Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 1458–1467

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Multimodal Machine Learning for Mental Disorder Detection: A Scoping Review

Thuy Trinh Nguyen[a,*], Viet Hoang-Quoc Pham[b], Duc-Trong Le[b], Xuan-Son Vu[c], Fani Deligianni[a], Hoang D. Nguyen[d]

[a]*School of Computing Science, University of Glasgow, United Kingdom*
[b]*University of Engineering and Technology, Vietnam National University, Vietnam*
[c]*Department of Computing Science, Umeå University, Sweden*
[d]*School of Computer Science and Information Technology, University College Cork, Ireland*

Abstract

Recent advancements in machine learning and multimedia technologies have paved new ways for automatic medical diagnosis. In mental health, multimodal inputs such as visual and audible sensing data are promising to investigate the underlying mechanisms of many conditions, such as depression and bipolar disorders. With the increasing burden on healthcare systems, timely diagnosis of mental diseases using multiple modalities might benefit millions of people worldwide. This scoping review provides an exploratory overview of recent multimodal machine learning approaches for mental disorder screening. We also discuss a generalised end-to-end multimodal machine learning pipeline for future research and development of multimodal disease detection.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Keywords: Multimodal machine learning; mental disorder diagnosis; depression; stress disorders; bipolar disorders

1. Introduction

One in ten people is susceptible to mental health issues with the most prevalent disorders being depression, anxiety and bipolar disorders [27]. Despite the persistent and high-risk nature of mental disorders, a large number of cases can easily go undetected due to social stigma and lack of accessibility to resources. Given that these conditions necessitate frequent attention and monitoring, there is an emerging need for accessible methods that can be integrated easily into the patients' daily lifestyles.

Artificial intelligence (AI) has been studied for decades for its role in healthcare decision-making monitoring and assistance. With the development of mobile technologies and unobtrusive sensors, AI can easily collect and analyse a

* Corresponding author. Tel.: +64-452-280-034
E-mail address: 2718725n@student.gla.ac.uk

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems
10.1016/j.procs.2023.10.134

large amount of unconventional data that are beyond clinical context to make health-related predictions using social media behaviour [34], phone calls [14] and wearable sensors [39]. Leveraging these sources simultaneously will be helpful to shorten the process of detecting symptoms of mental disorders. The use of AI, therefore, can potentially address issues of conventional mental disorder diagnostic methods. With the possibility of mobile application integration, AI mental disorder detection solutions can provide rapid screening to overcome delayed diagnosis caused by the shortage of healthcare clinics [26]. While the traditional mental disorder diagnosis methods often involve a high level of subjectivity, which have shown an alarming rate of false diagnoses [23], AI-based measures can make this process more objective and explainable.

The past decade has witnessed a growing body of automated mental disorder detection literature. Previous studies reveal some underlying mechanisms of mental disorders through how patients interact in different senses. For instance, factors in the visual domain such as less emotional expressivity and fidgeting eye movements have shown to be effective discriminant information to detect depression [8]. Similarly, several approaches utilise auditory modality features including shortened speech and lengthened pause duration to successfully identify depressed individuals [22]. The uniqueness of how mental disorder patients process different modalities sparks interest in information integration to possibly obtain a holistic view of complementary symptoms and the compounding impact of multimodal data [20]. Fortunately, recent advancements in multi-aspect sensing technologies pave the way for multimodal mental disorder recognition [15] that could overcome limitations of the unimodal-based prediction and enhance the joint analysis of multimodalities.

This paper employs a systematic approach to explore the effectiveness of multimodal machine learning (MMML) in mental disorder detection, as well as to highlight current trends and future directions for the field. We focus our study on common mental disorders including depression, stress- and bipolar disorders.

2. Background

A mental disorder is a syndrome characterised by clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning [5]. Mental disorders are leading contributors to the global health-related burden, especially in the context of the COVID-19 epidemic, the consequences of mental illness were becoming even more serious [32]. Mental disorders can be categorised into several common categories, such as depression, stress- and bipolar disorders. In efforts to reduce the harmful effects of mental illness, early detection and diagnosis play a vital role, which will help patients start an early and better treatment based on the symptoms [30]. With great progress in recent years, machine learning shows great potential in enabling speedy and scalable analysis of complex data, thereby opening up opportunities to aid in the diagnosis and treatment of mental disorders [33].

Multimodal machine learning aims to build models that can process and combine information from multiple modalities. Each modality in the multimodal model will handle a kind of data such as text, audio, and image. This field has great potential and is being strongly researched and applied in the fields [6]. And multimodal machine learning has the advantage of simultaneously processing data coming from multiple sources in different formats to combine useful information for making final predictions. Therefore, this paper aims to investigate recent multimodal approaches for the intelligent detection of mental disorders.

3. Methodology

This section describes the search strategy of our survey paper.

3.1. Search strategy

This work focuses on reviewing and outlining trends of high impact papers in MMML for mental disorders detection. Hence, we define our search around several high-quality machine learning and AI venues namely JCAI, AAAI, IEEE, ACM Multimedia. The list of keywords used for the search query is as follows:

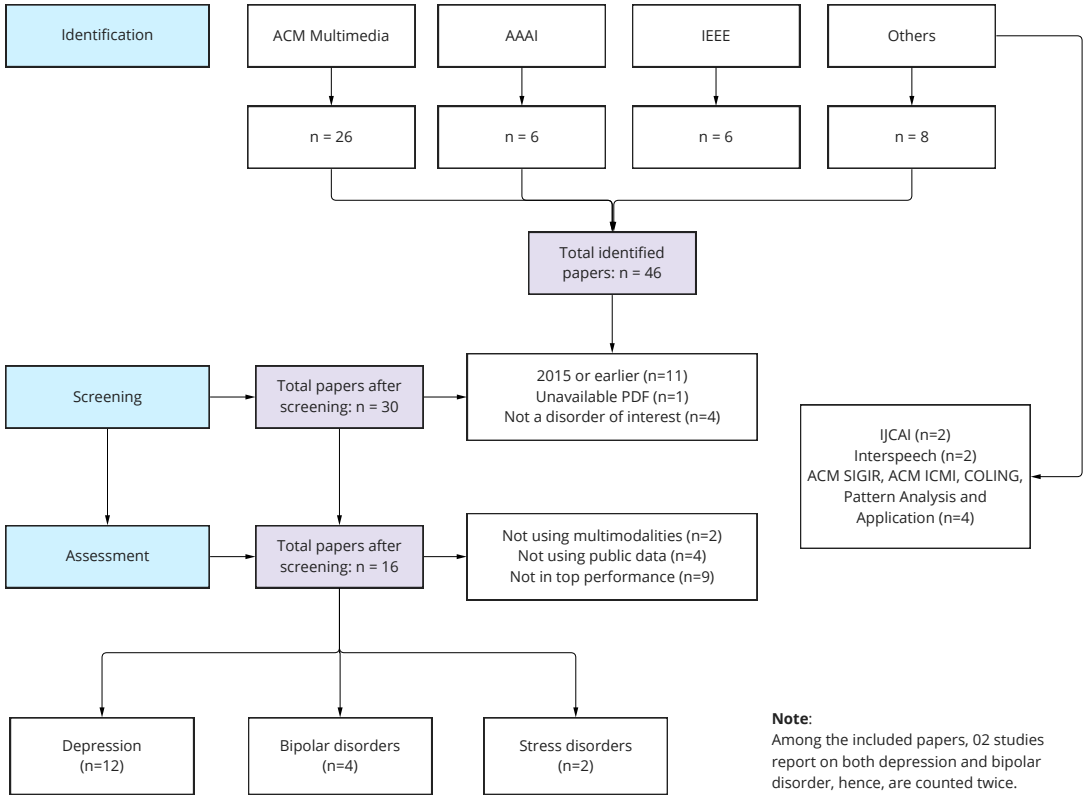

Fig. 1: Search results

- Multimodal machine learning: Multimodal* OR cross-modal* or cross-domain OR audiovisual OR fusi* OR ((text* OR lingu* OR semantic*) AND (audio OR vocal) AND (video OR vis* OR fac*))
- Mental disorders: depress* OR stress* OR bipolar* OR mental*
- Detection (optional): detect* OR identif* OR predict* OR classif* OR recogn* OR tackl*

3.2. Exclusion criteria

To narrow down the list of literature, we apply several exclusion criteria as follows: (i) the paper was published earlier than 2015; (ii) the paper is not related to a mental disorder or a type of distress (i.e., depression, stress and bipolar disorders); (iii) the paper does not propose an MMML solution; (iv) the paper does not use public datasets; (v) the paper does not belong to the top 3 performing papers for each identified dataset. The summary of search results will be provided in the next section.

4. Results

After the search, we identify 16 papers that match the scope of this paper. Figure 1 summarises the filter process and its results. This section analyses the search results in terms of datasets, performance and MMML approaches.

Mental disorders	Dataset	Modality		Train (size)	Dev (size)	Test (size)	Imbalance rate in train (%)
		Classic	Others				
Depression	E-DAIC [11]	A V T		163 ≈ 60%	56 ≈ 20%	56 ≈ 20%	22.70
	DAIC-WOZ [17]	A V T	physiological data	107 ≈ 55%	35 ≈ 20%	47 ≈ 25%	28.04
	Twitter depression [34]	T I		2243 ≈ 80%	561 ≈ 20%	-	50
	Well-being [24]	A V T		24 ≈ 70%	11 ≈ 30%	-	50
	(Private dataset) [3]	A T		47 ≈ 80%	12 ≈ 20%	-	-
	(Private dataset) [42]	A V T I	keystrokes mouse operations	26 ≈ 95%	1 ≈ 5%	-	-
Stress disorders	MuSE [19]	A V T I	physiological data	27 ≈ 95%	1 ≈ 5%	-	-
	Ulm-TSST [35]	A V T	ECG, RESP and BPM signals	41 ≈ 60%	14 ≈ 20%	14 ≈ 20%	-
Bipolar disorders	BDC [10]	A V		104 ≈ 50%	60 ≈ 25%	54 ≈ 25%	39.42

Table 1: Summary of mental disorders datasets. A, V, T, I denote the use of audio, video, text, and image modality respectively.

4.1. Report on datasets and performance

Dataset and performance summaries are included in Table 1 and 2 respectively. Of the common mental disorders covered by this survey, depression was the one with the largest number of articles as well as the number of datasets used in experiments and evaluation.

Depression. In 12 articles related to depression, 4 datasets were used including DAIC-WOZ [17], E-DAIC [11], Twitter Depression Dataset [34], and Well-being [24]. DAIC-WOZ is a multimodal dataset containing audio-video recordings of interviews conducted by virtual interviewer Ellie for psychological distress conditions and E-DAIC is an extension of DAIC-WOZ. Well-being is a non-clinical dataset containing facial expressions, body motion information, gestures, and audio recordings for mental distress recognition. Self-evaluation questionnaires were employed in these datasets. Twitter Depression Dataset contains collected tweets (both images and text) and their label based on text patterns (e.g., diagnosed depression) from Twitter users. Regarding means of the diagnosis, DAIC, E-DAIC and Well-being, ground truths were assigned using self-evaluation questionnaires. In the Twitter depression dataset, users were assigned as depressed based on the strict text pattern "(I'm/I was/ I am/ I've been) diagnosed depression" in their posts. Performance-wise, [25] and [41] show the best performance on the E-DAIC and DAIC-WOZ respectively for regression task, while [43] achieve the highest F1-score on Twitter depression dataset for classification.

Stress Disorders. In stress disorders detection, there are total 3 articles, in which [4] and [7] used MuSE dataset [19] that were designed for stress detection and its relation to human emotion and the other conduct experiment on Ulm-TSST dataset (Muse-Stress sub-challenge of MuSe 2021) [35]. The labelled in MuSE dataset were assigned as self-report annotations by participants, while Ulm-TSST were labelled by three annotators using the RAWW method. In stress disorders, [7] obtain the SOTA result of F1-score of 89.3% on MuSE dataset.

Bipolar Disorders. Bipolar Disorder Corpus (BDC) [10] in the AVEC2018 [1] is used in 4 articles which were annotated for bipolar disorder states as well as the Young Mania Rating Scale (YMRS) scores by psychiatrists [9]. [9] outperform others on the BDC dataset in terms of Unweighted Average Recall (UAR).

As can be seen that most of the datasets have a rather limited number of samples, there is only one dataset having more than 1000 samples in the training set. Apart from the datasets used in the Audio/Visual Emotion Challenge and Workshop (AVEC) such as E-DAIC, DAIC-WOZ, BDC and Ulm-TSST, the other datasets are evaluated by using the

Dataset	Disease	Paper	Performance
E-DAIC	Depression	[38]	RMSE: 5.22†
		[40]	RMSE: 4.48†
		[25]	RMSE **4.28**†
DAIC-WOZ	Depression	[29]	RMSE: 4.99†
		[31]	RMSE: 4.81†
		[16]	RMSE: 4.27†
		[41]	RMSE: **3.28**†
Twitter Depression	Depression	[4]	F1: 84.2%§
		[34]	F1: 85.0%§
		[18]	F1: 90.0%§
		[43]	F1: **91.2%**§
Well-being	Depression	[9]	F1: 87%§
Ulm-TSST	Stress disorders	[13]	CCC: 0.66†
MuSE	Stress disorders	[4]	F1: 84.9%§
		[7]	**F1: 89.3%**§
BDC	Bipolar disorders	[2]	UAR: 61.65%§
		[40]	UAR: 70.90%§
		[36]	UAR: 72.09%§
		[9]	UAR: **88.36%**§

Table 2: Summary of performance on datasets. Because the results reported in the articles are inconsistent, this table only aggregates the results on the most commonly used metrics for each dataset. The † symbol denotes regression tasks; whereas, The § symbol denotes classification tasks.

cross-validation method such as k-fold cross-validation or leave-one-out cross-validation, hence, there is no actual test set on those and the size of training and development sets are estimated.

4.2. Comparison unimodal vs multimodal

To capture the performance gap of using extra modalities, we select the best performing model of single modality as the unimodal baseline to compare against the proposed multimodal approach of each paper. An average will be applied if the performance on development and test sets is available.

Out of 16 papers that we review, 15 papers benchmark their multimodal approach against a unimodal method. The results show that in all 15 papers, multimodal approaches enhance performance. On average, there is an improvement of 7.9% in F1 (8 out of 15 papers), 5% in UAR (3 out of 15 papers). In particular, on the BDC dataset, multimodal solutions consistently outperform unimodal base models by a significant 19.7% (F1) [40] and 12.17% (UAR) [36]. In the depression recognition task, a hierarchical recall model utilising audio, text and video obtains a 12% increase in F1 compared to its textual method [29] on DAIC-WOZ. In general, multimodal approaches have shown that they can significantly improve outcomes compared to methods using individual modalities.

5. Discussion

We propose an end-to-end multimodal pipeline for mental disorder detection in Figure 2. The following sections will provide a detailed discussion on each of the five stages.

5.1. Multimodal Data Preprocessing

Data collection is not compulsory and can rely on data mining if suitable secondary data such as social behaviour is available. Primary data collection is required to build one or more feature sets from scratch.

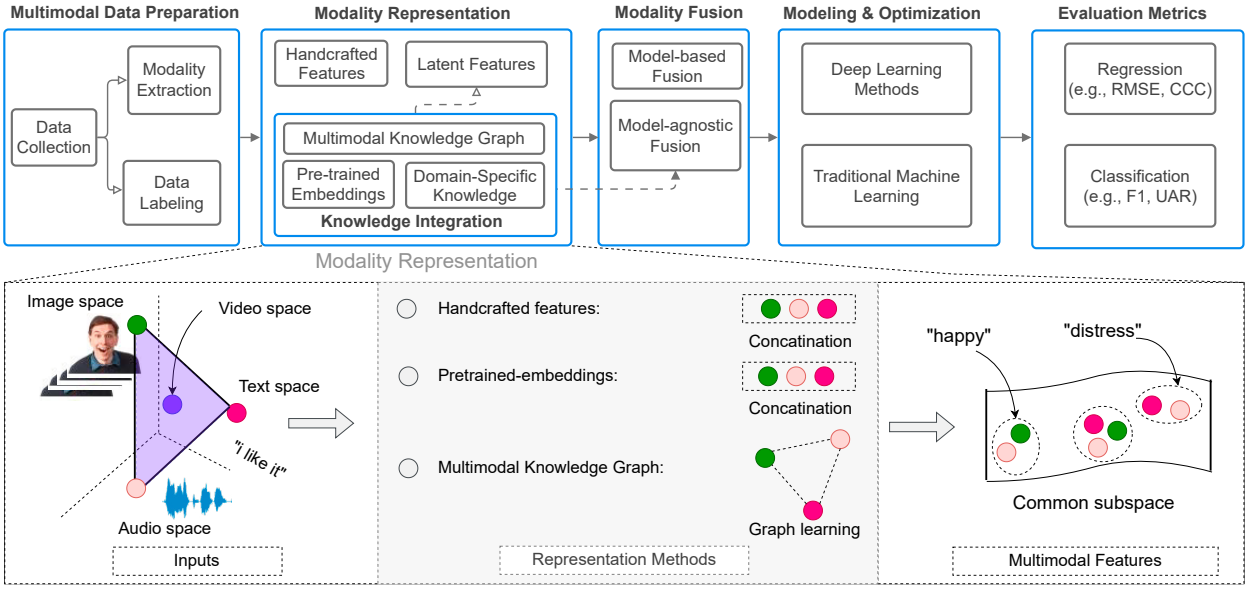

Fig. 2: Multimodal machine learning pipeline for detection of mental disorders

Data Labelling. To collect ground truths, the most widely adopted options are: (1) clinical assessment in [10], (2) self-assessment in [11, 17] and (3) self-interpreted (e.g., social behaviour data mining) in [34].

While clinical assessment labels are the most reliable, the other methods are more popular due to their convenience. One should consider their resources and desired sample size when selecting the label source. On the other hand, the format of ground truths depends on the objective task of classification or regression.

It is recommended to define the primary task clearly as different objectives can limit the choice of suitable model architectures. For some datasets, the conversion from regression to classification is possible when the labels are given thresholds to become categorical data. While this allows some studies to report their performance in both tasks [40, 29] for instance, in DAIC-WOZ, apart from the original PHQ-8 scores ranging from 0 to 24, a threshold of 10 is used to classify depressed individuals.

Modality extraction. Some popular methods for modality data preprocessing are:

- **Audio**: OpenSMILE and COVAREP are mainstream toolkits in most studies for extracting auditory features such as MFCC, GeMAPs, pitch and voice segmentation [37, 36]
- **Text**: a variety of extraction methods are in place for textual modality, including audio transcription using speech recognition [36, 40] and topic-related, semantic or handcrafted data [34]
- **Vision**: To extract facial expressions and eye movements from video and images using toolkits, such as OpenFace [29, 16] and Face++ [36].

5.2. *Representation*

Representation for multimodal learning is a crucial task in enabling the utilization of ubiquitous multimodal data. Here we review different methods in representing multimodal data for the later process of detecting mental disorders.

Multimodal data are composed of multiple modes, in which, each mode possesses a different form of information. Typically, there are three ways to represent multimodal data: (1) feature-level concatenation, (2) joint feature learning, and (3) graph-based representation.

Feature-level concatenation is a popular method for representing multimodal data in the task (such as in [2]). In this approach, the features are extracted from each unimodal data, then they are fused into a single feature vector. The advantage of this approach is that each mode can proceed independently to exploit meaningful information.

However, this approach has some disadvantages. First, the fused feature vector may be too large to be processed by certain machine learning algorithms. Second, the fused feature vector may not be discriminative enough to distinguish different modes.

Joint feature learning method concurrently learns features from all unimodal data. It is an effective representation method since features are jointly learned from all the modes and thus can exploit the complementary information in different modes. Most of the deep learning-based methods employ this representation approach (e.g., [25]). However, it is more difficult to optimise the model since the number of trained parameters are normally high and the task is more data-hungry than the first approach.

Graph-based representation are getting more attention recently [41]. In this approach, the multimodal data is constructed as a multimodal graph, where the nodes represent each unimodal data, and the edges represent their interactions. This approach has many advantages over other approaches in learning capabilities. For instance, [41] showed that the method could help to learn meta-paths across modalities to model the interplay between them. However, this approach has some drawbacks such as the graph construction being computationally expensive and it is also a data-hungry method.

5.3. Knowledge Integration

The modelling of multiple modalities only manifests the local knowledge, which is strongly dependent on the training data. The integration of global knowledge would be informative to reveal underlying patterns in a generic way. It becomes more significant since the expert or prior knowledge plays an important role in the context of investigating mental health problems. Generally, there are several trends in integrating such knowledge as follows:

Pre-trained embeddings. The primary purpose is to exploit the semantic relationship of textual modalities, e.g., words or tokens. There is a prominent selection of BERT [12], a Transformer-based model. For examples, [4] and [37] utilise the BERT-Base variant with 768-dimensional hidden states as the shared text encoder while [28] consider the BERT-large model to extract contextual information of transcribed transcripts from the E-DAIC corpus. Recently, [40] apply doc2vec to infer the document fixed-length representation for transcribed text. [43] investigate several pre-trained models including BERT, RoBERTa and XLNet with the BART summarization model [21] to summarise a large number of tweets associated with each user. As a limitation, pre-trained embeddings need to be fine-tuned to cover more textual features in the complex nature of data, e.g., social posts.

Domain-specific knowledge. It provides a faithful foundation to extract informative signals which support the reasoning process in mental disorder detection tasks. As a typical case of depression detection, domain-specific features including depression symptoms and antidepressants are explored to determine depression-related users. [43] and [40] analyse Twitter tweets with nine depression symptoms in DSM-IV criteria and antidepressant medicine names built from Wikipedia to diagnose depression. Regardless of carrying meaningful factors, this approach is domain-dependent and costly to build respective vocabularies.

Multimodal knowledge graph. There is an interesting trend to exploit knowledge graph [41]. Instead of building from existing corpus, It is constructed via leveraging the relatedness among multimodal entities. Although this approach may have a remarkable computational cost, it enhances the exploration of cross-modality patterns.

5.4. Modality Fusion

Due to the heterogeneity of multiple data streams, joining information from two or more modalities to predict an outcome measure has been one of the original focuses of multimodal machine learning.

5.4.1. Model-agnostic fusion

Model-agnostic or model-free approaches are ones that do not directly rely on the architecture of a specific machine learning model which can be categorised as early (feature-level) and late (decision-level) fusion.

Early fusion. The first attempt at multimodal fusion is an early fusion that combines the modalities at the feature level. The vast majority of identified papers employ early fusion. The most basic form of early fusion, concatenation, is employed widely across different mental disorders including stress detection task [37], and bipolar disorders classification [40, 2]. A variation of concatenation in early fusion is weighted average [38].

Rather than approaching feature-level fusion as a pure concatenation of independent components, incorporating the correlation among modalities could enrich the overall performance of the architecture. [39] applies multidimensional projection fusion using group sparse canonical correlation analysis on EEG and eye movement which maximise the cross-correlation between two input streams and achieve the highest accuracy on anxiety detection. Similarly, [25] obtains the winning performance in AVEC2019 on Extended DAIC by the proposal of a multi-layer attention fusion technique that also captures the focus of input features.

Late fusion. In contrast to early fusion, late fusion or decision-based fusion integrates the outcomes of modality-wise predictions. Late fusion is employed in the majority of papers on a variety of fusing mechanisms including (1) simple voting in [7] that allows a drastic improvement of 33.9% accuracy compared to early fusion technique in the same study, (2) winner-take-all voting [31] and (3) learned classifiers such as LSTM fusion classifier [13] and simple feed-forward neural network [9].

5.4.2. Model-based fusion

One innovative model-based method employing a combination of deep learning techniques, namely TCN and Knowledge graph, however, achieves state-of-the-art performance in depression diagnosis on DAIC-WOZ dataset with F1 95.4% for the classification task, RMSE 3.28 and MAE 2.62 for the regression task [41].

An idea of layer-by-layer fusion via a hierarchical fusion structure to classify different levels of bipolar disorders using tree-based method [36]. While this novel technique achieves a 7.4% UAR improvement compared to the audiovisual baseline with an ordinary fusion method on the development set, the model suffers from overfitting as performance drops greatly on the test set.

Model-based fusion with LSTM is proven more effective compared to other model-independent methods for depression detection with an 11% improvement in F1 [29]. This could be explained by the ability to train multimodal representation and fusion in parallel with LSTM-based fusion. Despite its promising results, the model-dependent fusion technique is the least popular multimodal fusion method among the identified studies, which indicates the underexplored potential of this approach.

5.5. Modeling & Optimization

Towards the objective of tackling the primary task, i.e., classification or regression, that is predefined in the preparation step, various backbone models are employed to facilitate the representation learning of multimodal data.

Depending on the primary task (i.e., classification or regression), there are two main streams of modeling & optimization which are *traditional machine learning* and *deep learning*.

Traditional machine learning. This is a straightforward solution for exploring handcrafted features for detecting mental disorders. In the direction of tree-based methods, [31, 9] propose three random forest models for the three different modalities, i.e., audio, video and text whilst [36] present a hierarchical recall model built on a gradient boosting decision tree (GBDT). In [16], several regression models are considered namely random forest (RF), stochastic gradient descent (SGD), support vector regression (SVR).

Deep learning. This group of methods creates a more effective means to leverage underlying latent relationships among different modalities for the mental problem detection task. It flexibly supports various fusion strategies. For early fusion, [18, 40] employ deep neural networks while [37] propose a Transformer-based model architecture. There are some efforts to exploit convolutional neural networks [2], recurrent neural networks (RNNs) [38, 4]. Noticeably, attention-based models [25, 41] enable cross-modality exploration via model-based fusion. For late fusion, RNN-based methods are overwhelming others [7, 13]. Overall, the use of deep learning methods is a promising approach to modelling multimodal data for the mental disorder detection problem.

To infer adequate parameter settings, models are optimised for the respective task. For regression, the minimization of mean square errors is favoured [31]. For classification, the cross-entropy objective function is often chosen [41].

5.6. Evaluation Metrics

Evaluation methods are important metrics used to summarise the system's performance on different tasks. They are also used to compare different systems on the same task. In mental disorder detection, we have observed the use

of the following evaluation metrics depending on corresponding tasks. In the case of regression, root-mean-square-error (RMSE) and Concordance Correlation Coefficient (CCC) are used in most papers such as in [38, 40, 41]. CCC is used commonly in emotion recognition tasks. It measures the similarity between the true emotion with predicted emotion degree. For the classification task, F1 and Unweighted Average Recall (UAR) are used in different works. For instance, UAR is used in [40, 9] due to the sample class ratio of BDC dataset is highly imbalanced of 39.42% as Table 1.

We have discussed common evaluation metrics across all surveyed papers. Despite the fact that there is no perfect evaluation metrics, however, we observed some important insights regarding this problem as follow. First, despite applying ML methods in medical domains, there was not any study that tried to follow medical-related metrics to report the performance. Second, the evaluation metrics used in the studies were mainly focused on the predictions, which might not be the best criterion to measure the performance of the proposed methods. Third, most of the work does not employ statistical tests in their reported performances. These shortcomings are important to be addressed in future work.

6. Conclusion and Future Directions

In this survey, we provide a scoping review of multimodal machine learning approaches in mental disorder detection from high impact venues. Moreover, we propose an end-to-end pipeline for multimodal machine learning tasks. This serves as an important source of information for developers and researchers to advance multimodal artificial intelligence for mental health. We outline a list of popular datasets and modalities. Finally, we propose an end-to-end multimodal pipeline for the investigated task.

The use of multiple modalities is promising yet to be further explored. Based on our review, we suggest the following future directions of research:

One of the most challenging tasks with multimodal machine learning is the harmonisation of different input modalities. Future developments of this field, therefore, could lie in the following research directions:

1. **Multidimensional fusion** is the next focus of fusion technique. Firstly, due to the increased representation complexity and number of modalities involved, dealing with high-dimensional fusion is expected as a part of enhanced information integration. Secondly, capturing the correlation among modalities is desired to better assist classifiers rather than simple concatenation.
2. **Co-learning**. Future studies are expected to be more widely adopted since this method is especially suitable when there are limited resources, which is a common trait of mental health datasets with small sample sizes and often involve self-reported metrics (i.e., a high chance of having incomplete data). This technique can enrich all modalities by transferring knowledge from one domain to another.

As technologies for human interaction evolve, more modalities can be captured for us to understand the mechanisms of mental disorders more clearly to help ease the burden of these conditions at an early stage.

References

[1] , 2018. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. AVEC .
[2] Abaeikoupaei, N., Al Osman, H., 2020. A multi-modal stacked ensemble model for bipolar disorder classification. IEEE TAC .
[3] Aloshban, N., Esposito, A., Vinciarelli, A., 2021. Language or paralanguage, this is the problem: Comparing depressed and non-depressed speakers through the analysis of gated multimodal units. Interspeech , 2496–2500.
[4] An, M., Wang, J., Li, S., Zhou, G., 2020. Multimodal topic-enriched auxiliary learning for depression detection, in: COLING.
[5] Arbanas, G., 2015. Diagnostic and statistical manual of mental disorders (dsm-5). Alcoholism and psychiatry research 51, 61–64.
[6] Baltruaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE TPAMI 41, 423–443.
[7] Bara, C.P., Papakostas, M., Mihalcea, R., 2020. A deep learning approach towards multimodal stress detection., in: AffCon@ AAAI, pp. 67–81.
[8] Bourke, C., Douglas, K., Porter, R., 2010. Processing of facial emotion expression in major depression: a review. Australian & New Zealand Journal of Psychiatry 44, 681–696.
[9] Ceccarelli, F., Mahmoud, M., 2021. Multimodal temporal machine learning for bipolar disorder and depression recognition. Pattern Analysis and Applications , 1–12.

[10] Ciftçi, E., Kaya, H., Güleç, H., Salah, A.A., 2018. The turkish audio-visual bipolar disorder corpus. ACII Asia , 1–6.
[11] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G.M., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D.R., Wood, R., Xu, Y., Rizzo, A.A., Morency, L.P., 2014. Simsensei kiosk: a virtual human interviewer for healthcare decision support, in: AAMAS.
[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, ACL. pp. 4171–4186.
[13] Duong, A.Q., Ho, N.H., Yang, H.J., Lee, G.S., Kim, S.H., 2021. Multi-modal stress recognition using temporal convolution and recurrent network with positional embedding, in: MuSe, pp. 37–42.
[14] Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E.M., Winther, O., Bardram, J.E., Kessing, L.V., 2016. Voice analysis as an objective state marker in bipolar disorder. Translational psychiatry 6, e856–e856.
[15] Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K.J., Tørresen, J., 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing 51, 1–26.
[16] Gong, Y., Poellabauer, C., 2017. Topic modeling based multi-modal depression detection, in: AVEC, pp. 69–76.
[17] Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D.R., Rizzo, A.A., Morency, L.P., 2014. The distress analysis interview corpus of human and computer interviews, in: LREC.
[18] Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., Chen, Z., 2019. Cooperative multimodal approach to depression detection in twitter, in: AAAI, pp. 110–117.
[19] Jaiswal, M., Aldeneh, Z., Bara, C.P., Luo, Y., Burzo, M., Mihalcea, R., Provost, E.M., 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. ICASSP , 7415–7419.
[20] Lahat, D., Adali, T., Jutten, C., 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE 103, 1449–1477.
[21] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: COLING, pp. 7871–7880.
[22] Marazziti, D., Consoli, G., Picchetti, M., Carlini, M., Faravelli, L., 2010. Cognitive impairment in major depression. European journal of pharmacology 626, 83–86.
[23] Mitchell, A.J., Vaze, A., Rao, S., 2009. Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet 374, 609–619.
[24] Orton, I., 2020. Vision based body gesture meta features for affective computing. ArXiv abs/2003.00809.
[25] Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R., 2019. Multi-level attention network using text, audio and video for depression prediction, in: AVEC, pp. 81–88.
[26] Ricky, C., O'Donnell Siobhan, M.N., et al., 2017. Factors associated with delayed diagnosis of mood and/or anxiety disorders. Health promotion and chronic disease prevention in Canada: research, policy and practice 37, 137.
[27] Ritchie, H., Roser, M., 2020. Mental health. our world in data. 2018.
[28] Rodrigues Makiuchi, M., Warnita, T., Uto, K., Shinoda, K., 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection, in: AVEC, pp. 55–63.
[29] Rohanian, M., Hough, J., Purver, M., et al., 2019. Detecting depression with word-level multimodal fusion., in: INTERSPEECH.
[30] S., S., Raj, J.S., 2021. Analysis of deep learning techniques for early detection of depression on social media network - a comparative study.
[31] Samareh, A., Jin, Y., Wang, Z., Chang, X., Huang, S., 2018. Predicting depression severity by multi-modal feature engineering and fusion, in: AAAI.
[32] Santomauro, D.F., Herrera, A.M.M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D.M., Abbafati, C., Adolph, C., Amlag, J.O., Aravkin, A.Y., et al., 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. The Lancet 398, 1700–1712.
[33] Shatte, A.B.R., Hutchinson, D.M., Teague, S.J., 2019. Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine 49, 1426 – 1448.
[34] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.S., Zhu, W., 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution., in: IJCAI.
[35] Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E.M., Cambria, E., Zhao, G., Schuller, B.W., 2021. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. MuSe .
[36] Xing, X., Cai, B., Zhao, Y., Li, S., He, Z., Fan, W., 2018. Multi-modality hierarchical recall based on gbdts for bipolar disorder classification, in: AVEC, pp. 31–37.
[37] Yao, Y., Papakostas, M., Burzo, M., Abouelenien, M., Mihalcea, R., 2021. Muser: Multimodal stress detection using emotion recognition as an auxiliary task. arXiv preprint arXiv:2105.08146 .
[38] Yin, S., Liang, C., Ding, H., Wang, S., 2019. A multi-modal hierarchical recurrent neural network for depression detection, in: AVEC, pp. 65–71.
[39] Zhang, X., Pan, J., Shen, J., Din, Z.U., Li, J., Lu, D., Wu, M., Hu, B., 2020a. Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection. IEEE TAC .
[40] Zhang, Z., Lin, W., Liu, M., Mahmoud, M., 2020b. Multimodal deep learning framework for mental disorder recognition, in: FG, IEEE.
[41] Zheng, W., Yan, L., Gou, C., Wang, F.Y., 2020. Graph attention model embedded with multi-modal knowledge for depression detection, in: ICME, pp. 1–6.
[42] Zhou, D., Luo, J., Silenzio, V.M., Zhou, Y., Hu, J., Currier, G., Kautz, H., 2015. Tackling mental health by integrating unobtrusive multimodal sensing, in: AAAI.
[43] Zogan, H., Razzak, I., Jameel, S., Xu, G., 2021. Depressionnet: learning multi-modalities with user post summarization for depression detection on social media, in: ACM SIGIR, pp. 133–142.