MULTIMODAL MACHINE LEARNING IN MEDICAL SCREENINGS

THUY TRINH NGUYEN

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Master of Science by Research

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

SEPTEMBER 2023

Abstract

The healthcare industry, with its high demand and standards, has long been considered a crucial area for technology-based innovation. However, the medical field often relies on experience-based evaluation. Limited resources, overloading capacity, and a lack of accessibility can hinder timely medical care and diagnosis delivery. In light of these challenges, automated medical screening as a decision-making aid is highly recommended. With the increasing availability of data and the need to explore the complementary effect among modalities, multimodal machine learning has emerged as a potential area of technology. Its impact has been witnessed across a wide range of domains, prompting the question of how far machine learning can be leveraged to automate processes in even more complex and high-risk sectors.

This paper delves into the realm of multimodal machine learning in the field of automated medical screening and evaluates the potential of this area of study in mental disorder detection, a highly important area of healthcare. First, we conduct a scoping review targeted at high-impact papers to highlight the trends and directions of multimodal machine learning in screening prevalent mental disorders such as depression, stress, and bipolar disorder. The review provides a comprehensive list of popular datasets and extensively studied modalities. The review also proposes an end-to-end pipeline for multimodal machine learning applications, covering essential steps from preprocessing, representation, and fusion, to modelling and evaluation. While cross-modality interaction has been considered a promising factor to leverage fusion among multimodalities, the number of existing multimodal fusion methods employing this mechanism is rather limited. This study investigates multimodal fusion in more detail through the proposal of Autofusion, an autoencoder-infused fusion technique that harnesses the cross-modality interaction among different modalities. The technique is evaluated on DementiaBank's Pitt corpus to detect Alzheimer's disease, leveraging the power of cross-modality interaction. Autofusion achieves a promising performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. The technique consistently outperforms all unimodal methods by an average of 5.24% across all metrics. Our method consistently outperforms early fusion and late fusion. Especially against the late fusion hard-voting technique, our method outperforms by an average of 20% across all met-

rics. Further, empirical results show that the cross-modality interaction term enhances the model performance by 2-3% across metrics. This research highlights the promising impact of cross-modality interaction in multimodal machine learning and calls for further research to unlock its full potential.

Keywords: multimodal machine learning, automated medical screening, mental disorder detection, Alzheimer's disease detection

Acknowledgements

I would like to express my utmost gratitude towards my supervisors, Dr. Harry Nguyen and Dr. Fani Deligianni, who have guided me through this journey and supported me since my first day. I am deeply indebted to Alex and Ashley, who have placed their belief in me and uplifted my spirit through all the ups and downs. To my family, my grandparents, parents, sister, and nephews, many thanks to you for being my forever mental support.

Declaration

I hereby declare that, except where specific references are made to the work of others, the contents of this document are original and have not been submitted, in whole or in part, for consideration for any other degree or qualification, in this or any other university. This thesis is the result of my own work under the supervision of Dr. Harry Nguyen and Dr. Fani Deligianni. Nothing included is the outcome of work done in collaboration, except where otherwise indicated within the text. The following publications serve as a foundation for the research contributions.

Publications:

- ***Nguyen, T.T.**, Pham, V.H.Q., Le, D.T., Vu, X.S, Deligianni, F., and Nguyen, H.D., 2023. Multimodal machine learning in mental disorder detection: A scoping review. In Knowledge-Based and Intelligent Information and Engineering Systems: 27th International Conference, KES 2023, Athens, Greece, September 6-8, 2023, Proceedings. Springer Berlin Heidelberg.*

  Contribution: Lead author. Thuy Trinh Nguyen proposed the idea of the paper, defined the search strategy and exclusion criteria, summarised the search outcome, analysed trends and patterns, and prepared the manuscript. This work is the foundation for Chapter 3.

- *Hoang, T., **Nguyen, T.T.** and Nguyen, H.D., 2022. Unified tensor network for multimodal dementia detection. AAAI Workshop 2022. In Multimodal AI in healthcare: A paradigm shift in health intelligence (pp. 409-416). Cham: Springer International Publishing.*

  Contribution: Thuy Trinh Nguyen was responsible for data exploration, designed and executed the experiments, and prepared the manuscript. This work contributes to the understanding of DementiaBank's Pitt Corpus, which is used in Chapter 4 for evaluation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The past decades have seen tremendous expansion in technology and artificial intelligence in particular. Almost all spheres of life now include intelligence in a revolutionary way thanks to machine learning. Machine learning may boost the effectiveness and automation of the process with each stream of sensory retrieval. Some real-world applications for computer vision technologies include traffic management [1] and manufacturing defect detection [2]. Applications of natural language processing, another crucial field of machine learning, range from smart assistants [3] to sentiment analysis [4]. Now that technology has a significant impact across a variety of domains, we may raise the question of how far machine learning might be used to automate processes in even more complicated and high-risk sectors.

The medical industry has long been perceived as a high-risk sector requiring the utmost precision for effective treatment. Most medical cases are handled individually as a case-by-case evaluation based on the healthcare professional's experience. However, the experience-based and dependent-on-people aspect of the healthcare sector leads to capacity overload due to the scarcity of medical personnel and prolonged diagnosis periods. Certain populations cannot obtain appropriate and timely medical care in some circumstances due to poor accessibility and inadequate resources [5]. The lack of explainability and objectivity in screening also questions the effectiveness and scalability of traditional diagnostic methods [6]. As the world has observed the impact of technology-based solutions across numerous sectors in the past decade, researchers are increasingly interested in applying machine learning to healthcare challenges.

With the rise of technology-driven advancements in society, machine learning-based algorithms have been applied in healthcare as a screening aid or a second opinion for medical staff. Accumulating different sources of information, machine learning can leverage this in-

formation, discover patterns, and offer an appropriate screening result. Automated medical screenings can benefit both patients and healthcare professionals. For patients, the following advantages exist or are significantly enhanced by healthcare automation:

- Receive a collaborative screening outcome from technological analysis and experts' opinions; hence, face fewer chances of misdiagnosis [7]
- Obtain an objective outcome with plausible explainability [8]
- Bear potentially shorter wait time for patient sorting and screening [9]
- Experience better privacy, especially for persons with socially stigmatised conditions or who have trouble communicating due to social anxiety or a fear of medical professionals and hospitals [10]
- Receive continuous medical attention despite geographical and temporal constraints (i.e., increased engagement) [11]

Medical professionals gain access to technology's comprehensive perspective on patients' medical histories, symptoms, and other monitoring data as a helpful reference for the final decision-making. Given the disproportionate specialist-to-patient ratio, automation can help reduce burnout among healthcare workers [12].

Regarding machine learning-based methods for this task, multimodal machine learning is a promising subset for utilising and harmonising multiple data streams for medical screenings and decision-making aids. A modality is a data stream, with some of the most common modalities being media input types such as audio, video, text, and image or tabular types such as metadata and time series data. In healthcare applications, multimodal machine learning can incorporate data from various sources, such as medical imaging, electronic health records, and patient-tracking data, to better comprehend the disease and its progression. ***Multimodal fusion***, or the process of integrating diverse data sources and exploiting their complementary effects, is regarded as the most critical factor in the success of multimodal machine learning in the medical field; among the challenges accompanying this machine learning scheme.

Applied machine learning has been used to detect a wide range of medical problems ranging from diabetic retinopathy [13], lung cancer [14], respiratory conditions such as COVID-19 [15], and various mental disorders such as depression [16] and bipolar disorder [17]. Among these applications of automated medical screenings, the group of mental disorders remains one of the most prevalent conditions and obtain suitable features for multimodal machine learning.

Our conversation will revolve around the crucial topic of mental health, which is a significant area of concern within the medical field. This is primarily because mental disorders are highly prevalent, with a reported 10% of the population suffering from common mental disorders [18]. However, despite their prevalence, mental disorders are often misdiagnosed or go undetected due to their social stigma, hidden nature, and limited accessibility in certain countries. For instance, a study on soldiers' attitudes towards technology-based approaches to mental health care indicates that one-third of those who were reluctant to engage in face-to-face counselling were open to trying at least one technology-based approach for mental health care [10]. These findings suggest that technology-based approaches can be a viable solution to overcome obstacles to accessing mental health care. Furthermore, mental disorders can lead to and worsen other health problems. For instance, chronic stress is a significant risk factor for hypertension and cardiovascular disease [19]. Lastly, mental disorders possess distinctive characteristics, which make them particularly advantageous for machine learning and multimodal machine learning studies. Studies have shown that individuals with mental disorders process different modalities in unique ways, highlighting the potential use of multimodal machine learning to harness this discriminatory information and evaluate the complementary impact of multimodal data [20].

This thesis intends to validate the application of multimodal machine learning in mental disorders and design an implementation centred on multimodal fusion for a type of mental disease detection.

1.2 Thesis Statement

This thesis studies the application of multimodal machine learning in medical screenings, validates this approach in mental disorder detection, and proposes the Autofusion technique, a potential autoencoder-integrated fusion that embraces the cross-modality interaction for multimodal machine learning.

1.3 Thesis Questions

This thesis seeks to answer the following research questions by investigating the background of multimodal machine learning, automated medical screening, and the influence of multimodal fusion implementation:

- **RQ1:** What are the current methods and fusion techniques to integrate multimodalities (e.g., video, audio, text) in medical screenings, specifically mental disorder detection?

- **RQ2:** Considering the current state of multimodal fusion, early and late fusion techniques are among the most popular. These fusion options, however, lack the consideration of cross-modality interaction among different modalities. How to design a neural network framework to efficiently fuse multimodal data representations with cross-modality interactions?
  - **RQ2a:** How effective is this framework compared to single-modality models?
  - **RQ2b:** How effective is this framework compared to existing multimodal fusion techniques?
  - **RQ2c:** How does the cross-modality interaction impact the overall performance?

1.4 Contributions

This thesis contributes in three distinct ways. The thesis begins by providing the audience with a background of the problems and potential of multimodal machine learning in the context of automated medical screening.

Second, we provide a scoping overview of multimodal machine learning applications in a specific area of healthcare that is mental disorder detection. This review analyses patterns across articles from high-impact venues. To facilitate the use of multimodal machine learning in the assessment of mental disorders, this research presents an end-to-end pipeline.

Finally, this thesis proposes a novel fusion mechanism with an emphasis on cross-modality interaction to further enhance the performance of the classification network. Autofusion consistently outperforms both existing fusion methods and single-modality models. This enhancement is an exciting result of the thesis, with the potential to boost the effectiveness of multimodal machine learning in other settings.

This thesis may serve as a resource for researchers looking to include a multimodal machine learning model in their work on medical screening in general and mental disorder screening specifically. Researchers may build their process using the end-to-end framework and use autofusion as their multimodal fusion module to improve the performance of their model.

The thesis demonstrates to professionals the value of incorporating multimodal machine learning and automated medical screening into user-friendly applications to address the limitations of existing medical screening techniques. It is time for people with mental illness to have some relief from the monetary, emotional burdens and the social stigma that come with getting a proper diagnosis. While this automated approach is still being tested, it has the potential to become a valuable resource that aids healthcare professionals in making well-informed decisions quickly and efficiently.

1.5 Thesis Structure

The remaining three chapters of the thesis outline a progression of learning from theoretical multimodal machine learning in medical applications to the practical utilisation of multimodal fusion techniques to improve performance. The following are the chapters:

- **Chapter 2: Background.** The chapter provides an overview of automated medical screening, especially in mental disorder detection, its effects, and an evaluation of conventional diagnostic techniques. The background also sheds light on the development of machine learning, the characteristics of multimodal machine learning, and the opportunities and challenges associated with this topic.

- **Chapter 3: Multimodal Machine Learning in Mental Disorder Detection - A Scoping Review.** The chapter examines influential papers on multimodal machine learning for detecting three prevalent mental disorders, namely depression, stress, and bipolar disorders. This chapter also presents an end-to-end multimodal pipeline for multimodal machine learning implementation.

- **Chapter 4: Autofusion - Multimodal Machine Learning in Dementia Detection.** Chapter 4 examines the use of multimodal machine learning in screening Alzheimer's disease, the most common cause of dementia cases globally. This chapter proposes the Autofusion technique that leverages cross-modality interaction and autoencoder incorporation as a viable alternative to unimodal and existing fusion methods.

- **Chapter 5: Discussion.** This chapter discusses significant observations and lessons learned from prior sections. Chapter 5 provides a reflection on the proposed pipelines, models, and their effectiveness. This chapter addresses the potential and ramifications of multimodal machine learning in medical screening.

- **Chapter 6: Conclusion and Future Direction.** This chapter summarises the major contributions of the thesis, emphasises significant experiment results, and addresses the stated research questions. Finally, it concludes the thesis by discussing prospective work opportunities and recommendations.

Chapter 2

Background

This chapter provides fundamental background on two important subjects that are automated medical screening and multimodal machine learning. We first discuss the feasibility of automated medical screening and the medical conditions relevant to the thesis. Subsequently, the fundamentals of machine learning will be discussed, followed by the requirements, opportunities, and obstacles of multimodal machine learning.

2.1 Medical Screening

2.1.1 Automated Medical Screening

Healthcare has long been considered a high-risk industry that requires extreme accuracy for optimal treatment. The vast majority of medical issues are treated individually based on the professional's experience. However, the shortage of medical professionals and requirements for professional expertise result in capacity overflow and extended diagnosis wait time. In the field of heavy experience reliance, it raises the question of objectivity and explainability for fully-human diagnosis.

As technological advancements proliferate, machine learning-based algorithms have been used in healthcare as a decision-making aid for medical professionals. Machine learning may exploit information from many sources, detect trends, and provide meaningful explanations. With the contribution of technology, patients can enjoy the following improvements in healthcare processes:

- Receive a more objective and explainable diagnosis (i.e., lower the likelihood of misdiagnosis) with the collaboration between technical analysis and experts' opinions

- Experience shorter wait time for diagnosis thanks to the involvement of technology in the pre-screening process
- Experience better privacy for people with mental or communication barriers
- Enhance engagement despite resource and accessibility limitations

As a result, automated medical screening is urgently necessary. To cater for this need, there has been a growing body of literature on machine learning-based screening methods for medical problems ranging from both chronic conditions such as lung cancer [14] to acute ones such as COVID-19 [15].

Among the areas of the medical domain, we will focus the discussion on a crucial subset: mental health. There are a few reasons for this selection. Firstly, mental disorders are highly prevalent conditions. It is estimated that about one in every ten people worldwide suffers from at least one type of mental disorder [21]. Depression is ranked by WHO as the single most significant contributor to global disability (7.5% of all years lived with disability in 2015), while anxiety disorders are ranked 6th (3.4%) [22]. Despite their prevalence, mental disorders are easily misdiagnosed or undetected due to their social stigma, hidden nature, and limited accessibility in many countries. This point will be discussed further in the next section. Moreover, mental disorders tend to create and worsen other health problems. For instance, chronic stress is identified as a significant risk factor for hypertension and cardiovascular disease [19]. Last but not least, mental disorders possess traits that are particularly useful for machine learning and multimodal machine learning studies. Some research revealed the uniqueness of how mental disorder patients process different modalities, encouraging the use of multimodal machine learning to employ this discriminant information and study the complementary impact of multimodal data [20]. A discussion of prevalent mental disorders and automated screening in their detection will be provided in Chapter 3.

2.1.2 Automated Screening in Mental Disorders

According to WHO, there was a significant 25% spike in anxiety and depression cases globally in the first year of the Covid-19 pandemic, with the primarily affected groups being young people and women [23]. Mass quarantine, emotional and financial losses are the main attributes of this downfall in mental well-being [24]. With an escalated self-harming rate, this was a wake-up call for the world to acknowledge the importance of mental health.

A worldwide conversation on mental health has just lately gained attention, but COVID-19 is merely the tip of the iceberg. Mental disorders have been a silent but detrimental part of many human lives. With mental illness, there is a higher chance of early school abandonment, a lesser chance of finding full-time employment, and a worse quality of life overall [25].

Despite the adverse effects of mental disorders, between 76 and 85% of the patients in developing countries are not being treated [21]. A few reasons contribute to this lack of treatment. Firstly, the widespread misunderstanding and stigma surrounding the topic of mental illnesses are among the top barriers that prevent help-seeking. While the concept of mental illnesses has changed over time, the stigma toward their patients remains strong, especially among low- and middle-income countries (LMICs) [26]. More concerned, stigma could negatively affect care-seeking and treatment engagement. Secondly, since face-to-face examination remains the primary mode of diagnosis, the accessibility of mental disorder detection is limited, especially in developing countries. This shows the scarcity, inequity, and inefficiency of resource distribution and availability for mental health in developing countries. Whilst the primary form of diagnosis remains through psychiatrist evaluation, the substantial lack of mental health personnel in South Africa, for instance, at only 0.08 to 0.89 per 100,000 uninsured population is particularly alarming [27]. In the last few years, when the peak of COVID-19, a high-impact prominent source of psychological distress, occurred, many LMICs recognised the need to address the state of their people's mental health and introduce innovative means such as digital technology solutions to ease social stigma, identify individuals at risk through social media footprint, and assist front line workers [28].

This thesis focuses on common conditions such as depression, bipolar-, stress disorder, and Alzheimer's disease. Table 2.1 summarises the key characteristics of each mental disorder.

(1) Depressive Disorders

Currently, the official definition of depression is still debated among psychiatrists. Within the framework of this article, we refer to the depressive disorder as a disorder which can be characterised by sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, feelings of tiredness, and poor concentration [29].

Prevalence According to the World Health Organisation (WHO), depression is a common illness, affecting an estimated 3.8% of the global population, including 5.0% of adults and 5.7% of adults over the age of 60 [30]. Although depression affects people of all ages, those who are impoverished and unemployed or experiencing critical life events such as the death of a loved one, a relationship break-up, physical illness, and substance-related issues are at a higher risk of becoming depressed.

Symptoms There are three criteria of physical symptoms for major depressive disorders listed in the DSM-IV, which are sleep disturbance, appetite disturbance, and fatigue or loss of energy [31]. About 50% of depressed patients report pain, and many types of pain occur more frequently in people with depression than in those without [32]. A change in thinking is a crucial aspect of depression. A person suffering from depression will likely have poor self-perception and feel unlovable and worthless. Pessimism about themselves, the present, and

Disorder	Symptom	Assessment scale
Depressive disorder	- Loss of interest or pleasure - Feelings of guilt or low self-worth - Disturbed sleep or appetite - Feelings of tiredness - Poor concentration	- Hamilton Depression Rating Scale (HDRS) - Montgomery and Åsberg Depression Rating Scale (MADRS) - Beck Depression Inventory (BDI or BDI-II) - Patient Health Questionnaire (PHQ-9)
Bipolar disorder	- Fluctuation of energy - Sudden mood swing - Excessive impulsive behavior - Greatly elevated mood - Impatience - Decreasing desire for sleep	- Hypomania Checklist 32 (HCL-32) - Young Mania Rating Scale (YMRS) - Altman Self-Rating Mania Scale (ASRM)
Stress disorder	- Emotional distress - Muscular ache and tension - Over arousal - Elevated blood pressure	- Perceived Stress Scale (PSS) - Depression, Anxiety, and Stress Scale (DASS)
Dementia	- Significant memory loss - Poor judgement - Loss of memory of recently learned information - Shortened attention span - Hallucinations and delusions	- National Institute of Neurological Disorders and Stroke - Alzheimer Disease and Related Disorders (NINCDS – ADRDA) criteria - Alzheimer's Disease Assessment Scale (ADAS)

Table 2.1: List of concerned mental disorders

the future is the obvious manifestation of these people. People with the depressive disorder frequently have difficulties concentrating and making simple decisions. When depression is mild to severe, some people have suicidal thoughts. Amongst all types of depression disorders, major depressive disorder is the most prevalent, affecting approximately 15–17% of the population and showing a high suicide risk rate equivalent to around 15% [33].

Assessment Traditionally, psychiatrists can use a wide variety of depression disorders assessment scales. These scales are divided into three categories: clinician-rated measures, patient self-report scales, and scales that include both administrations.

Each clinician-rated scale has a unique collection of psychometric qualities, including items, scales, and dimensionality. The most common is the Hamilton Depression Rating Scale (HDRS) [34]. It is a multiple-item questionnaire addressing depression indicators to measure the severity of the condition. The result shall be divided into different categories like no depression (HDRS 0-7), mild depression (HDRS 8-12), less than major depression (HDRS 13-17), major depression (HDRS 18-29) and more than major depression (30+). Another scale is the Montgomery and Åsberg Depression Rating Scale (MADRS) [35], which consists of a clinical interview and ten items covering major depressive symptoms. The MADRS appears to be a uni-dimensional scale and is more oriented towards psychic than somatic aspects of depression [36].

Some measurements method could be completed by the patients themselves, such as the Beck Depression Inventory (BDI - with the updated version BDI-II) [37], and Patient Health Questionnaire-9 (PHQ-9) [38]. PHQ-9 is a self-administered Primary Care Evaluation of Mental Disorders (PRIME-MD) version. This scale contains nine questions about the patient's experience within the last two weeks. The questions include the amount of interest in daily activities, feelings of sadness or depression, sleep, energy levels, food choices, self-perception, capacity to focus, function rapidity, and suicidal thoughts.

Although there are different scales of diagnosing depressive disorders, each method, when used independently, does not cover all the depressive items. For example, the HDRS-17 version, the MADRS, and the BDI do not contain symptoms of atypical depression (e.g., hypersomnia, weight or hunger gain). The BDI does not include symptoms of motor retardation or anxiety. Motor retardation is not assessed in the MADRS. The discrepancy among evaluation scales further highlights the lack of objectivity in traditional diagnosis.

In the past decades, using the benefits of technology-based methods, the research community has been intrigued by automated depression detection for more robust and explainable solutions. Brain signals EEG-based approach is often used with convolutional neural networks [39]. Another direction is to explore behavioural factors such as audiovisual cues from interview speech and facial expressions [40]. The combination of different data streams has shown promising results, which paves the way for multimodal machine learning in depression detection.

The Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ) [41] is one of the most prominent corpora in detecting depression with multiple data streams. More studies have been released to encourage the collection of modern multimodal datasets in detection, such as the Wellbeing dataset [42] and the depression dataset extracted from social media [16]. A more in-depth review of multimodal machine learning in depression detection will be included in Chapter 3.

(2) Bipolar Disorders

The Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-5) identifies three distinct subtypes of the condition, as follows: bipolar disorder is a chronic mental disorder that involves significant fluctuations in mood state and energy [43]. [44] defines bipolar disorder as a manic-depressive illness or manic-depressive psychosis, characterised by sudden swings in mood and a person's ability to function without a seemingly justifying cause. The classification is as follows:

- *Type I Bipolar Disorder*: It is diagnosed when (1) there is a mixed episode combination of excitatory symptoms such as overconfidence, grandiosity, chattiness, excessive im-

pulsive behaviour, impatience, decreased desire for sleep, and greatly elevated mood, or (2) when there is at least one episode of depression followed by at least one sudden manic episode.

- *Type II Bipolar Disorder*: It is identified when at least one severe depressive episode is followed by at least one spontaneous hypomanic episode. This is incredibly challenging to appropriately diagnose since it is difficult to distinguish this condition from recurrent unipolar depression in depressed people.
- *Cyclothymic Disorder*: It is diagnosed with hypomanic and depressed symptoms that do not match depressive episode criteria.

Prevalence According to WHO, 40 million people experienced bipolar disorder in 2019 [45]. Often diagnosed in the employed population (i.e., young adulthood), bipolar disorder infers substantial economic loss to society [46]. The life expectancy of bipolar patients decreases substantially from 8 to 12 years [47], further demonstrating this disorder's adverse impact. Bipolar disorder requires both acute management (e.g., mood stabilisers and antipsychotics) and chronic management, including a combination of pharmacological, psychological, and lifestyle approaches [48]. This mental condition, therefore, poses a tremendous burden on individuals' lives.

Symptoms Bipolar disorder is defined by the alternation of a depressive state and a manic state when there is a simultaneous presence of depressive and manic symptoms (i.e., one is neither wholly depressed nor completely in mania) with the predominance of irritability, anxiety, and restlessness [44]. DSM-5 mentions excitatory symptoms, including energy fluctuation, excessive impulsive behaviours, significantly elevated mood, and disturbed sleep pattern [49]. While effective symptomatology is not the most visible manifestation, behavioural repercussions are more objective and relevant for diagnostic purposes.

Assessment Like depressive disorder, clinical-based and self-reported assessment scales contribute to the bipolar disorder diagnosis process. Some examples of these scales are the Young Mania Rating Scale (YMRS), Altman Self-Rating Mania Scale (ASRM), and the Hypomania Checklist (HCL-32).

YMRS is an 11-item diagnostic scale to assess manic symptoms, which is generally based on the subject's 48 hours report of their condition [50]. 4 of 11 questions are scored from 0-8 (i.e., Irritability, Speech, Content, and Disruptive-Aggressive behaviours, while the other seven are graded from 0-4. YMRS scores are used in the Turkish Audio-Visual Bipolar Disorder Corpus as a source for labels [51].

ASRM is a 5-question scale that assesses an individual's (1) mood, (2) self-confidence, (3) sleep disturbance, (4) speech, and (5) activity over a week to detect bipolar disorder [52]. While this assessment is practical and quick to determine the existence of bipolar conditions,

it is challenging to identify which type of bipolar disorder is associated with the case. ASRR cutoff score at five has shown an optimal combination of sensitivity and specificity of 85.5% and 87.3%, respectively.

HCL-32 targets identifying hypomanic traits within patients of major depressive disorder to screen for bipolar disorder conditions [53]. The questionnaire helps discriminate against individuals with bipolar and depressive disorder. There are 32 items listed in question 3 in which the individual is asked to describe how they felt when they were in an up-mood state.

Despite their ease of use and popularity, phenomenology-based diagnosis methods such as assessment scales are controversial for bipolar disorders. Substantial variations between the DSM-5 and ICD-10 definitions of bipolar disorder imply that certain people will be labelled with bipolar disorder under one system but not the other. Furthermore, the inability to confirm a diagnosis of bipolar disorder may be unavoidable until a full-blown episode of mania or hypomania has occurred, even though many patients will begin their disease with a bout of depression and may have had another hypomanic episode that might not meet criteria for the duration of symptoms. Bipolar disorder is misdiagnosed as a depressive disorder in various cases. As a result, an inherent diagnostic delay could require 8 to 10 years to reach an accurate diagnosis and treatment [54]. To bridge this gap of delay and ineffectiveness of bipolar condition detection, novel methods are highly demanded to track the patients' mental states consistently to avoid unobserved episodes and carefully consider multiple factors around the patients before reaching a diagnosis.

Technology-based methods, therefore, are researched to add value to the traditional diagnosis of bipolar disorders. One approach to automated bipolar detection focuses on investigating the time-based social footprint and behaviours. For instance, [55] reports that automatic smartphone sensing via Social Rhythm Metric (SRM) is a viable indicator for bipolar individuals as it tracks the continuity of social behaviours that could reveal the entire series of the patient's episodes. EEG-based machine learning methods also suggest promising results for automated bipolar screening [56]. Regarding diagnosis via behavioural information, the study of audiovisual cues [17] is receiving growing attention. Automated bipolar detection, hence, is a potential area of study.

(3) Stress Disorders

A stress response is a natural reaction to threats and changing environments. Stress exists in two forms which are acute and chronic stress. According to DSM-IV, a person suffers from acute stress disorder if they have experienced or witnessed a life-threatening or traumatic event that incurs extreme fear, helplessness, or terror [31]. A profound example of acute stress disorder is a post-traumatic stress disorder. On the other hand, chronic stress is caused by long-standing pressures and expectations. There can be a progression of acute stress

becoming chronic if the stressors are repeated to a certain extent. This thesis covers this topic as generalised stress-related disorders rather than investigating each condition individually.

Prevalence Stress disorder is among the most prevalent mental disorders worldwide. According to the World Health Organisation, chronic stress can trigger pre-existing health conditions and encourage increased consumption of substances [57]. More alarmingly, it can worsen other medical conditions, including hypertension, heart disease [19], anxiety and depressive disorder [58].

Symptoms Stress, whether chronic or acute, has a mental and physical impact on its victims. According to [59], common symptoms include emotional distress, headache, muscular soreness, and elevated blood pressure. Patients may also develop sleep deprivation and altered appetite due to stress disorder [57]. Stress symptoms, if they continue, can impose significant difficulties and diversions on the patient's everyday functioning.

Assessment Similar to most mental disorders, traditional stress disorder diagnosis mainly relies on self-reported measures. Examples include the Perceived Stress Scale (PSS) and the Depression, Anxiety, and Stress Scale (DASS).

PSS 10-item and 14-item scales are developed to assess the stress level of an individual based on life events within the previous month. It is further reported that the 10-item version demonstrated superior psychometric traits compared to the 14-item measure. [60].

DASS has two versions the original 42-item and the compressed 21-item. This self-reported questionnaire is designed to measure the magnitude of three mental disorders with the DASS-stress subcategory focusing on tension and irritability [61].

In addition to validated scales, technology-based methods employing biomarkers and physiological metrics have also been explored to enhance the objectivity of the diagnosis. Cortisol levels, for example, are a prominent sign of stress disorder. Studies suggest that cortisol sensing at point-of-care is gaining more attention, specifically the salivary cortisol test strips, thanks to their portability, low cost, and fast analysis time [62]. Wearable technology improvements have increased the use of heart electrical activity (ECG) and brain activity (EEG) in stress detection. [63] reports ECG improves the efficacy of stress detection when combined with emotion identification using facial expressions. These promising applications point to the potential and viability of automated medical screenings.

(4) Alzheimer's Disease

Dementia is a general term for clinical memory loss and cognitive deterioration. Alzheimer's disease (AD) accounts for the majority of dementia cases. AD is a degenerative cognitive disease that, when becomes prominent, leaves its patients dependent and requiring around-

the-clock care [64]. There are three primary stages of AD which are preclinical AD, mild cognitive impairment (MCI) and Alzheimer's dementia [65].

Prevalence Currently, there are more than 55 million people who suffer from dementia, over 60% of whom live in low-and middle-income countries [66]. This number is expected to increase by 10 million cases annually. [67] suggests that AD adversely affects its patients' spirituality, religiosity, and quality of life. This means that patients with AD require specialised care and support to maintain their well-being. Caregivers of AD patients are also impacted by the condition, and it is recommended that they seek counselling and support for their physical health [68].

Symptoms Individuals with AD often exhibit both behavioural and psychological symptoms that can vary in severity depending on the progression of the condition. The most notable symptoms of AD are related to cognitive deficiencies, such as memory loss, poor judgement, and a short attention span. What sets AD apart from natural ageing is that the cognitive decline is more severe and interferes with the patient's daily activities, work, or social interaction [69]. Other neuropsychiatric symptoms, including apathy, aggression, and psychosis, are also commonly observed in AD [70]. Apathy-type symptoms are linked to cognitive and motivational decline and can present as a loss of interest, unresponsiveness, and passivity [71]. Physical agitation and verbal aggression symptoms can also manifest in patients with AD, and the severity of these symptoms is often correlated with the progression of the disease. Late-stage AD patients may also experience psychosis symptoms, such as hallucinations and delusions.

Assessment Early AD assessments focus on clinical criteria for diagnosis. National Institute of Neurological Disorders and Stroke – Alzheimer Disease and Related Disorders (NINCDS – ADRDA) criteria is the prevailing diagnostic benchmark in research [72]. The criteria outline different groups of clinical symptoms for each AD subcategory: (1) Probable AD dementia, (2) Possible AD dementia, and (3) Probable or possible AD dementia with evidence of the AD pathophysiological process.

A different approach to diagnosing AD involves the use of evaluation scales. One such scale is the Alzheimer's Disease Assessment Scale (ADAS), which is comprised of 21 items categorised into three parts: (1) short neuropsychological tests for patients, (2) clinician's rating based on observations, and (3) an interview with the patient's caregiver [73]. ADAS focuses on a wide range of symptoms associated with AD, including cognitive issues such as memory loss and language difficulties, as well as non-cognitive symptoms like agitation and psychotic patterns.

The complex nature of electronic health records has led to the development of technology-based solutions that utilise multimodal data to identify the underlying mechanisms of AD dementia and aid in decision-making. Over the past decade, high-tech clinical support meth-

ods have been extensively explored and have proven their potential in accurately predicting and shortening the diagnosis of AD. Various methods have been studied and evaluated for automated AD screening, including brain activity EEG-focused diagnosis [74], brain scans MRI-based deep learning networks [75, 76], and multimodal machine learning techniques [77, 78]. Technology such as machine learning also provides a confidence score for each projection and can explain the top contributing factors for the diagnosis, making it a valuable tool.

From the previous discussion of prevalent mental disorders, there is a rise in the development of mental health monitoring technologies. With the help of sensor technologies and connectivity, data from various sources such as social networks, smartphones, wearable sensors, and neuroimaging technology can be easily collected. Utilising this vast amount of data can prove to be highly beneficial in detecting mental disorders. Chapter 3 will revisit the subject of automated mental disorder diagnosis using multimodal machine learning, a mighty stream of machine learning.

2.2 Multimodal Machine Learning

2.2.1 Machine Learning

Machine learning is a branch of artificial intelligence (AI) that studies and implements systems that can learn and improve from data without being explicitly programmed. The study of machine learning is believed to have emerged in the 1950s, with one of the first well-known self-learning applications being Samuel's game checkers; machine learning brings great potential for a wide range of applications [79] including:

- Computer vision: Object recognition, object detection
- Speech recognition and generation
- Semantic analysis, natural language processing and information retrieval
- Prediction: Classification, analysis, diagnosis

The world has seen the footprints of machine learning in almost every aspect of life, from daily activities with autonomous vehicles guided by the latest computer vision technologies to high-expertise functions such as medical diagnosis. As the years go by, there has been significant progress in popular machine learning algorithms.

Supervised machine learning. This scheme provides input-output pairs that train the model to separate the data for classification problems or fit the data for regression ones [80]. Supervised machine learning is one of the most classic algorithm types; however, not all real-world problems have definite labels (e.g., financial fraud).

Unsupervised machine learning. Data without labels is the fundamental feature that distinguishes unsupervised machine learning from the previous scheme. This algorithm type is often used for clustering problems where the main objective is to group data points with the same characteristics for dimension reduction purposes to concentrate the key features.

Semi-supervised machine learning. Being a hybrid of the two aforementioned machine learning types, labelled and unlabeled data are provided. Since non-labelled data is widely available, semi-supervised machine learning enhances flexibility and accuracy in fields where labelled data is scarce.

Reinforcement learning. Establishing itself as a central pillar of machine learning, reinforcement learning focuses on an environment-based approach. Reinforcement learning allows the agent and the machine to efficiently interact through rewards and penalties to guide the algorithm [81].

As society evolves and data flows through multiple channels, the enormous amount of data from various streams provides an excellent opportunity for further analysis to support the machine learning goal. Multimodal machine learning, therefore, was introduced to utilise the unprecedented stream of diverse inputs.

2.2.2 Multimodal Machine Learning: Opportunities and Challenges

The world is designed for humans to experience with different senses. Seeing, hearing, touching, smelling, and tasting contribute to our perception of life. Fascinatingly, while these senses are often connected to generate a complete understanding of something, even its fragmented pieces can infer the whole picture. For instance, we know a rose when we see one, but the smell of a rose can also distinguish it from other flowers; when we hear the sound "meow", we know it is a cat without even looking, as well as when we touch its fur or claws. Different ways of experiencing are forms of *modalities*. The concept of integrating multiple facets of a subject transcends human natural senses. This idea has been investigated and adopted in machine learning, now known as multimodal machine learning.

Multimodal machine learning is a machine learning stream involving multiple data streams with unique features. While information from a single modality can be minor or discreet, it can be valuable in a multimodal setting where we could study the complementary effects of each modality and obtain a robust final decision. Multimodal machine learning aims to build models to process and combine information from multiple modalities. Each modality in the multimodal model will handle a different kind of data, such as tabular data (e.g., demographic) and multimedia data (e.g., text, audio, and image). Although multimodal machine learning is not a traditional concept, it has been strongly researched and applied in the field due to its potential. With multimodal machine learning advancements in autonomous vehicles [82] and multi-sensory healthcare monitoring [83], this stream of study has shown its potential.

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. That can shorten the prediction time and increase the robustness of the model. As this scheme works with multi-structure and multi-source data, it provides opportunities for studying and incorporating the correlations among modalities, which could highlight an enriched connection among the data points and reveal meaningful insights.

In its primary nature, multimodal machine learning involves combining different data sources in varied formats. Due to the heterogeneity of multiple data streams, however, there remain major challenges for any projects that take on this machine learning camp. [84] identified

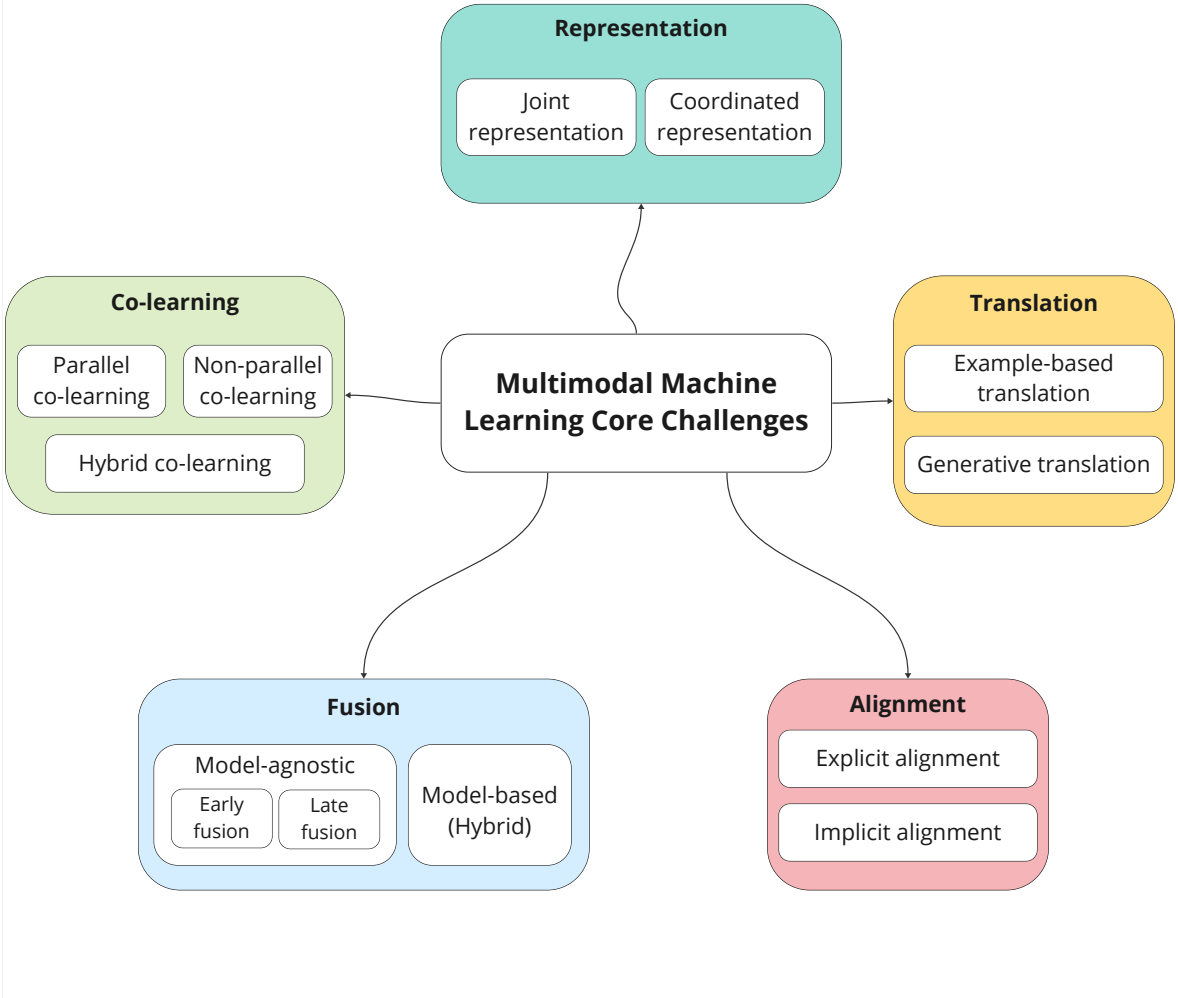

Figure 2.1: Multimodal machine learning core challenges

five main challenges of multimodal machine learning: representation, fusion, alignment, translation, and co-learning. Figure 2.1 summarises the core challenges of multimodal machine learning.

(1) Representation

As data for machine learning problems could come from various sources in various formats, representation for multimodal learning is a crucial task in enabling the utilisation of ubiquitous multimodal data. Often, visual modality is formatted as pixels, audio modality is represented as signals and linguistic modality is symbolic with texts and meanings. Representation in multimodal machine learning is the first task to transform modalities into an informative and workable representation to later be combined with other modalities. The area of unimodal representation has received enormous attention for decades with the rise of high-impact AI/ML conferences for representation learning such as NIPS [1], ICML [2], and ICLR [3]. To determine what characterises a good representation, [85] identifies 10 traits, which include smoothness, sparsity, and coherence (i.e., temporal and spatial), among other qualities.

In terms of modality-specific representations, a great body of literature has been built to reflect the contemporary movements within this field of study. With the advances in the computer vision field, most images are represented as the learned output of convolutional neural networks (CNN) [86]. For audio input, recently, Mel Frequency Cepstral Coefficients (MFCCs) have been used in many studies to represent audio samples [87]. For the text domain, apart from the traditional bag-of-words, n-grammes, and count vectorisers, transformer feature families such as Bidirectional Encoder Representations from Transformers (BERTs) are well utilised to represent textual data [88].

The field of multimodal representation, however, is underexplored. [84] attempts to classify representations for multimodalities into two groups. Figure 2.2 illustrates these types of multimodal representation.

Joint representations are achieved by applying one or more functions to unimodal representations to project them on a shared space. Equation 2.1 explains joint representations in mathematical terms. The function or collection of functions for joining representations is flexible; they can range from concatenation to deep neural networks, probabilistic models, and recurrent neural networks.

$$x_m = f(m_1, m_2, \ldots, m_n) \tag{2.1}$$

[1] Neural Information Processing Systems
[2] International Conference on Machine Learning
[3] International Conference on Learning Representations

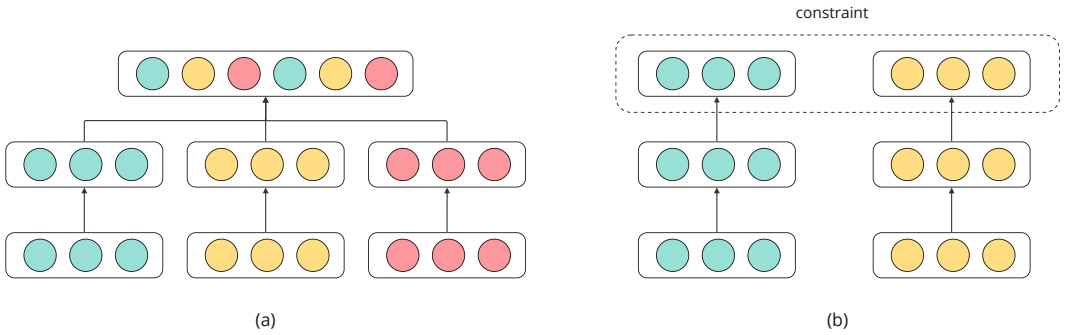

Figure 2.2: Overview of multimodal representation: (a) Joint representation and (b) Coordinated representation

where:

- x_m: Multimodal representation
- $m_1, m_1, \ldots, m_n$: Single representations from n modalities
- $f()$: Function that uses unimodal representations as inputs to produce a multimodal representation

One popular example of constructing joint multimodal representations is using neural networks. Since neural networks are hypothesised to generate abstract features from inputs [85], they are commonly employed as an intermediate or final layer of the data representation process. Unimodal data are fed through layers of neural networks before being used for the final objective function (e.g., classification and regression), which indicates a close relationship between multimodal representation and fusion. Joint representations are suitable for cases where all modalities must be present for the process as they are projected in a common space. This type of representation also allows researchers to work with two or more data streams simultaneously.

Coordinated representations are independent projections of each unimodal on their own space. Equation 2.2 expressed this concept mathematically in an example of two modalities.

$$f(m_1) \sim g(m_2) \tag{2.2}$$

where:

- $m_1, m_1, \ldots, m_n$: Single representations from n modalities
- $f()$ and $g()$: Projection functions for each modality

The main difference between coordinated and joint representations is the composition of the working space. While joint representations combine single representations onto a unified space, coordinated representations maintain each modality in its own independent but coordinated space via some constraints [89]. Based on the constraint types, coordinated representations can be grouped into cross-modal similarity and cross-modal correlation-based. *Cross-modal similarity* methods minimise the distance between similar semantics and maximise that between dissimilar semantics or objects. For instance, the goal would be for the representation of the word "dog" to have closer proximity to an image of a dog compared to that of a non-animal image. Several widely used constraints for similarity models include Euclidean distance [90], cross-modal ranking, and the visual-semantic embedding model (DeViSe) [91]. While cross-modal similarity models focus on the inter-modality similarity distance, *cross-modal correlation-based* models aim to maximise the correlation among modalities via learning.

(2) Translation

Translation in a multimodal machine learning context means translating data from one modality to another or finding the equivalent (i.e., coordinated) form of an entity across modalities. For example, provided an image of an object or event is provided, a parallel version of this in the linguistic domain, such as image captioning, could be helpful for various applications. Although all multimodal machine learning problems require translation among modalities, multimodal translation is the core idea of image captioning, video description, and speech synthesis. Generally, multimodal translation is categorised into example-based and generative models [84]. Figure 2.3 demonstrates these types of multimodal translation.

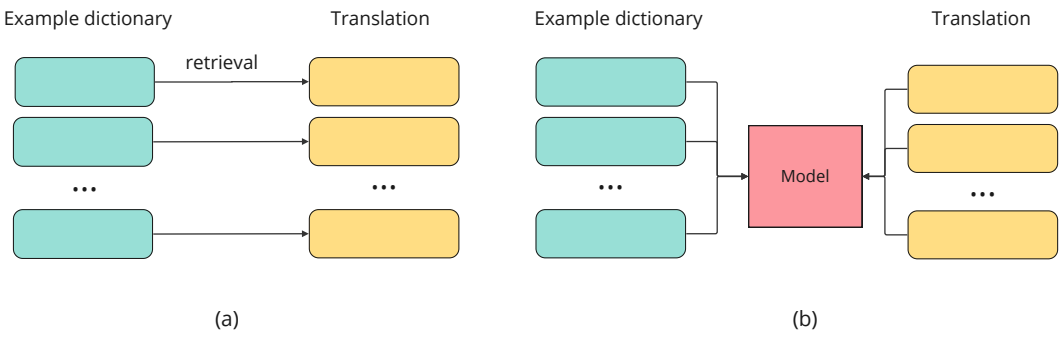

Figure 2.3: Overview of multimodal translation: (a) Example-based translation, and (b) Generative translation

Example-based translation generates its output via a dictionary. While this method is straightforward to implement and has been the foundation of early multimodal translation works in visual speech synthesis [92] and image-to-text description [93], it is inherently limited due to the dictionary size. The example-based translation is often one-sided because only the unimodal representation from the side of the translation input is required. To improve the performance of this method, some studies promote the concept of coordinated semantic space, where there are representations of both sides of the translation [94]. The bi-directional quality of the semantic space allows translation to be more accurate compared to models using direct retrieval.

Generative translation constructs models that learn to produce the translation without direct data retrieval. This technique does not require preparation for a dictionary; therefore, it is less restricted and potentially produces a more flexible output than its counterpart, which is an example-based method. However, evaluating generative models is challenging due to the sheer number of acceptable possibilities. Distinctive camps of generative translation are template-based, continuous, and machine translation-influenced. *Template-based* approaches predetermine a set of template structures for the models to fill in the blanks with corresponding information and return an output in the desired format of the destination modality. For linguistic-related translation, this technique is also referred to as "grammar-based" translation since the input and output involve sentence formation. Generative models have been seen in a variety of translations, such as visual abstraction (i.e., text-to-image synthesis) [95]. While template-based models provide a formula for the model to adapt to, the creativity of these models is substantially limited. *Continuous translation* learns from existing examples to continue the translation. More commonly, this method is utilised in temporal translation [96].

(3) Alignment

Multimodal alignment concerns the cross-modal correlation among modalities to align the content of one modality with another. Specifically, alignment is the task of matching corresponding elements of the same event in all modalities. The demand for alignment problems emerges from synchronising instruction video captions step-by-step, aligning the movie to the original script, and retrieving video segments based on text cues. Multimodal alignment requires models to recognise similarities between modalities while maintaining long-range dependencies. The likelihood of multiple optimal outputs poses a challenge for data annotation and ground truth selection. Based on the two main groups of tasks, multimodal alignment can be classified into explicit and implicit alignment. Figure 2.4 illustrates the two alignment types.

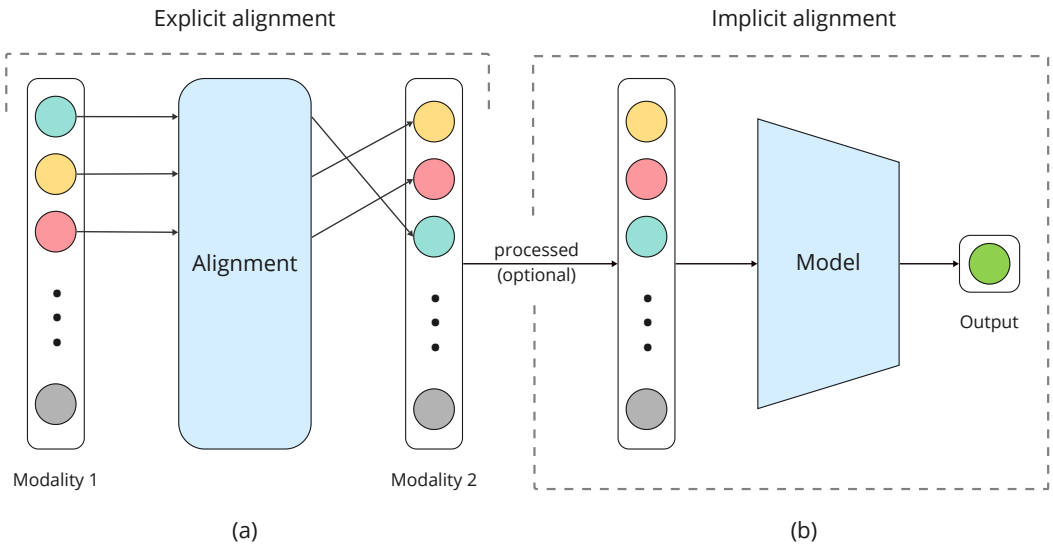

Figure 2.4: Overview of multimodal alignment: (a) Explicit alignment and (b) Implicit alignment

Explicit alignment refers to the direct mapping of elements from at least two modalities regarding the same instance. Both supervised and unsupervised algorithms could be applied for explicit alignment. The early days of explicit alignment featured mostly *unsupervised* techniques using graphical models and dynamic programming [97]. Under *supervised learning*, given the increasing availability of predefined aligned datasets, deep learning methods are deployed widely to align modalities. For instance, in the previous example, [98] used aligned movie scenes with their original books. [99] applied a combination of CNN and long-short-term memory (LSTM) to generate a detailed description (i.e., referring expression) that applies to a section of the photo.

Implicit alignment serves as the intermediate step for other downstream tasks. This precursor learns a latent representation of the alignment during initial training steps to improve performance in image captioning, transcription matching, and speech recognition. [84] suggests that applying implicit alignment before multimodal translation could benefit generative autoencoder models. The attention mechanism is incorporated into many models as it targets local areas that are useful for tasks such as visual-textual question answering using images [100] and videos [101].

In general, while multimodal alignment is facing a lack of specialised datasets due to difficulties in dataset annotation and ground truth selection, this area of study could solve cross-modality alignment and facilitate other multimodal processes as an intermediate step.

(4) Fusion

Joining information from two or more modalities to predict an outcome measure has been one of the original focuses of multimodal machine learning. Fusion is one of the most studied areas within multimodal machine learning in various applications, including affective computing [102], and semantic image segmentation [103].

We suggest a few aspects that require thorough consideration to address multimodal fusion, including the fusion elements, level of fusion, and fusion methods.

- *What to fuse?* For multimodal fusion to accomplish complementary effects, it often requires some treatments with single modalities. Due to the heterogeneity of multiple data streams, investigating the suitable feature sets for each modality and determining the optimal number of modality-specific features are some of the key issues for the pre-fusion stage. Representation, therefore, is tightly related to fusion. Also, since different modalities are captured in different formats and at different rates, multimodal alignment is often considered to synchronise corresponding elements of the modalities before single modalities are ready for multimodal fusion [104].

- *When to fuse?* A critical strategy for multimodal fusion is the level of fusion we must decide. Conventionally, there are (1) *feature-level* fusion (i.e., early fusion) that fuses information prior to model training, (2) *decision-level* fusion (i.e., late fusion) that combines the outputs from different modalities, and (3) *hybrid fusion*, which is a hybrid of the two methods. Each technique has its own advantages and disadvantages, depending on the application.

- *How to fuse?* Given the level of fusion, a variety of techniques can be used at the fusion stage. The following paragraphs will discuss these techniques in detail.

While there are various ways to form a multimodal discussion, we will follow the categorisation by model relation. Two types of multimodal fusion are (1) Model-agnostic fusion (including early and late fusion) and (2) Model-based fusion. The following section describes multimodal fusion techniques.

Model-agnostic approaches do not directly rely on the architecture of a specific machine learning model. This group of strategies includes early fusion (i.e., feature-level) and late fusion (i.e., decision-level) because they both concern non-model aspects of fusion. Figure 2.5 illustrates the mechanism of multimodal model-agnostic fusion techniques.

Early fusion. The most frequently adapted model-free fusion is early fusion. Early fusion methods gained popularity for a few reasons. Firstly, fusion at an initial stage gives an

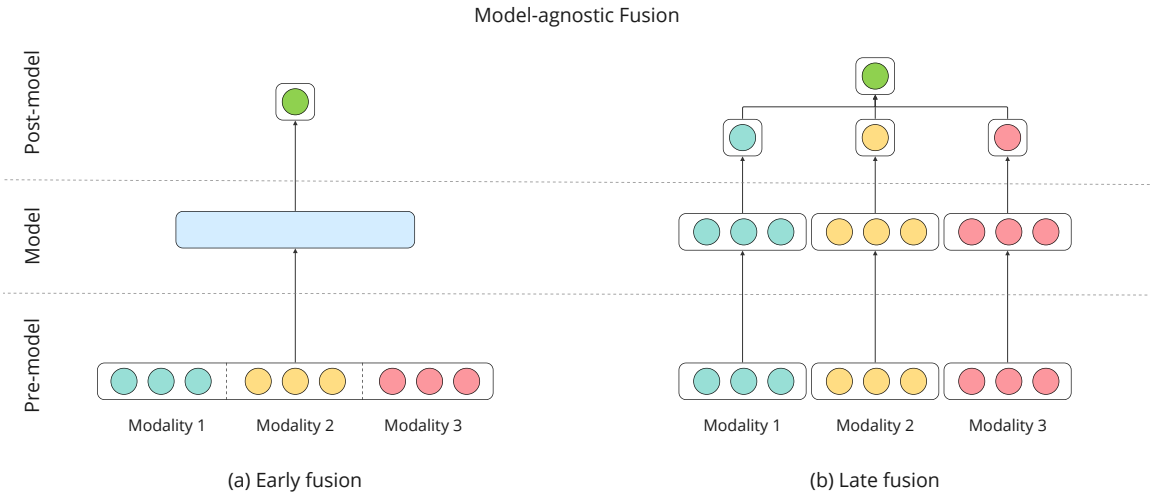

Figure 2.5: Overview of multimodal model-based fusion: (a) Early fusion, (b) Late fusion

opportunity for low-level features from modalities to interact and potentially provide cross-modality correlation. Secondly, since the fusion is performed prior to model training, only one model is required for fine-tuning. However, modality-specific features need to be represented in a similar format prior to fusion, which could put a toll on the multimodal representation step [104]. In terms of techniques, concatenation and weighted linear network [84] are frequently-visited techniques in early fusion.

Late fusion. On the contrary, late fusion integrates the outcomes of modality-wise predictions after model training. Late fusion allows for more flexibility and impacts training each modality separately. Also, it is suggested that decision-level fusion offers scalability regarding the number of modalities involved [84]. There are, however, some disadvantages to late fusion. Since this approach requires more than one model to be trained, the learning process of multiple models could be computationally expensive and time-consuming. Another drawback of late fusion is the possible missed opportunity for feature-level correlation among single modalities. Widely used techniques to fuse different output streams include soft- and hard-voting and learnt models. Specifically, soft voting refers to the weighted sum of predicted probabilities from single modality models, while hard voting often refers to majority voting.

Model-based. In terms of the fusion level, this group of methods is model-level. In contrast to model-agnostic approaches, model-based fusion incorporates the fusing step into the architecture of the models, which allows for more involvement during the fusion process. In the next paragraphs, we discuss popular models used in hybrid fusion, which are classic methods, graphical models, and neural networks. Figure 2.6 depicts the general mechanism of multimodal model-based fusion techniques.

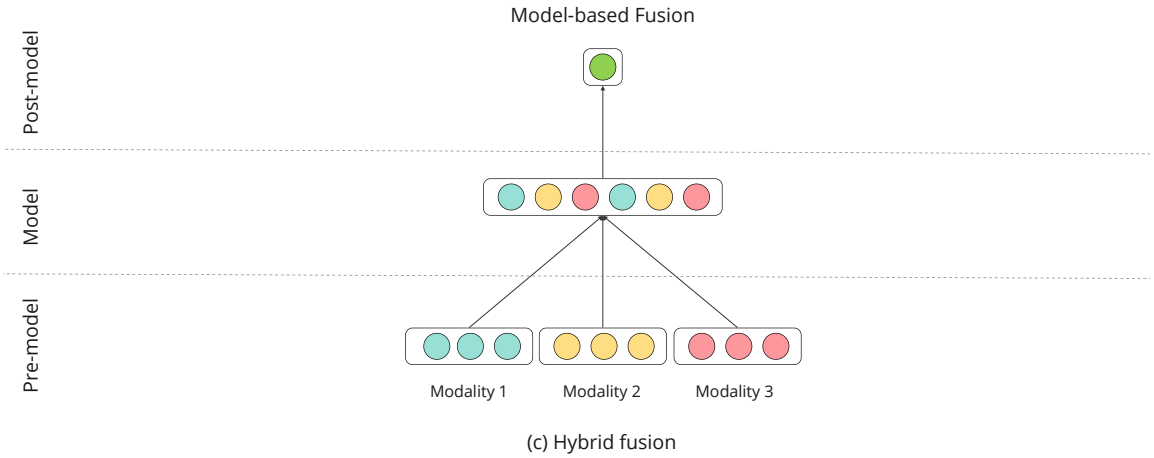

Figure 2.6: Overview of multimodal model-based fusion

Classic methods such as Support Vector Machine (SVM), have been widely used in a variety of problems, including multimodal affective computing [105], human activity recognition [106], artwork recognition [107], and medical screenings such as Alzheimer's detection [108]. These classic methods are variations of multiple kernel learning (MKL). Benefits of MKL include flexible kernel choice and ease of implementation thanks to the rise of advanced libraries.

Graphical models have received more attention for multimodal fusion in recent years. In the healthcare domain, [109] reports promising results on depression detection using graph attention fusion. [110] recognises speech emotion via audio and text using a hierarchical model. Graphical models are suitable for temporal data modelling tasks [84]. Further, the connection of entities within graph-based fusion offers possibilities for better cross-modality correlation and interpretability.

Neural networks, both shallow and deep, are deployed in the fusion stage of multimodal machine learning. The approach has been seen in news detection [111], audio-speech recognition for emotion recognition [112], medical condition diagnosis [113], and surveillance tracking [114] among other topics. Given its learning ability, the neural network fusion approach could learn a large amount of data and unveil complex patterns. Building an end-to-end model is another benefit of selecting neural network-based fusion over other options. This approach, however, struggles with interpretability and explainability due to its abstract nodes and layers. Another point to consider when adopting this technique is the multimodal corpus size since this data-hungry method could pose a challenge in the data collection and preparation.

(5) Co-learning

Co-learning is the challenge of knowledge transferring from one modality to another. Data availability and quality could be issues for multimodal machine learning, which inevitably leads to cases where the data is missing or noisy. To enhance the performance of multimodal models, co-learning is crucial for modalities to complement one another in such limited circumstances. This issue, however, can only be compensated during the training stage and is not relevant for the testing phase. Empirical evidence has shown that models with multimodal co-learning perform better than those without [115]. [116] proposes a holistic taxonomy on multimodal co-learning that discusses aspects ranging from the presence of modality, noisy modality, interpretability, and fairness. This section discusses co-learning in terms of data parallelism - a major slice of the topic. Figure 2.7 demonstrates different types of multimodal co-learning based on the data parallelism objective.

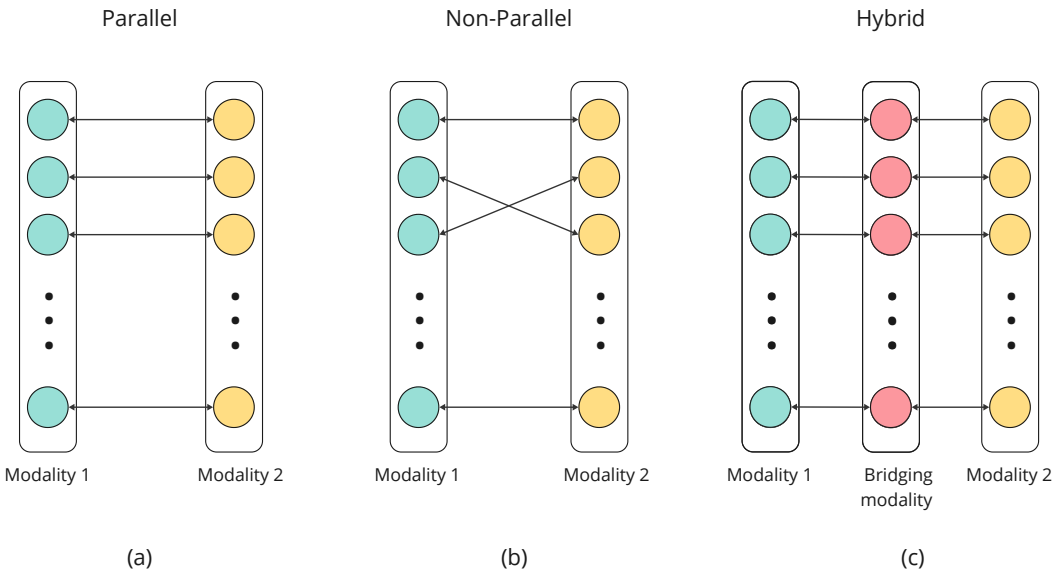

Figure 2.7: Overview of multimodal co-learning: (a) Parallel, (b) Non-parallel, and (c) Hybrid co-learning

Parallel co-learning are for strongly paired data, such as aligned audio, video, and transcripts. [84] introduces co-training and transfer learning for coordinated data. Co-training mainly refers to label management to generate labelled data and remove unreliable samples. For instance, [117] applies co-training to leverage learned information for proactive human-robot assembly. [118] uses visual modality (i.e., lip sync visual data and expression) to support audio modality. A second method for paired co-learning is transfer learning. In this method, the knowledge from a source domain is transferred to a target domain that is different but related [119]. Transfer learning has been adopted to share information across modalities for Covid-19 detection [120] and mental condition diagnosis [121]. The core idea

of transfer learning is for the model to successfully capture the transferable feature to facilitate learning in the second domain. One risk of this method is negative transfer, where signals from the source domain negatively affect learning in the target modality.

Non-parallel co-learning. Non-parallel data is weakly paired, often less expensive to collect and prepare than strongly paired data. Non-parallel data could also benefit from transfer learning. Conceptual grounding and zero-shot learning (ZSL) are unique techniques applied when it requires co-learning for non-parallel data. Conceptual grounding refers to forming semantic concepts based on other modalities apart from textual data [122]. This method is usually achieved with coordinated representation in the multimodal representation section. Several studies have used this method for non-parallel data [123] and reported good performance. ZSL is the task in which the model is required to classify non-observed concepts in prior training. In multimodal ZSL, the two modalities can support each other when the idea is seen in either. For instance, [124] introduces a ZSL model that recognises unseen classes in visual modality through information from unsupervised text corpora.

Hybrid co-learning constructs a shared modality to bridge between the original modalities. Bridging modalities offer more flexible co-learning where annotated data is limited or unavailable. This approach has been practised in domains such as machine translation [125] and bridged transliteration systems [126] where a shared language space is available.

Like its umbrella research topic, machine learning, multimodal machine learning has developed to a level where researchers are testing this concept for more high-risk applications. Given the myriad patient-related data, multimodal machine learning approaches have been used for various medical fields, including mental disorders [127] and cognitive impairments [128]. Despite the increasing popularity of multimodal machine learning techniques in the medical context, the overview picture of an end-to-end pipeline for multimodal machine learning approaches is rarely discussed, mainly when integrated datasets are involved. The following section will discuss the states of automated medical screening and examples of multimodal machine learning in this application.

2.3 Autoencoders

The concept of autoencoders was initially introduced in the 1980s as a multilayer network with a hidden layer [129]. These artificial neural networks use nonlinear transformations to reconstruct input data. Unlike supervised learning methods, autoencoders are unsupervised and not trained on labels. They use input data as the target output, making them an effective unsupervised learning method. The primary goal of autoencoders is to encode input into a meaningful latent representation and then decode that representation to accurately reconstruct the input data. The minimal example of an autoencoder is illustrated in Figure 2.8.

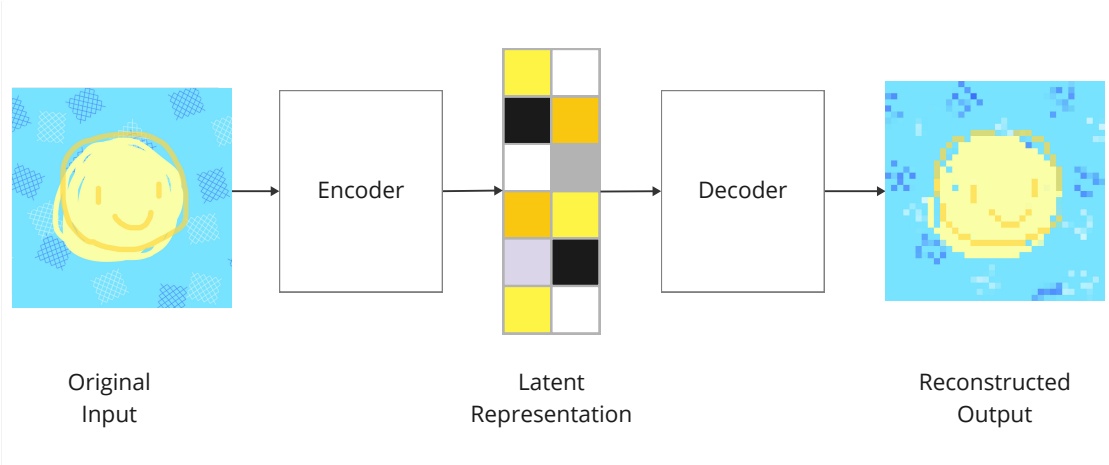

Figure 2.8: Simplified architecture of Autoencoder

2.3.1 Components of Autoencoders

Autoencoders consist of three primary components. The encoder learns the characteristics of the input data and compresses it into a latent representation using non-linear mapping. This latent representation, also known as the hidden layer, is the result of the encoder. Autoencoders generally come in two types, determined by the size of their latent representation: overcomplete and undercomplete autoencoders. Overcomplete autoencoders have hidden layers with more nodes than the input size, and they are used to generate new features from the input data. In contrast, undercomplete autoencoders have a hidden layer that is typically smaller than the input size. They are popular in the literature because of their dimensional reduction property. The decoder rebuilds the input from the latent representation to create a reconstructed input that closely resembles the original input. Figures 2.10 and 2.9 illustrate example architectures of overcomplete and undercomplete autoencoders, respectively.

When implementing autoencoders, the decoder is always present in the model. However, in representation learning, the focus is not on the output of the decoder but rather on its latent

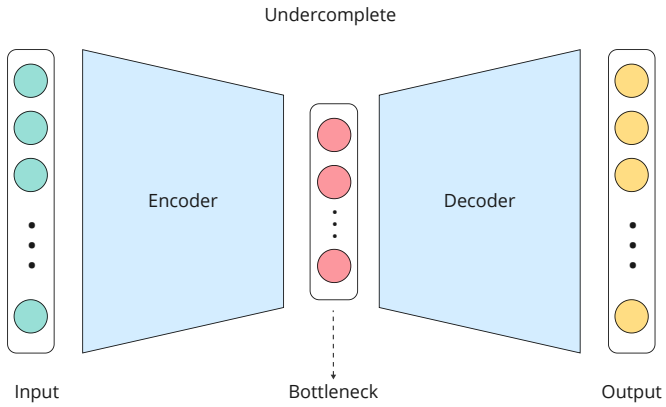

Figure 2.9: Undercomplete Autoencoder

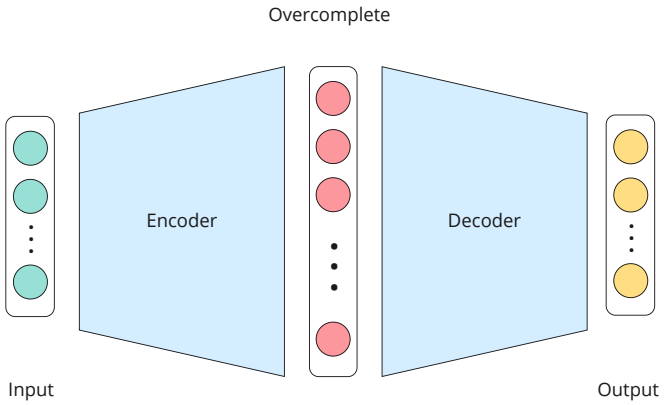

Figure 2.10: Overcomplete Autoencoder

representation. Typically, both the encoder and decoder are neural networks. Autoencoders can be trained end-to-end or gradually by adding layers; in the latter case, it makes them deep autoencoders.

2.3.2 Applications of Autoencoders

Historically, autoencoders were developed for nonlinear dimensionality reduction, e.g., an extension of principal component analysis (PCA) [130]. As time goes by, autoencoders are implemented to serve multiple purposes wherein various models of autoencoders emerge. Further, autoencoders have been employed to pre-train neural networks to provide better initialisation of parameters, hence, improving the main training process [131]. Starting as a form of representation learning, the role of autoencoders has expanded over time. This section discusses some applications of autoencoders.

(1) Dimensionality Reduction

One common use of autoencoders is to reduce the dimensionality of input data. The goal of dimensionality reduction is to learn a lower dimension manifold or the *intrinsic dimensionality space*. Dimensionality reduction, in general, can be divided into linear and non-linear types. For linear techniques, PCA and Linear Discriminant Analysis (LDA) are some prominent methods.

When handling complex data sources, however, non-linear techniques such as autoencoders may often achieve superior results. Undercomplete autoencoders are often used for this purpose, where the latent layer is designed to compress the input data into a smaller format while retaining important features. By limiting the number of nodes in the hidden layer to be smaller than the input layer, we can achieve this compression effect. [132] reports that autoencoder-based dimensionality reduction differs from its alternatives. Autoencoders can even detect repetitive structures in data, which can be useful in various applications.

(2) Denoising Data

Autoencoders have a valuable application in denoising input data. Although standard autoencoders also possess some denoising capabilities due to their selective extraction impact, denoising autoencoders exhibit a more pronounced effect. These autoencoders are a regularisation option that reconstructs a clean input version [133]. During the operation of a denoising autoencoder, the input data is partially corrupted by noises to encourage the autoencoder to learn important features of the input data.

Researchers have developed a novel technique to handle missing data using a specialized denoising autoencoder [134]. According to the study, the proposed autoencoder is more successful in predicting missing data than other reconstruction methods, such as PCA. Speech enhancement is another example of the application of denoising autoencoders. A study evaluated the use of denoising autoencoders based on noise reduction, speech distortion, and perceptual evaluation of speech quality (PESQ) and confirmed that increasing the depth of the autoencoder improves its performance [135]. Besides speech, the deep autoencoder approach has been used for image enhancement. For instance, [136] tailors a stacked-sparse denoising autoencoder to enhance images in natural low-light settings and using hardware-degraded equipment. Figure 2.11 illustrates the structure of a denoising autoencoder.

(3) Multimodal Machine Learning.

In the field of multimodal machine learning, autoencoders are utilised for various purposes, such as representation, fusion, and co-learning. Due to the limited availability of labelled

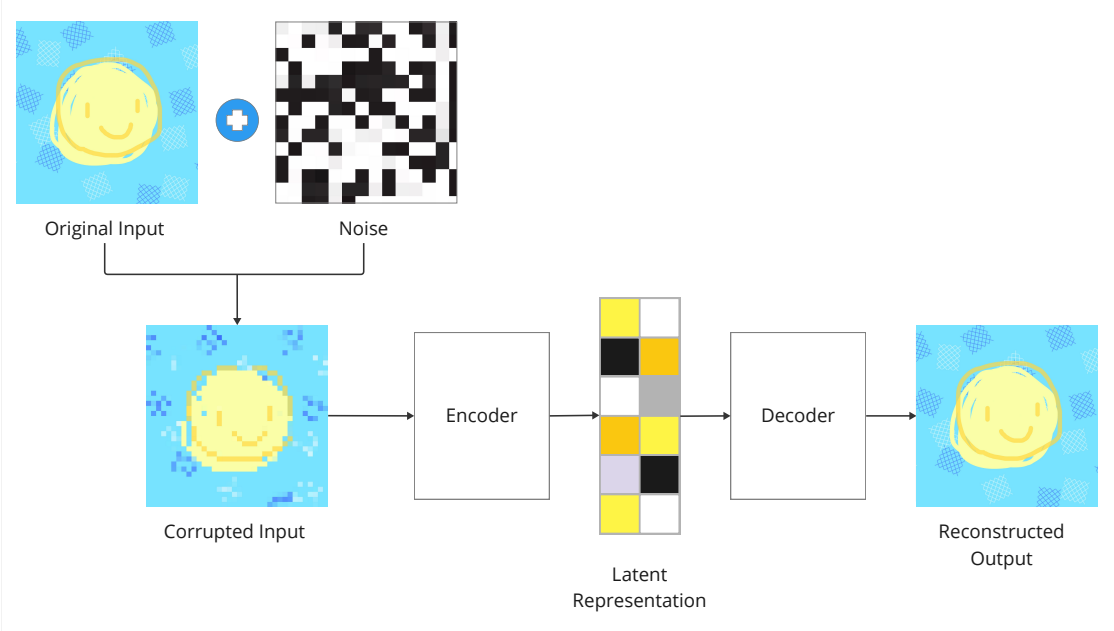

Figure 2.11: Example of a denoising autoencoder

multimodal data, autoencoders are often employed for pre-training representations [84]. Pre-training representations are considered one of the fundamental applications of autoencoders in multimodal machine learning.

Furthermore, it has been found that autoencoders are capable of capturing correlations between different modalities. This discovery has greatly benefited multimodal learning in various ways. For instance, some researchers have integrated autoencoders into their networks to reconstruct all modalities based on one modality. In fact, [137] have gone even further and used stacked denoising autoencoders to reconstruct both video and audio modalities even when only audio data is available. This study highlights two key advantages of using autoencoders for multimodal data. Firstly, they can identify the correlations between different modalities. Secondly, they provide a robust model for reconstructing missing data.

In the realm of co-learning with parallel data, transfer learning can be a useful method, according to a study by [84]. Another promising solution is using multimodal autoencoders to transfer information between modalities. However, despite their potential, autoencoders are not yet widely utilised in the field of multimodal machine learning. This area of study has the potential to bridge cross-modality gaps and improve collaborative learning among multimodal data.

In short, an autoencoder comprises an encoder and a decoder in which the former learns characteristics of the input data and compresses it into a latent representation via a non-linear mapping while the latter reconstructs the input only using that compact code through another transformation. Autoencoder is a widely-interested research topic and has shown effectiveness in various applications, such as denoising, multimodal fusion, and co-learning.

Chapter 3

Multimodal Machine Learning in Mental Disorder Detection: A Scoping Review

3.1 Introduction

It is estimated that one in ten people is vulnerable to mental health problems, with depression, anxiety, and bipolar disorders being the most common [18]. Unfortunately, many cases of mental illness remain undetected due to social stigma and limited access to resources. Given the ongoing and high-risk nature of these disorders, it is crucial to develop accessible methods that can be easily integrated into patients' daily routines for regular attention and monitoring.

There has been extensive research on the use of technology in healthcare decision-making, including the role of artificial intelligence (AI). With the advancement of mobile technologies, unobtrusive sensors, and the ability to collect data from unconventional sources such as social media behaviour, phone calls, and wearable sensors, AI is becoming increasingly useful in detecting symptoms of mental disorders [16]. This has the potential to address issues of conventional mental disorder diagnostic methods, which are often subjective and have a high rate of false diagnoses [138]. By using AI-based measures, this process can become more objective and explainable. Furthermore, integrating mobile applications with AI mental disorder detection solutions can provide rapid screening to overcome the delayed diagnosis caused by the shortage of healthcare clinics [139].

Over the past ten years, there has been a rise in the literature pertaining to automated mental disorder detection. Research has unveiled the ways in which patients interact in various senses and how certain visual factors, like reduced emotional expressivity and fidgety eye

movements, can be indicative of depression [140]. Additionally, auditory modality features such as shortened speech and longer pauses have been used successfully to identify those with depression [141]. 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 that could overcome limitations of the unimodal-based prediction and enhance the joint analysis of multimodalities.

In this study, we take a systematic approach to examine the potential of multimodal machine learning (MMML) for detecting mental disorders. We focus on prevalent mental disorders such as depression, stress, and bipolar disorders. Additionally, we aim to identify current trends and future directions in the field.

3.2 Background

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 [142]. 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 [143]. 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. This section provides some background on mental disorders and MMML.

Mental disorders are conditions that may cause significant disruption in an individual's cognition, emotion regulation, or behaviour, indicating a dysfunction in the psychological, biological, or developmental processes underlying mental functions [144]. These disorders are a major contributor to the global health burden, particularly in light of the COVID-19 pandemic, which has exacerbated the consequences of mental illness [142]. Depression, stress, and bipolar disorders are among the most common categories of mental disorders. Early detection and diagnosis are critical in mitigating the harmful effects of mental illness, as they enable patients to receive prompt and effective treatment based on their symptoms [143]. Machine learning has made significant progress in recent years, offering promising possibilities for fast and scalable analysis of complex data and providing opportunities to aid in the diagnosis and treatment of mental disorders.

Mental health monitoring technologies are increasingly developed. Data from social networks, smartphones, wearables, and neuroimaging can be easily collected thanks to the development of sensor technologies, connectivity, etc. Mining this large amount of data will

be extremely helpful in detecting mental disorders.

The potential of AI in healthcare is vast, with applications in speech recognition, computer vision, and natural language processing [145]. Multimodal machine learning is an exciting area of research that aims to combine information from multiple sources, including text, audio, and image, to make more accurate predictions. This approach has already shown great promise in the field of mental health, where early detection and diagnosis are critical for effective treatment. By analysing data from multiple modalities, multimodal machine learning can provide a more comprehensive understanding of an individual's mental health, leading to better outcomes and improved quality of life [84]. This paper explores the latest research in multimodal approaches for the intelligent detection of mental disorders.

3.3 Methodology

This section describes the search strategy of our survey paper.

3.3.1 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, and ACM Multimedia. The list of keywords used for the search query is as follows:

- 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*

3.3.2 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. A summary of search results will be provided in the next section.

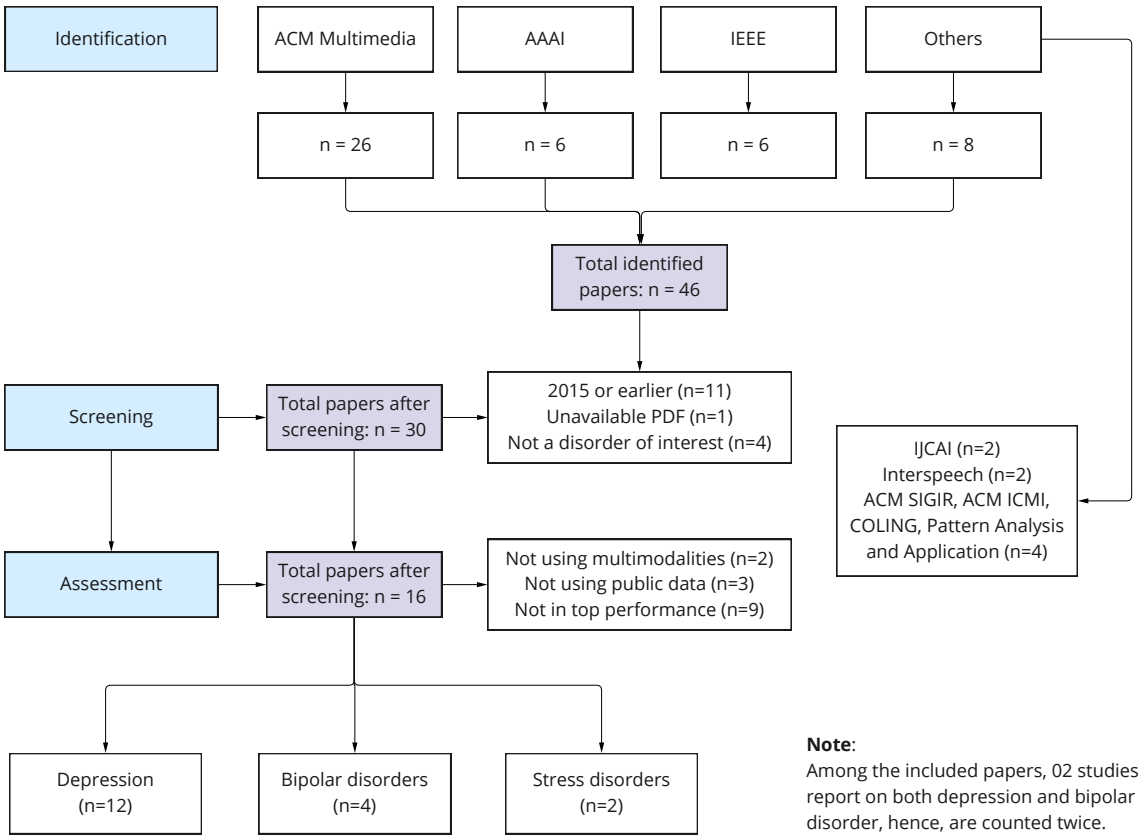

Figure 3.1: Search results

3.4 Results

After the search, we identified 16 papers that match the scope of this paper. Figure 3.1 summarises the filter process and its results. This section analyses the search results regarding datasets, performance and MMML approaches.

3.4.1 Datasets and Performance

Dataset and performance summaries are included in Table 3.1 and 3.2, respectively. Of the common mental disorders covered by this survey, depression had the largest number of articles and datasets used in experiments and evaluation.

Depression. In 12 articles related to depression, 4 datasets were used, including DAIC-WOZ [41], E-DAIC [146], Twitter Depression Dataset [16], and Well-being [147]. 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 ques-

Mental disorders	Dataset	Modality		Train (size)	Dev (size)	Test (size)	Imbalance rate in train (%)
		Classic	Others				
Depression	E-DAIC [146]	A V T		163 $\approx 60\%$	56 $\approx 20\%$	56 $\approx 20\%$	22.70
	DAIC-WOZ [41]	A V T	physiological data	107 $\approx 55\%$	35 $\approx 20\%$	47 $\approx 25\%$	28.04
	Twitter depression [16]	T I		2243 $\approx 80\%$	561 $\approx 20\%$	-	50
	Well-being [147]	A V T		24 $\approx 70\%$	11 $\approx 30\%$	-	50
Stress disorders	MuSE [148]	A V T I	physiological data	27 $\approx 95\%$	1 $\approx 5\%$	-	-
	Ulm-TSST [149]	A V T	ECG, RESP and BPM signals	41 $\approx 60\%$	14 $\approx 20\%$	14 $\approx 20\%$	-
Bipolar disorders	BDC [150]	A V		104 $\approx 50\%$	60 $\approx 25\%$	54 $\approx 25\%$	39.42

Table 3.1: Summary of mental disorders datasets. A, V, T, and I denote the use of audio, video, text, and image modality respectively.

tionnaires 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, [151] and [109] show the best performance on the E-DAIC and DAIC-WOZ, respectively, for regression task, while [152] achieves the highest F1-score on Twitter depression dataset for classification.

Stress Disorders. Regarding stress disorder detection, three articles have been published. Two of these, [153] and [154], utilised the MuSE dataset [148], which was specifically designed for stress detection and its relationship to human emotion. The labels in the MuSE dataset were assigned as self-report annotations by the participants. The third article conducted experiments on the Ulm-TSST dataset (Muse-Stress sub-challenge of MuSe 2021) [149], which were labelled by three annotators using the RAWW method. In terms of stress disorders, [154] achieved the state-of-the-art result with an F1-score of 89.3% on the MuSE dataset.

Bipolar Disorders. In the AVEC2018, four articles were identified that utilized the Bipolar Disorder Corpus (BDC) [150] to detect bipolar disorder states and Young Mania Rating Scale (YMRS) scores. The annotations were performed by psychiatrists. According to [17], their approach showed the best results in terms of Unweighted Average Recall (UAR) on the BDC dataset, outperforming the others.

Dataset	Disease	Paper	Performance
E-DAIC	Depression	[113]	RMSE: 5.22†
		[155]	RMSE: 4.48†
		[151]	RMSE **4.28**†
DAIC-WOZ	Depression	[156]	RMSE: 4.99†
		[157]	RMSE: 4.81†
		[158]	RMSE: 4.27†
		[109]	RMSE: **3.28**†
Twitter Depression	Depression	[153]	F1: 84.2%§
		[16]	F1: 85.0%§
		[159]	F1: 90.0%§
		[152]	F1: **91.2%**§
Well-being	Depression	[17]	F1: 87%§
Ulm-TSST	Stress disorders	[86]	CCC: 0.66†
MuSE	Stress disorders	[153]	F1: 84.9%§
		[154]	**F1: 89.3%**§
BDC	Bipolar disorders	[160]	UAR: 61.65%§
		[155]	UAR: 70.90%§
		[161]	UAR: 72.09%§
		[17]	UAR: **88.36%**§

Table 3.2: 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.

The majority of the datasets have a relatively limited number of samples, with only one dataset containing over 1000 samples in the training set. Apart from the datasets employed in the Audio/Visual Emotion Challenge and Workshop (AVEC), such as E-DAIC, DAIC-WOZ, BDC and Ulm-TSST, the other datasets are assessed through the implementation of cross-validation methods, namely k-fold cross-validation or leave-one-out cross-validation. Therefore, there is no actual test set for those datasets, and the size of the training and development sets is estimated.

3.4.2 Comparative Analysis of Unimodal and Multimodal Approaches

In order to capture the performance gap resulting from the use of additional modalities, we shall select the top-performing model of a single modality as the unimodal baseline. This will be compared against the proposed multimodal approach in each paper. If available, an average shall be applied to the performance of the development and test sets.

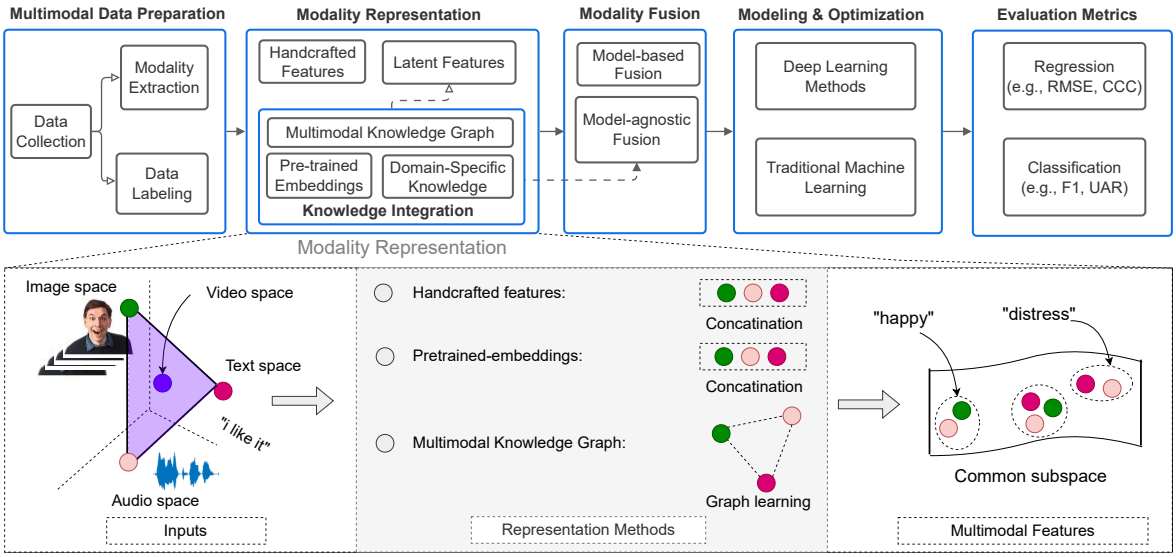

Figure 3.2: Multimodal machine learning pipeline for detection of mental disorders

Of the 16 papers we reviewed, 15 papers compared their multimodal approach to an unimodal method. The results indicate that in all 15 papers, the multimodal approach enhances performance. On average, there is an improvement of 7.9% in F1 (8 out of 15 papers) and 5% in UAR (3 out of 15 papers). Notably, on the BDC dataset, multimodal solutions consistently outperform unimodal base models by a significant 19.7% (F1) [155] and 12.17% (UAR) [161]. In the depression recognition task, a hierarchical recall model that uses audio, text, and video achieves a 12% increase in F1 compared to its textual method [156] on DAIC-WOZ. In general, multimodal approaches have shown that they can significantly improve outcomes compared to methods using individual modalities.

3.5 Discussion

We propose an end-to-end multimodal pipeline for mental disorder detection in Figure 3.2. The following sections will provide a detailed discussion of each of the five stages.

3.5.1 Multimodal Data Preprocessing

Data collection. When it comes to data collection, it's worth noting that while it's not always necessary, data mining can be a viable option if there is suitable secondary data available. That said, primary data collection is generally required to build feature sets from scratch.

Data labelling. To collect ground truths, the most widely adopted options are: (1) clinical assessment in [150], (2) self-assessment in [146, 41] and (3) self-interpreted (e.g., social behaviour data mining) in [16].

When selecting a label source, it's important to note that while clinical assessment labels are the most reliable, the other methods are more popular due to their convenience. Resources and desired sample size should be considered when making this decision. Additionally, it's recommended to define the primary task clearly, as different objectives can limit the choice of suitable model architectures. In some cases, conversion from regression to classification is possible when labels are given thresholds to become categorical data. This approach allows some studies to report their performance in both tasks [155, 156]. For instance, in DAIC-WOZ, a threshold of 10 is used to classify depressed individuals, in addition to the original PHQ-8 scores ranging from 0 to 24.

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

- **Audio**: OpenSMILE and COVAREP are mainstream toolkits in most studies for extracting auditory features such as MFCC, GeMAPs, pitch and voice segmentation [162, 161]
- **Text**: a variety of extraction methods are in place for textual modality, including audio transcription using speech recognition [161, 155] and topic-related, semantic or handcrafted data [16]
- **Vision**: To extract facial expressions and eye movements from video and images using toolkits, such as OpenFace [156, 158] and Face++ [161].

3.5.2 Representation

Representation for multimodal learning is a crucial task in enabling the utilization of ubiquitous multimodal data. Here we review different methods for 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 [160]). This approach involves extracting features from each unimodal data and then combining them into a single feature vector. The main advantage of this method is that each mode can operate independently to uncover significant information. However, there are a few drawbacks to this approach. First, the fused feature vector may be too large for certain machine learning algorithms to handle. Second, the fused feature vector may not be distinctive enough to differentiate between different modes.

Joint feature learning method concurrently learns features from all unimodal data. Using the feature-level concatenation method for representing multimodal data is a highly effective

approach as it allows for features to be learned jointly from all modes, which can then exploit the unique information present in each mode. Many deep learning-based methods utilise this approach, such as those discussed in [151]. However, it can be more challenging to optimise the model due to the higher number of trained parameters and greater data requirements associated with this approach.

Graph-based representation is getting more attention recently. When constructing multimodal data, a multimodal graph can be used where each unimodal data is represented by a node and their interactions by edges. This approach is advantageous in terms of learning capabilities as shown in [109], which demonstrated its ability to model interplay between modalities by learning meta-paths across them. However, it is worth noting that this approach can be computationally expensive for graph construction and also requires a lot of data.

3.5.3 Knowledge Integration

The modelling of multiple modalities only manifests the local knowledge, which is strongly dependent on the training data. To reveal underlying patterns in a more generic way, it would be informative to integrate global knowledge. This becomes even more significant when investigating mental health problems, as experts or prior knowledge can play a crucial role. Generally, there are several trends in integrating such knowledge, including:

Pre-trained embeddings. Exploiting the semantic relationship of textual modalities, such as words or tokens, is the primary purpose. A prominent selection is the BERT model, which is based on Transformers. For instance, the BERT-Base variant with 768-dimensional hidden states is used as the shared text encoder in [153] and [162], while the BERT-large model is considered in [88] to extract contextual information from transcribed transcripts in the E-DAIC corpus. Recently, [155] applied doc2vec to infer the document fixed-length representation of transcribed text. Several pre-trained models, including BERT, RoBERTa, and XLNet, are investigated in [152], along with the BART summarisation model [163], to summarise a large number of tweets associated with each user. However, pre-trained embeddings must be fine-tuned to cover more textual features in the complex nature of data, such as social posts.

Domain-specific knowledge. Extracting informative signals that aid in the detection of mental disorders is crucial, and the BERT model provides a reliable foundation for this purpose. In the case of depression detection, features specific to the domain, such as depression symptoms and antidepressants, have been explored to identify users who may be suffering from this condition. For example, studies like [152] and [155] have analysed Twitter tweets, using nine depression symptoms from the DSM-IV criteria and antidepressant medicine names

from Wikipedia to diagnose depression. Although this approach carries meaningful factors, it is dependent on the domain and can be costly to build respective vocabularies.

Multimodal knowledge graph. There is an interesting trend to exploit knowledge graph [109]. Instead of building from an 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.

3.5.4 Modality Fusion

Owing to the heterogeneity of many input streams, one of the primary focuses of multimodal machine learning was merging information from two or more modalities to predict an outcome measure.

Model-agnostic Fusion

Model-agnostic or model-free techniques, which may be divided into early (feature-level) and late (decision-level) fusion, are those that do not directly depend on the architecture of a particular machine learning model.

Early fusion. 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 [162], and bipolar disorders classification [155, 160]. A variation of concatenation in early fusion is weighted average [113].

Taking a different approach to feature-level fusion, researchers have found that incorporating the correlation among modalities can enhance the overall performance of the architecture. For example, [164] employs multidimensional projection fusion using group sparse canonical correlation analysis on EEG and eye movement. This approach maximizes the cross-correlation between two input streams and achieves the highest accuracy in anxiety detection. [151] proposes a multi-layer attention fusion technique that captures the focus of input features and achieves the winning performance on Extended DAIC.

Late fusion. On the other hand, decision-based fusion or late fusion integrates the outcomes of modality-wise predictions. Late fusion is commonly employed in numerous kinds of research through various fusing mechanisms, including simple voting [154] that allows a drastic improvement of 33.9% accuracy compared to early fusion technique in the same study, winner-take-all voting [157], as well as learned classifiers such as LSTM fusion classifier [86] and simple feed-forward neural network [17].

Model-based Fusion

A recent and innovative approach to depression diagnosis combines deep learning techniques, specifically TCN and Knowledge graph. This method has achieved state-of-the-art performance with an F1 score of 95.4% for the classification task, as well as RMSE and MAE values of 3.28 and 2.62, respectively, for the regression task [109].

Researchers have attempted to classify different levels of bipolar disorders using a tree-based method that involves layer-by-layer fusion through a hierarchical fusion structure, as described in [161]. Although this approach resulted in a 7.4% improvement in the UAR compared to the audiovisual baseline with an ordinary fusion method on the development set, the model, unfortunately, suffers from overfitting, causing a significant drop in performance on the test set.

Depression detection has been shown to be more effective using model-based fusion with LSTM, which resulted in an 11% improvement in F1 [156]. This approach allows for the training of multimodal representation and fusion simultaneously and is thought to be the reason for its success. However, despite its promising results, this technique is not widely used compared to other model-independent methods, indicating the potential for further exploration in this area.

3.5.5 Modeling & Optimization

In order to achieve the main goal, which is the classification or regression that has been predetermined during the preparation stage, different backbone models are utilised to assist with the learning of multimodal data representations. Depending on the primary task (i.e., classification or regression), there are two main streams of modelling & optimisation, which are *traditional machine learning* and *deep learning*.

Traditional machine learning. During the exploration phase of detecting mental disorders, various tree-based methods have been proposed. For instance, in [157, 17], three random forest models were introduced for the three different modalities, which are audio, video, and text. Additionally, in [161], a hierarchical recall model was created based on a gradient-boosting decision tree (GBDT). Furthermore, in [158], several regression models were considered, including random forest (RF), stochastic gradient descent (SGD), and support vector regression (SVR).

Deep learning. These methods effectively leverage underlying latent relationships among different modalities for the mental problem detection task. They provide flexible support for various fusion strategies. Deep neural networks are employed for early fusion in [159, 155], while [162] propose a Transformer-based model architecture. Convolutional neural networks are exploited in [160], and recurrent neural networks (RNNs) are used in [113, 153].

Attention-based models in [151, 109] enable cross-modality exploration via model-based fusion. For late fusion, RNN-based methods are overwhelmingly used in [154, 86]. Overall, using deep learning methods provides a promising approach to modelling multimodal data for detecting mental disorders.

When optimising models for a particular task, it's important to consider the appropriate parameter settings. In the case of regression, minimising mean square errors is typically preferred, as noted in [157]. Conversely, for classification tasks, the cross-entropy objective function is often the go-to choice, as pointed out in [109].

3.5.6 Evaluation Metrics

Evaluation methods play a crucial role in summarizing a system's performance across various tasks and comparing different systems on the same task. In the field of mental disorder detection, different evaluation metrics are utilised depending on the corresponding task. For regression tasks, most studies employ root-mean-square-error (RMSE) and Concordance Correlation Coefficient (CCC) [113, 155, 109]. Notably, CCC is frequently used in emotion recognition tasks as it quantifies the degree of similarity between the predicted emotion and the genuine emotion. On the other hand, F1 and Unweighted Average Recall (UAR) are typically used in classification tasks. For instance, [155, 17] utilize UAR due to the highly imbalanced sample class ratio of the BDC dataset, which is 39.42%, as shown in Table 3.1.

We have discussed common evaluation metrics across all surveyed papers. Although there are no perfect evaluation metrics, we observed some important insights regarding this problem as follows. 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.

3.6 Conclusion and Future Directions

Our scoping review delves into the diverse range of multimodal machine learning approaches to detect mental disorders from high-impact venues. To further promote the advancement of multimodal artificial intelligence for mental health, we propose an end-to-end pipeline for multimodal machine learning tasks. Our review includes a comprehensive list of popular datasets and modalities, and we outline a roadmap for implementing an end-to-end multimodal pipeline for the investigated task.

Although the use of multiple modalities is promising, there is still much to be explored in this field. Based on our review, we suggest that future research should focus on harmonising different input modalities, which remains one of the most challenging tasks in multimodal machine learning. Looking ahead, we believe that developments in this area could lead to significant breakthroughs in the detection and treatment of mental health disorders. Future directions would be:

1. **Multidimensional fusion** is the next focus of the fusion technique. With the increased complexity in representation and the number of modalities involved, high-dimensional fusion is expected to be a part of an enhanced information integration process. Additionally, it is desirable to capture the correlation among modalities to better assist classifiers instead of relying solely on simple concatenation.

2. **Co-learning**. It is anticipated that this area of study will become more popular in future studies, particularly in mental health datasets with small sample sizes that often involve self-reported metrics. This technique is especially suitable when resources are limited, and there is a high chance of incomplete data. By transferring knowledge from one domain to another, this technique can potentially enhance all modalities.

As the evolution of technologies for human interaction continues, we can capture more modalities and gain a clearer understanding of the mechanisms behind mental disorders. This can help alleviate the burden of these conditions at an earlier stage.

Chapter 4

Autofusion - Multimodal Machine Learning in Dementia Detection

4.1 Introduction

On a worldwide scale, around 55 million people suffer from dementia, which is increasing by 10% annually [66]. Accounting for 60-80% of dementia cases, Alzheimer's disease (AD) is a degenerative cognitive condition that not only interferes with the patient's daily activities, work routine, and social interaction, but it can result in immobility and organ failure in later stages [64]. Since AD becomes more prevalent with age, prompt identification is critical [165]. Many AD instances, however, have not been recognised until later stages. For starters, early indicators of Alzheimer's disease may be misconstrued as the natural ageing process. Second, the diagnostic process entails multiple clinical tests that need substantial knowledge, making the operation more costly [64]. Due to the adverse impact of AD on its patients' memory, mental cognition, and daily routine, timely detection is crucial to appropriate treatment and intervention. As a result, there is an urgent need for alternative approaches to supplement established diagnostic procedures.

Artificial Intelligence (AI) has been used for a variety of medicinal reasons. Various AI-based studies have successfully attempted to identify mental diseases such as depression [17] and bipolar disorders [160]. Similarly, earlier research has reported on the efficacy of AI and machine learning approaches in the identification of Alzheimer's disease using speech [166], and medical visual data such as brain scans [78]. Because Alzheimer's disease patients interpret information differently than healthy people, there is increased interest in information integration to examine the compounding influence of multimodalities. Furthermore, provided the availability of continuous multi-aspect monitoring, current advancements in non-invasive sensor technologies have opened the road for multimodal machine learning

for AD identification. Multimodal machine learning has the potential to give AD patients increased accessibility and functionality.

In recent years, there has been a growing interest in the early detection of Alzheimer's disease (AD) due to the realisation that it may be necessary to detect AD pathology decades before a clinical diagnosis of dementia is made. While certain biomarkers are accurate diagnostic methods for AD, there is a need for alternative tools that are less invasive and more cost-effective for AD screening and diagnostics. With the proliferation of technologies that enable personal health monitoring in daily life, there is a possibility of developing tools to predict AD based on the processing of behavioural signals. This active research area has been applied to medical screening to provide a rapid and effective diagnosis method [77, 167]. Due to the heterogeneity of data streams, the multimodal data fusion mechanism has been one of the main issues with multimodal machine learning. Existing approaches, however, often opt for traditional multimodal fusion techniques that neglect the inter- and intra-modal interactions among modalities.

To address this issue, this paper proposes Autofusion, a multimodal machine learning network to detect AD based on text and audio data by incorporating an autoencoder-based fusion technique. The first element of the network is an autoencoder that captures the reconstruction of crucial information on one dimension. In contrast, the second aspect is motivated by the tensor fusion technique to highlight the inter-relationship aspect of multimodal fusion. Extensive experiments conducted on DementiaBank's Pitt Corpus have shown the potential of our proposed method.

4.2 Background

4.2.1 Alzheimer's Disease

Dementia is a mental disorder with the primary characteristics of progressive deterioration of cognitive functions. According to the World Health Organisation, dementia is the seventh leading cause of death for people and one of the primary reasons for disability and dependency among the elderly [168]. The number of dementia patients is growing rapidly and is projected to be 135 million people by 2050 [169]. While dementia is not the same as normal biological ageing, the risk of dementia for people aged 85 years old or older can reach 30% [170].

Dementia is an umbrella term for the clinical syndrome of memory loss and cognitive decline. Alzheimer's disease is the most common cause of dementia [171]. AD is a neurodegenerative disorder involving gradual brain damage. Clinical symptoms of AD include progressive dementia, confusion, motor skill, and memory loss [172], which could result in

patients being bed-bound, dependent and requiring around-the-clock assistance. The progression of AD can be characterised by three main phases that correspond to their accelerating burden: preclinical AD, mild cognitive impairment (MCI) and Alzheimer's dementia [65]. While there is no current cure for AD, early detection is crucial for effective medical support and timely intervention.

The diagnostic process for AD has become less invasive with the use of diagnostic imaging [173]. Other blood test-related methods have been explored and have shown promising results in small-scale testing environments [174]. Therefore, innovations for AD detection are needed to be highly accurate, less invasive, more accessible, and reasonably priced. High-tech clinical support methods have been experimented with extensively in the past decade and have shown their potential in high-accuracy prediction and shortened diagnosis without being invasive [78]. Machine learning has also proven to be useful in providing a confidence score for each prediction and explaining the top contributing factors for the diagnosis. With the multi-aspect nature of electronic health records, multimodal data can be a useful tool to identify the underlying mechanisms of AD dementia and leverage large-scale data for decision-making.

4.2.2 Multimodal Machine Learning in Dementia Detection

Thanks to recent advancements in medical devices, the collection of medical processes and electronic health records (EHR) has become more extensive and accessible. By leveraging EHR data, it is now possible to identify people at risk of dementia [175]. For example, laboratory tests can measure levels of cognitive decline, vital signs can track changes in overall health, and medications can identify potential side effects of dementia. Longitudinal clinical EHR data can track the progression of AD over time, leading to more timely and accurate diagnosis and treatment for patients.

In recent years, machine learning and deep learning techniques have been utilized to develop automated diagnostic systems for various diseases. These technologies have proven reliable in predicting degenerative conditions [176]. As scientists continue to explore the potential of multimodal machine learning, it is clear that this area of study can play a crucial role in healthcare innovation. By analysing data from multiple streams and medical data, it can help identify patterns and aid in clinical decision-making. Utilising automated detection schemes to analyse EHR data has the potential to improve the quality of life for those with dementia and their families. Multimodal machine learning may be the key to bringing together seemingly disparate perspectives and creating a comprehensive approach to detecting AD.

Dataset	Samples	Variables	Modality		
			Media	Demographic	Health
Pit Corpus	550	17	A T	age, gender, education	disease diagnosed, mms, cdr, hamilton
ADDReSS	156	-	A T	age, gender	disease diagnosed, mms, cdr, hamilton
Carolina Collection	600	-	A T V	age range, gender, occupation, education	disease diagnosed

Table 4.1: Summary of Alzheimer's disease multimodal datasets. A, T, V denote the use of audio, text, and video modality respectively.

(1) Alzheimer's Disease Datasets

Prior to creating any machine learning models, it is necessary to collect and explore data thoroughly. To help with this, Table 4.1 provides a summary of AD datasets with multiple modalities. The table includes the name of each dataset, as well as the number of samples and variables in each one.

DementiaBank's Pitt Corpus

The Pitt corpus from DementiaBank was created as part of a 5-year study for the Alzheimer Research program at the University of Pittsburgh in the 1980s [177]. The corpus contains 551 recorded interviews and transcripts from 292 patients who were evaluated for probable Alzheimer's disease. Its purpose was to investigate the interaction patterns exhibited by patients with dementia and Alzheimer's disease. The study was conducted by the School of Medicine at the University of Pittsburgh.

The study involved three different groups of participants, including an elderly control group, individuals with probable or possible Alzheimer's disease, and individuals diagnosed with other types of dementia. Data collection took place on an annual basis, and participants had to be at least 44 years old with no history of nervous system disorders. Additionally, all participants had to score at least 10 on the initial Mini-Mental State Exam (MMSE) in order to provide consent to join the study. The Pitt corpus captures a range of demographic data such as age, sex, and education status, as well as medical records like neuropsychological exam results and diagnosis. The corpus also includes audio and transcript data from interviews where participants described the Cookie Theft image. Due to its large sample size and longitudinal nature, the Pitt corpus is considered one of the most valuable datasets for multimodal machine learning used in early Alzheimer's disease detection. However, it is important to note that while the size of the corpus is commendable, the data is imbalanced with 309 audio samples from 193 patients with probable Alzheimer's disease and 242 recorded interviews from 99 control individuals.

ADReSS Dataset The ADReSS Dataset (Alzheimer's Disease Recognition through Spontaneous Speech) was introduced in the ADReSS challenge of INTERSPEECH 2020 [178]. This dataset is a balanced subset of DementiaBank's Pitt Corpus and consists of 156 speech samples, each with accompanying metadata such as age, sex, and MMSE score. The challenge baseline is established over a stratified train-test split of around 70-30. The ADReSS Challenge defines two prediction tasks: (1) AD prediction, which involves binary classification of AD and non-AD individuals, and (2) MMSE prediction, which targets a regression problem to predict MMSE score. The dataset's balanced nature and availability of metadata make it a well-suited resource for machine learning model training and evaluation.

Corolina Collection Corpus The Carolinas Conversation Collection (CCC) was collected in the early 2000s as a digital corpus of audio and video recordings of natural conversations between healthcare professionals and patients with chronic diseases, including Alzheimer's disease (AD) [179]. The CCC corpus contains 200 conversations with the non-AD group of patients and 400 conversations for the AD group, each with accompanying demographic and clinical information such as age, sex, occupation, disease, and level of education. The patients, all aged over 65, were interviewed by gerontology and linguistics students or researchers at least twice a year and assigned a unique alias to protect their identity.

There are potential areas of study that can be explored through the CCC, including the effects of AD on communication, how healthcare providers can better support their patients' communication needs, and the creation of new interventions to help those with chronic illnesses improve their communication skills.

(2) Alzheimer's Disease Multimodal Machine Learning Performance

In this section, we summarise the performance of multimodal machine learning on the discussed AD datasets. Table 4.2 provides the performance evaluation of machine learning-based models using the identified datasets for automated AD screening tasks.

Despite multimodal machine learning achieving great performance across datasets: 88% F1 for Pitt Corpus and 92% accuracy for ADReSS dataset [167], and 90% F1 for Carolina Collection Corpus, this group of techniques is still under-explored for AD detection. Even for those multimodal papers whose performances are promising, their fusion techniques are limited at hard voting (i.e., late fusion) [167] and simple concatenation early fusion [185]. Achieving a potential result using model-based fusion, [182] explores joint fusion using a simple feed-forward neural network combining text and image modalities.

In terms of classification/regression models, the majority of papers focus on traditional machine learning techniques, including Support Vector Machine (SVM) [180, 181, 183], Decision Tree [178, 184], linear regression [185]. [167, 182] opt for a designed neural network to classify AD patients.

Dataset	Paper	Modality	Fusion	Model	Performance
Pitt Corpus	[167]	A T	Late	NN	ACC: 88.0% § F1: 88.0% §
	[180]	A	-	SVM	ACC: 77.0% §
	[181]	A	-	SVM	ACC: 86.04% §
ADDReSS	[167]	A T	Late	NN	ACC: 92.0% §
	[178]	T	-	LDA, DT	ACC: 77.0% § F1: 77.0% § RMSE: 4.38 †
	[182]	T I	Join	NN	ACC: 81.0% § F1: 80.0% §
	[183]	T	-	SVM	ACC: 85.4% § RMSE: 4.56 †
Carolina Collection	[184]	A	-	DT	ACC: 86.5% §
	[185]	A	Early	LR	ACC: 90.0% § F1: 90.0% §

Table 4.2: Performance evaluation of machine learning-based models for Alzheimer's disease detection. 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. A, T, and I denote the use of audio, text, and image modality respectively.

Based on the analysis of performance evaluation for AD detection, it is apparent that there is a significant need for automated detection models. The initial results indicate ample opportunities for growth and improvement in this area.

4.3 Approach

This section describes the benchmarking dataset, proposed pipeline, and the multimodal machine learning architecture employed to detect AD through audio and text modalities.

4.3.1 Dataset: DementiaBank's Pitt Dataset

We select DementiaBank's Pitt corpus [177] as our benchmarking dataset. The dataset has audio recordings and text transcripts of an image description task in which the participants are required to describe the activities in the Cookie Theft image. The Pitt corpus is selected for benchmarking since it is one of the largest corpus there is for AD detection tasks. Additionally, due to the gradual degradation of AD, longitudinal studies are impactful; hence, the Pitt corpus is our dataset choice.

To identify labels for the experiment, we selected individuals with a primary diagnosis code of 800 as the control group and those with a diagnostic code of 100 as patients with AD dementia. Diagnostic codes were assigned via clinical assessment. For our experiments, in total, we use 552 samples which include 243 speech samples from 99 control healthy subjects and 309 speech samples from 194 AD subjects. Table 4.3 provides essential statistics of the data.

	Sample	Individual	Sex		Entry age			
			male (1)	female (0)	(45, 55]	(55, 70]	(70, 85]	(85, 100]
Control	243	99	41	58	15	64	20	0
Dementia	309	194	68	126	8	81	100	5
Total	552	293	109	184	23	145	120	5

Table 4.3: Dataset statistics

Of the 293 individuals included in the dataset, their entry age range varies between 46 and 88 years old. It is noteworthy that age data is approximate because (1) the study only records the age upon entry of an individual but not updated age for each participating year, and (2) no date of birth is given. On average, the mean age is 67.4 years old. In terms of participants' genders, 37.2% (109 out of 293) of the participants are male, while 62.8% (184 out of 293) are female.

4.3.2 Benchmarking Baselines

We compare the results of our approach to the benchmarks set out in [167]. This paper was submitted for the INTERSPEECH ADReSS Challenge 2020. Three factors led to the selection of this paper. First, the experiment's source code is made public, allowing for the potential reproducibility of the findings. Second, this article compares results using the DementiaBank Pitt Corpus. Finally, this paper has the highest reported performance on DementiaBank's Pitt Corpus among those meeting the above requirements.

The source code of the paper is accessible on Github [1]. While the article is also assessed using the ADReSS Dataset and DementiaBank's Pitt Corpus, the Pitt corpus is our primary focus. [167] presents a baseline employing late fusion (hard voting) with accuracy, recall, precision, and F1 scores of 0.88, 0.82, 0.92, and 0.88, respectively.

We made an effort to replicate the experiments using the source code; however, the outcome is different from what was originally published. There are a few theories as to why this occurs. Firstly, the random elements used in the experiments, such as the random seed number, are not specified in the source code. Secondly, although both papers perform 10-fold cross-validation on the dataset, there is no provided information on fold index (i.e., sample indexes in each fold), which implies different fold splits and contributes to the performance disparity. Last but not least, while all experiments in the benchmarking study are trained and assessed on the ADReSS Dataset before being retrained on DementiaBank's Pitt Corpus, only the Pitt corpus is used in our paper for training and evaluation. Consequently, this enhances the possibility of a data leak.

4.3.3 Framework

This section describes the proposed framework to detect dementia using audio and transcripts. Regarding model features, the proposed framework employs those engineered by [167] to be valid for benchmarking. The framework is inspired by the potential cross-modality interaction and denoising effects of autoencoders in multimodal machine learning; thus, the input features will be compressed using an autoencoder, and its bottleneck will be used for the primary classification. Figure 4.1 illustrates the detailed architecture of the network.

[1] https://github.com/ wazeerzulfikar/alzheimers-dementia

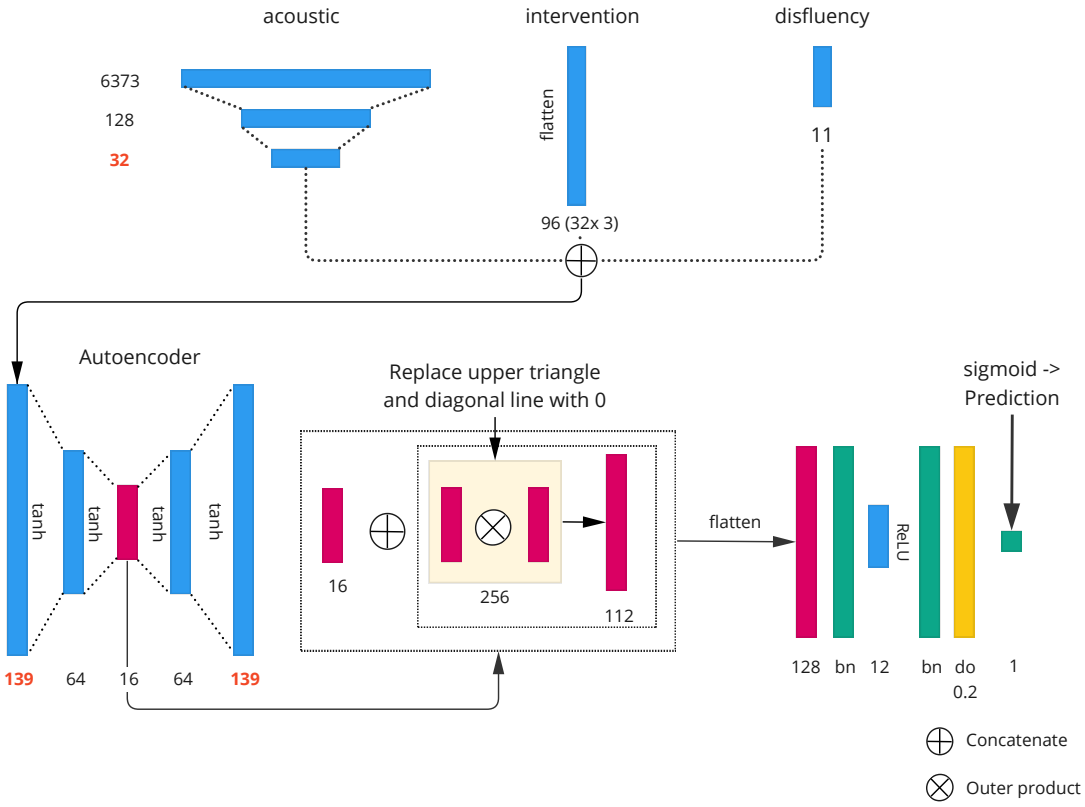

Figure 4.1: Autofusion architecture

Feature Engineering

The processing of the audio and transcript files from the DementiaBank Pitt dataset results in three distinct feature types. This approach is similar to the methodology used in [167] that explored multimodal analysis. Table 4.4 presents general statistics on the generated feature sets. Specifically, the features are as follows:

- *Acoustic*: We applied ComParE 2013 feature set [186] using the OpenSMILE toolkit to extract 6,373 features from audio files. These features include MFCC, low-level descriptors (LLDs), and other statistical features. We apply a linear network to reduce the dimensionalities of acoustic data.
- *Disfluency*: Disfluency is a set of 11 hand-crafted features from the transcript, including word rate and pause rates, to reflect any discontinuity in the speech.
- *Intervention*: The intervention feature set provides more context for the transcript by extracting a sequence of speakers to identify whether it is the interviewer or the participant speaking. The intervention feature set is padded or truncated to a fixed length of 32 with 3 channels of one-hot values (i.e., 1 or 0).

Feature Transformation

As different feature sets have different value ranges, some normalisation operation is conducted to ensure homogeneity in the input scale. Hence, a standard scaler is applied to address the varied ranges. Specifically, we implemented the *StandardScaler* function from the scikit-learn library to achieve this transformation, which standardises features by removing the mean and scaling to unit variance. The formula for the standard scaler is summarised in Equation 4.1.

$$x_{standardised} = \frac{(x - \bar{x})}{s} \tag{4.1}$$

where:

- $x_{standardised}$: Standardised score
- $\bar{x}$: Mean of training samples
- s: Standard deviation of training samples

The mean and standard deviation of the training data are stored to transform the validation set.

Autofusion Mechanism

Autoencoders have been studied for their reconstructive and denoising qualities. Their usage has been practised in the field of multimodal machine learning. The addition of autoencoders can encourage interactions among modalities [84, 137]. We use a modified version of the bottleneck's outer product with itself as the input of the classification network. This component shall be referred to as the cross-modality interaction or the interaction term interchangeably in the subsequent sections. For a detailed discussion on the use of autoencoders in multimodal machine learning, refer to Section 2.3.2.

After feature transformations, three feature sets become inputs of the autofusion network for multimodal fusion. Different from traditional early and late fusion, the autofusion mechanism focuses on studying the interactions among modalities.

First, three feature sets are concatenated and fed through an autoencoder with an undercomplete architecture. The bottleneck of the autofusion network is computed as follows:

Feature	Source	Extraction	Dimension
Acoustic	Audio	ComParE 2013 feature set using OpenSmile toolkit	[1, 6373]
Disfluency	Transcript	Handcrafted features including word rate, pause rates of various kinds, and intervention rate	[1, 11]
Intervention	Transcript	One-hot encoded sequence to indicate the speaker (i.e., interviewer or participant)	[1, 32, 3]

Table 4.4: Statistics of generated features

$$h_0 = x_a \oplus x_d \oplus x_i \tag{4.2}$$

$$h_1 = \sigma\left(W_0 h_0\right) \tag{4.3}$$

$$x_{bottleneck} = \sigma\left(W_i h_i\right) \tag{4.4}$$

where:

- x_a, x_d, x_i: Input feature acoustic, disfluency, and intervention respectively
- $\oplus$: Concatenation
- h: Input
- W: Weight
- σ: Activation function
- $x_{bottleneck}$: Bottleneck vector

Classification Network

The classification network comprises two main components: the autofusion bottleneck and its linearly transformed version. The two inputs are combined using the outer product function, creating the interaction term. The calculation of the interaction term can be found in Equation 4.5.

$$h_{interaction} = x_{bottleneck} \otimes x_{bottleneck} \tag{4.5}$$

Equation 4.6 and 4.7 describe the calculation of the network inputs of the first layer h_0 and the following layers h_l.

$$h_0 = h_{interaction} + x_{bottleneck} \tag{4.6}$$

$$h_l = \sigma\left(W_{(l-1)}\, h_{(l-1)}\right) \tag{4.7}$$

where:

- $x_{bottleneck}$: Bottleneck vector from autofusion
- $h_{interaction}$: Interaction term among feature sets
- $\otimes$: Outer product
- h: Input
- W: Weight
- σ: Activation function
- l: Layer number [0, number of layers]

Objective Function and Evaluation

We used Mean Square Error (MSE) loss to monitor the autoencoder training process and Binary Cross Entropy (BCE) loss for classification. The formulas of MSE and BCE losses are summarised in Equation 4.8 and 4.9, respectively.

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.8}$$

where:

- $\mathcal{L}_{MSE}$: Mean Square Error (MSE) loss
- N: Total number of samples
- y: Label
- $\hat{y}$: Prediction

$$\mathcal{L}_{BCE} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{4.9}$$

where:

- $\mathcal{L}_{BCE}$: Binary Cross Entropy (BCE) loss
- y: Label
- $\hat{y}$: Prediction

To calculate the total loss of the network, we applied the loss formula in Equation 4.10 with a 0.1 loss factor for autoencoder loss and a 0.9 loss factor for classification loss.

$$\mathcal{L}_{total} = (\lambda_{MSE} \times \mathcal{L}_{MSE}) + (\lambda_{BCE} \times \mathcal{L}_{BCE}) \tag{4.10}$$

where:

- $\mathcal{L}_{total}$: Total loss function
- λ_{MSE}: Loss factor of autofusion network ranged [0.0, 1.0]
- λ_{BCE}: Loss factor of classification network ranged [0.0, 1.0]
- $\mathcal{L}_{MSE}$: MSE loss from autoencoder network
- $\mathcal{L}_{BCE}$: BCE loss from classification network

4.4 Experiments

This section discusses experiment settings and results. The first part provides information on computational resources, frameworks, running parameters, and evaluation metrics. The second half of the section presents the results of key experiments and additional ablation study.

4.4.1 Experiment Settings

This section describes the experiment settings, resources and evaluation metrics of the experiments.

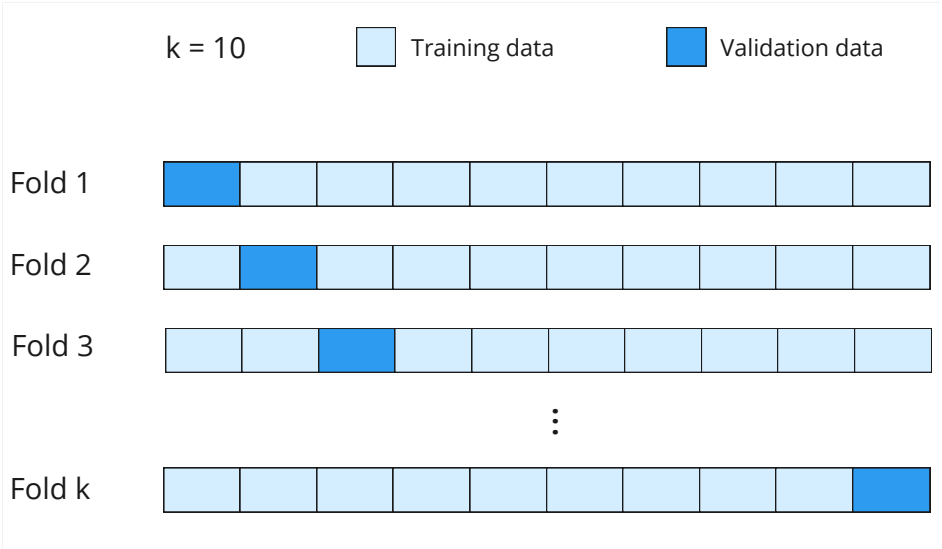

Figure 4.2: Example of k-fold cross-validation

Cross Validation

Cross-validation is a resampling method used to ensure the generalisation of models [187]. In classification problems, overfitting the training data is a major concern, as it can affect the model's ability to make accurate predictions on unseen data. The use of cross-validation addresses this issue by evaluating the model on a test dataset that is drawn from the same population as the training dataset. In our experiments, we utilised *k-fold cross-validation*, where k denotes the number of splits on the dataset. For each k, the dataset is randomly divided into k subsets, with the model being trained on all but one subset, which is used for validation. This approach ensures that no two test sets have overlapping elements, which minimises the risk of data leakage. Figure 4.2 illustrates the case of k-fold cross-validation with $k = 10$. This technique is particularly effective for datasets with limited samples. We applied stratified 10-fold cross-validation to all of our experiments. This technique provides that the validation set reflects the class ratio of the dataset, maintaining a consistent balance between positive and negative samples. Stratified k-fold is a variation of the k-fold technique that reflects the class ratio of the dataset, ensuring fair classification.

Resources

This section describes the computational power used for the experiments and the framework for the experiment setup. Our proposed model is implemented using PyTorch. The detailed settings are as follows:

- Computational Power:
  - Linux Ubuntu 22.04

  - CPU: Intel Core i9 10900X 10 Cores
  - RAM: 256 GB
  - GPU: Nvidia 3090 24 GB VRam
- Framework: PyTorch version 1.13.1

Running Parameters

In our experiments, we ran 300 epochs and utilised the Adam optimizer for all networks. The autoencoder training network was implemented with a $\tanh$ activation, while the classification network used ReLU. Additionally, we applied a learning rate decay mechanism with an initial learning rate of $lr_{init} = 5 \times 10^{-4}$ that decreases by 20% every 100 epochs, as shown in Equation 4.11, to fine-tune training and prevent local minima. This approach has been proven effective in other studies and allowed us to achieve accurate results while avoiding common machine learning issues, like overfitting.

$$lr_e = lr_{init} \times 0.2^{\frac{e}{100}} \tag{4.11}$$

where:

- lr_{init}: Initial learning rate ($= 5 \times 10^{-4}$)
- e: Current epoch
- $\frac{e}{100}$: Integer quotient of current epoch number divide by 100

Details of experiment running parameters are as follows:

- Number of epochs: 300
- Optimizer: Adam
- Number of folds: 10
- Activation: $tanh$ (for autoencoder), $ReLU$ (for classification)
- Learning rate: Initial learning rate 5×10^{-4}, decay 20% every 100 epochs
- Loss factors: 0.1 for MSE loss, 0.9 for classification loss

Evaluation Metrics

To ensure a fair comparison with other papers, we decided on a few evaluation metrics, which are: accuracy, recall, precision, and F1 score.

Accuracy is one of the most popular metrics that have been used in various papers. Equation 4.12 describes the formula for accuracy.

$$Accuracy = \frac{TP + TN}{N} \tag{4.12}$$

where:

- TP: Number of true positive samples
- TN: Number of true negative samples
- N: Total number of samples

When evaluating the performance of a model, accuracy is a commonly used metric that calculates the percentage of correct predictions over the total number of samples. While this measure is easy to understand, it can be problematic when dealing with imbalanced data. If one class dominates the dataset, accuracy can create a bias towards that class, which does not reflect the minority class's true performance. Additionally, accuracy does not account for the certainty or confidence of each prediction, which can be important in certain scenarios.

Recall measures how well the model accurately predicts a positive sample among all positive samples. The formula for recall is summarised in Equation 4.13.

$$Recall = \frac{TP}{TP + FN} \tag{4.13}$$

where:

- TP: Number of true positive samples
- FN: Number of false negative samples

Recall is often used in situations in which spotting positives is the main goal. For example, this metric is useful for fraud detection models since the frequency of successfully identifying fraud is important.

Precision measures the probability that the positive samples are predicted correctly. The formula for precision can be seen in Equation 4.14.

$$Precision = \frac{TP}{TP + FP} \tag{4.14}$$

where:

- TP: Number of true positive samples
- FP: Number of false positive samples

Precision provides insight into positive class; hence, the ratio of true positives and true negatives is highly concerned.

F1 Score is the weighted average of precision and recall. Equation 4.15 calculates F1 score, as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.15}$$

where:

- $precision$: Precision score
- $recall$: Recall score

The advantage of using the F1 score is that it takes into account the distribution of the data since it is the harmony between precision and recall. However, since it is a combined metric, the F1 score could be difficult to interpret. Often F1 score is used in cases where predicting false negatives has a significant impact. For medical screening, falsely identifying that the individuals do not carry the disease when they actually do could cause serious problems due to the possibility of late treatment or disease transmission.

4.4.2 Results

We evaluate unimodal methods and different fusion alternatives on DementiaBank's Pitt dataset. From the beginning of the paper, we aim to investigate and answer the research questions as follows:

- **RQ2a:** How effective is this framework compared to single-modality models?
- **RQ2b:** How effective is this framework compared to existing multimodal fusion techniques?
- **RQ2c:** How does the cross-modality interaction impact the overall performance?

In this section, we answer these research questions using experiment outcomes. Table 4.5 summarises the results of both unimodal and multimodal experiments. These results are reported as the average performance of 10-fold cross-validation.

Type	Experiment	Accuracy	Recall	Precision	F1
Unimodal	Acoustic	72.01 (±0.07)	76.61 (±0.08)	74.33 (±0.06)	75.35 (±0.06)
	Intervention	70.82 (±0.07)	82.26 (±0.15)	74.08 (±0.12)	75.92 (±0.04)
	Disfluency	77.70 (±0.07)	80.58 (±0.07)	80.93 (±0.09)	80.30 (±0.05)
Multimodal	Late fusion	56.87 (±0.03)	60.84 (±0.06)	61.56 (±0.02)	61.09 (±0.04)
	Early fusion	78.07 (±0.06)	81.22 (±0.07)	80.19 (±0.06)	80.53 (±0.05)
	Autofusion	**79.89** (±0.08)	83.85 (±0.07)	**81.72** (±0.09)	**82.47** (±0.06)
	Autofusion (no interaction)	77.91 (±0.08)	**84.15** (±0.06)	78.62 (±0.08)	81.12 (±0.06)

Table 4.5: Experiment results

(1) Comparison with unimodal baselines

We conducted experiments to compare our proposed technique against three single-input models, including acoustic, intervention, and disfluency models. These experiments were carried out using the same network settings as those of unimodal baselines described in [167]. Autofusion achieved a performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. It consistently outperformed all unimodal methods by an average of 5.24%. Specifically, our proposed method surpassed acoustic, intervention, and disfluency unimodal baselines with an average improvement of 7.4%, 6.2%, and 2.1% across all metrics, respectively.

In terms of accuracy, autofusion achieves a higher accuracy of 79.9% compared to the accuracy of 72%, 70.8%, and 77.7% achieved by the acoustic, intervention, and disfluency models correspondingly. Regarding F1 performance, autofusion achieves an F1 score of 82.5%, surpassing the acoustic model by 7.1%, intervention model by 6.6%, and disfluency model by 2.2%. Given the imbalanced nature of the Pitt corpus with 309 AD and 243 control samples, F1 is a more robust and reasonable metric compared to other evaluation metrics in this case.

In addition to the evaluation metrics, the stability of the models is another noteworthy factor. Stability refers to the consistency of a model's performance across different fold splits (i.e., standard deviation across folds). Most experiments report a variation of 0.03 - 0.09% across 10 folds, except for intervention models, which have a variation of 0.15% in recall and 0.12% in precision. Autofusion shows relative robustness compared to other single-input models as it deviates marginally from the unimodal averages.

Given the model performance and stability, this conclusion responds to **RQ2a** and confirms that our proposed autofusion technique outperforms unimodal baselines in all metrics and retains comparable robustness. The result corresponds to our findings in Section 3.4.2, where multimodal machine learning models consistently surpass single-input models in detecting other mental disorders, including depression, stress, and bipolar disorders. This demon-

strates that multimodal machine learning models are a highly effective method for automated medical screenings.

(2) Comparison with other fusion baselines

We contrast the autofusion method against the other fusion alternative methods, which are early and late fusion. For the early fusion technique, three input streams are concatenated and fed into the same classification network architecture as the autofusion technique. Autofusion outperforms early fusion with an average of 2% improvement overall metrics. Specifically, the performance increases from 78.07% to 79.89% in accuracy, from 81.22 to 83.85% in recall, from 80.19% to 81.72% in precision, and from 80.63% to 82.47% in F1 with the addition of autofusion mechanism prior to classification. Regarding the stability of the two fusion models, autofusion shows a greater standard deviation of 0.02% in accuracy, 0.03% in precision, and 0.01% in F1 compared to early fusion. As early fusion and autofusion share a similar classification architecture, the main difference lies in the fusion mechanism of concatenation versus autoencoder-incorporated fusion to capture the interaction effect.

As per the guidelines in [167], we opt for the hard-voting late fusion technique. This mechanism ensembles the mode of base predictions from discreet models. Compared to the late fusion technique, we observe a notable improvement of 20% average across all metrics. Although the three single-input models perform well independently, their combination through late fusion does not work as effectively. This could be due to the fact that late fusion ignores the relationships between modalities and their representations in earlier stages, resulting in a potential disconnect between predictions. In cases where the models are unstable, non-overlapping predictions may occur, resulting in a lower-performing majority vote outcome. Regarding the stability of the two models, the late fusion model establishes a smaller variation across folds than the autofusion model. The standard deviations of the late fusion model vary from 0.02 to 0.06%, while those of autofusion range from 0.06 to 0.09%.

Based on the experiment results, it is evident that the proposed autofusion method outperforms other fusion alternatives, including early and late fusion. This finding responds to **RQ2b** and highlights the promise of Autofusion technique that offers significant improvements compared to conventional fusion methods.

(3) Additional ablation studies

In this part, we discuss additional experiments to confirm the impact of certain components. The inputs of the classification network comprise two elements: (1) the autoencoder latent representation (i.e., bottleneck) and (2) an interaction term which is a form of outer product between the bottleneck and itself. The use of the outer product is inspired by Tensor Fusion Network in which the outer product contributes to the model acquiring both intra-modality and inter-modality dynamics [188]. This section of the thesis aims to address **RQ2c** to understand the impact of the cross-modality interaction term.

When we omit the interaction term, the results reduce by roughly 2% from 79.89% to 77.91% in accuracy, 3% from 81.72% to 78.62% in precision, and 1.4% from 82.47% to 81.12% in F1. One exception is that by removing the interaction term, the model presents a 0.03% increase from 83.85% to 84.15% in recall. In terms of stability, as there is a minimal difference of 0.01% in recall and precision standard deviations between experiments across 10 folds, the two models are similarly stable. While the improvement provided by the interaction term remains modest, it shows potential in assisting the autofusion technique. With further exploration and fine-tuning in this direction, the interaction term may be a breakthrough for autofusion specifically and multimodal fusion in general.

(4) Explainability of results

The disfluency model, one of the three single-modality models employed in detecting Alzheimer's Disease (AD), has been found to demonstrate the highest accuracy and F1 scores, with the former being 77.7% and the latter being 80.3%. This performance is indeed noteworthy, given that it surpasses the performance of the acoustic and intervention models by 3-5%.

The disfluency model is particularly intriguing due to its reliance on a set of 11 handcrafted features derived from the transcript. As elucidated in section 4.3.3, the purpose of this feature set is to reflect any speech discontinuity that may be indicative of AD. These features have been designed to capture speech impediments, such as stuttering and slurring, through attributes such as varying word rates and pause rates.

The empirical results indicate that the disfluency model outperforms the other two models by a considerable margin, suggesting that speech discontinuity has the potential to serve as a valuable predictor of AD. Semantic characteristics have been proved to be potential indicators of other disorders. For instance, research has shown that individuals with schizophrenia exhibit less cohesive speech than healthy individuals [189]. Similarly, speech disfluency has been identified as a promising indicator for identifying individuals with social anxiety disorder, particularly in terms of speech duration, jitter, and shimmer [190], in line with observations made by psychologists.

This finding from the study not only highlights the effectiveness of the disfluency model in detecting AD, but also provides a basis for further research into the relationship between speech patterns and other disorders.

Chapter 5

Discussion

The work investigates the impact of multimodal machine learning on automated medical screenings and introduces a novel multimodal fusion method for identifying mental disorders. To achieve this goal, we conducted a scoping review of high-impact articles on multimodal machine learning in the detection of mental disorders, which is a crucial aspect of automated medical screening. Further, we propose Autofusion, a novel multimodal fusion approach that utilises the cross-modality interaction between audio and transcript data to identify Alzheimer's disease. Autofusion was evaluated and compared against other unimodal models and fusion alternatives. This chapter offers an overview of the previous chapters' discoveries and presents insights into how the findings could propel the field forward.

5.1 The States of Multimodal Machine Learning in Automated Medical Screenings

As we delve into the world of medical screenings, it becomes clear that relying solely on healthcare professionals' judgment is no longer the most efficient approach. With limited human resources and overwhelming demand, patients often wait for extended periods for a diagnosis. This creates an unbridged gap that can be filled with innovative technologies such as machine learning. By utilising these tools, we can enhance the traditional diagnostic methods and provide more effective, time-saving, and cost-efficient diagnoses. The application of technology in general and machine learning in particular in the medical field has yielded remarkable results in identifying a diverse range of medical conditions. From diabetic retinopathy [13], various types of breast cancers [191, 14], respiratory-related conditions [15] to various chronic mental disorders such as depression [192] and stress disorder [86], machine learning has shown significant potential in improving screening accuracy. The

advancements in this technology have proven beneficial in improving patient care and treatment plans, leading to more favourable health outcomes.

In the field of medical diagnosis and treatment, utilising multiple sources of data to gain a deeper understanding of diseases and their progression is becoming increasingly important. By integrating data from various sources, such as medical imaging, electronic health records, and patient-tracking data, multimodal machine learning is a promising approach to providing a more comprehensive and accurate diagnosis. One of the key challenges of this approach is the process of integrating diverse data sources and exploiting their complementary effects, known as multimodal fusion. However, through this approach, with the potential to save time and costs, multimodal machine learning can be employed for better overall outcomes for patients and the medical field.

Mental health is a crucial aspect of healthcare that requires special attention. Mental disorders are highly prevalent conditions that affect a significant portion of the population [18]. Unfortunately, social stigma, limited accessibility, and the hidden nature of mental conditions often lead to misdiagnosis or underdiagnosis. Moreover, mental disorders can contribute to and exacerbate other health issues, such as hypertension and cardiovascular disease [19]. Multimodal machine learning is especially suitable for mental disorder screening, as recent research has revealed the uniqueness of how mental disorder patients process different modalities [20]. This indicates the urgency and potential of multimodal machine learning in mental disorder detection. To understand the states of multimodal machine learning in mental disorder detection, we conduct a scoping review to study articles from high-impact venues for trends and techniques in detecting depression, stress, and bipolar disorder. From the identified papers, we pinpoint several key findings to address ***RQ1: "What are the current methods and fusion techniques to integrate multimodalities (e.g., video, audio, text) in medical screenings, specifically mental disorder detection?"***.

Multimodal datasets for mental disorders. The availability of datasets pertaining to the diagnosis of mental health disorders is significantly constrained, with the vast majority of datasets containing fewer than 300 samples. This can be attributed to the considerable difficulties associated with the diagnosis of mental disorders, as well as the costs incurred in the management, collection, and cleaning of data. Fortunately, there is a positive trend emerging with respect to the use of diverse modalities in datasets apart from classic media input (i.e., audio, video, text, and image) to improve data availability. Specifically, 30% of the identified datasets offer the addition of either medical-related data or behavioural patterns. Furthermore, the increasing use of wearable and tracking devices is an optimistic sign to enhance data collection and facilitate progress in multimodal machine learning for medical screenings.

Impact of the multimodal approach. Multimodal machine learning has emerged as a robust and effective approach in the realm of medical research. Our review is proof of multimodal machine learning potential and its effectiveness in handling medical-related problems. Our review of 16 papers found that multimodal approaches outperformed unimodal methods in 15 cases, with performance differences ranging from 7.9 to 19.7%. Overall, using multimodal approaches significantly improves outcomes compared to individual modalities. With the increasing availability of medical data, this exciting development is certain to lead to more effective multimodal machine learning technology in medical-related problems.

Multimodal machine learning trends. While reviewing high-impact articles, we identify outstanding trends in multimodal machine learning. Regarding the modelling and optimisation step, besides the use of traditional methods such as SVM, and random forest, deep learning techniques are gaining popularity in detecting mental disorders. Convolutional neural networks are exploited in [160], and recurrent neural networks (RNNs) are employed in [154, 153]. Several studies utilise attention-based models to explore cross-modality relationships [151, 109]. While the modelling stage witnesses a wide variety of techniques, it has been observed that traditional fusion methodologies continue to be heavily favoured for multimodal fusion. In fact, more than 75% of existing studies incorporate early and late fusion, which are recognised as model-agnostic techniques. Though these methods can be conveniently implemented, they often fall short of effectively capturing the intricate relationships present between diverse modalities [84]. As such, it may be beneficial to consider exploring novel fusion approaches to achieve more optimal outcomes.

Potential impact of cross-modality interaction. Among challenges regarding multimodal machine learning, cross-modality interaction is an area of study about the inter- and intra-interactions among modalities. Empirical evidence indicates that models with multimodal interactions perform better than those without [115]. Cross-modality interaction is particularly relevant in the realm of mental health research, where data may be incomplete or limited. Despite its potential benefits, this topic has yet to be widely adopted, as none of the identified papers applies cross-modality interaction techniques in their studies. However, as medical-related data is expensive to collect and the data sample size is limited, methods of capturing multimodal interactions, including parallel learning, transfer learning, and autoencoders, are expected to play an influential role.

Implications. The review presents the audience with a scoping review to explore mental disorder detection as a potential area for multimodal machine learning. This scoping review is designed to be an effective first glance into popular multimodal datasets, current state-of-the-art performances, and current states and trends of a subset of automated medical screening, mental disorder detection specifically. This paper, as an entry point, can equip practitioners with the necessary understanding of the field, follow the end-to-end multimodal framework for experiments and motivate them to further investigate.

5.2 Autofusion: A Cross-Modality Interaction Focused Multimodal Fusion Approach

As the world continues to evolve with more data being collected, multimodal machine learning is becoming increasingly popular. However, there are still five main challenges in this area of research, including representation, translation, alignment, co-learning, and fusion. Multimodal fusion is particularly important for combining different modalities, extracting valuable compounding effects, and capturing the correlation among modalities. This information is essential for unlocking the potential of multimodal machine learning and setting it apart from traditional single-modality methods.

Multimodal fusion can be classified into two distinct approaches: model-free and model-based. The former, which includes early and late fusion techniques, is model-agnostic and pertains to non-model aspects of fusion. Early fusion is often preferred due to its ability to enable interaction between low-level features from different modalities, potentially leading to cross-modality correlation. Late fusion integrates the predictions of different modalities after model training, allowing for more flexibility and the capacity to impact the training of each modality separately. However, the implementation of late fusion necessitates training more than one model, which can be computationally expensive and time-consuming. Model-based techniques, in contrast to model-agnostic ones, involve integrating the fusion step into the model's architecture, thereby affording greater control during the fusion process.

Despite its importance, the range of multimodal fusion techniques is highly limited. This research proposes Autofusion, a novel autoencoder-integrated approach to multimodal fusion. Autoencoders have been employed in studies, and in some studies, autoencoders have been used for cross-modality interaction as their selective compressing nature can potentially extract key information and correlation of modalities. This work applies Autofusion in detecting Alzheimer's disease (AD), one of the most prevalent degenerative cognitive disorders. We evaluate the model on DementiaBank's Pittsburgh corpus, which is among some largest multimodal AD datasets that offer audio and transcript samples. Our proposed framework achieves a promising performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. We will discuss key observations and implications of the study in the following sections.

Comparative analysis of unimodal and multimodal approaches. To address ***RQ2a: "How effective is this framework compared to single-modality models?"***, extensive research has demonstrated that multimodal approaches tend to outperform unimodal models. This research confirms that the combination of modalities consistently produces superior results, as evidenced by the superior performance of both early fusion benchmark and Autofusion methods compared to single-modality techniques. Specifically, the Autofusion model can

improve performance by a significant margin of 2.2 to 7.1% relative to various unimodal approaches. Given the vast potential of multimodal machine learning in the domain of automated medical screening, it is probable that this field will witness an increase in research activities in the coming years.

Impact of cross-modality interaction in multimodal fusion. To address ***RQ2b: "How effective is this framework compared to existing multimodal fusion techniques"***, compared to other fusion alternatives, Autofusion outperforms early and late fusion considerably across all evaluation metrics. This finding corresponds to the effectiveness of previous literature on model-based fusion techniques. The interaction effect generated from the autoencoder-infused component of its model is a major contributor to its success. Previous studies have suggested the use of autoencoders portrays the potential impact of cross-modality interaction terms. When used in conjunction with multimodal fusion, autoencoders can further enhance the correlation among participating modalities. To respond to ***RQ2c: "How does the cross-modality interaction impact the overall performance?"***, empirical results show that the interaction term improves the overall performance by 2-3% of the model. Although modest, the improvement provided by the interaction term has the potential to further assist the Autofusion technique in this research. Therefore, continued exploration and refinement in this area could lead to significant advancements in multimodal fusion techniques. One observation on the infusion of autoencoders, however, is that the Autofusion model stability can be improved as the standard deviations across folds exhibit a slight increase compared to early and late fusion.

Implications. The recent research study has substantial implications for machine learning and medical research. The study introduces a new approach to multimodal fusion by incorporating autoencoders, which can be further explored to improve the current state-of-the-art. The network architecture proposed in this study can serve as a valuable reference and benchmark for developing solutions for DementiaBanks's Pitt corpus. Furthermore, the study confirms that automated medical screening using multimodal machine learning can effectively leverage the complementary effect of different data streams. This has the potential to revolutionise chronic disease detection, as the condition is notoriously difficult to diagnose and often takes years to identify. Healthcare professionals may find periodic tracking and task-directed voice recording to be useful decision-making aids, and this research may pave the way for more efficient and effective medical screening in the future.

In summary, this section discusses the findings and implications in key chapters of this work. The following conclusion section will provide a summary of the overall research and its contributions. We will also address the limitations of the research and discuss directions for further study.

Chapter 6

Conclusion

This thesis investigates multimodal machine learning, examines the current state of automated medical screening, and designs an implementation centred on multimodal fusion for mental disorder detection, a critical area of medical screening.

6.1 Summary of Contributions

This thesis contributes in several distinct ways.

Firstly, this thesis presents a scoping review comprised of papers from high-impact venues, including AAAI, IEEE, ACM Multimedia, and JCAI, to study the current states of multimodal machine learning in mental disorder detection and fusion techniques to integrate multimodalities. The scoping review highlights the potential of the multimodal machine learning approach in mental disorder detection specifically and automated medical screening in general. The review includes a comprehensive list of popular datasets and frequently visited modalities for depression, anxiety, and bipolar disorder. We then propose an end-to-end pipeline for multimodal machine learning applications that cover recommendations over essential steps from multimodal data collection and representation to multimodal fusion, modelling, and evaluation. The scoping review highlights the potential yet under-explored use of cross-modality interaction, which is a novel direction for multimodal fusion, one of the core problems of multimodal machine learning.

Secondly, this thesis studies the use of cross-modality interaction in multimodal fusion and validates the concept in the area of mental disorder detection. We study multimodal fusion in more detail by proposing the Autofusion mechanism and evaluating the framework on DementiaBank's Pitt corpus to detect Alzheimer's disease. Autofusion achieves a promising performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. The technique consistently outperforms all unimodal methods by an average of 5.24%

across all metrics. Compared to other fusion alternatives, Autofusion is recorded to perform better than existing fusion methods (i.e., early fusion and late fusion). Especially against the late fusion hard-voting technique, our method shows a significant difference of 20% on average across metrics. Further, we capture the impact of the cross-modality interaction term as its incorporation enhances the model performance by 2-3% across metrics. The proposed approach harnesses the power of the interaction effect as a leveraging element to achieve superior results over unimodal baselines and common fusion alternatives. This research highlights the potential of autoencoder usage in multimodal machine learning and calls for further research to unlock its full potential.

6.2 Limitations and Future Study

Due to resource limitations, several aspects of this thesis require further exploration. Specifically, concerning the scoping review, there are opportunities to expand the scope in various ways.

1. While the proposed pipeline is based on consistent trends identified in previous research papers, it is crucial to validate this framework with empirical results. To achieve this, we intend to conduct further experimental analyses in future studies, where the pipeline can be utilised and benchmarked across different datasets to ensure its validity.

2. There is immense potential in broadening the review to encompass other medical conditions. For example, anxiety-, substance-related, and mood disorders can be included in an extended version. In general, the development of multimodal machine learning has the potential to improve diagnosis times and employ under-utilised medical data in the domain of chronic conditions.

3. The scoping review can serve as a starting point for a systematic review that covers a broader range of papers in addition to existing high-impact venues. The papers can be sourced by query search from reliable corpora such as PubMed and Scorpus, while the selection process will adhere to systematic review standards. Alternatively, incorporating research from psychology and healthcare journals could be a valuable addition to the existing body of research, providing insights into the practical applications of machine learning and technology as observed by healthcare professionals and psychologists.

Further study can also be pursued in various directions with regards to the Autofusion technique.

1. One aspect that could be explored is benchmarking the technique on other datasets to validate it further and gain a better understanding of its behaviours. Potential candidates for this include the ADreSS dataset from INTERSPEECH in 2020 [178] and the ADreSSo dataset from 2021 [193], as they are balanced and cleaned versions of the Pitt corpus, which despite its generous sample size, is imbalanced.

2. Apart from dataset choice, another possible approach for further study is experimenting with a broader selection of modalities. Currently, the Autofusion model uses classic media input streams such as audio and transcript samples. However, medical data such as physiology features and tracking data could also be incorporated into the model to widen the range of processing methods used. This could involve more advanced model architecture, including handling signals and scan visualisation.

In conclusion, this study emphasizes the potential application of multimodal machine learning in the automated medical screening process. With the goal of achieving a resource-efficient and comprehensible healthcare-related diagnosis system, multimodal machine learning, in particular, and technology, in general, would become a promising component to support decision-making in medical screening.

Bibliography

[1] D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera, "Online incremental machine learning platform for big data-driven smart traffic management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4679–4690, 2019.

[2] Y. Fu, A. R. Downey, L. Yuan, T. Zhang, A. Pratt, and Y. Balogun, "Machine learning algorithms for defect detection in metal laser-based additive manufacturing: A review," *Journal of Manufacturing Processes*, vol. 75, pp. 693–710, 2022.

[3] A. de Barcelos Silva, M. M. Gomes, C. A. da Costa, R. da Rosa Righi, J. L. V. Barbosa, G. Pessin, G. De Doncker, and G. Federizzi, "Intelligent personal assistants: A systematic literature review," *Expert Systems with Applications*, vol. 147, p. 113193, 2020.

[4] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2020.

[5] M. Tuczyńska, M. Matthews-Kozanecka, and E. Baum, "Accessibility to non-covid health services in the world during the covid-19 pandemic," *Frontiers in public health*, vol. 9, p. 760795, 2021.

[6] K. Bilton and C. Zaslawski, "Reliability of manual pulse diagnosis methods in traditional east asian medicine: a systematic narrative literature review," *The Journal of Alternative and Complementary Medicine*, vol. 22, no. 8, pp. 599–609, 2016.

[7] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated lung nodule detection and classification using deep learning combined with multiple strategies," *Sensors*, vol. 19, no. 17, p. 3722, 2019.

[8] J. Wang, J. Ji, M. Zhang, J.-W. Lin, G. Zhang, W. Gong, L.-P. Cen, Y. Lu, X. Huang, D. Huang *et al.*, "Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening," *JAMA Network Open*, vol. 4, no. 5, pp. e218 758–e218 758, 2021.

[9] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103792, 2020.

[10] J. A. Wilson, K. Onorati, M. Mishkind, M. A. Reger, and G. A. Gahm, "Soldier attitudes about technology-based approaches to mental health care," *CyberPsychology & Behavior*, vol. 11, no. 6, pp. 767–769, 2008.

[11] D. Ganesh, G. Seshadri, S. Sokkanarayanan, P. Bose, S. Rajan, and M. Sathiyanarayanan, "Automatic health machine for covid-19 and other emergencies," in *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2021, pp. 685–689.

[12] P. J. Bridgeman, M. B. Bridgeman, and J. Barone, "Burnout syndrome among healthcare professionals," *The Bulletin of the American Society of Hospital Pharmacists*, vol. 75, no. 3, pp. 147–152, 2018.

[13] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[14] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, "Explainable machine learning for lung cancer screening models," *Applied Sciences*, vol. 12, no. 4, p. 1926, 2022.

[15] G. Marques, D. Agarwal, and I. de la Torre Díez, "Automated medical diagnosis of covid-19 through efficientnet convolutional neural network," *Applied soft computing*, vol. 96, p. 106691, 2020.

[16] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution." in *IJCAI*, 2017, pp. 3838–3844.

[17] F. Ceccarelli and M. Mahmoud, "Multimodal temporal machine learning for bipolar disorder and depression recognition," *Pattern Analysis and Applications*, pp. 1–12, 2021.

[18] H. Ritchie and M. Roser, "Mental health. our world in data. 2018," 2020.

[19] T. G. Pickering, "Mental stress as a causal factor in the development of hypertension and cardiovascular disease," *Current hypertension reports*, vol. 3, no. 3, p. 249, 2001.

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

[21] W. H. Organization *et al.*, "About mental disorders," World Health Organization. Regional Office for the Eastern Mediterranean, Tech. Rep., 2019.

[22] W. H. Organization, "Depression and other common mental disorders: global health estimates," Technical documents, 2017.

[23] Who, "Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide," 2022.

[24] K. Usher, J. Durkin, and N. Bhullar, "The covid-19 pandemic and mental health impacts," *International journal of mental health nursing*, vol. 29, no. 3, p. 315, 2020.

[25] C. M. Doran and I. Kinchin, "A review of the economic impact of mental illness," *Australian Health Review*, vol. 43, no. 1, pp. 43–48, 2017.

[26] F. Mascayano, J. Toso-Salman, Y. C. S. Ho, S. Dev, T. Tapia, G. Thornicroft, L. J. Cabassa, A. Khenti, J. Sapag, S. J. Bobbili *et al.*, "Including culture in programs to reduce stigma toward people with mental disorders in low-and middle-income countries," *Transcultural psychiatry*, vol. 57, no. 1, pp. 140–160, 2020.

[27] S. Docrat, D. Besada, S. Cleary, E. Daviaud, and C. Lund, "Mental health system costs, resources and constraints in south africa: a national survey," *Health policy and planning*, vol. 34, no. 9, pp. 706–719, 2019.

[28] L. Kola, B. A. Kohrt, C. Hanlon, J. A. Naslund, S. Sikander, M. Balaji, C. Benjet, E. Y. L. Cheung, J. Eaton, P. Gonsalves *et al.*, "Covid-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health," *The Lancet Psychiatry*, vol. 8, no. 6, pp. 535–550, 2021.

[29] W. H. Organization *et al.*, "Depression and other common mental disorders: global health estimates," World Health Organization, Tech. Rep., 2017.

[30] "Depression," https://www.who.int/news-room/fact-sheets/detail/depression, sep 13 2021.

[31] American-Psychiatric-Association *et al.*, "Diagnostic and statistical manual of mental disorders, edn 4 (dsm-iv) washington," *DC, USA: American-Psychiatric-Association*, 1994.

[32] C. Katona, R. Peveler, C. Dowrick, S. Wessely, C. Feinmann, L. Gask, H. Lloyd, A. C. de C Williams, and E. Wager, "Pain symptoms in depression: definition and clinical significance," *Clinical Medicine*, vol. 5, no. 4, p. 390, 2005.

[33] L. Orsolini, R. Latini, M. Pompili, G. Serafini, U. Volpe, F. Vellante, M. Fornaro, A. Valchera, C. Tomasetti, S. Fraticelli *et al.*, "Understanding the complex of suicide in depression: from research to clinics," *Psychiatry investigation*, vol. 17, no. 3, p. 207, 2020.

[34] J. B. Williams, "A structured interview guide for the hamilton depression rating scale," *Archives of general psychiatry*, vol. 45, no. 8, pp. 742–747, 1988.

[35] S. Montgomery, "Montgomery and asperg depression-rating-scale (madrs)," *German Version by Neumann NU, Schulte RM. Erlangen, Perimed Fachbuch-Verlagsgesellschaft GmbH*, 1989.

[36] A. Galinowski and P. Lehert, "Structural validity of madrs during antidepressant treatment," *International Clinical Psychopharmacology*, vol. 10, no. 3, pp. 157–161, 1995.

[37] A. T. Beck, R. A. Steer, G. K. Brown *et al.*, *Beck depression inventory*. Harcourt Brace Jovanovich New York:, 1987.

[38] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[39] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated eeg-based screening of depression using deep convolutional neural network," *Computer methods and programs in biomedicine*, vol. 161, pp. 103–113, 2018.

[40] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, 2015.

[41] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. R. Traum, A. A. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *LREC*, 2014.

[42] I. J. Orton, "Vision based body gesture meta features for affective computing," *arXiv preprint arXiv:2003.00809*, 2020.

[43] I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," *The Lancet*, vol. 387, no. 10027, pp. 1561–1572, 2016.

[44] G. Perrotta, "Bipolar disorder: definition, differential diagnosis, clinical contexts and therapeutic approaches," *J Neuroscience and Neurological Surgery*, vol. 5, no. 1, 2019.

[45] "Mental disorders," https://www.who.int/news-room/fact-sheets/detail/mental-disorders, jun 8 2022.

[46] H. H. Gardner, N. L. Kleinman, R. A. Brook, K. Rajagopalan, T. J. Brizee, and J. E. Smeeding, "The economic impact of bipolar disorder in an employed population from an employer perspective," *Journal of Clinical Psychiatry*, vol. 67, no. 8, pp. 1209–1218, 2006.

[47] L. V. Kessing, E. Vradi, and P. K. Andersen, "Life expectancy in bipolar disorder," *Bipolar disorders*, vol. 17, no. 5, pp. 543–548, 2015.

[48] J. R. Geddes and D. J. Miklowitz, "Treatment of bipolar disorder," *The lancet*, vol. 381, no. 9878, pp. 1672–1682, 2013.

[49] J. B. Williams and M. First, "Diagnostic and statistical manual of mental disorders," in *Encyclopedia of social work*, 2013.

[50] R. Young, J. Biggs, V. Ziegler, and D. Meyer, "Young mania rating scale," *Journal of Affective Disorders*, 2000.

[51] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, "The turkish audio-visual bipolar disorder corpus," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[52] E. G. Altman, D. Hedeker, J. L. Peterson, and J. M. Davis, "The altman self-rating mania scale," *Biological psychiatry*, vol. 42, no. 10, pp. 948–955, 1997.

[53] J. Angst, R. Adolfsson, F. Benazzi, A. Gamma, E. Hantouche, T. D. Meyer, P. Skeppar, E. Vieta, and J. Scott, "The hcl-32: towards a self-assessment tool for hypomanic symptoms in outpatients," *Journal of affective disorders*, vol. 88, no. 2, pp. 217–233, 2005.

[54] R. M. Hirschfeld, L. Lewis, L. A. Vornik *et al.*, "Perceptions and impact of bipolar disorder: how far have we really come? results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder," *Journal of Clinical Psychiatry*, vol. 64, no. 2, pp. 161–174, 2003.

[55] S. Abdullah, M. Matthews, E. Frank, G. Doherty, G. Gay, and T. Choudhury, "Automatic detection of social rhythms in bipolar disorder," *Journal of the American Medical Informatics Association*, vol. 23, no. 3, pp. 538–543, 2016.

[56] S. Yasin, S. A. Hussain, S. Aslan, I. Raza, M. Muzammel, and A. Othmani, "Eeg based major depressive disorder and bipolar disorder detection using neural networks:

A review," *Computer Methods and Programs in Biomedicine*, vol. 202, p. 106007, 2021.

[57] W. H. Organization, "Stress," Feb 2023. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/stress#

[58] J. Herbert, "Fortnightly review: Stress, the brain, and mental illness," *Bmj*, vol. 315, no. 7107, pp. 530–535, 1997.

[59] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 2016.

[60] R. S. Reis, A. Hino, and C. Añez, "Perceived stress scale," *J. health Psychol*, vol. 15, no. 1, pp. 107–114, 2010.

[61] L. Parkitny and J. McAuley, "The depression anxiety stress scale (dass)," *Journal of physiotherapy*, vol. 56, no. 3, p. 204, 2010.

[62] M. Yamaguchi, S. Yoshikawa, Y. Tahara, D. Niwa, Y. Imai, and V. Shetty, "Point-of-use measurement of salivary cortisol levels," in *SENSORS, 2009 IEEE*. IEEE, 2009, pp. 343–346.

[63] S. Tivatansakul and M. Ohkura, "Improvement of emotional healthcare system with stress detection from ecg signal," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6792–6795.

[64] A. Association *et al.*, "2021 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 17, no. 3, pp. 327–406, 2021.

[65] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine *et al.*, "Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[66] W. H. Organization, "Dementia," Mar 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[67] Y. Kaufman, D. Anaki, M. Binns, and M. Freedman, "Cognitive decline in alzheimer disease: Impact of spirituality, religiosity, and qol," *Neurology*, vol. 68, no. 18, pp. 1509–1514, 2007.

[68] M. S. Mittelman, D. L. Roth, O. J. Clay, and W. E. Haley, "Preserving health of alzheimer caregivers: impact of a spouse caregiver intervention," *The American Journal of Geriatric Psychiatry*, vol. 15, no. 9, pp. 780–789, 2007.

[69] O. L. López and S. T. DeKosky, "Clinical symptoms in alzheimer's disease," *Handbook of clinical neurology*, vol. 89, pp. 207–216, 2008.

[70] X.-L. Li, N. Hu, M.-S. Tan, J.-T. Yu, and L. Tan, "Behavioral and psychological symptoms in alzheimer's disease," *BioMed research international*, vol. 2014, 2014.

[71] M. Ota, N. Sato, Y. Nakata, K. Arima, and M. Uno, "Relationship between apathy and diffusion tensor imaging metrics of the brain in alzheimer's disease," *International journal of geriatric psychiatry*, vol. 27, no. 7, pp. 722–726, 2012.

[72] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[73] R. C. Mohs and L. Cohen, "Alzheimer's disease assessment scale (adas)," *Psychopharmacol Bull*, vol. 24, no. 4, pp. 627–628, 1988.

[74] S. Bhat, U. R. Acharya, N. Dadmehr, and H. Adeli, "Clinical neurophysiological and automated eeg-based diagnosis of the alzheimer's disease," *European neurology*, vol. 74, no. 3-4, pp. 202–210, 2015.

[75] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of medical systems*, vol. 44, pp. 1–16, 2020.

[76] U. R. Acharya, S. L. Fernandes, J. E. WeiKoh, E. J. Ciaccio, M. K. M. Fabell, U. J. Tanik, V. Rajinikanth, and C. H. Yeong, "Automated detection of alzheimer's disease using brain mri images–a study with various feature extraction techniques," *Journal of Medical Systems*, vol. 43, pp. 1–14, 2019.

[77] T. Hoang, T.-T. Nguyen, and H. D. Nguyen, "Unified tensor network for multimodal dementia detection," in *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 2022, pp. 409–416.

[78] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici *et al.*, "A deep learning model to

predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[79] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[80] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021.

[81] M. A. Wiering and M. Van Otterlo, "Reinforcement learning," *Adaptation, learning, and optimization*, vol. 12, no. 3, p. 729, 2012.

[82] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 537–547, 2020.

[83] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervasive and Mobile Computing*, vol. 51, pp. 1–26, 2018.

[84] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, pp. 423–443, 2019.

[85] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[86] A.-Q. Duong, N.-H. Ho, H.-J. Yang, G.-S. Lee, and S.-H. Kim, "Multi-modal stress recognition using temporal convolution and recurrent network with positional embedding," in *MuSe*, 2021, pp. 37–42.

[87] V. Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[88] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *AVEC*, 2019, pp. 55–63.

[89] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[90] T. Zhang and J. Wang, "Collaborative quantization for cross-modal similarity search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2036–2045.

[91] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[92] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[93] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012, pp. 359–368.

[94] X. Jiang, F. Wu, Y. Zhang, S. Tang, W. Lu, and Y. Zhuang, "The classification of multi-modal data with hidden conditional random field," *Pattern Recognition Letters*, vol. 51, pp. 63–69, 2015.

[95] C. L. Zitnick and D. Parikh, "Bringing semantics into focus using visual abstraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3009–3016.

[96] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2405–2413.

[97] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Aligning plot synopses to videos for story-based retrieval," *International Journal of Multimedia Information Retrieval*, vol. 4, pp. 3–16, 2015.

[98] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[99] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[100] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *International conference on machine learning*. PMLR, 2016, pp. 2397–2406.

[101] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[102] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[103] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[104] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[105] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 508–513.

[106] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," *Pattern recognition*, vol. 47, no. 5, pp. 1800–1812, 2014.

[107] Z. Liao, L. Gao, T. Zhou, X. Fan, Y. Zhang, and J. Wu, "An oil painters recognition method based on cluster multiple kernel learning algorithm," *IEEE Access*, vol. 7, pp. 26 842–26 854, 2019.

[108] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal alzheimer's disease classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 984–990, 2013.

[109] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Graph attention model embedded with multi-modal knowledge for depression detection," in *ICME*, 2020, pp. 1–6.

[110] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, p. 107316, 2021.

[111] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.

[112] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," *arXiv preprint arXiv:1907.03196*, 2019.

[113] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *AVEC*, 2019, pp. 65–71.

[114] N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," *Expert Systems with Applications*, vol. 211, p. 118523, 2023.

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

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

[117] S. Li, P. Zheng, J. Fan, and L. Wang, "Toward proactive human–robot collaborative assembly: A multimodal transfer-learning-enabled action prediction approach," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 8, pp. 8579–8588, 2021.

[118] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[119] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[120] 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. 149 808–149 824, 2020.

[121] B. Cheng, M. Liu, H.-I. Suk, D. Shen, D. Zhang, and A. D. N. Initiative, "Multimodal manifold-regularized transfer learning for mci conversion prediction," *Brain imaging and behavior*, vol. 9, pp. 913–926, 2015.

[122] M. Baroni, "Grounding distributional semantics in the visual world," *Language and Linguistics Compass*, vol. 10, no. 1, pp. 3–13, 2016.

[123] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2016, pp. 160–170.

[124] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," *Advances in neural information processing systems*, vol. 26, 2013.

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

[126] M. M. Khapra, A. Kumaran, and P. Bhattacharyya, "Everybody loves a rich cousin: An empirical study of transliteration through bridge languages," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 420–428.

[127] D. Zhou, J. Luo, V. M. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. Kautz, "Tackling mental health by integrating unobtrusive multimodal sensing," in *AAAI*, 2015.

[128] B. Naik, A. Mehta, and M. Shah, "Denouements of machine learning and multimodal diagnostic classification of alzheimer's disease," *Visual Computing for Industry, Biomedicine, and Art*, vol. 3, no. 1, pp. 1–18, 2020.

[129] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[130] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[131] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, 2006.

[132] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[133] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.

[134] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 202–208.

[135] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[136] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[137] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[138] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *The Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[139] C. Ricky, M. N. O'Donnell Siobhan *et al.*, "Factors associated with delayed diagnosis of mood and/or anxiety disorders," *Health promotion and chronic disease prevention in Canada: research, policy and practice*, vol. 37, no. 5, p. 137, 2017.

[140] C. Bourke, K. Douglas, and R. Porter, "Processing of facial emotion expression in major depression: a review," *Australian & New Zealand Journal of Psychiatry*, vol. 44, no. 8, pp. 681–696, 2010.

[141] D. Marazziti, G. Consoli, M. Picchetti, M. Carlini, and L. Faravelli, "Cognitive impairment in major depression," *European journal of pharmacology*, vol. 626, no. 1, pp. 83–86, 2010.

[142] D. F. Santomauro, A. M. M. Herrera, J. Shadid, P. Zheng, C. Ashbaugh, D. M. Pigott, C. Abbafati, C. Adolph, J. O. Amlag, A. Y. Aravkin *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[143] S. S. and J. S. Raj, "Analysis of deep learning techniques for early detection of depression on social media network - a comparative study," 2021.

[144] G. Arbanas, "Diagnostic and statistical manual of mental disorders (dsm-5)," *Alcoholism and psychiatry research*, vol. 51, pp. 61–64, 2015.

[145] M. I. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, pp. 255 – 260, 2015.

[146] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. M. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. R. Traum, R. Wood, Y. Xu, A. A. Rizzo, and L.-P. Morency, "Simsensei kiosk: a virtual human interviewer for healthcare decision support," in *AAMAS*, 2014.

[147] I. Orton, "Vision based body gesture meta features for affective computing," *ArXiv*, vol. abs/2003.00809, 2020.

[148] M. Jaiswal, Z. Aldeneh, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations," *ICASSP*, pp. 7415–7419, 2019.

[149] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," *MuSe*, 2021.

[150] E. Ciftçi, H. Kaya, H. Güleç, and A. A. Salah, "The turkish audio-visual bipolar disorder corpus," *ACII Asia*, pp. 1–6, 2018.

[151] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *AVEC*, 2019, pp. 81–88.

[152] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Depressionnet: learning multi-modalities with user post summarization for depression detection on social media," in *ACM SIGIR*, 2021, pp. 133–142.

[153] M. An, J. Wang, S. Li, and G. Zhou, "Multimodal topic-enriched auxiliary learning for depression detection," in *COLING*, 2020, pp. 1078–1089.

[154] C.-P. Bara, M. Papakostas, and R. Mihalcea, "A deep learning approach towards multimodal stress detection." in *AffCon@ AAAI*, 2020, pp. 67–81.

[155] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *FG*. IEEE, 2020, pp. 344–350.

[156] M. Rohanian, J. Hough, M. Purver *et al.*, "Detecting depression with word-level multimodal fusion." in *INTERSPEECH*, 2019, pp. 1443–1447.

[157] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Predicting depression severity by multi-modal feature engineering and fusion," in *AAAI*, vol. 32, no. 1, 2018.

[158] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *AVEC*, 2017, pp. 69–76.

[159] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *AAAI*, vol. 33, no. 01, 2019, pp. 110–117.

[160] N. Abaeikoupaei and H. Al Osman, "A multi-modal stacked ensemble model for bipolar disorder classification," *IEEE TAC*, 2020.

[161] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality hierarchical recall based on gbdts for bipolar disorder classification," in *AVEC*, 2018, pp. 31–37.

[162] Y. Yao, M. Papakostas, M. Burzo, M. Abouelenien, and R. Mihalcea, "Muser: Multimodal stress detection using emotion recognition as an auxiliary task," *arXiv preprint arXiv:2105.08146*, 2021.

[163] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *COLING*, 2020, pp. 7871–7880.

[164] X. Zhang, J. Pan, J. Shen, Z. U. Din, J. Li, D. Lu, M. Wu, and B. Hu, "Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection," *IEEE TAC*, 2020.

[165] S. A. Dhedhi, D. Swinglehurst, and J. Russell, "'timely'diagnosis of dementia: what does it mean? a narrative analysis of gps' accounts," *BMJ open*, vol. 4, no. 3, p. e004439, 2014.

[166] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[167] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity," *arXiv preprint arXiv:2009.00700*, 2020.

[168] W. H. Organization *et al.*, "World health organization fact sheet—dementia," 2021.

[169] J. Vrijsen, T. Matulessij, T. Joxhorst, S. E. de Rooij, and N. Smidt, "Knowledge, health beliefs and attitudes towards dementia and dementia risk reduction among the dutch general population: a cross-sectional study," *BMC public health*, vol. 21, no. 1, pp. 1–11, 2021.

[170] T. A. Widiger, P. T. Costa, A. P. Association *et al.*, *Personality disorders and the five-factor model of personality*. JSTOR, 2013.

[171] J. Garre-Olmo, "Epidemiology of alzheimer's disease and other dementias," *Revista de neurologia*, vol. 66, no. 11, pp. 377–386, 2018.

[172] Z. S. Khachaturian, "Diagnosis of alzheimer's disease," *Archives of neurology*, vol. 42, no. 11, pp. 1097–1105, 1985.

[173] J. Weller and A. Budson, "Current understanding of alzheimer's disease diagnosis and treatment," *F1000Research*, vol. 7, 2018.

[174] H. Dong, J. Li, L. Huang, X. Chen, D. Li, T. Wang, C. Hu, J. Xu, C. Zhang, K. Zen *et al.*, "Serum microrna profiles serve as novel biomarkers for the diagnosis of alzheimer's disease," *Disease markers*, vol. 2015, 2015.

[175] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, "Machine learning for dementia prediction: A systematic review and future research directions," *Journal of medical systems*, vol. 47, no. 1, p. 17, 2023.

[176] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens, "Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, 2020.

[177] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[178] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[179] C. Pope and B. H. Davis, "Finding a balance: The carolinas conversation collection," 2011.

[180] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, "Semantic feature extraction using sbert for dementia detection," *Brain Sciences*, vol. 12, no. 2, p. 270, 2022.

[181] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, pp. 44–55, 2015.

[182] I. Krstev, M. Pavikjevikj, M. Toshevska, and S. Gievska, "Multimodal data fusion for automatic detection of alzheimer's disease," in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design: 13th International Conference, DHM 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II*. Springer, 2022, pp. 79–94.

[183] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.

[184] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *arXiv preprint arXiv:1811.09919*, 2018.

[185] S. Nasreen, J. Hough, M. Purver *et al.*, "Detecting alzheimer's disease using interactional and acoustic features from spontaneous speech." Interspeech, 2021.

[186] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[187] D. Berrar, "Cross-validation." 2019.

[188] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[189] K. Bar, V. Zilberstein, I. Ziv, H. Baram, N. Dershowitz, S. Itzikowitz, and E. V. Harel, "Semantic characteristics of schizophrenic speech," *arXiv preprint arXiv:1904.07953*, 2019.

[190] V. Silber-Varod, H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, "Do social anxiety individuals hesitate more? the prosodic profile of hesitation disfluencies in social anxiety disorder individuals," *Speech Prosody 2016*, pp. 1211–1215, 2016.

[191] M. Panagopoulou, M. Karaglani, V. G. Manolopoulos, I. Iliopoulos, I. Tsamardinos, and E. Chatzaki, "Deciphering the methylation landscape in breast cancer: diagnostic and prognostic biosignatures through automated machine learning," *Cancers*, vol. 13, no. 7, p. 1677, 2021.

[192] H. Park, M. W. R. Jung, and U. Oh, "Apd: An autoencoder-based prediction model for depression diagnosis," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2021, pp. 376–379.

[193] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *arXiv preprint arXiv:2104.09356*, 2021.