using machine learning for prostate cancer detection," *Analyst*, vol. 146, no. 13, pp. 4195–4211, 2021.

[99] M. Nahiduzzaman, M. Tasnim, N. T. Newaz, M. S. Kaiser, and A. Mahmud, "Machine learning based early fall detection for elderly people with neurological disorder using multimodal data fusion," in *Proc. Int. Conf. Brain Informat.* Cham, Switzerland: Springer, 2020, pp. 204–214.

[100] S. El-Sappagh, T. Abuhmed, S. M. R. Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, Oct. 2020.

[101] T. Abuhmed, S. El-Sappagh, and J. M. Alonso, "Robust hybrid deep learning models for Alzheimer's progression detection," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106688.
[102] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "A multimodal feature fusion-based method for individual depression detection on Sina Weibo," in *Proc. IEEE 39th Int. Perform. Comput. Commun. Conf. (IPCCC)*, Nov. 2020, pp. 1–8.
[103] S. El-Sappagh, H. Saleh, R. Sahal, T. Abuhmed, S. M. R. Islam, F. Ali, and E. Amer, "Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data," *Future Gener. Comput. Syst.*, vol. 115, pp. 680–699, Feb. 2021.
[104] M. S. Hussain, R. A. Calvo, and F. Chen, "Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference," *Interacting Comput.*, vol. 26, no. 3, pp. 256–268, 2013.
[105] C. Chen, G. Du, D. Tong, G. Lv, X. Lv, R. Si, J. Tang, H. Li, H. Ma, and J. Mo, "Exploration research on the fusion of multimodal spectrum technology to improve performance of rapid diagnosis scheme for thyroid dysfunction," *J. Biophotonics*, vol. 13, no. 2, Feb. 2020, Art. no. e201900099.
[106] X. Zhou, Q. Lin, Y. Gui, Z. Wang, M. Liu, and H. Lu, "Multimodal MR images-based diagnosis of early adolescent attention-deficit/hyperactivity disorder using multiple kernel learning," *Frontiers Neurosci.*, vol. 15, Sep. 2021, Art. no. 710133.
[107] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity," 2020, *arXiv:2009.00700*.
[108] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020.
[109] E. Martinez-Ríos, L. Montesinos, and M. Alfaro-Ponce, "A machine learning approach for hypertension detection based on photoplethysmography and clinical data," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105479.
[110] T. Cai, H. Ni, M. Yu, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. C. Wong, "DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102522.
[111] R. P. Thati, A. S. Dhadwal, and P. Kumar, "A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms," *Multimedia Tools Appl.*, vol. 82, pp. 1–34, Apr. 2022.
[112] A. Dong, Z. Li, M. Wang, D. Shen, and M. Liu, "High-order Laplacian regularized low-rank representation for multimodal dementia diagnosis," *Frontiers Neurosci.*, vol. 15, Mar. 2021, Art. no. 634124.
[113] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 173–183, Jan. 2018.
[114] B. Lazli and A. Mohamed, "Computer-aided diagnosis system of Alzheimer's disease based on multimodal fusion: Tissue quantification based on the hybrid fuzzy-genetic-possibilistic model and discriminative classification based on the SVDD model," *Brain Sci.*, vol. 9, no. 10, p. 289, Oct. 2019.
[115] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2411–2422, Oct. 2019.
[116] Y. Shi, H.-I. Suk, Y. Gao, S.-W. Lee, and D. Shen, "Leveraging coupled interaction for multimodal Alzheimer's disease diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 186–200, Jan. 2020.
[117] N. E. Benzebouchi, N. Azizi, A. S. Ashour, N. Dey, and R. S. Sherratt, "Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis," *J. Experim. Theor. Artif. Intell.*, vol. 31, no. 6, pp. 841–874, Nov. 2019.
[118] G. A. Tadesse, H. Javed, N. L. N. Thanh, H. D. H. Thi, L. V. Tan, L. Thwaites, D. A. Clifton, and T. Zhu, "Multi-modal diagnosis of infectious diseases in the developing world," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2131–2141, Jul. 2020.
[119] W. Lin, Q. Gao, M. Du, W. Chen, and T. Tong, "Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104478.
[120] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, and M. J. Fulham, "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2014.
[121] A. Naglah, F. Khalifa, R. Khaled, A. A. K. A. Razek, and A. El-Baz, "Thyroid cancer computer-aided diagnosis system using MRI-based multi-input CNN model," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1691–1694.
[122] Y. Zhang, S. Wang, K. Xia, Y. Jiang, and P. Qian, "Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion," *Inf. Fusion*, vol. 66, pp. 170–183, Feb. 2021.
[123] R. Mokni, N. Gargouri, A. Damak, D. Sellami, W. Feki, and Z. Mnif, "An automatic computer-aided diagnosis system based on the multimodal fusion of breast cancer (MF-CAD)," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102914.
[124] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, and P. Zhang, "Alzheimer's disease diagnosis via multimodal feature fusion," *Comput. Biol. Med.*, vol. 148, Jan. 2022, Art. no. 105901.
[125] F. Nan, S. Li, J. Wang, Y. Tang, J. Qi, M. Zhou, Z. Zhao, Y. Yang, and P. Yang, "A multi-classification accessment framework for reproducible evaluation of multimodal learning in Alzheimer's disease," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Sep. 6, 2022, doi: 10.1109/TCBB.2022.3204619.
[126] M. Abdelaziz, T. Wang, and A. Elazab, "Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks," *J. Biomed. Informat.*, vol. 121, Sep. 2021, Art. no. 103863.
[127] V. P. Jayachitra, S. Nivetha, R. Nivetha, and R. Harini, "A cognitive IoT-based framework for effective diagnosis of COVID-19 using multimodal data," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102960.
[128] Y. Cai, M. Landis, D. T. Laidley, A. Kornecki, A. Lum, and S. Li, "Multi-modal vertebrae recognition using transformed deep convolution network," *Computerized Med. Imag. Graph.*, vol. 51, pp. 11–19, Jul. 2016.
[129] L. Marais, J. A. Louw, J. Badenhorst, K. Calteaux, I. Wilken, N. van Niekerk, and G. Stein, "AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–8.
[130] A. Luu, J. Leistico, T. Miller, S. Kim, and J. Song, "Predicting TCR-epitope binding specificity using deep metric learning and multimodal learning," *Genes*, vol. 12, no. 4, p. 572, Apr. 2021.
[131] S. Walter, S. Gruss, M. Kächele, F. Schwenker, P. Werner, A. Al-Hamadi, A. Andrade, G. Moreira, and H. Traue, "Data fusion for automated pain recognition," in *Proc. 9th Int. Conf. Pervasive Comput. Technol. Healthcare*, 2015, pp. 261–264.
[132] K. Ivanov, Z. Mei, M. Penev, L. Lubich, O. O. Mumini, S. V. Nguyen Van, Y. Yan, and L. Wang, "Identity recognition by walking outdoors using multimodal sensor insoles," *IEEE Access*, vol. 8, pp. 150797–150807, 2020.
[133] N. S. Jong, A. G. S. de Herrera, and P. Phukpattaranont, "Multimodal data fusion of electromyography and acoustic signals for Thai syllable recognition," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1997–2006, Jun. 2021.
[134] O. B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, and C. B. Amar, "Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning," *Neurocomputing*, vol. 220, pp. 98–110, Jan. 2017.
[135] M. Khezri, M. Firoozabadi, and A. R. Sharafat, "Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 149–164, Nov. 2015.
[136] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2069–2072.
[137] P. Sobecki, R. Jóźwiak, K. Sklinda, and A. Przelaskowski, "Effect of domain knowledge encoding in CNN model architecture—A prostate cancer study using mpMRI images," *PeerJ*, vol. 9, Mar. 2021, Art. no. e11006.
[138] D. Wang, R. Zhang, J. Zhu, Z. Teng, Y. Huang, F. Spiga, M. H.-F. Du, J. H. Gillard, Q. Lu, and P. Liò, "Neural network fusion: A novel CT-MR aortic aneurysm image segmentation method," *Proc. SPIE*, vol. 10574, pp. 542–549, Mar. 2018.
[139] J. Dutta, D. Chakraborty, and D. Mondal, "Multimodal segmentation of brain tumours in volumetric MRI scans of the brain using time-distributed U-Net," in *Computational Intelligence in Pattern Recognition*. Singapore: Springer, 2020, pp. 715–725.

[140] J. Liu, H. Liu, S. Gong, Z. Tang, Y. Xie, H. Yin, and J. P. Niyoyita, "Automated cardiac segmentation of cross-modal medical images using unsupervised multi-domain adaptation and spatial neural attention structure," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102135.

[141] E. Şener, E. U. Mumcuoglu, and S. Hamcan, "Bayesian segmentation of human facial tissue using 3D MR-CT information fusion, resolution enhancement and partial volume modelling," *Comput. Methods Programs Biomed.*, vol. 124, pp. 31–44, Feb. 2016.

[142] B. Wang, J. Yang, H. Peng, J. Ai, L. An, B. Yang, Z. You, and L. Ma, "Brain tumor segmentation via multi-modalities interactive feature learning," *Frontiers Med.*, vol. 8, p. 341, May 2021.

[143] R. Su, J. Liu, D. Zhang, C. Cheng, and M. Ye, "Multimodal glioma image segmentation using dual encoder structure and channel spatial attention block," *Frontiers Neurosci.*, vol. 14, Oct. 2020, Art. no. 586197.

[144] F. Tang, S. Liang, T. Zhong, X. Huang, X. Deng, Y. Zhang, and L. Zhou, "Postoperative glioma segmentation in CT image using deep feature fusion model guided by multi-sequence MRIs," *Eur. Radiol.*, vol. 30, no. 2, pp. 823–832, Feb. 2020.

[145] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," *Med. Image Anal.*, vol. 43, pp. 98–111, Jan. 2017.

[146] J. Morano, Á. S. Hervella, N. Barreira, J. Novo, and J. Rouco, "Multimodal transfer learning-based approaches for retinal vascular segmentation," 2020, *arXiv:2012.10160*.

[147] B. Gutiérrez-Becker, D. Mateus, L. Peter, and N. Navab, "Guiding multimodal registration with learned optimization updates," *Med. Image Anal.*, vol. 41, pp. 2–17, Oct. 2017.

[148] T. Kim and B. Lee, "Multi-attention multimodal sentiment analysis," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 436–441.

[149] F. Segovia, J. Ramírez, D. Castillo-Barnes, D. Salas-Gonzalez, M. Gómez-Río, P. Sopena-Novales, C. Phillips, Y. Zhang, and J. M. Górriz, "Multivariate analysis of dual-point amyloid PET intended to assist the diagnosis of Alzheimer's disease," *Neurocomputing*, vol. 417, pp. 1–9, Dec. 2020.

[150] Y. Peng, X. Zhang, Y. Li, Q. Su, S. Wang, F. Liu, C. Yu, and M. Liang, "MVPANI: A toolkit with friendly graphical user interface for multivariate pattern analysis of neuroimaging data," *Frontiers Neurosci.*, vol. 14, p. 545, Jul. 2020.

[151] X.-A. Bi, R. Cai, Y. Wang, and Y. Liu, "Effective diagnosis of Alzheimer's disease via multimodal fusion analysis framework," *Frontiers Genet.*, vol. 10, p. 976, Oct. 2019.

[152] S. Halfon, M. Doyran, B. Türkmen, E. A. Oktay, and A. A. Salah, "Multimodal affect analysis of psychodynamic play therapy," *Psychotherapy Res.*, vol. 31, no. 3, pp. 313–328, Apr. 2021.

[153] M. Sato, E. U. Mumcuoglu, and S. Hamcan, "Development of novel deep multimodal representation learning-based model for the differentiation of liver tumors on B-mode ultrasound images," *J. Gastroenterol. Hepatol.*, vol. 37, no. 4, pp. 678–684, 2022.

[154] D. Bethge, P. Hallgarten, O. Özdenizci, R. Mikut, A. Schmidt, and T. Grosse-Puppendahl, "Exploiting multiple EEG data domains with adversarial learning," 2022, *arXiv:2204.07777*.

[155] V. Lopes, A. Gaspar, L. A. Alexandre, and J. Cordeiro, "An AutoML-based approach to multimodal image sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–9.

[156] Y. Qu, C. Deng, W. Su, Y. Wang, Y. Lu, and Z. Chen, "Multimodal brain MRI translation focused on lesions," in *Proc. 12th Int. Conf. Mach. Learn. Comput.*, Feb. 2020, pp. 352–359.

[157] M. Blendowski, L. Hansen, and M. P. Heinrich, "Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101822.

[158] Y. Wang, J. Zhang, M. Cavichini, D.-U.-G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, "Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework," *IEEE Trans. Image Process.*, vol. 30, pp. 3167–3178, 2021.

[159] J. Kim and B. Lee, "Identification of Alzheimer's disease and mild cognitive impairment using multimodal sparse hierarchical extreme learning machine," *Hum. Brain Mapping*, vol. 39, no. 9, pp. 3728–3741, 2018.

[160] L. Tavabi, A. Poon, A. S. Rizzo, and M. Soleymani, "Computer-based PTSD assessment in VR exposure therapy," in *Proc. Int. Conf. Hum.-Comput. Interact.* Copenhagen, Denmark: Springer, 2020, pp. 440–449.

[161] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.

[162] D. Silva, S. Leonhardt, and C. H. Antink, "Copula-based data augmentation on a deep learning architecture for cardiac sensor fusion," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2521–2532, Jul. 2021.

[163] S. Ding, H. Huang, Z. Li, X. Liu, and S. Yang, "SCNET: A novel UGI cancer screening framework based on semantic-level multimodal data fusion," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 1, pp. 143–151, Jan. 2021.

[164] C. D. Heath, H. Venkateswara, T. McDaniel, and S. Panchanathan, "Using multimodal data for automated fidelity evaluation in pivotal response treatment videos," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[165] D. Kelly, J. Condell, K. Curran, and B. Caulfield, "A multimodal smartphone sensor system for behaviour measurement and health status inference," *Inf. Fusion*, vol. 53, pp. 43–54, Jan. 2020.

[166] H. Balabin, C. T. Hoyt, C. Birkenbihl, B. M. Gyori, J. Bachman, A. T. Kodamullil, P. G. Plöger, M. Hofmann-Apitius, and D. Domingo-Fernández, "STonKGs: A sophisticated transformer trained on biomedical text and knowledge graphs," *Bioinformatics*, vol. 38, no. 6, pp. 1648–1656, Mar. 2022.

[167] S. H. Lee, "Natural language generation for electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–7, Nov. 2018.

[168] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2443–2449.

[169] C. Cangea, P. Velickovic, and P. Lio, "XFlow: Cross-modal deep neural networks for audiovisual classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3711–3720, Sep. 2020.

[170] H. M. Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proc. 7th Joint Conf. Lexical Comput. Semantics*, 2018, pp. 225–234.

[171] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 24–38, Jan. 2006.

[172] J. Nitsch, J. Nieto, R. Siegwart, M. Schmidt, and C. Cadena, "Object classification based on unsupervised learned multi-modal features for overcoming sensor failures," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4369–4375.

[173] K. K. Parida, N. Matiyali, T. Guha, and G. Sharma, "Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3251–3260.

[174] M. Altmann, P. Ott, N. C. Stache, and C. Waldschmidt, "Multi-modal cross learning for an FMCW radar assisted by thermal and RGB cameras to monitor gestures and cooking processes," *IEEE Access*, vol. 9, pp. 22295–22303, 2021.

[175] L. Specia, L. Barrault, O. Caglayan, A. Duarte, D. Elliott, S. Gella, N. Holzenberger, C. Lala, S. J. Lee, J. Libovicky, P. Madhyastha, F. Metze, K. Mulligan, A. Ostapenko, S. Palaskar, R. Sanabria, J. Wang, and R. Arora, "Grounded sequence to sequence transduction," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 577–591, Mar. 2020.

[176] G. van Tulder and M. de Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 638–648, Feb. 2019.

[177] X. Liu, M. Wang, Z.-J. Zha, and R. Hong, "Cross-modality feature learning via convolutional autoencoder," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1s, pp. 1–20, Jan. 2019.

[178] S. Vashishtha and S. Susan, "Inferring sentiments from supervised classification of text and speech cues using fuzzy rules," *Proc. Comput. Sci.*, vol. 167, pp. 1370–1379, Mar. 2020.

[179] D. Liparas, Y. HaCohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "News articles classification using random forests and weighted multimodal features," in *Proc. Inf. Retr. Facility Conf.* Copenhagen, Denmark: Springer, 2014, pp. 63–75.

[180] N. Poh, J. Kittler, and A. Rattani, "Handling session mismatch by fusion-based co-training: An empirical study using face and speech multimodal biometrics," in *Proc. IEEE Symp. Comput. Intell. Biometrics Identity Manage. (CIBIM)*, Dec. 2014, pp. 81–86.

[181] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Inf. Fusion*, vol. 51, pp. 259–270, Nov. 2019.

[182] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "SpeakingFaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, p. 3465, 2021.

[183] G. Paraskevopoulos, P. Pistofidis, G. Banoutsos, E. Georgiou, and V. Katsouros, "Multimodal classification of safety-report observations," *Appl. Sci.*, vol. 12, no. 12, p. 5781, Jun. 2022.

[184] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 649–665.

[185] D. Bricher and A. Müller, "Using multimodal contextual process information for the supervised detection of connector lock events," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2020, pp. 123–134.

[186] L. Mathur and M. J. Matarić, "Introducing representations of facial affect in automated multimodal deception detection," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 305–314.

[187] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. G. A. Perera, M. Pandey, and J. J. Corso, "Multimedia event detection with multimodal feature fusion and temporal concept localization," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 49–69, Jan. 2014.

[188] G. Kim, J. G. Choi, M. Ku, H. Cho, and S. Lim, "A multimodal deep learning-based fault detection model for a plastic injection molding process," *IEEE Access*, vol. 9, pp. 132455–132467, 2021.

[189] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–15.

[190] P. Chakravarty, J. Zegers, T. Tuytelaars, and H. Van Hamme, "Active speaker detection with audio-visual co-training," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 312–316.

[191] S. Yang, G. Li, and Y. Yu, "Relationship-embedded representation learning for grounding referring expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2765–2779, Aug. 2021.

[192] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, F. M. Bianchi, G. Moser, R. Jenssen, and S. N. Anfinsen, "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 12, 2022, doi: 10.1109/TNNLS.2022.3172183.

[193] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, "Bridge correlational neural networks for multilingual multimodal representation learning," 2015, *arXiv:1510.03519*.

[194] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *J. Artif. Intell. Res.*, vol. 44, pp. 179–222, May 2012.

[195] M. Li, "Research on extraction of useful tourism online reviews based on multimodal feature fusion," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–16, Sep. 2021.

[196] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 815–827, Mar. 2018.

[197] G. Cai and G. Lv, "Heterogeneous transfer with deep latent correlation for sentiment analysis," in *Proc. 10th Int. Symp. Comput. Intell. Design*, Dec. 2017, pp. 252–256.

[198] M. Zhou, R. Cheng, Y. Jae Lee, and Z. Yu, "A visual attention grounding neural model for multimodal machine translation," 2018, *arXiv:1808.08266*.

[199] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 798–810, Feb. 2022.

[200] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 803–814, Mar. 2017.

[201] S. Wang and W. Guo, "Sparse multigraph embedding for multimodal feature representation," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1454–1466, Jun. 2017.

[202] X. Cheng, Y. Zheng, J. Zhang, and Z. Yang, "Multitask multisource deep correlation filter for remote sensing data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3723–3734, 2020.

[203] X. Zhang and X. Liu, "Interference signal recognition based on multimodal deep learning," in *Proc. 7th Int. Conf. Dependable Syst. Appl. (DSA)*, Nov. 2020, pp. 311–312.

[204] J. O. Bastida, A.-J. Gallego, and A. Pertusa, "Multimodal object recognition using deep learning representations extracted from images and smartphone sensors," in *Proc. Iberoamer. Congr. Pattern Recognit.* Cham, Switzerland: Springer, 2018, pp. 521–529.

[205] S. Moon, S. Kim, and H. Wang, "Multimodal transfer deep learning with applications in audio-visual recognition," 2014, *arXiv:1412.3121*.

[206] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5185–5189.

[207] M. Song, Y. Liu, and X. F. Liu, "Semi-supervised 3D shape recognition via multimodal deep co-training," *Comput. Graph. Forum*, vol. 39, pp. 279–289, Oct. 2020.

[208] L. Gao and L. Guan, "A discriminative vectorial framework for multi-modal feature representation," *IEEE Trans. Multimedia*, vol. 24, pp. 1503–1514, 2022.

[209] L. Gao, R. Zhang, L. Qi, E. Chen, and L. Guan, "The labeled multiple canonical correlation analysis for information fusion," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 375–387, Feb. 2018.

[210] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[211] M.-S. Dao, N.-T. Nguyen, R. U. Kiran, and K. Zettsu, "Insights from urban sensing data: From chaos to predicted congestion patterns," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2020, pp. 661–668.

[212] P. H. Silva, E. Luz, L. A. Zanlorensi, D. Menotti, and G. Moreira, "Multimodal feature level fusion based on particle swarm optimization with deep transfer learning," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2018, pp. 1–8.

[213] Y. Wei, X. Wang, W. Guan, and L. Nie, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image. Process.*, vol. 29, pp. 1–14, 2020.

[214] A. Sellami, F.-X. Dupe, B. Cagna, H. Kadri, S. Ayache, T. Artieres, and S. Takerkart, "Mapping individual differences in cortical architecture using multi-view representation learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[215] H. Pang, C. Zhang, F. Wang, J. Liu, and L. Sun, "Towards low latency multi-viewpoint 360° interactive video: A multimodal deep reinforcement learning approach," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 991–999.

[216] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang, "Multi-level multimodal common semantic space for image-phrase grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12476–12486.

[217] A. Solomon, B. Shapira, and L. Rokach, "Predicting application usage based on latent contextual information," *Comput. Commun.*, vol. 192, pp. 197–209, Aug. 2022.

[218] M. Y. Seker, A. Ahmetoglu, Y. Nagai, M. Asada, E. Oztop, and E. Ugur, "Imitation and mirror systems in robots through deep modality blending networks," *Neural Netw.*, vol. 146, pp. 22–35, Feb. 2022.

[219] L. Cong, H. Liang, P. Ruppel, Y. Shi, M. Görner, N. Hendrich, and J. Zhang, "Reinforcement learning with vision-proprioception model for robot planar pushing," *Frontiers Neurorobot.*, vol. 16, Mar. 2022, Art. no. 829437.

[220] S. Palaskar, R. Sanabria, and F. Metze, "Transfer learning for multimodal dialog," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101093.

[221] Y. Yu, Z. Ji, J. Guo, and Z. Zhang, "Zero-shot learning via latent space encoding," *IEEE Trans. Cybern.*, vol. 49, no. 10, pp. 3755–3766, Oct. 2019.

[222] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations of satellite image time series," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Würzburg, Germany: Springer, 2019, pp. 306–321.

[223] Y. Liu, N. R. Pal, A. R. Marathe, and C.-T. Lin, "Weighted fuzzy Dempster–Shafer framework for multimodal information integration," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 338–352, Feb. 2018.

[224] H. Wen, J. Ding, W. Jin, Y. Wang, Y. Xie, and J. Tang, "Graph neural networks for multimodal single-cell data integration," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 4153–4163.

[225] S. Vachmanus, A. A. Ravankar, T. Emaru, and Y. Kobayashi, "Multimodal sensor fusion-based semantic segmentation for snow driving scenarios," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16839–16851, Aug. 2021.

[226] G. A. Florea and R.-C. Mihailescu, "Multimodal deep learning for group activity recognition in smart office environments," *Future Internet*, vol. 12, no. 8, p. 133, Aug. 2020.

[227] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, p. 679, 2018.

[228] M. Gjoreski, V. Janko, G. Slapničar, M. Mlakar, N. Reščič, J. Bizjak, V. Drobnič, M. Marinko, N. Mlakar, M. Luštrek, and M. Gams, "Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors," *Inf. Fusion*, vol. 62, pp. 47–62, Oct. 2020.

[229] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in *Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 158–165.

[230] C. Tian, Y. Yuan, and X. Lu, "Deep temporal architecture for audiovisual speech recognition," in *Proc. CCF Chin. Conf. Comput. Vis.* Singapore: Springer, 2017, pp. 650–661.

[231] F. M. Castro, M. J. Marín-Jiménez, and N. Guil, "Empirical study of audio-visual features fusion for gait recognition," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2015, pp. 727–739.

[232] A. Wijekoon, N. Wiratunga, and K. Cooper, "Heterogeneous multi-modal sensor fusion with hybrid attention for exercise recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[233] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences," *Expert Syst. Appl.*, vol. 139, Jan. 2020, Art. no. 112829.

[234] F. M. Castro, M. J. Marín-Jiménez, and N. Guil, "Multimodal features fusion for gait, gender and shoes recognition," *Mach. Vis. Appl.*, vol. 27, no. 8, pp. 1213–1228, Nov. 2016.

[235] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image Vis. Comput.*, vol. 65, pp. 76–86, Sep. 2017.

[236] L. Meng, J. Pang, Z. Wang, R. Xu, and D. Ming, "The role of surface electromyography in data fusion with inertial sensors to enhance locomotion recognition and prediction," *Sensors*, vol. 21, no. 18, p. 6291, Sep. 2021.

[237] H. Zou, J. Yang, H. P. Das, H. Liu, Y. Zhou, and C. J. Spanos, "WiFi and vision multimodal learning for accurate and robust device-free human activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–8.

[238] M. Muaaz, A. Chelli, A. A. Abdelgawwad, A. C. Mallofre, and M. Patzold, "WiWeHAR: Multimodal human activity recognition using Wi-Fi and wearable sensing modalities," *IEEE Access*, vol. 8, pp. 164453–164470, 2020.

[239] O. Caglayan, R. Sanabria, S. Palaskar, L. Barraul, and F. Metze, "Multimodal grounding for sequence-to-sequence speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8648–8652.

[240] K. Audhkhasi, A. Rosenberg, G. Saon, A. Sethy, B. Ramabhadran, S. Chen, and M. Picheny, "Recent progress in deep end-to-end models for spoken language processing," *IBM J. Res. Develop.*, vol. 61, nos. 4–5, pp. 1–2, 2017.

[241] O. Banos, A. Calatroni, M. Damas, H. Pomares, D. Roggen, I. Rojas, and C. Villalonga, "Opportunistic activity recognition in IoT sensor ecosystems via multimodal transfer learning," *Neural Process. Lett.*, vol. 53, no. 5, pp. 3169–3197, Oct. 2021.

[242] W. Thomason and R. A. Knepper, "Recognizing unfamiliar gestures for human–robot interaction through zero-shot learning," in *Proc. Int. Symp. Exp. Robot.* Cham, Switzerland: Springer, 2016, pp. 841–852.

[243] Y. Zhao, S. Guo, Z. Chen, Q. Shen, Z. Meng, and H. Xu, "Marfusion: An attention-based multimodal fusion model for human activity recognition in real-world scenarios," *Appl. Sci.*, vol. 12, no. 11, p. 5408, May 2022.

[244] P. Augustyniak and G. Ślusarczyk, "Graph-based representation of behavior in detection and prediction of daily living activities," *Comput. Biol. Med.*, vol. 95, pp. 261–270, Apr. 2018.

[245] M. S. Hussain, H. Monkaresi, and R. A. Calvo, "Combining classifiers in multimodal affect detection," in *Proc. 10th Australas. Data Mining Conf.*, vol. 134, 2012, pp. 103–108.

[246] C. D. D. Monteiro, F. Shipman, and R. Gutierrez-Osuna, "Comparing visual, textual, and multimodal features for detecting sign language in video sharing sites," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 7–12.

[247] C.-L. Chou, "Multimodal presentation attack detection based on mouth motion and speech recognition," in *Proc. Int. Conf. Secur. Intell. Comput. Big-Data Services*. Cham, Switzerland: Springer, 2019, pp. 290–298.

[248] H. Ponce, L. Martinez-Villasenor, and J. Nunez-Martinez, "Sensor location analysis and minimal deployment for fall detection system," *IEEE Access*, vol. 8, pp. 166678–166691, 2020.

[249] H.-H. Huang, M. Fukuda, and T. Nishida, "An investigation on the effectiveness of multimodal fusion and temporal feature extraction in reactive and spontaneous behavior generative RNN models for listener agents," in *Proc. 7th Int. Conf. Hum.-Agent Interact.*, Sep. 2019, pp. 89–96.

[250] J. Archila, A. Manzanera, and F. Martínez, "A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106607.

[251] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112793.

[252] W. Wang, B. Chen, P. Xia, J. Hu, and Y. Peng, "Sensor fusion for myoelectric control based on deep learning with recurrent convolutional neural networks," *Artif. Organs*, vol. 42, no. 9, pp. E272–E282, Sep. 2018.

[253] H. Ng, T. T. V. Yap, H. L. Tong, C. C. Ho, L. K. Tan, W. X. Eng, S. K. Yap, and J. H. Soh, "Action classification on the Berkeley multimodal human action dataset (MHAD)," *J. Eng. Appl. Sci.*, vol. 12, no. 3, pp. 520–526, 2017.

[254] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Cognitive load measurement in a virtual reality-based driving system for autism intervention," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 176–189, Apr. 2017.

[255] K. Jun, S. Lee, D.-W. Lee, and M. S. Kim, "Deep learning-based multimodal abnormal gait classification using a 3D skeleton and plantar foot pressure," *IEEE Access*, vol. 9, pp. 161576–161589, 2021.

[256] D. Zhang, S. Li, Q. Zhu, and G. Zhou, "Modeling the clause-level structure to multimodal sentiment analysis via reinforcement learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 730–735.

[257] D. Giritlioğlu, B. Mandira, S. F. Yilmaz, C. U. Ertenli, B. F. Akgür, M. Kınıklıoğlu, A. G. Kurt, E. Mutlu, Ş. C. Gürel, and H. Dibeklioğlu, "Multimodal analysis of personality traits on videos of self-presentation and induced behavior," *J. Multimodal User Interface*, vol. 15, no. 4, pp. 337–358, Dec. 2021.

[258] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. J. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 316–329, Jul. 2018.

[259] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[260] A. Sorrentino, L. Fiorini, I. Fabbricotti, D. Sancarlo, F. Ciccone, and F. Cavallo, "Exploring human attitude during human–robot interaction," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 195–200.

[261] C. Xia, K. Chen, K. Li, and H. Li, "Identification of autism spectrum disorder via an eye-tracking based representation learning model," in *Proc. 7th Int. Conf. Bioinf. Res. Appl.*, Sep. 2020, pp. 59–65.

[262] Z. Zhang and D. Wang, "Remote sensing and time series data fused multimodal prediction model based on interaction analysis," in *Proc. 3rd Int. Conf. Video Image Process.*, Dec. 2019, pp. 190–194.

[263] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.

[264] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Magnetic and radar sensing for multimodal remote health monitoring," *IEEE Sensors J.*, vol. 19, no. 20, pp. 8979–8989, Oct. 2019.

[265] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. A. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, pp. 443–455, 2018.

[266] R. Li, J. Yang, L. Li, F. Shen, T. Zou, H. Wang, X. Wang, J. Li, C. Deng, X. Huang, C. Wang, Z. He, F. Lu, L. Zeng, and H. Chen, "Integrating multilevel functional characteristics reveals aberrant neural patterns during audiovisual emotional processing in depression," *Cerebral Cortex*, vol. 32, no. 1, pp. 1–14, Nov. 2021.

[267] K. A. Araño, C. Orsenigo, M. Soto, and C. Vercellis, "Multimodal sentiment and emotion recognition in hyperbolic space," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115507.

[268] H. Xue, X. Yan, S. Jiang, and H. Lai, "Multi-tensor fusion network with hybrid attention for multimodal sentiment analysis," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Dec. 2020, pp. 169–174.

[269] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE Multimedia*, vol. 27, no. 1, pp. 37–48, Jan. 2019.

[270] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, and S. Feng, "Chinese multimodal emotion recognition in deep and traditional machine leaming approaches," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–6.

[271] P. Shah, P. P. Raj, P. Suresh, and B. Das, "Contextually aware multimodal emotion recognition," in *Proc. Int. Conf. Recent Trends Mach. Learn., IoT, Smart Cities Appl.* Singapore: Springer, 2021, pp. 745–753.

[272] P. Bota, C. Wang, A. Fred, and H. Silva, "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?" *Sensors*, vol. 20, no. 17, p. 4723, Aug. 2020.

[273] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *J. Med. Biol. Eng.*, vol. 40, no. 2, pp. 149–157, Apr. 2020.

[274] B. Xing, H. Zhang, K. Zhang, L. Zhang, X. Wu, X. Shi, S. Yu, and S. Zhang, "Exploiting EEG signals and audiovisual feature fusion for video emotion recognition," *IEEE Access*, vol. 7, pp. 59844–59861, 2019.

[275] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, and T. Yang, "HEU emotion: A large-scale database for multimodal emotion recognition in the wild," *Neural Comput. Appl.*, vol. 33, no. 14, pp. 8669–8685, Jul. 2021.

[276] G. Boateng and T. Kowatsch, "Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning," in *Proc. Companion Publication Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 12–16.

[277] D. Li, Z. Wang, C. Wang, S. Liu, W. Chi, E. Dong, X. Song, Q. Gao, and Y. Song, "The fusion of electroencephalography and facial expression for continuous emotion recognition," *IEEE Access*, vol. 7, pp. 155724–155736, 2019.

[278] G. Boateng, "Towards real-time multimodal emotion recognition among couples," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 748–753.

[279] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, Nov. 2021.

[280] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, and W. Minker, "An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild," 2020, *arXiv:2010.03692*.

[281] F. Qi, X. Yang, and C. Xu, "Zero-shot video emotion recognition via multimodal protagonist-aware transformer network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1074–1083.

[282] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Trans. Affect. Comput.*, early access, Jul. 14, 2021, doi: 10.1109/TAFFC.2021.3097002.

[283] A. Bhatti, B. Behinaein, D. Rodenburg, P. Hungler, and A. Etemad, "Attentive cross-modal connections for deep multimodal wearable-based emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2021, pp. 1–5.

[284] S. Zhang, B. Li, and C. Yin, "Cross-modal sentiment sensing with visual-augmented representation and diverse decision fusion," *Sensors*, vol. 22, no. 1, p. 74, Dec. 2021.

[285] Y. C. Yoon, "Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation," *IEEE Access*, vol. 10, pp. 64516–64524, 2022.

[286] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, N. Rohleder, H. Baumeister, and B. W. Schuller, "An evaluation of speech-based recognition of emotional and physiological markers of stress," *Frontiers Comput. Sci.*, vol. 3, p. 107, Dec. 2021.

[287] H. Miao, Y. Zhang, D. Wang, and S. Feng, "Multi-output learning based on multimodal GCN and co-attention for image aesthetics and emotion analysis," *Mathematics*, vol. 9, no. 12, p. 1437, Jun. 2021.

[288] H. Pham, T. Manzini, P. Pu Liang, and B. Poczos, "Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis," 2018, *arXiv:1807.03915*.

[289] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[290] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 97–104.

[291] R. Walambe, P. Nayak, A. Bhardwaj, and K. Kotecha, "Employing multimodal machine learning for stress detection," *J. Healthcare Eng.*, vol. 2021, pp. 1–12, Oct. 2021.

[292] Z. Zhao and K. Wang, "Unaligned multimodal sequences for depression assessment from speech," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 3409–3413.

[293] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening," *Frontiers Neurosci.*, vol. 8, p. 94, May 2014.

[294] T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, "A multimodal deep learning method for Android malware detection using various features," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 3, pp. 773–788, Mar. 2019.

[295] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–6.

[296] D. Gibert, C. Mateu, and J. Planes, "HYDRA: A multimodal deep learning framework for malware classification," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101873.

[297] J. M. H. Jiménez and K. Goseva-Popstojanova, "Using four modalities for malware detection based on feature level and decision level fusion," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.* Cham, Switzerland: Springer, 2020, pp. 1383–1396.

[298] I. Lamiche, G. Bin, Y. Jing, Z. Yu, and A. Hadid, "A continuous smartphone authentication method based on gait patterns and keystroke dynamics," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 11, pp. 4417–4430, Nov. 2019.

[299] S. Belhia and A. Gafour, "Feature level fusion in multimodal biometrie identification," in *Proc. 2nd Int. Conf. Innov. Comput. Technol.*, Sep. 2012, pp. 418–423.

[300] E.-S. M. El-Alfy and G. M. BinMakhashen, "Improved personal identification using face and hand geometry fusion and support vector machines," in *Proc. Int. Conf. Networked Digit. Technol.* Berlin, Germany: Springer, 2012, pp. 253–261.

[301] J. Lee, S. Qu, Y. Kang, and W. Jang, "Multimodal machine learning for display panel defect layer identification," in *Proc. 32nd Annu. SEMI Adv. Semiconductor Manuf. Conf. (ASMC)*, May 2021, pp. 1–7.

[302] C. Wang, M. Zhang, F. Shi, P. Xue, and Y. Li, "A hybrid multimodal data fusion-based method for identifying gambling websites," *Electronics*, vol. 11, no. 16, p. 2489, Aug. 2022.

[303] B. R. Naidu and M. S. P. Babu, "Biometric authentication data with three traits using compression technique, HOG, GMM and fusion technique," *Data Brief*, vol. 18, pp. 1976–1986, Jun. 2018.

[304] P. Baynath, K. M. S. Soyjaudah, and M. H.-M. Khan, "Machine learning algorithm on keystroke dynamics fused pattern in biometrics," in *Proc. Conf. Next Gener. Comput. Appl. (NextComp)*, Sep. 2019, pp. 1–6.

[305] R. Sindhuja and S. Srinivasan, "Efficient fusion based multi-modal biometric authentication system using machine learning," in *Electronic Systems and Intelligent Computing*. Singapore: Springer, 2020, pp. 119–131.

[306] B. Arjun and H. Prakash, "Feature level fusion of seven neighbor bilinear interpolation data sets of finger vein," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 1531–1536, Apr. 2020.

[307] R. Leonardo, A. Hu, M. Uzair, Q. Lu, I. Fu, K. Nishiyama, S. M. Subrahmannian, and D. Ravichandran, "Fusing visual and textual information to determine content safety," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 2026–2031.

[308] L. Lu, X. Zhang, and X. Xu, "Fusion of face and visual speech information for identity verification," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 502–506.

[309] E. S. M. El-Alfy and A. A. AlHasan, "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm," *Future Generat. Comput. Syst.*, vol. 64, pp. 98–107, Nov. 2016.

[310] S. Li, B. Zhang, L. Fei, and S. Zhao, "Joint discriminative feature learning for multimodal finger recognition," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107704.

[311] A. Rahman, M. E. H. Chowdhury, A. Khandakar, S. Kiranyaz, K. S. Zaman, M. B. I. Reaz, M. T. Islam, M. Ezeddin, and M. A. Kadir, "Multimodal EEG and keystroke dynamics based biometric system using machine learning algorithms," *IEEE Access*, vol. 9, pp. 94625–94643, 2021.

[312] S. Prabu, M. Lakshmanan, and V. N. Mohammed, "A multimodal authentication for biometric recognition system using intelligent hybrid fusion techniques," *J. Med. Syst.*, vol. 43, no. 8, pp. 1–9, Aug. 2019.

[313] L. Chen, "A multimodal biometric recognition system based on finger snapping and information fusion," in *Proc. 5th Int. Conf. Inf. Sci., Comput. Technol. Transp. (ISCTT)*, Nov. 2020, pp. 84–100.

[314] M. H. Safavipour, M. A. Doostari, and H. Sadjedi, "A hybrid approach to multimodal biometric recognition based on feature-level fusion of face, two irises, and both thumbprints," *J. Med. Signals Sensors*, vol. 12, no. 3, pp. 177–191, 2022.

[315] A. A. Joseph, A. N. H. Lian, K. Kipli, K. L. Chin, D. A. A. Mat, C. S. C. Voon, D. C. S. Ngie, and N. S. Song, "Person verification based on multimodal biometric recognition," *Pertanika J. Sci. Technol.*, vol. 30, no. 1, pp. 161–183, Nov. 2021.

[316] C.-L. Chou, "Presentation attack detection based on score level fusion and challenge-response technique," *J. Supercomput.*, vol. 77, no. 5, pp. 4681–4697, May 2021.

[317] F. Farid, M. Elkhodr, F. Sabrina, F. Ahamed, and E. Gide, "A smart biometric identity management framework for personalised IoT and cloud computing-based healthcare services," *Sensors*, vol. 21, no. 2, p. 552, Jan. 2021.

[318] Q. D. Tran and P. Liatsis, "RABOC: An approach to handle class imbalance in multimodal biometric authentication," *Neurocomputing*, vol. 188, pp. 167–177, Jan. 2016.

[319] B. A. El-Rahiem, F. E. A. El-Samie, and M. Amin, "Multimodal biometric authentication based on deep fusion of electrocardiogram (ECG) and finger vein," *Multimedia Syst.*, vol. 28, no. 4, pp. 1325–1337, Aug. 2022.

[320] A. Gona and M. Subramoniam, "Convolutional neural network with improved feature ranking for robust multi-modal biometric system," *Comput. Electr. Eng.*, vol. 101, Jul. 2022, Art. no. 108096.

[321] S. Venkatraman, "Transforming grid to cloud services for multimodal biometrics," *Int. J. Comput. Sci. Eng.*, vol. 13, no. 1, pp. 1–12, 2016.

[322] J. Li, H. Peng, H. Hu, Z. Luo, and C. Tang, "Multimodal information fusion for automatic aesthetics evaluation of robotic dance poses," *Int. J. Social Robot.*, vol. 12, no. 1, pp. 5–20, Jan. 2020.

[323] M. Bednarek, P. Kicki, and K. Walas, "On robustness of multi-modal fusion—Robotics perspective," *Electronics*, vol. 9, no. 7, p. 1152, 2020.

[324] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. André, "Exploring a model of gaze for grounding in multimodal HRI," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 247–254.

[325] S. Li, P. Zheng, J. Fan, and L. Wang, "Toward proactive human–robot collaborative assembly: A multimodal transfer-learning-enabled action prediction approach," *IEEE Trans. Ind. Electron.*, vol. 69, no. 8, pp. 8579–8588, Aug. 2022.

[326] L. C. Gussen, M. Ellerich, and R. H. Schmitt, "Prediction of perceived quality through the development of a robot-supported multisensory measuring system," *Proc. CIRP*, vol. 84, pp. 368–373, Jan. 2019.

[327] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using LDA," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 3943–3948.

[328] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11205–11212, Oct. 2022.

[329] M. Costanzo, G. De Maria, G. Lettera, C. Natale, and D. Perrone, "A multimodal perception system for detection of human operators in robotic work cells," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 692–699.

[330] M. Jayaratne, D. Alahakoon, D. De Silva, and X. Yu, "Bio-inspired multisensory fusion for autonomous robots," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2018, pp. 3090–3095.

[331] J.-B. Delbrouck and S. Dupont, "Modulating and attending the source image during encoding improves multimodal translation," 2017, *arXiv:1712.03449*.

[332] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," 2017, *arXiv:1706.00932*.

[333] W. Xia, Y. Yang, and J.-H. Xue, "Unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement," *Neural Netw.*, vol. 131, pp. 50–63, Jan. 2020.

[334] D. Elliott and Á. Kádár, "Imagination improves multimodal translation," 2017, *arXiv:1705.04350*.

[335] J. Hitschler, S. Schamoni, and S. Riezler, "Multimodal pivots for image caption translation," 2016, *arXiv:1601.03916*.

[336] C. Liu, F. Sun, C. Wang, F. Wang, and A. Yuille, "MAT: A multimodal attentive translator for image captioning," 2017, *arXiv:1702.05658*.

[337] R. Bibi, Z. Mehmood, R. M. Yousaf, T. Saba, M. Sardaraz, and A. Rehman, "Query-by-visual-search: Multimodal framework for content-based image retrieval," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5629–5648, Nov. 2020.

[338] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," 2016, *arXiv:1609.03976*.

[339] X. Li, Z. Tang, W. Chen, and L. Wang, "Multimodal and multi-model deep fusion for fine classification of regional complex landscape areas using ZiYuan-3 imagery," *Remote Sens.*, vol. 11, no. 22, p. 2716, 2019.

[340] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, 2020.

[341] Z. Yang, Z. Gui, H. Wu, and W. Li, "A latent feature-based multimodality fusion method for theme classification on web map service," *IEEE Access*, vol. 8, pp. 25299–25309, 2020.

[342] G. Machado, E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama, and J. A. D. Santos, "AiRound and CV-BrCT: Novel multiview datasets for scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 488–503, 2021.

[343] S. Kumar, B. Banerjee, and S. Chaudhuri, "Improved landcover classification using online spectral data hallucination," *Neurocomputing*, vol. 439, pp. 316–326, Jun. 2021.

[344] H. V. Le, T. Murata, and M. Iguchi, "Deep modular multimodal fusion on multiple sensors for volcano activity recognition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2018, pp. 602–617.

[345] Y. Yin, A. Tran, Y. Zhang, W. Hu, G. Wang, J. Varadarajan, R. Zimmermann, and S.-K. Ng, "Multimodal fusion of satellite images and crowdsourced GPS traces for robust road attribute detection," in *Proc. 29th Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2021, pp. 107–116.

[346] T. A. G. da Costa, R. I. Meneguette, and J. Ueyama, "Providing a greater precision of situational awareness of urban floods through multimodal fusion," *Expert Syst. Appl.*, vol. 188, Feb. 2022, Art. no. 115923.

[347] K. Perifanos and D. Goutsos, "Multimodal hate speech detection in Greek social media," *Multimodal Technol. Interact.*, vol. 5, no. 7, p. 34, Jun. 2021.

[348] F. Huang, X. Zhang, J. Xu, Z. Zhao, and Z. Li, "Multimodal learning of social image representation by exploiting social relations," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1506–1518, Mar. 2021.

[349] Z. Wang, Z. Yin, and Y. A. Argyris, "Detecting medical misinformation on social media using multimodal deep learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2193–2203, Jun. 2021.

[350] T. Deshpande and N. Mani, "An interpretable approach to hateful meme detection," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 723–727.

[351] P. Du, X. Li, and Y. Gao, "Employ multimodal machine learning for content quality analysis," in *Proc. IEEE 4th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Jun. 2020, pp. 2658–2661.

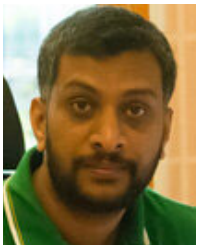
[352] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah, "Multimodal analysis of disaster tweets," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 94–103.

[353] A. Azri, C. Favre, N. Harbi, J. Darmont, and C. Noûs, "Rumor classification through a multimodal fusion framework and ensemble learning," *Inf. Syst. Frontiers*, vol. 2022, pp. 1–16, Aug. 2022.

[354] R. Rivas, S. Paul, V. Hristidis, E. E. Papalexakis, and A. K. Roy-Chowdhury, "Task-agnostic representation learning of multimodal Twitter data for downstream applications," *J. Big Data*, vol. 9, no. 1, pp. 1–19, Dec. 2022.
[355] Q. Yang, A. Farseev, S. Nikolenko, and A. Filchenkov, "Do we behave differently on Twitter and Facebook: Multi-view social network user personality profiling for content recommendation," *Frontiers Big Data*, vol. 5, Aug. 2022, Art. no. 931206.
[356] A. Azri, C. Favre, N. Harbi, J. Darmont, and A. Noûs, "Monitor: A multimodal fusion framework to assess message veracity in social networks," in *Proc. Eur. Conf. Adv. Databases Inf. Syst.* Cham, Switzerland: Springer, 2021, pp. 73–87.
[357] G. Melotti, C. Premebida, and N. Goncalves, "Multimodal deep-learning for object recognition combining camera and LiDAR data," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions (ICARSC)*, Apr. 2020, pp. 177–182.
[358] S. Richoz, L. Wang, P. Birch, and D. Roggen, "Transportation mode recognition fusing wearable motion, sound, and vision sensors," *IEEE Sensors J.*, vol. 20, no. 16, pp. 9314–9328, Apr. 2020.
[359] K. Das, M. Abouelenien, M. Burzo, and R. Mihalcea, "Towards imbalanced multiclass driver distraction identification," in *Proc. Int. FLAIRS Conf.*, vol. 35, 2022, pp. 1–4.
[360] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114693.
[361] A. R. Aftab, M. von der Beeck, and M. Feld, "You have a point there: Object selection inside an automobile using gaze, head pose and finger pointing," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 595–603.
[362] A. Gomaa, G. Reyes, and M. Feld, "ML-PersRef: A machine learning-based personalized multimodal fusion approach for referencing outside objects from a moving vehicle," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 318–327.
[363] P.-Y. Huang, F. Liu, S. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," in *Proc. 1st Conf. Mach. Transl.*, vol. 2, 2016, pp. 639–645.
[364] V. Kniaz, V. Knyaz, and V. Mizginov, "Synthesis and visualization of photorealistic textures for 3D face reconstruction of prehistoric human," in *Proc. 30th Int. Conf. Comput. Graph. Mach. Vis.*, Dec. 2020, pp. 1–9.
[365] H. Lin, F. Meng, J. Su, Y. Yin, Z. Yang, Y. Ge, J. Zhou, and J. Luo, "Dynamic context-guided capsule network for multimodal machine translation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1320–1329.
[366] X. Liu, J. Zhao, S. Sun, H. Liu, and H. Yang, "Variational multimodal machine translation with underlying semantic alignment," *Inf. Fusion*, vol. 69, pp. 73–80, May 2021.
[367] J. Im, W. Cho, and D.-S. Kim, "Cross-active connection for image-text multimodal feature fusion," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2021, pp. 343–354.
[368] R. Hinami, J. Liang, S. Satoh, and A. Hauptmann, "Multimodal co-training for selecting good examples from webly labeled video," 2018, *arXiv:1804.06057*.
[369] M. Elhoseiny, J. Liu, H. Cheng, H. Sawhney, and A. Elgammal, "Zero-shot event detection by multimodal distributional semantic embedding of videos," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1–15.
[370] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3210–3223, Feb. 2020.
[371] A. Yazici, M. Koyuncu, T. Yilmaz, S. Sattari, M. Sert, and E. Gulen, "An intelligent multimedia information system for multimodal content extraction and querying," *Multimedia Tools Appl.*, vol. 77, no. 2, pp. 2225–2260, Jan. 2018.
[372] E. Dumont and G. Quenot, "A local temporal context-based approach for TV news story segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 973–978.
[373] É. Dumont and G. Quénot, "Automatic story segmentation for TV news video using multiple modalities," *Int. J. Digit. Multimedia Broadcast.*, vol. 2012, pp. 1–11, Jan. 2012.
[374] J. Ortega-Bastida, A. J. Gallego, J. R. Rico-Juan, and P. Albarrán, "A multimodal approach for regional GDP prediction using social media activity and historical information," *Appl. Soft Comput.*, vol. 111, Nov. 2021, Art. no. 107693.
[375] A. Choube and M. Soleymani, "Punchline detection using context-aware hierarchical multimodal fusion," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 675–679.
[376] K. Miyazawa, Y. Kyuragi, and T. Nagai, "Simple and effective multimodal learning based on pre-trained transformer models," *IEEE Access*, vol. 10, pp. 29821–29833, 2022.
[377] M. Brousmiche, J. Rouat, and S. Dupont, "Audio-visual fusion and conditioning with neural networks for event recognition," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.
[378] F. Cricri, M. Roininen, S. Mate, J. Leppanen, I. D. D. Curcio, and M. Gabbouj, "Multi-sensor fusion for sport genre classification of user generated mobile videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.
[379] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1801–1810.
[380] F. Haider, F. A. Salim, D. B. W. Postma, R. van Delden, D. Reidsma, B.-J. van Beijnum, and S. Luz, "A super-bagging method for volleyball action recognition using wearable sensors," *Multimodal Technol. Interact.*, vol. 4, no. 2, p. 33, Jun. 2020.
[381] C.-Y. Chiu, P.-C. Lin, S.-Y. Li, T.-H. Tsai, and Y.-L. Tsai, "Tagging web-cast text in baseball videos by video segmentation and text alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 999–1013, Jul. 2012.
[382] K. Aryafar and A. Shokoufandeh, "Multimodal music and lyrics fusion classifier for artist identification," in *Proc. 13th Int. Conf. Mach. Learn. Appl.*, Dec. 2014, pp. 506–509.
[383] J. Quilingking Tomas, R. A. S. Jamilla, K. S. Lopo, and C. E. Camba, "Multimodal emotion detection model implementing late fusion of audio and lyrics in Filipino music," in *Proc. 3rd Int. Conf. Comput. Big Data*, Aug. 2020, pp. 78–84.
[384] B. Marenco, M. Fuentes, F. Lanzaro, M. Rocamora, and A. A. Gómez, "A multimodal approach for percussion music transcription from audio and video," in *Proc. Iberoamerican Congr. Pattern Recognit.* Cham, Switzerland: Springer, 2015, pp. 92–99.
[385] N. Henderson, J. Rowe, L. Paquette, R. S. Baker, and J. Lester, "Improving affect detection in game-based learning with multimodal data fusion," in *Proc. Int. Conf. Artif. Intell. Educ.* Cham, Switzerland: Springer, 2020, pp. 228–239.
[386] B. Jeong, J. Lee, H. Kim, S. Gwak, Y. K. Kim, S. Y. Yoo, D. Lee, and J.-S. Choi, "Multiple-kernel support vector machine for predicting internet gaming disorder using multimodal fusion of PET, EEG, and clinical features," *Frontiers Neurosci.*, vol. 16, p. 963, Jun. 2022.
[387] B.-T. Zhang, "Teaching an agent by playing a multimodal memory game: Challenges for machine learners and human teachers," in *Proc. AAAI Spring Symp., Agents Learn Hum. Teachers*, vol. 144, 2009, pp. 1–6.
[388] P. Salve, P. Yannawar, and M. Sardesai, "Multimodal plant recognition through hybrid feature fusion technique using imaging and non-imaging hyper-spectral data," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1361–1369, Jan. 2022.
[389] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, "Soybean yield prediction from UAV using multimodal data fusion and deep learning," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111599.
[390] X. Wang, S. Chen, and J. Su, "Automatic mobile app identification from encrypted traffic with hybrid neural networks," *IEEE Access*, vol. 8, pp. 182065–182077, 2020.
[391] N. Bendre, K. Desai, and P. Najafirad, "Generalized zero-shot learning using multimodal variational auto-encoder with semantic concepts," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1284–1288.
[392] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.
[393] Q. Shi, J. Fan, Z. Wang, and Z. Zhang, "Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108837.
[394] J. Wagner, F. Lingenfelser, and E. André, "The social signal interpretation framework (SSI) for real time signal processing and recognition," in *Proc. Interspeech*, Aug. 2011, pp. 831–834.

[395] G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Deep canonical time warping for simultaneous alignment and representation learning of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1128–1138, May 2018.
[396] M. A. Khan, I. Ashraf, M. Alhaisoni, R. Damaševičius, R. Scherer, A. Rehman, and S. A. C. Bukhari, "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," *Diagnostics*, vol. 10, no. 8, p. 565, Aug. 2020.
[397] S. Al-Azani and E.-S.-M. El-Alfy, "Using feature-level fusion for multimodal gender recognition for opinion mining videos," in *Proc. 8th Int. Conf. Modeling Simulation Appl. Optim. (ICMSAO)*, Apr. 2019, pp. 1–6.
[398] X. Shen, F. Xiao, Y. Wu, C. Li, and H. Dai, "A multimodal semantic model of packaging sorting field based on deep learning," in *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*. Cham, Switzerland: Springer, 2020, pp. 64–76.
[399] C. Yan, Y. Li, Y. Wan, and Z. Zhang, "Joint image-text representation learning for fashion retrieval," in *Proc. 12th Int. Conf. Mach. Learn. Comput.*, Feb. 2020, pp. 412–417.
[400] T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor, "Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in E-commerce," 2016, *arXiv:1611.09534*.
[401] M. Ajith and M. Martínez-Ramón, "Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data," *Appl. Energy*, vol. 294, Jan. 2021, Art. no. 117014.
[402] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Pedestrian trajectory prediction with structured memory hierarchies," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2018, pp. 241–256.
[403] K. P. Seng, L.-M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.
[404] C. Hu, Q. Li, Z. Zhang, K.-H. Chang, and R. Zhang, "A multimodal fusion framework for brand recognition from product image and context," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–4.
[405] W. Wang and C. Li, "A multimodal quality inspection system based on 3D, hyperspectral, and X-ray imaging for onions," Amer. Soc. Agricult. Biol. Eng., St. Joseph, MI, USA, Tech. Rep., 2014.
[406] M. Carvalho, R. Cadene, D. Picard, L. Soulier, and M. Cord, "Images and recipes: Retrieval in the cooking context," in *Proc. IEEE 34th Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2018, pp. 169–174.
[407] M. Henke, A. Junker, K. Neumann, T. Altmann, and E. Gladilin, "A two-step registration-classification approach to automated segmentation of multimodal images for high-throughput greenhouse plant phenotyping," *Plant Methods*, vol. 16, no. 1, pp. 1–10, Dec. 2020.
[408] Z. Yang, L. Pinto-Alva, F. Dernoncourt, and V. Ordonez, "Backpropagation-based decoding for multimodal machine translation," *Frontiers Artif. Intell.*, vol. 4, pp. 1–11, Jan. 2022.
[409] G. A. Pinheiro, J. L. F. Da Silva, and M. G. Quiles, "SMICLR: Contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning," *J. Chem. Inf. Model.*, vol. 62, no. 17, pp. 3948–3960, Sep. 2022.
[410] M. P. S. Gôlo, A. F. Araújo, R. G. Rossi, and R. M. Marcacini, "Detecting relevant app reviews for software evolution and maintenance through multimodal one-class learning," *Inf. Softw. Technol.*, vol. 151, Nov. 2022, Art. no. 106998.
[411] L. Parcalabescu, N. Trost, and A. Frank, "What is multimodality?" 2021, *arXiv:2103.06304*.
[412] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 1–14.
[413] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Appl. Sci.*, vol. 12, no. 3, p. 1353, Jan. 2022.
[414] A. Degas, M. R. Islam, C. Hurter, S. Barua, H. Rahman, M. Poudel, D. Ruscio, M. U. Ahmed, S. Begum, M. A. Rahman, S. Bonelli, G. Cartocci, G. Di Flumeri, G. Borghini, F. Babiloni, and P. Aricó, "A survey on artificial intelligence (AI) and explainable AI in air traffic management: Current trends and development with future research trajectory," *Appl. Sci.*, vol. 12, no. 3, p. 1295, Jan. 2022.
[415] S. S. Sheuly, M. U. Ahmed, and S. Begum, "Machine-learning-based digital twin in manufacturing: A bibliometric analysis and evolutionary overview," *Appl. Sci.*, vol. 12, no. 13, p. 6512, Jun. 2022.

ARNAB BARUA received the M.Sc. degree (thesis) in computer science and technology from Xidian University, China, in 2020. He is currently pursuing the Ph.D. degree with the Computer Science and Software Engineering Division, Artificial Intelligence and Intelligent Systems Group, Mälardalen University. Before his doctoral studies, he joined as a Research Engineer with the KIOS Research & Innovation Center of Excellence, University of Cyprus. He is also a member of the Computer Science and Software Engineering Division, Artificial Intelligence and Intelligent Systems Group, Mälardalen University. He has been involved in research and development, since 2017. His research interests include multimodal machine learning, deep learning, image processing, and healthcare applications involving artificial intelligence. He is also involved in the European Union funded project named FITDrive. As part of his research activities, he is also working on developing multimodal alignment methods using machine learning. He is also involved in teaching. In future, he wants to involve more in research and development activities for industrial applications.



MOBYEN UDDIN AHMED received the master's degree, in May 2004, and the Ph.D. degree, in 2011.

He is currently a Professor in artificial intelligence with the Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Engineering (IDT), Mälardalen University. He has been strongly involved in research and education in the field of artificial intelligence (AI) and machine learning (ML). Since completion of his Ph.D. degree, he has been awarded external funding mainly from VR, Vinnova, H2020, and KKS for several externally funded research projects as a Principal Investigator (PI) and the Co-PI/Subproject Leader with MDU, for example H2020 Projects: "ARTIMATION: Transparent Artificial Intelligence and Automation to Air Traffic Management Systems;" "FitDrive: Monitoring devices for overall FITness of DRIVErs;" and "SimuSafe: Simulator of Behavioural Aspects for Safer Transport." A bilateral project between Italy and Sweden funded by VR "BRAINSAFEDRIVE: A Technology to detect Mental States during Drive for improving the Safety of the road," and several national projects: "DIGICOGS," "AUTOMAD," "InVIP," and "ThirdEye." He has been also involved in many other national and international projects, such as Adapt2030, CPMXai, xApp, ecare@home, ESS-H, ESS-H+ SafeDriver, PainOut, VDM, Prompt, and FutureE. He is one of the Principal Investigator of the Research Profile Embedded Sensor Systems for Health Plus (ESS-H+) at MDU. He has contributed to designing, coordinating and/or teaching approximately 20 courses both for regular university students and industrial professionals related to AI and ML.

Dr. Ahmed has been active in the research community and has served as a steering committee member, the program chair, the co-chair, and an organizer of international conferences and workshops.

SHAHINA BEGUM received the Ph.D. degree in computer science/artificial intelligence from Mälardalen University (MDU), Sweden, in 2011. She is currently a Professor in AI. She is also the Deputy Leader of the Artificial Intelligence and Intelligent Systems Research Group, MDU. Her research interests include developing intelligent systems in industrial and healthcare applications. She received a Swedish Knowledge Foundation's Prospect individual grant for prominent young researchers, in 2011. She has been listed amongst the 100 most relevant researchers in sustainable method development for industrial applications by the Royal Swedish Academy of Engineering Sciences 2020. She has been the Principal Applicant and the Project Manager for a number of research projects at MDU. She is the main responsible for a number of courses both campus-based and online professional courses in AI with MDU both for regular students and industrial professionals.