HAL open science

A review of deep learning-based information fusion techniques for multimodal medical image classification

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, Gwenolé Quellec

▶ To cite this version:

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, et al.. A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine, 2024, 177, pp.108635. 10.1016/j.compbiomed.2024.108635 . hal-04580775

HAL Id: hal-04580775
https://imt-atlantique.hal.science/hal-04580775v1

Submitted on 29 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Computers in Biology and Medicine 177 (2024) 108635

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A review of deep learning-based information fusion techniques for multimodal medical image classification

Yihao Li [a,b], Mostafa El Habib Daho [a,b,*], Pierre-Henri Conze [a,c], Rachid Zeghlache [a,b], Hugo Le Boité [d,e], Ramin Tadayoni [f,e], Béatrice Cochener [a,b,g], Mathieu Lamard [a,b], Gwenolé Quellec [a]

[a] LaTIM UMR 1101, Inserm, Brest, France
[b] University of Western Brittany, Brest, France
[c] IMT Atlantique, Brest, France
[d] Sorbonne University, Paris, France
[e] Ophthalmology Department, Lariboisière Hospital, AP-HP, Paris, France
[f] Paris Cité University, Paris, France
[g] Ophthalmology Department, CHRU Brest, Brest, France

ARTICLE INFO

Keywords:
Multimodality fusion
Deep learning
Medical image classification
Computer-aided diagnosis

ABSTRACT

Multimodal medical imaging plays a pivotal role in clinical diagnosis and research, as it combines information from various imaging modalities to provide a more comprehensive understanding of the underlying pathology. Recently, deep learning-based multimodal fusion techniques have emerged as powerful tools for improving medical image classification. This review offers a thorough analysis of the developments in deep learning-based multimodal fusion for medical classification tasks. We explore the complementary relationships among prevalent clinical modalities and outline three main fusion schemes for multimodal classification networks: input fusion, intermediate fusion (encompassing single-level fusion, hierarchical fusion, and attention-based fusion), and output fusion. By evaluating the performance of these fusion techniques, we provide insight into the suitability of different network architectures for various multimodal fusion scenarios and application domains. Furthermore, we delve into challenges related to network architecture selection, handling incomplete multimodal data management, and the potential limitations of multimodal fusion. Finally, we spotlight the promising future of Transformer-based multimodal fusion techniques and give recommendations for future research in this rapidly evolving field.

1. Introduction

1.1. Context

In recent years, the field of medical image analysis has seen a surge in efforts to apply deep learning-based methods to the classification of various diseases, notably related to the brain [1–3], breasts [4–6], prostate [7–9] and eyes [10–12]. The ability to accurately classify and diagnose diseases from medical images has the potential to revolutionize healthcare by improving diagnostic accuracy, reducing human error, and enabling more personalized treatment planning. This trend has highlighted the need for robust and efficient methods for analyzing medical images across multiple imaging modalities.

With advances in medical image acquisition systems, many new imaging modalities have been developed to diagnose patients [13–15], resulting in larger and more diverse datasets. An imaging modality alone does not often provide all the information needed to ensure accurate clinical diagnosis. Therefore, clinicians increasingly base their diagnosis on images obtained from a variety of sources: a combination of abundant information can be used in clinical practice with more confidence. Following this trend and to improve diagnosis results, artificial intelligence-based classification models are increasingly being developed by combining data from multiple sources to take advantage of both redundancies and complementarities across modalities.

Several surveys have been conducted in recent years to analyze the trends in the application of multimodality in various fields [16, 17], among which the field of medicine is gaining a great deal of attention. In medicine, several such survey papers focus on specific image analysis tasks: image fusion [14], image synthesis [18], image

* Corresponding author at: University of Western Brittany, Brest, France.
E-mail address: mostafa.elhabibdaho@univ-brest.fr (M. El Habib Daho).

https://doi.org/10.1016/j.compbiomed.2024.108635
Received 5 October 2023; Received in revised form 18 March 2024; Accepted 18 May 2024
Available online 22 May 2024
0010-4825/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

segmentation [19], or image registration [20]. However, medical image classification was never addressed in a comprehensive manner. A few surveys target specific fields such as neurology [21] or oncology [22], but they do not provide a comprehensive discussion of how multimodal fusion may be applied to other fields. To fill this gap, we propose to review deep learning-based information fusion techniques for multimodal medical classification across all medical fields. We restrict this analysis to classification, given the sufficient coverage of other analysis tasks [14,18–20]. We include in the scope of classification methods any method assigning class labels, possibly with probabilities, to a patient or a region of interest in the patient, regardless of the application (diagnosis, prognosis, risk estimation, etc.). Throughout our review, we summarize and discuss the advantages and disadvantages of various information fusion methods that can be applied to various organs and imaging modalities. As the first review to examine the use of deep learning in multimodal medical classification, this paper aims to guide future investigations into medical diagnosis using multiple imaging modalities.

1.2. Traditional methods

Information fusion not based on deep learning strategies, relying on traditional image processing and machine learning, has been reviewed in a previous survey [23]. We summarize hereafter the main developments and highlight the benefits of non-deep learning-based information fusion.

Input fusion is the most commonly used strategy among traditional methods. It involves the fusion of images from various modalities into structured data and fuses them into different categories depending on the fusion domain: spatial fusion [20,24–27], frequency fusion [28–33] and sparse representation [34–36]. In spatial fusion, multimodal images are combined at the pixel level, but this approach often leads to spectral degradation [37] and color distortion [38]. Frequency fusion, which involves transforming the input image into the frequency domain, is more complex and results in limited spatial resolution [39]. Sparse representation, on the other hand, can be sensitive to registration errors and lacks attention to details [38].

Other strategies include intermediate and output fusion, which do not require registration of the input images. Intermediate fusion involves extracting features from different imaging modalities, concatenating them, and feeding them into a classifier, generally a support vector machine (SVM), for diagnosis [40–42]. This approach requires extensive testing and rich domain knowledge for feature extraction and selection. On the other hand, output fusion involves stacking the data results from unimodal models and combining them [43]. While this approach circumvents the need for early integration, it presents its own set of challenges. Individual models in output fusion may be heavily influenced by their respective modality-specific idiosyncrasies, potentially introducing biases into the final combined output. Moreover, if these unimodal models yield correlated or redundant information, the utility of stacking them diminishes, as it might not deliver significant additional value.

Traditional methods typically involve complex pre-processing steps paired with relatively simple model structures. Such a combination frequently leads to information loss during feature extraction, thereby complicating efforts to fully leverage the synergies between various imaging modalities.

Besides requiring domain knowledge, these traditional multimodal fusion approaches do not fully utilize the complementarity between multimodal features. These limitations highlight the need for more advanced techniques, such as deep learning-based multimodal fusion methods, able to overcome the challenges faced by traditional methods. Deep learning network architectures offer complex models that can explore more possibilities for multimodal fusion. Furthermore, various end-to-end models significantly reduce the amount of domain knowledge required for diagnosis purposes, albeit at the cost of interpretability [44].

1.3. Development trends

Recognizing the potential of deep learning-based methods for multimodal medical image classification, researchers have increasingly focused on this area. In order to obtain more accurate diagnoses, multimodal medical image analysis have also become a growing trend. Fig. 2 shows the number of publications about multimodal medical classification each year, which was queried on February 27, 2024, on PubMed. As illustrated by the figure, the number of papers has increased yearly from 2016 to 2023, indicating that multimodal medical classification tasks based on deep learning have gained greater attention in recent years. Furthermore, we report the number of publications on different organs in multimodal diagnosis tasks in Fig. 3. We found that brain-related publications currently account for a substantial portion of multimodal studies. This is due to the disclosure of many large multimodal image datasets on the brain. On the other hand, not many studies were conducted on other organs, except whenever a public dataset was released. This finding motivated us to focus the review on studies performed on public datasets from various organs. One advantage is to allow direct quantitative comparisons between methods.

1.4. Paper selection

In our initial literature search, we identified a total of 14 public multimodal image datasets. These datasets are detailed in Section 2.2, with a summary presented in Table 1. The methodology for finalizing the list of papers for this review was as follows:

1. For each of the 14 datasets, we conducted a search on PubMed for publications that mentioned the dataset name, coupled with any of the following terms: (multimodality), (multimodal), (multi-modal), (multiparametric), or (multi-parametric).
2. We then concatenated the 14 resulting lists.
3. Based on the abstracts, we handpicked articles that addressed multimodal information fusion through deep learning methods.

Notably, there are gaps in the availability of public multimodal datasets focused on classification tasks for certain organs—namely, the breast, lung, prostate, kidneys, larynx, heart, and liver, even though they are frequently discussed in the multimodal medical image analysis literature. To ensure a comprehensive review, we expanded our scope to include 19 pertinent articles that target these organs but utilize private datasets. This brought our final tally to 114 publications.

1.5. Highlights

Through our examination of the deep learning-based multimodal image classification literature (overview presented in Fig. 1), we propose in this paper an updated taxonomy for multimodal information fusion. As discussed in Section 1.2 and other surveys [13,16,45], multimodal fusion methods are traditionally classified as *input fusion*, *intermediate fusion* or *output fusion*, based on the stage of information fusion in the classification pipeline, as in Fig. 4(a). Note that some publications refer to input fusion as early fusion, while intermediate fusion may be considered as feature-level fusion, and output fusion is equivalent to decision-level fusion or late fusion [10,16,45]. Our analysis points to intermediate fusion as the prevailing category at present. To grant readers a more in-depth understanding of multimodal deep learning networks, we further segment intermediate fusion into *single-level fusion*, *hierarchical fusion*, and *attention-based fusion*, as illustrated in Fig. 4(b). The proposed taxonomy is detailed and discussed in Section 4.1: it covers the majority of the current multimodal classification network architectures, providing insight into their stages and styles of information fusion.

In this paper, we present the following contributions:

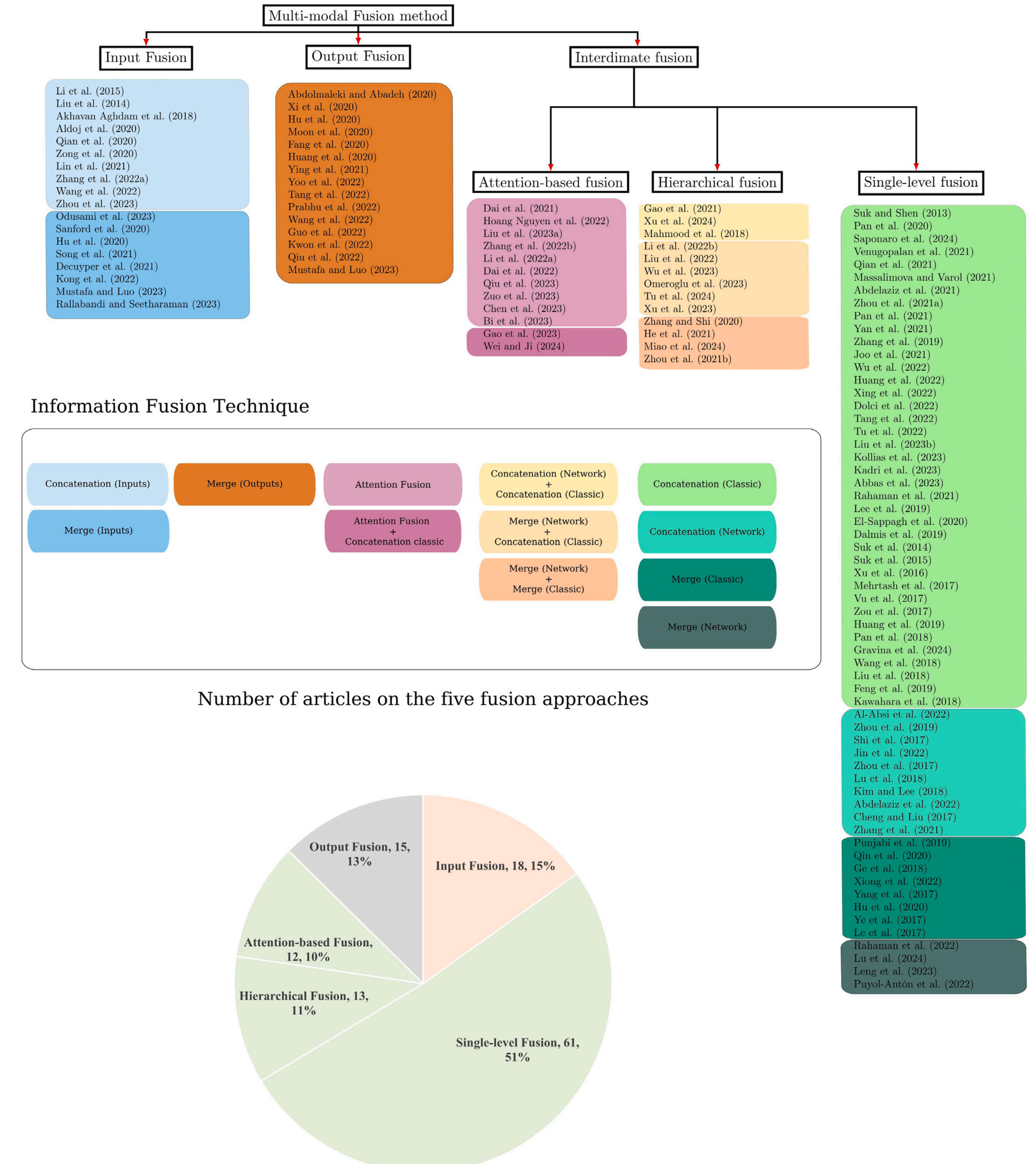

Fig. 1. Overview and proportion of deep learning-based information fusion techniques for multimodal medical image classification presented in this paper.

(1) *Identify the process of medical multimodal classification.*
The methodological approach of deep learning-based multimodal classification can divided into four steps: data processing, deep learning network, multimodal information fusion, and the final classification algorithm. Crucially, the classification of multimodal information fusion methods hinges on the sequential positioning of these stages. Propose network architectures apt for generic multimodal fusion classification endeavors.

(2) *Propose network architectures for generic multimodal fusion classification tasks.*
In order to address medical classification tasks involving different organs and imaging modalities, we summarized five strategies of

Table 1
A list of multimodal image datasets. The list is sorted by the number of publications on PubMed (Keywords: dataset name AND 'multimodal'). Details about imaging modalities are given in Table 2.

Dataset	Year	Modalities	Body Organ(s)	Medical Diagnosis	EHR
ADNI[a]	(2004–2009) (2010–2016) (2016–2022) (2023–2027)	sMRI, fMRI, PET	Brain	Alzheimer's Disease	Available
BraTS[b]	(2012–2023) yearly	MRI (TI, T2, T1c, FLAIR)	Brain	Brain Tumor	N/A
TCIA[c]	2014	CT, MRI, PET, US, etc.	Brain, Breast, Lung, Kidney, Head–Neck, Liver, Pancreas, etc.	Common Cancer Disease	Available
OASIS[d]	2007	MRI, PET	Brain	Alzheimer's Disease	Available
SPC[e]	2018	Dsc, Clinical Image, Metadata	Skin	Skin Lesion	Available
TCGA[f]	2006	Pathological data, Genomic data	Brain, Lung, etc.	Common Cancer Disease	Available
ABIDE[g]	2012	sMRI, fMRI	Brain	Autism Spectrum Disorder (ASD)	Available
ADHD-200[h]	2011	sMRI, fMRI	Brain	Attention Deficit Hyperactivity Disorder (ADHD)	Available
COBRE[i]	2012	sMRI, fMRI	Brain	Schizophrenia	Available
GAMMA[j]	2021	OCT, Fundus Image	Eye	Glaucoma	N/A
CPM-RadPath[k]	2019, 2020	MRI (TI, T2, T1c, FLAIR)	Brain	Brain Tumor	N/A
ISIT-UMR[l]	2019	White Light RGB, Narrow Band Imaging (NBI)	Digestive Tract	Gastrointestinal Lesions	N/A
MRNet[m]	2018	MRI (TI, T2)	Knee	Knee Injuries	N/A
CTU-UHB[n]	2014	CT (FHR, UC)	Uterus	Fetal Distress Diagnosis	Available

[a] https://adni.loni.usc.edu/
[b] http://braintumorsegmentation.org/
[c] https://www.cancerimagingarchive.net/
[d] https://www.oasis-brains.org/
[e] https://derm.cs.sfu.ca/Welcome.html
[f] https://www.cancer.gov/ccg/research/genome-sequencing/tcga
[g] http://fcon_1000.projects.nitrc.org/indi/abide/
[h] http://fcon_1000.projects.nitrc.org/indi/adhd200/
[i] https://fcon_1000.projects.nitrc.org/indi/retro/cobre.html
[j] https://aistudio.baidu.com/aistudio/competition/detail/90/0/introduction
[k] https://zenodo.org/records/3718894
[l] http://www.depeca.uah.es/colonoscopy_dataset/
[m] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002699
[n] https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/

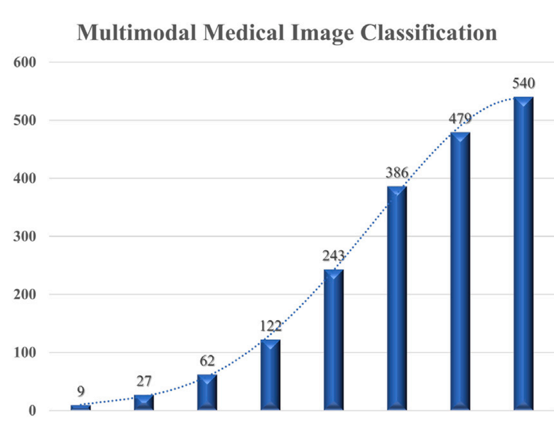

Fig. 2. Number of publications on medical multimodal image classification. Per-year statistics obtained using PubMed from 2016 to 2023.

multimodal fusion: input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion. These fusion methods can be applied to any multimodal classification problem in medicine, allowing for greater flexibility and potential for improved results.

(3) *Present the prevailing challenges and predict future trends.* While it is apparent that multimodal fusion is still in its early stages, our paper analyzes specific challenges tied to this domain and predicts future trends in the field.

The remainder of the paper is organized as follows: Section 2 describes commonly used multimodal data for medical multimodal classification tasks and their publicly available datasets. Section 3 describes the multimodal medical image classification task process mentioned in contribution 1. A review of papers implementing each of the five fusion strategies of contribution 2 is presented in Section 4. The purpose of Section 5 is to discuss the existing problems and to make predictions regarding future fusion methods in contribution 3. Finally, Section 6 contains our concluding comments. A list of frequently used abbreviations throughout the paper is shown in Table 6.

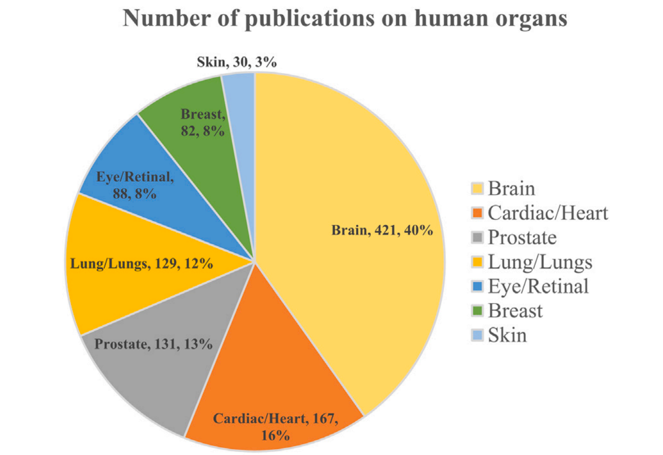

Fig. 3. Number of publications dealing with medical multimodal image classification on human organs, from 2016 to 2023. Tags: organ, number of publications, percentage.

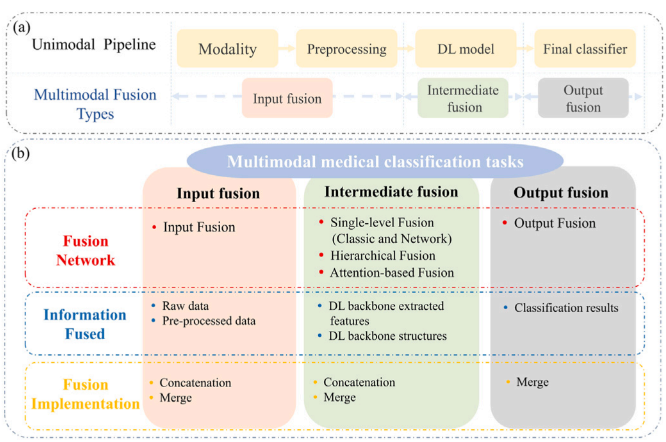

Fig. 4. (a) Unimodal classification task flow and different types of multimodal fusion based on the level in which they perform information fusion. (b) Information fusion networks for the three types of multimodal fusion, inputs to information fusion, and the implementation of information fusion.

2. Multimodal medical images

2.1. Imaging modalities

For medical diagnosis purposes, each imaging modality has its own characteristics and information. Different medical imaging modalities use different frequency bands of the electromagnetic spectrum in order to screen and diagnose different medical conditions in the human body [14]. There are different wavelengths and frequencies associated with each imaging modality, as well as different characteristics (structure, function, etc.) [46]. Furthermore, medical imaging modalities can be classified as invasive or non-invasive. Invasive methods involve inserting an object into the body through an incision or needle injection in order to examine an organ, while non-invasive methods utilize some form of radiation or sound [14]. Table 2 shows some modalities that appear in multimodal medical image datasets.

Due to their complementary nature, there has been a significant focus on the following combinations of modalities targeting various diseases: (1) multi-parametric MRI (TI, T2, T1C, and FLAIR) [54–58], (2) MRI and PET [59–62], (3) PET and CT [63], (4) multi-view ultrasound (US B-mode, US color Doppler) [4,6], (5) Color Fundus Photographs (CFP) and Optical Coherence Tomography (OCT) [10,64,65], (6) Dermatoscope (Dsc) and Clinical Images [66–68], and (7) combined diagnosis of Image Data and Clinical Data [69–72]. The complementary relationships between these modalities will be briefly discussed.

Neurology and neurosurgery frequently use MRI. Different MRI images can be obtained by changing the factors affecting the magnetic resonance (MR) signal, and these different images are referred to as sequences. Depending on the sequence used, the behavior of tumors may vary, and it is essential to use multiple sequences to accurately determine tumor location and size [73]. T1-weighted (T1) and T2-weighted (T2) MRIs are the most common MRI sequences. Tomographic anatomical maps can be observed with the T1 sequence, and the T2 sequence clearly shows the location and size of the lesion [74]. The Fluid Attenuated Inversion Recovery (Flair) sequence provides better visualization of the area around the tumor site, making it easier to detect the tumor's boundaries [75]. Furthermore, contrast-enhanced T1-weighted (T1c) sequences can be used to detect intra-tumor conditions and distinguish tumors from non-tumorigenic lesions [76]. T2 and Flair are suitable for detecting tumors with peritumoral edema, while T1 and T1c are suitable for detecting tumors without peritumoral edema [19].

Diffusion-weighted imaging (DWI) is another useful sequence designed to detect the random movements of water protons. Therefore, DWI sequence is a highly sensitive method for detecting acute strokes [77]. An increased apparent diffusion coefficient (ADC) value with lower signals of DWI images could reveal the fast diffusion of water molecules [78]. In addition to using multiple sequences, co-diagnosis using structural MRI (sMRI) and functional MRI (fMRI) is becoming increasingly popular [79,80]. fMRI measures the small changes in blood flow that occur with brain activity. This test can be used to determine which parts of the brain are performing critical functions and to determine the effects of strokes and other diseases on the brain [81].

The combination of PET and MRI, PET and CT has been recognized as a valuable method for screening and diagnosing various diseases [82–86]. The PET scan is preceded by the administration of a radioactive agent to the patient. This allows doctors to determine the metabolic processes in which the brain tissue is involved [48]. Compared to other imaging methods such as CT and MRI, PET has a high sensitivity and can detect lesions even if MRI/CT does not yet show abnormalities. PET also has high specificity, making it possible to determine whether a tumor is malignant based on its metabolism at the time of MRI/CT detection [87]. However, because PET scan lacks information about organ anatomy, they should be conducted in conjunction with CT/MRI scans [79]. Indeed, the combination of PET and MRI/CT scans provides structural and functional information related to various diseases, improving the effectiveness of diagnosis. Fig. 5 shows the images of PET, CT, and MRI, as well as several sequences of MRI.

Availability, low cost, and safety make ultrasonography the most widely used clinical diagnostic tool, with applications ranging from breast cancer diagnosis to cervical lymph node detection. Conventional B-mode imaging is used to examine abnormal masses in tissues, Color Doppler imaging shows the distribution of blood vessels within tissues [88], while Strain Elastography (SE) is a qualitative technique and provides information on the relative stiffness between one tissue and another. For example, the combined use of Conventional B-mode imaging and Color Doppler is common in identifying cervical lymph nodes [89], diagnosing breast cancer [4,6], and so forth [90–92]. Fig. 6 shows the US images of Conventional B-mode, Color Doppler, and Strain Elastography.

In the diagnosis of ophthalmic diseases, CFP and OCT are the two most cost-effective methods [10]. These imaging modalities provide prominent biomarkers that can be used to identify glaucoma suspects, such as the vertical cup-to-disk ratio (vCDR) on fundus images and the retinal nerve fiber layer thickness (RNFL) on an OCT image. A more accurate and reliable diagnosis, compared to a single modality, is often achieved by taking both screenings in clinical practice [64]. Fig. 7 shows the images of CFP and OCT.

Table 2
Some examples of imaging modalities and organs found in the multimodal medical image analysis literature.

Modalities	Body organs examined	Invasive/ Non-invasive	Description
Magnetic Resonance Image (MRI)	Brain, Prostate, Breast, etc.	Non-Invasive	In addition to high spatial resolution and exquisite soft tissue contrast, MRI can also display dynamic physiologic changes in three dimensions [47].
Positron Emission Tomography (PET)	Brain, Prostate, Breast, etc.	Invasive	The PET provides information about the organs' activity, as well as its sugar use as energy [48].
Computed Tomography (CT)	Lung, Bone, Oral, etc.	Non-Invasive (harmful)	CT is an excellent tool for detecting bone, joint, and soft tissue lesions that may affect bone, joints, or soft tissues [49].
Ultrasound (US)	Abdomen, Breast, etc.	Non-Invasive	In addition to showing the activity and function of certain organs in the body, US can also identify whether a tissue or organ contains fluid or gas [50].
Optical Coherence Tomography (OCT)	Eye, Heart	Non-Invasive	Biological tissues can be visualized in high-resolution with OCT scanning in two-dimensional or three-dimensional modes [51].
Dermatoscope (Dsc)	Skin	Non-Invasive	Dsc allows better visualization of subsurface structures and improved identification of skin diseases [52].
Color Fundus Photographs (CFP)	Eye	Non-Invasive	CFP monitors the progression of eye disorders using color photographs taken with a fundus camera [53].

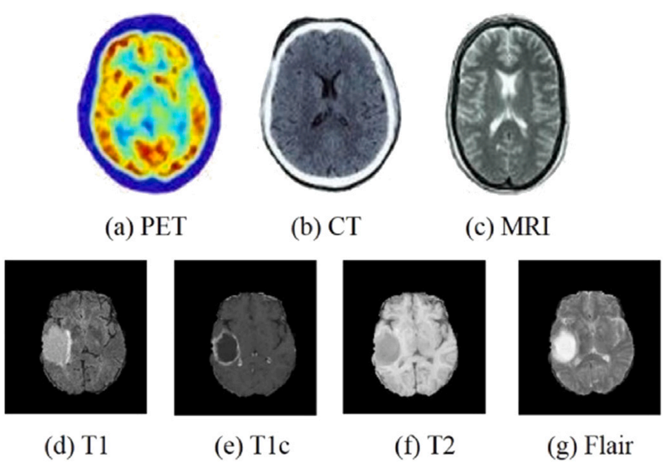

Fig. 5. (a)–(c) are the images of PET, CT, and MRI. (d)–(g) are the different sequences of MRI.
Source: Images from Zhou et al. [19]

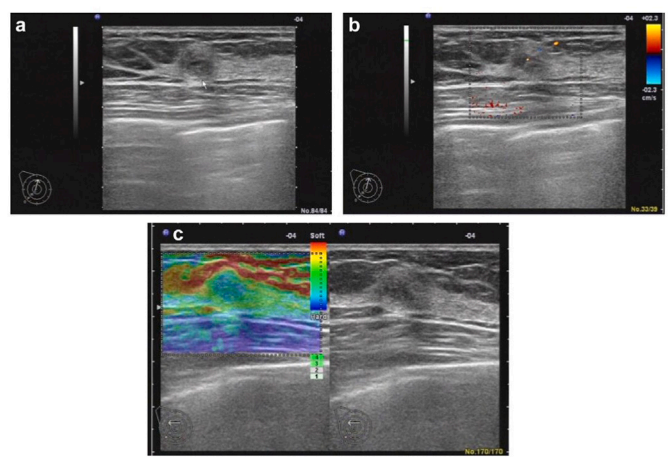

Fig. 6. Breast ultrasound images of a 37-y-old woman with fibroadenoma. (a) Conventional B-mode, (b) Color Doppler, and (c) Strain elastography (SE) image [93].

In the diagnosis of skin cancer, a combination of dermoscopic and clinical images is often used [66]. The clinical image is captured using a digital camera and shows the visualized feature in different views and lighting conditions. On the other hand, dermoscopic images provide a clear view of the skin's subsurface structures and are obtained using a specific skin imaging technique in contact with the skin [94]. Fig. 8 shows examples of the dermoscopic and clinical images.

In addition to multimodal image combinations, clinical information regarding the patient's medical history and symptoms can significantly contribute to the diagnosis of the disease. These data may contain implicit features that may improve the model's classification performance. Electronic Health Records (EHR) are commonly used to detect brain diseases by integrating image analysis features [70,71]. Similarly, skin cancer detection also relies heavily on metadata [66,69].

Fig. 7. Images of CFP and OCT from GAMMA challenge [64].

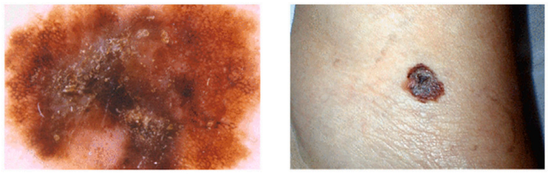

Fig. 8. Dermoscopic and clinical image. Image from public datasets SPC [67].

2.2. Multimodal image datasets

In multimodal medical diagnosis, multimodal datasets are particularly valuable for testing various networks and developing fusion methods. However, the privacy and cost of medical images often make obtaining more comprehensive multimodal datasets challenging for researchers. Fortunately, there are several freely available multimodal datasets. These datasets provide information regarding the diagnosis of diseases at various locations in the body, as well as the analysis of various multimodal combinations. These datasets are expected to contribute to the analysis of fusion methods and serve as a foundation for the future development of multimodal fusion methods.

Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multicenter longitudinal study to discover clinical, imaging, genetic, and biochemical biomarkers for Alzheimer's disease (AD). ADNI has three stages: ADNI 1 included 400 subjects diagnosed with mild cognitive impairment (MCI), 200 subjects with early AD, and 200 elderly control subjects [95]; ADNI 2 added new participant groups: 150 elderly controls, 100 EMCI subjects, 150 late mild cognitive impairment (LMCI) subjects, and 150 mild AD patients [96]; ADNI 3 added hundreds of new MCI subjects, mild AD subjects, and elderly controls [97].

The MRI Brain Tumor Segmentation (BraTS) challenge has been held since 2012 and currently includes classification tasks in addition to tumor segmentation [98]. Each subject has four MRI modalities (T1, T1C, T2, and T2 FLAIR), human annotation of tumor segmentation, and tumor grade.

The Cancer Imaging Archive (TCIA) is a large-scale public database containing medical images of common tumors (lung cancer, prostate cancer, etc.) and corresponding clinical information (treatment protocol, genetics, pathology, etc.) [99].

Open Access Series of Imaging Studies (OASIS) seeks to make neuroimaging datasets freely accessible to the scientific community [100]. OASIS-3 contains 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42–95 years [101].

Seven-point Criteria Evaluation Database (SPC) provides a database for evaluating computerized image-based prediction of the 7-point malignancy checklist for skin lesions. The dataset contains more than 2000 clinical and dermoscopy color images and structured metadata for training and evaluating computer-aided diagnosis systems [67].

As part of the Cancer Genome Atlas (TCGA), an internationally recognized cancer genomics project, more than 20,000 primary cancer samples and matched normal samples were molecularly characterized [102,103].

The Autism Brain Imaging Data Exchange (ABIDE) initiative now includes two large-scale collections, ABIDE I and ABIDE II, whose ultimate goal is to facilitate discovery science and comparative analysis across samples. ABIDE I contains 1112 datasets, including 539 from individuals with ASD and 573 from typical controls (ages 7–64 years, median 14.7 years across groups) [104]. ABIDE II contains 1114 datasets from 521 individuals with ASD and 593 controls (age range: 5–64 years) [105].

ADHD-200 Sample is a grassroots initiative that aims to improve scientific understanding of the neural basis of ADHD through the implementation of open data sharing and discovery-based research methods [106].

The Center for Biomedical Research Excellence (COBRE) is providing raw anatomical and functional magnetic resonance imaging data from 72 patients with schizophrenia and 75 healthy controls (ages ranging from 18 to 65 in each group) [107].

The Glaucoma Grading from Multimodality Images (GAMMA) Challenge is intended to facilitate the development of fundus and OCT-based glaucoma grading [108]. GAMMA contains 2D fundus images and 3D OCT images of 300 patients.

Computational Precision Medicine: Radiology-Pathology Challenge on Brain Tumor Classification 2020 (CPM-RadPath) is a brain tumor classification challenge. There are 221 cases in the training dataset, each with a paired radiology and digital pathology image. Within the 221 cases, there are 54, 34, and 133 cases for lower grade astrocytoma, IDH-mutant, oligodendroglioma, 1p/19q codeletion, and glioblastoma and diffuse astrocytic glioma with molecular features of glioblastoma, IDH-wildtype, respectively [109,110]. The CPM-RadPath 2020 challenge also contains 35 and 73 validation and testing sets, respectively. Each patient contains multiple MRI sequences: T1, post-contrast T1-weighted (T1Gd), T2, and FLAIR.

ISIT-UMR is a dataset for the classification of gastrointestinal lesions in regular colonoscopy. The dataset consists of 76 polyps with white light and NBI videos from the same polyp [111].

The MRNet dataset consists of 1,370 knee MRI exams performed at Stanford University Medical Center between January 1, 2001, and December 31, 2012. There were 1,104 (80.6%) abnormal exams in the dataset, with 319 anterior cruciate ligament (ACL) tears and 508 meniscal tears [112].

CTU-CHB Intrapartum Cardiotocography is a database containing 552 cardiac tomography recordings from the Czech Technical University (CTU) in Prague and the University Hospital in Brno (UHB). As part of each CT, a fetal heart rate time series (FHR), as well as a uterine contraction (UC) signal, are recorded [113].

The previously mentioned datasets provide valuable resources for developing and testing multimodal fusion methods. They contain images of different medical modalities of the same patient, as well as images of different patients. Access to these datasets is available upon request and at no cost. In this review, we summarize the fusion methods presented in 53 articles that use ADNI, 11 articles that use TCIA, 7 articles that use BraTS (2015, 2017, 2019 and 2021 editions), 7 article that uses OASIS, 4 articles that use COBRE, 4 articles that use SPC, 4 articles that use ABIDE, 3 articles that use ADHD-200, 2 articles that use CPM-RadPath (2020 edition), 2 articles that use GAMMA, 2 articles that use MRNet, 1 article that uses TCGA, 1 article that uses CTU-UHB and 1 article that uses ISIT-UMR. As mentioned earlier, 19 papers discussed in this review are not based on public datasets.

3. Multimodal classification pipeline

Multimodal fusion of biomedical data using deep learning remains an evolving field. The terminology used to describe fusion methods often varies between publications, leading to ambiguity. For instance, terms like input, intermediate, and output fusion are commonplace, but their interpretations may differ. To bring clarity and standardization to the multimodal classification area, we adopt the five-stage pipeline proposed in Sleeman et al. [114], referenced in Table 3. This pipeline offers a structured approach to encapsulate all medical multimodal classification tasks. Within this section, we elucidate each of these stages, detailing their definitions and the methodologies for their implementation. Subsequently, based on the sequence and structure of the information fusion stage paired with the deep learning (DL) backbone stage, we categorize multimodal fusion techniques into five distinct strategies in Section 4.

Table 3
Multimodal classification pipeline.

Stage	Description
Data preprocessing	The initial step of the classification task is to perform operations such as registration, denoising, and data augmentation on the raw data.
DL backbone	Extraction of high-dimensional features of data by the deep learning network structure.
Information fusion	Fusion of multimodal data/features by different methods.
Final classifier	The final stage of generating classification results from multimodal data.
Model evaluation	Different metrics are used to evaluate the performance of multimodal models.

3.1. Pre-processing

Image pre-processing is crucial for multimodal medical classification tasks, as it enhances DL network efficiency and effectiveness in extracting features. Pre-processing techniques, such as image registration, cropping, denoising, resampling, intensity normalization, regions-of-interest (ROI) extraction [115–117], and feature selection [4,79,118,119], prepare the data for more accurate and efficient analysis by DL models.

To further improve the performance of these models, data augmentation techniques play an essential role in the pre-processing pipeline. For example, data augmentation helps prevent overfitting [120] using methods like random cropping, flipping, and rotation during training. In addition, increasing the training dataset's diversity improves the model's generalization capabilities.

Considering the large volumes of data generated by multimodal medical images, it is noteworthy that only a small fraction is relevant to diagnosing diseases. Therefore, feature selection emerges as a crucial pre-processing step, aiming to reduce data dimensionality while retaining pertinent information. Common feature selection methods include manual selection [121–123] and Principal Component Analysis (PCA) [124–126].

Another critical aspect of pre-processing multimodal medical images is image registration. It involves aligning images from different modalities (e.g., MRI, CT, and PET) into a common coordinate system, enabling the accurate matching of corresponding anatomical structures across image types [20,127]. Such alignment facilitates comprehensive data analysis and becomes particularly critical for input-level fusion, where combining complementary information from different modalities requires proper alignment [10].

Image registration in this context presents several challenges. A significant one is the lack of ample training datasets for supervised deep learning. Another is defining accurate similarity measures, especially with the varied appearance of different modalities. The registration process can be further complicated when trying to align images from different patients or even the same patient over time due to factors like changes in anatomy and metabolic processes.

In tackling these challenges, deep similarity metrics have shown promise, especially in traditional frameworks. While multimodal registration has seen advancements, direct transformation prediction lags behind, especially compared to single-modality methods. One innovative solution is using Generative Adversarial Networks (GANs) to make multimodal images more consistent.

3.2. Information fusion

A key component of multimodal image classification is information fusion. Based on the level at which information is fused, information fusion can be divided into input fusion, intermediate fusion, and output fusion. And there are two ways to achieve fusion [114], namely concatenation and merge. Concatenation involves the concatenation of data from different modalities into a single tensor for the next step. Merge involves complex calculations such as adding data from different modalities, and the final result is a smaller amount of data. Fig. 9 illustrates the two types of fusion. Our study focuses on the fusion of different medical imaging modalities, and in Section 4, we will examine the different fusion methods in greater detail.

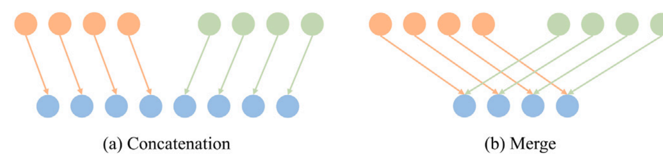

Fig. 9. Two types of fusion. Orange and green: data of different modalities. Blue: the output fused data.

3.3. Deep learning backbone

DL backbones are used to extract high-dimensional features of modalities during the classification process. Over recent years, several high-performing network architectures have emerged, including AlexNet [128], VGG [129], GoogLeNet [130], ResNet [131], DenseNet [132], AE [133–135], ViT [136], and others, providing state-of-the-art performance in classification. A summary of the common architectures for DL is presented in Table 4. DL has developed rapidly due to several factors, including the development of hardware devices like graphics processing units (GPUs) and tensor processing units (TPUs), which have greatly improved the training speed of DL networks. Additionally, publicly available datasets such as ImageNet [137] have facilitated the training and testing of various models. Furthermore, DL is capable of learning advanced features directly from data without requiring extensive expertise or prior experience, making it easily adaptable across various domains.

In input fusion, a single backbone can extract features from fused modalities. However, in other fusion schemes such as intermediate or output fusion, multiple DL backbones may be used to extract features from different modalities. In current multimodal fusion research, Convolutional Neural Networks (CNN) are the preferred choice of the majority of researchers due to their effectiveness in feature extraction from medical images. Many pre-trained models have already been tested on large datasets, making them suitable for use in medical imaging research. In the articles analyzed, CNNs were used in 65 articles, Fully Connected Neural Networks (FCNN) in 10 articles, Auto-Encoders (AE) in 8 articles, and Transformers in 6 articles.

Table 4
Some common architectures of deep neural networks. Different architectures are more suitable for different types of data.

Architecture	Description
Fully Connected Neural Network (FCNN)	FCNN are the most traditional deep neural networks. Every neuron in a layer is connected to every neuron in the layer below it [138].
Convolutional Neural Network (CNN)	CNN can model spatial structures, such as images or volumes. Convolutional kernels model local information by sliding over input data [138].
Autoencoders (AE)	By compressing and reconstructing the input data, AE learns low-dimensional encoding. There are different types of layers, such as convolutional and fully connected [139].
Transformer	Transformer is a model that uses a multi-headed attention mechanism. Feature extraction is solely based on attention [140].

3.4. Final classifier

Multimodal classification employs a final classifier to generate the classification results based on multimodal features or multiple independent classification results, depending on the employed fusion scheme. In DL networks, the Fully Connected (FC) layer [10,63,65,79,80,141] is often used as the final classifier. Other methods, such as SVM [124,142], Random Forest [5], and Score Merge [143,144] can also be used as final classifiers.

3.5. Evaluation metrics

Evaluation metrics for multimodal fusion tasks are similar to those used in unimodal classification tasks. Commonly used indicators for assessing the performance of multimodal fusion methods and DL networks in the context of medical classification tasks include True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These indicators can be used to calculate several performance metrics, such as sensitivity, specificity, accuracy, precision, and F1 score, among others. Additionally, AUC and Kappa are commonly used metrics to evaluate medical classification tasks.

- Accuracy (ACC) = $\frac{TP+TN}{TP+TN+FP+FN}$
- Sensitivity (SEN) = $\frac{TP}{TP+FN}$
- Specificity (SPEC) = $\frac{TN}{TN+FP}$
- F1 Score = $\frac{2\times TP}{2\times TP+FP+FN}$
- Positive Predictive Value (PPV) = $\frac{TP}{TP+FP}$
- Negative Predictive Value (NPV) = $\frac{TN}{TN+FN}$
- Area Under the receiver operating characteristic Curve (AUC)
- Cohen's Kappa (Kappa) = $\frac{p_0-p_e}{1-p_e}$

where p_0 is the accuracy and p_e the sum of the products of the actual and predicted numbers corresponding to each category, divided by the square of the total number of samples.

4. Multimodal classification networks

4.1. Information fusion taxonomy for multimodal image classification

The positions of pre-processing and the final classifier are fixed during the process of multimodal classification. Based on the number and sequence of DL backbones and information fusion step, multimodal DL network architectures can be categorized into five types: input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion, as shown in Fig. 10. As explained hereafter, single-level, hierarchical, and attention-based fusion are sub-categories of intermediate fusion. These categories describe how the network processes and combines the input modalities to produce classification results.

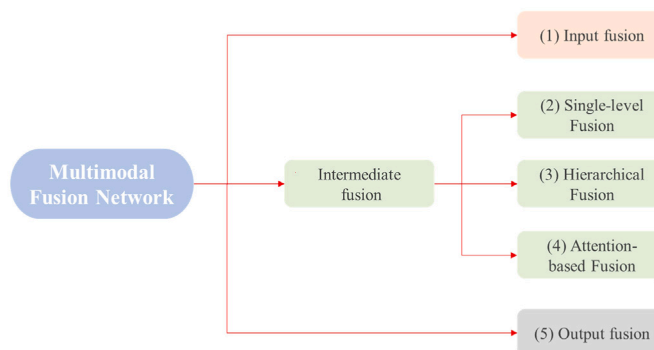

Fig. 10. Five types of multimodal fusion networks.

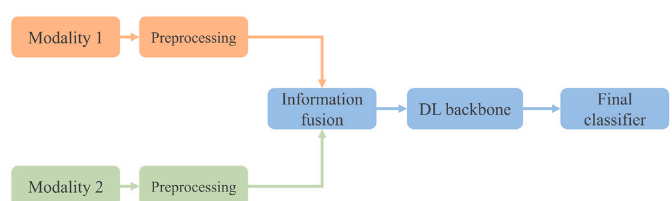

Fig. 11. Input fusion process diagram. Information fusion: Concatenation/Merge (Inputs).

(1) **Input fusion** can also be referred to as input-level fusion, where the information fusion phase precedes the DL backbone. Concatenation and Merge are two methods of information fusion. For the concatenation method, data of different modalities are used as different channels of the input. In the merge approach, data is fused at the pixel or voxel level, and the merged images are used as inputs for the DL classifier. The process diagram for input fusion is shown in Fig. 11.

(2) **Single-level fusion** involves information fusion after the DL backbone but before the final classifier. As part of a single-level fusion, the features extracted by the DL backbone are fused only once at some point before the classifier is applied. It can be divided into two types: Classic Fusion and Network Fusion, depending on the network structure. In Classic Fusion, high-dimensional features are extracted from different modalities using different DL classifiers and then merged or concatenated. This is the most common network structure in intermediate fusion, so we call it *Classic*. Fig. 12 illustrates the process

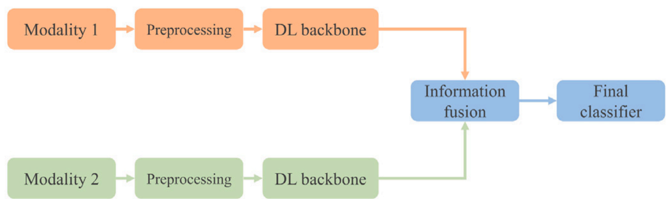

Fig. 12. Single-level fusion process diagram. Information fusion: Concatenation/Merge (Classic).

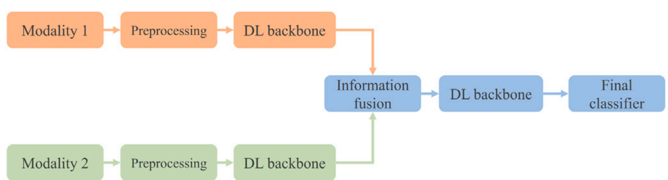

Fig. 13. Single-level fusion process diagram. Information fusion: Concatenation/Merge (Network).

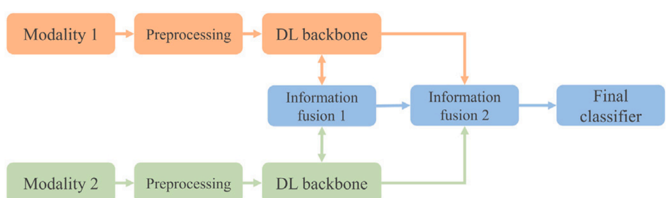

Fig. 14. Hierarchical fusion process diagram. Information fusion 1: Concatenation/Merge (Network). Information fusion 2: Concatenation/Merge (Classic).

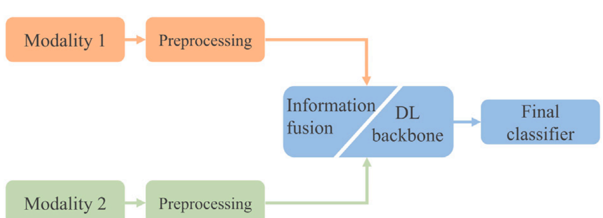

Fig. 15. Attention-based fusion process diagram.

diagram of classic fusion. In network fusion, the intermediate features of different modalities are first extracted using DL classifiers, followed by the extraction of high-level features of the fused modalities using additional DL backbones. Fig. 13 shows the process diagram of network single-level fusion.

(3) **Hierarchical fusion** is an improvement over single-level fusion. In this approach, DL backbone extracts features from the data of different modalities, while features from each level are then fused at the network level by concatenation or merging. Additionally, further feature fusion is performed following the DL backbone. This allows for more complex feature combinations to be learned, improving classification accuracy. The process diagram for output fusion is shown in Fig. 14.

(4) The emergence of Transformers has led to the development of **Attention-based fusion** as a new network architecture. Through its unique DL backbone, this architecture is able to extract features and implement feature fusion based on the attention relationship between different modalities. Fig. 15 illustrates the process of attention-based fusion. A more detailed analysis of the network architecture is presented in Section 4.5.

(5) **Output fusion**, also known as decision-level fusion or late fusion, involves the use of DL backbones to extract high-dimensional

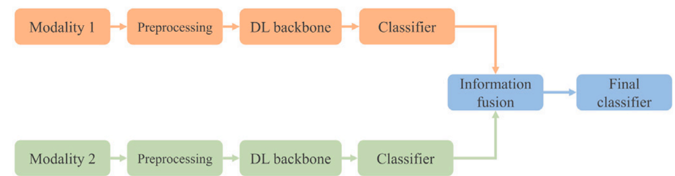

Fig. 16. Output fusion process diagram. Information fusion: Merge (Outputs).

features from different modalities of data. The extracted features are then used to generate separate classification results for each modality. These results are then combined using a fusion technique, such as majority voting or averaging, to produce a final classification result. The process diagram for output fusion is depicted in Fig. 16.

Recent years have seen a growing trend toward the use of deep learning networks in multimodal fusion research. Fig. 1 illustrates the distribution of five fusion strategies in the scope of the study. In contrast to traditional methods, single-level fusion is the most commonly used method in DL multimodal fusion, followed by input and output fusion. Hierarchical fusion and attention-based fusion are also gaining attention and present great potential for research. These more recent fusion methods offer more complex ways of combining modalities, enabling deep learning networks to learn more powerful representations of multimodal data.

4.2. Input fusion networks

Input fusion combines data from multiple modalities into a single feature tensor fed into the deep neural network as an input. Input fusion typically involves the fusion of modalities with similar structures, making implementation relatively straightforward. Some modalities can be acquired together at the time of clinical photography (e.g., CT and PET). In many cases, these modalities have the same voxels and spacing after data processing, making obtaining registered multimodal data easy. Furthermore, the majority of input fusion tasks do not require re-modeling, only modifying the input part of the unimodal model to achieve multimodality. Fusion can be accomplished in three ways: concatenating or merging multimodal medical images, extracting high-dimensional features from multimodal images, and then fusing them.

(1) The registered multimodal data are fed into the DL classifier as input for different channels to obtain classification results, which is the most common input fusion approach. Fig. 17 illustrates this typical input fusion network architecture used in the research of Aldoj et al. [145], Lin et al. [146], Zong et al. [147], Zhou et al. [148]. Aldoj et al. [145] proposed a semi-automatic method for the classification of prostate cancer without feature selection. Several combinations of 3D volumes (e.g., ADC, DWI, and T2) are utilized as inputs of the CNN network. Each sequence is considered an input channel; the output is the classification of significant versus nonsignificant lesions. Lin et al. [146] employed MRI and PET to diagnose Alzheimer's disease. PET and MRI are used as two channels for the input of the CNN classification network, based on an ROI crop model to learn a classifier and fuse different features from MRI and PET. Zong et al. [147] concatenated T2, ADC and DWI for tumor foci classification using an end-to-end CNN network. In order to diagnose triple-negative breast cancer, Zhou et al. [148] concatenated manually segmented multiparametric MRI images (PEI, DWI) into a CNN network. Despite the ease of implementing this fusion architecture, it has some limitations with regard to the modal data requirements. For instance, the registration performance of different modal data can influence the classification results. Moreover, this approach is not suitable for fusing heterogeneous data, such as 3D medical images and 1D clinical records, which have different characteristics and dimensions.

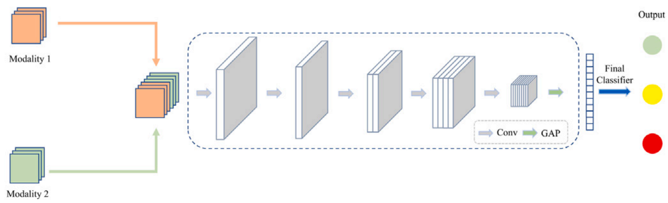

Fig. 17. Schematic diagram of the network architecture for input fusion. Information fusion method: Concatenation (Inputs).

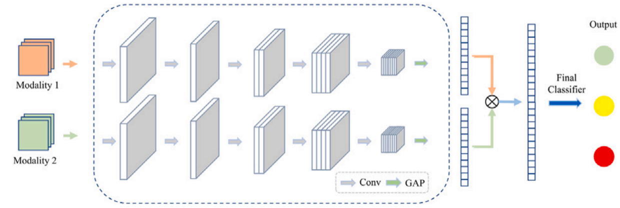

Fig. 18. Schematic diagram of the network architecture for classic single-level fusion. Information fusion method: Concatenation (Classic).

(2) The merging of images is another input fusion method in addition to concatenation. Various image modalities are fused at the pixel or voxel level in order to create a new fused image that is used for classification [1,149,150]. Song et al. [149] proposed an effective multimodal image fusion method for Alzheimer's disease diagnosis using MRI and PET. Through registration and mask coding, they were able to fuse gray matter (GM) and 18-fluorodeoxyglucose positron emission tomography (FDG-PET) images to create a new imaging modality called "GM-PET". In the resultant composite image, the GM area is clearly highlighted, allowing AD diagnosis to be made while maintaining both the contour and metabolic characteristics of the subject's brain tissue. They then fed the fused images to the CNN for classification. The GM region cropped from the MRI image is mapped onto the PET image, resulting in the fusion of PET and MRI data in Kong et al. [1] research. In addition to providing anatomical and metabolic information about the brain, the fusion modality also allows the viewer to focus on the main features of the brain by reducing the visual noise. Rallabandi and Seetharaman [150] employed a fusion approach integrating images from MRI and PET for the diagnosis of Alzheimer's disease. The fusion process involved applying two-dimensional Fourier and discrete wavelet transform (DWT) to combine MRI and PET images. Subsequently, the MR-PET fused image was reconstructed using inverse Fourier and DWT methods. The benefit of fused images is that they contain a wealth of medical information, but the process of generating them often requires an extensive amount of prior medical knowledge.

(3) Some studies have performed input fusion after extracting features from multimodal images instead of performing a direct fusion of medical images [59,124]. Li et al. [124] used PCA to extract features from MRI, PET, and cerebrospinal fluid (CSF) and then concatenated these features into the Restricted Boltzmann Machine (RBM) network for the diagnosis of Alzheimer's disease. Liu et al. [59] manually extracted features from MRI and PET and then used stacked auto-encoder (SAE) to classify the concatenated multimodal features in order to diagnose Alzheimer's disease. The architecture of extracting features and combining them can solve the problem of multimodal heterogeneity. However, PCA-based or manual feature extraction requires prior knowledge and does not fully utilize image information.

In input fusion, fused data is used in single-branch feature extraction, and the network architecture design significantly reduces network parameters and deployment difficulties. However, due to the fusion of the data at the input level, the complementary information from the different modalities is not utilized to the fullest extent possible.

4.3. Single-level fusion networks

The single-level fusion process uses different DL backbones to extract features from different modalities separately, followed by an information fusion process before making the final decision. Based on the position of information fusion within the network architecture, it can be divided into classic fusion structures and network fusion structures.

(1) The most common single-level fusion architecture is to extract features from multimodal data by using different branches, then fuse these features and feed them to the final classifier [8,55,56,63,69,72,121,133,134,142,151–157]. A schematic diagram of its network architecture is shown in Fig. 18. After preprocessing the data, the architecture [121] extracted low-level 3D features from fMRI and sMRI to classify Attention Deficit Hyperactivity Disorder (ADHD) automatically. As soon as the features are concatenated, softmax classifiers are used to differentiate ADHD cases from typically developing children (TDC) cases. In order to diagnose breast cancer, Joo et al. [152] fused MRI (T1, T2) and clinical information. Two 3D ResNet-50 were used to extract features from contrast-enhanced T1 subtraction MR images and T2 MR images, while the FC layer provided clinical inputs. For the prediction of pathological complete response, the outputs of each 3D ResNet-50 and FC layer were concatenated, and the final FC layer with sigmoid activation function was used. Likewise, Yap et al. [69] employed ResNet and FC layers to extract features from DSC, Clinical Image, and metadata, then applied FC layers for skin lesion classification. Aside from these methods of concatenating modal features, complex computations can also be used to merge features. Xiong et al. [155] used visual field (VF) and OCT for the diagnosis of glaucoma. VFNet and OCTNet were used to extract features from the VF and OCT modes, respectively. A weighted average was used to obtain an aggregated representation from bimodal features using an attention module. Each modal feature was assigned a weight using a fully connected layer, followed by a sigmoid function to calculate a scalar value (0–1) indicating the feature's relative contribution to the aggregate representation. To aggregate all features, a global average pooling layer was also used. The results of glaucoma diagnosis were predicted using three FC layers and a softmax layer. For CT and PET modalities, Qin et al. [63] extracted features using CNN networks, merged the features using gated multimodal units (GMU), and classified lung cancer using FC layers. GMU, unlike the widely used connection operation, allows for the learning of intermediate representations of multimodality features using hidden structures and gate controls, thus enabling the prediction layer to assign weights more effectively to intrinsically associated features.

(2) Two stages can be described as the single-level fusion architecture for network fusion. The first stage involves extracting single-level features separately from different modalities using DL backbones, followed by the second stage of information fusion which involves utilizing an additional DL backbone to extract high-level features from the fused features [122,123,158–163]. Lastly, the extracted high-level features are used in the final classification process. Fig. 19 illustrates a typical network fusion architecture. Cheng and Liu [159] used cascaded CNN for the multimodal fusion of MRI and PET to diagnose Alzheimer's disease. They proposed a 2D CNN to combine the multimodality features and make the final classification. After 3D CNN output features are flattened to one dimension, the 1D feature vectors of MRI and PET are combined to produce a two-dimensional feature map for 2D CNN analysis. Kim and Lee [123] developed a multimodal architecture for combining MRI, PET, and CSF features. Each modality's individual representation of high-level features is calculated using the stacked sparse extreme learning machine auto-encoder (sELM-AE). Another stacked sELM-AE is used to get the joint features from the high-level MRI, PET, and CSF features. The kernel-based extreme learning machine classifies the joint feature representation. With multimodality

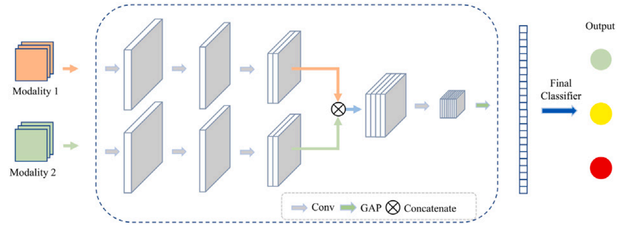

Fig. 19. Schematic diagram of the network architecture for single-level network fusion. Information fusion method: Merge (Network).

neuroimaging and genetic data, Zhou et al. [164] proposed a three-stage deep feature learning DNN framework for Alzheimer's disease classification. Each modality's latent representation is learned in the first stage, then each pair of modalities' joint latent representations are learned in the second stage, and in the third stage, each pair of modalities' joint latent representations are used to create the classification model. Rahaman et al. [160] classified schizophrenia using sMRI, fMRI, and single nucleotide polymorphisms. The latent representations for the static functional network connectivity (sFNC), sMRI, and single nucleotide polymorphism (SNP) are learned using an autoencoder, multi-layered perceptron, and bi-directional long short-term memory (LSTM). The Multimodal Bottleneck Attention Module performs the fusion of the embeddings and then sends the combined embeddings to a variational autoencoder for encoding, followed by a Softmax layer for classification.

The single-level fusion method is currently used to merge multiple medical modalities for classification tasks and can be applied to the fusion of different medical modalities. The method does not require a specific format for the data as it extracts features from modalities using different branches and fuses data at a high-dimensional feature level. In this regard, single-level fusion is a suitable solution for unregistered or different dimensional data. Due to the fact that information fusion occurs only at the end of the network architecture, single-level fusion is not capable of analyzing low-dimensional features jointly.

4.4. Hierarchical fusion networks

Hierarchical fusion extends single-level fusion further in order to further exploit the complementary information between multimodal data. The hierarchical fusion process involves the fusion of different dimensional features and the classification of these jointly represented features through the process of fusion [10,57,58,126,141,165–170]. There are two ways to implement hierarchical fusion: by using additional branches for multimodal feature fusion or by using fusion blocks for joining features from different modalities.

(1) The common hierarchical fusion architecture involves extracting different modalities via different branches and simultaneously combining multimodal features of different dimensions via another parallel branch. Finally, the high-dimensional features from the fusion branch and each modal branch are combined for classification. Fig. 20 shows a typical network architecture for hierarchical fusion, Zhou et al. [126], Zhang and Shi [166], Li et al. [10] utilized this network architecture. Zhou et al. [126] utilized three sparse-response Deep Belief Network (DBN) branches to extract features from PET/MRI modalities, fuse them, and then employed an Extreme Learning Machine (ELM) to classify the fused features for brain diseases. Zhang and Shi [166] used a deep multi-modal fusion network (DMFNet) to fuse PET and MRI data for the diagnosis of Alzheimer's disease. Three branches are present in DMFNet, two of which extract features from the MRI and PET scans, respectively. A channel attention model is used to extract the features from each branch and merge the reweighted feature maps. In the third branch, fused features are further extracted. Li et al. [10] used a three-branch CNN network to combine 2D fundus images with

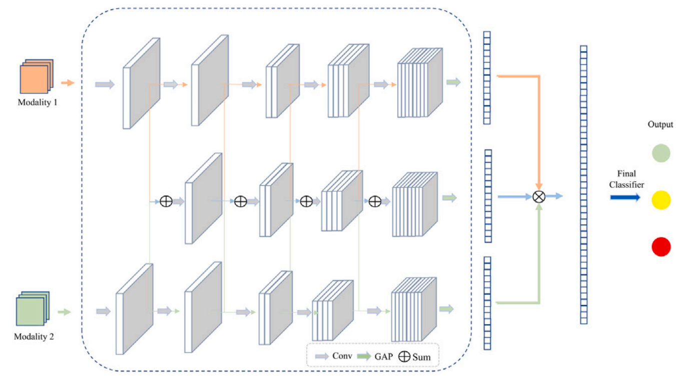

Fig. 20. Schematic diagram of the network architecture for hierarchical fusion. Information fusion method: Merge (Network) and Concatenation (Classic).

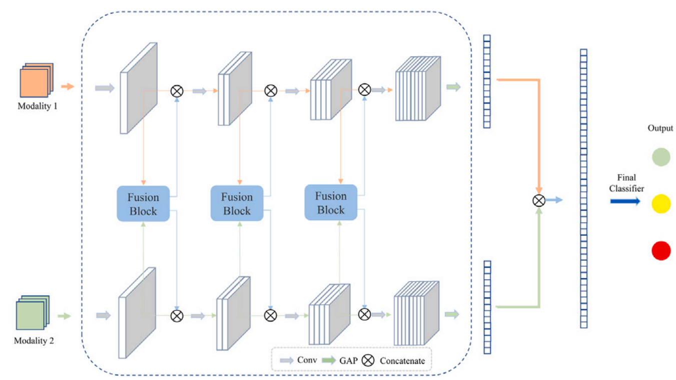

Fig. 21. Schematic diagram of another network architecture for hierarchical fusion. Information fusion method: Merge (Network) and Concatenation (Classic).

3D OCT images in order to classify glaucoma and diabetic retinopathy. The fusion of 2D and 3D data features on the fusion pointers was achieved by changing the dimensionality of the data features using a transformation layer.

(2) Hierarchical fusion can also be structured in another way by extracting features using different branches and fusing them in different dimensions by using fusion blocks, which are then returned to each modality branch for further fusion. The design of such a network structure can reduce the number of model parameters while fusing features at multiple levels. Fig. 21 illustrates a typical network architecture, Gao et al. [171], He et al. [141] utilized this network architecture. To classify brain diseases, Gao et al. [171] proposed a pathwise transfer deep convolution network that gradually learned and combined the multi-level and multimodal features of MRI and PET. The pathwise transfer blocks are designed to fully utilize complementary information from different imaging modalities. Pathwise transfer blocks are used to communicate information across PET and MRI, which helps to improve the classification model's performance. He et al. [141] proposed a multimodal MRI hierarchical-order multimodal interaction fusion network (HOMIF) to diagnose gliomas. There are two branch networks for each modality, several multimodal interaction modules with different scales and orderings, diverse learning constraints, and a predictive subnet in the framework. Each branch network has three CNN blocks with multiscale inputs and an arm with diverse high-order multimodal interaction (HOMI) modules to integrate and interact deeply with the multiscale features.

The multi-level feature fusion allows hierarchical fusion to explore more fully the complex and complementary information between modalities. Learning the synergy of multimodal data while maintaining the features of the modalities improves the model's classification performance [166]. However, as it involves the fusion of low-dimensional features, the registration of multimodal data may affect the classification performance of hierarchical fusion.

4.5. Attention-based fusion networks

As attentional mechanisms [172] have been proposed and developed, more and more studies are beginning to incorporate attentional mechanisms into network architectures. Some of the network architectures mentioned above also included attention mechanisms in order to enhance the performance of the models. Zhang and Shi [166] added attention modules to reweight the modal features. Xing et al. [136] use a vision transformer (ViT) to extract the modal features and fuse them. These studies, however, only operate on unimodal modalities and do not utilize the attention mechanism for multimodal interactions. Recently, some studies have used the attention mechanism to extract and combine features [61,68,173–181]. This network architecture is called attention-based fusion, which is not related to any of the previous fusion architectures.

In the study of Dai et al. [173], they propose TransMed, which combines CNN and transformer to capture high-level cross-modalities and low-level features. First, TransMed sends the multimodal images to CNN, where they are processed as sequences, then transformers learn the relationships between them and predict the end result. TransMed is more efficient and accurate than existing multimodal fusion methods because it effectively models the global features of multimodal images.

Attention-based Hierarchical Multimodal Fusion (AHM-Fusion) is a novel fusion module Qiu et al. [174] designed. The system includes both an early feature guidance module and a late feature fusion module, capturing deep interaction information between different multimodal features. In the early stage of feature aggregation, the early feature guidance module is used to capture multimodal interactions. To obtain classification results, late feature fusion modules based on attention mechanisms are used. Through cascading double attention layers in the late feature fusion module, the deep interaction information is further captured. Then, they used a gating-based attention mechanism to decrease the impact of insignificant features in each modality.

Zhang et al. [175] proposed a multimodal Medical Information Fusion (MMIF) framework that combines the Category Constrained-Parallel ViT framework (CCPViT) and the multimodal Representation Alignment Network (MRAN) as backbones, enabling the modeling of images and texts as unimodal features, as well as cross-modal features. CCPViT is proposed as a tool for learning key features of different modalities and for solving unaligned multimodal tasks. hen in MRAN, Cross-attention was used to cascade encoded images and decoded texts to explore deep-level interactive representations of cross-modal data, assisting with modal alignment and identifying abnormalities. MMIF is an image-text foundation modeling that could contribute to a much higher-precision classification model when compared with unimodal models.

Multimodal Mixing Transformer (3MT) was presented Liu et al. [176] as a novel technique to classify diseases. Based on neuroimaging data, gender, age, and the Mini-Mental State Examination (MMSE), They tested it for Alzheimer's Disease classification. Multimodal information is incorporated through a Cascaded Modality Transformers architecture with cross-attention. Different embedding layers are used to obtain Key (K) and Value (V) from imaging features and clinical data. K and V are then placed into a cross-attention layer with a latent code known as Query (Q). 3MT allows mixing an unlimited number of modalities and formats and full data utilization.

Zuo et al. [179] has introduced a novel Swapping Bi-Attention Mechanism (SBM) designed for the diagnosis of Alzheimer's disease through the amalgamation of structural-functional brain images. The proposed model capitalizes on the transformer's bi-attention mechanism to explore mutually beneficial information inherent in both structural and functional images. Historically, transformers have been investigated solely within the confines of single-modality brain regions, neglecting the potential for leveraging complementary information across modalities. SBM, however, implements token exchange between the two modalities and adaptive fusion of intermediate features during the application of the Transformer for feature extraction from diverse modalities. This process facilitates the collaborative exchange of information embedded in bimodal images, enhancing the transparency of feature alignment and fusion. The incorporation of SBM results in a more lucid understanding of the influence of different modalities on feature extraction and multimodal fusion.

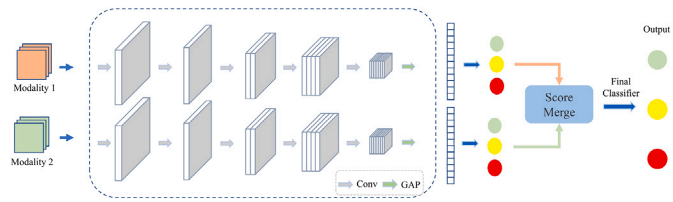

Fig. 22. Schematic diagram of the network architecture for output fusion. Information fusion method: Merge (Outputs).

Research is increasingly incorporating attention mechanisms, particularly Transformer structures, into multimodal classification tasks. While performing cross-modal attention computation, a multi-level fusion of multimodal features is achieved. Furthermore, the Transformer structure is well-suited for joining modalities of different dimensions. Nevertheless, Transformer research in medical tasks is still in its infancy, and various studies tend to focus on solving particular problems, making it difficult to conclude a general multimodal classification architecture. A further important point to be noted is that while the success of Transformer is accompanied by pre-training on large datasets, the number of samples in medical datasets is often not sufficient to achieve the good training effect of Transformer. As a result, it is recommended that Transformer and CNN are used together in a hybrid fashion.

4.6. Output fusion networks

In output fusion, each modality uses a separate DL backbone to extract features and make decisions, and the results are merged into one final decision. Fig. 22 shows a typical network architecture for output fusion. The final Classifier of decision fusion can be achieved by simple operations [182,183] such as voting, weighting, and averaging, or by classifiers [11,60,184–186] such as SVM, extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), Categorical Boosting (CatBoost), Decision Tree, and K-nearest neighbor (KNN). Moon et al. [182] used unweighted average, weighted average, weighted voting, and stacking to fusion the classification results from different modalities of the US to identify breast tumors. Guo et al. [183] applied a linear weighted module to assemble the predicted probabilities of the pre-trained models based on the 4 MRI modalities for the classification of gliomas. In order to achieve the diagnosis of early glottic cancer, Kwon et al. [185] used decision trees to combine the classification results from the sound data and the image data. Abdolmaleki and Abadeh [184] used SVM, KNN, and linear discriminant analysis (LDA) to fuse the classification results of fMRI and sMRI to diagnose ADHD.

The output fusion process involves combining unimodal results from different modalities. As a result, it is relatively easy to implement and generally does not require additional training. It is, however, difficult to exploit the complementary information between different modalities because there is no feature fusion. Furthermore, output fusion may not improve classification performance if there are large differences in classification performance between different modalities.

Table 5
Comparison of the results of different fusion methods on ADNI dataset. In the multi-classification task, 3 classes is NC vs. MCI vs. AD and 4 classes is NC vs. ncMCI vs. cMCI vs. AD. Unit:%.

Research	Year	Fusion Methods	Dataset	NC vs. AD	NC vs MCI	Multi-classification
Liu et al. [59]	2015	Input Fusion	331 subjects: 77 NC, 102 ncMCI, 67 cMCI, 85 AD	ACC: 91.40 SEN: 92.32 SPE: 90.42	ACC: 82.10 SEN: 60.00 SPE: 92.32	4 Classes ACC: 53.79 SEN: 52.14 SPE: 86.98
Song et al. [149]	2021	Input Fusion	381 subjects: 126 NC, 160 MCI, 95 AD	ACC: 94.11 SEN: 93.33 SPE: 94.27	ACC: 85.00 SEN: 84.69 SPE: 85.60	3 Classes ACC: 71.52 SEN: 55.67 SPE: 83.40
Kong et al. [1]	2022	Input Fusion	370 subjects: 130 NC, 129 MCI, 111 AD	ACC: 93.21 SEN: 91.43 SPE: 95.42	ACC: 86.52 SEN: 94.34 SPE: 81.64	3 Classes ACC: 87.67
Suk et al. [142]	2014	Single-level Fusion	398 subjects: 101 NC, 128 ncMCI, 76 cMCI, 93 AD	ACC: 95.35 SEN: 94.65 SPE: 95.22	ACC: 85.67 SEN: 95.37 SPE: 65.87	–
Shi et al. [122]	2017	Single-level Fusion	202 subjects: 52 NC, 56 ncMCI, 43 cMCI, 51 AD	ACC: 97.13 SEN: 95.93 SPE: 98.53	ACC: 87.24 SEN: 97.91 SPE: 67.04	4 Classes ACC: 57.00 SEN: 53.65 SPE: 85.05
Zhang et al. [188]	2019	Single-level Fusion	392 subjects: 101 NC, 200 MCI, 91 AD	ACC: 98.47 SEN: 96.58 SPE: 95.39	ACC: 85.74 SEN: 90.11 SPE: 91.82	–
Abdelaziz et al. [189]	2022	Single-level Fusion	959 subjects: 264 NC, 273 ncMCI, 204 cMCI, 218 AD	ACC: 98.24 SEN: 98.82 SPE: 97.52	ACC: 94.59 SEN: 90.26 SPE: 96.98	–
Zhang and Shi [166]	2020	Hierarchical Fusion	500 subjects: 163 NC, 113 ncMCI, 105cMCI, 119 AD	ACC: 95.21 SEN: 93.56 SPE: 97.48	–	4 Classes ACC: 86.15
Fang et al. [60]	2020	Output Fusion	398 subjects: 101 NC, 204 MCI, 93 AD	ACC: 99.27 SEN: 95.89 SPE: 98.72	ACC: 90.35 SEN: 88.36 SPE: 92.56	–

5. Discussion

5.1. Which fusion method is the best?

The choice of a fusion method is crucial when dealing with multimodal medical classification problems. Fortunately, many fusion architectures have been evaluated on the same dataset: ADNI. Based on quantitative results reported by the authors, comparisons between fusion architectures is possible to some extent. We consider studies performed on ADNI where MRI and PET are used to diagnose Alzheimer's disease. There are three stages in the progression of Alzheimer's disease: normal cognition (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). In spite of the fact that MCI does not significantly interfere with daily activities, a high risk of AD progression has been consistently demonstrated in patients with MCI [187]. MCI subjects can be classified into MCI converters (cMCI) and MCI non-converters (ncMCI) to predict the transition risk of MCI. Table 5 reports the results of the different fusion methods obtained by their authors for different classification tasks. The experiments were not replicated: when comparing these results, it should be noted that each paper relies on a different subset of patients, although the number of subjects was similar.

In general, we believe that deep multi-level fusion can better exploit the synergy of multimodal data to produce better classification results. This is further supported by the results in Table 5. Compared with the input fusion, the single-level fusion has a more robust feature fusion, which improves the overall ACC of the middle fusion. Hierarchical fusion utilizing multi-level feature fusion did not significantly improve the performance of dichotomous classification but performed well for four-class classifications. Generally, a complex model does not improve performance much when applied to a simple classification task. The more complex the network, the better it is at solving complex classification problems. When the number of categories for multi-category classification increases from two to four, the hierarchical fusion classification accuracy improves significantly. Last but not least, we note that the output fusion achieves excellent results on NC versus AD classification, thanks to the pre-training of different modal branches. With output fusion, DL backbones can be pre-trained on a large number of unimodal datasets and then fine-tuned on the multimodal datasets. Similar results are reported on other datasets. ABIDE data was combined with sMRI and fMRI to diagnose autism spectrum disorders. It was found that the hierarchical fusion [80] result was 87.2%, which was better than the input fusion [79] result of 65.5%. Rahaman et al. [160] used the COBRE dataset for the diagnosis of schizophrenia, and the accuracy of input fusion, output fusion, and single-level fusion was 70%, 78%, and 95%, respectively. Based on the GAMMA dataset, Li et al. [10] achieved 63% accuracy in input fusion, 72% accuracy in single-level fusion, and 80% accuracy in hierarchical fusion when they used the same dataset for glaucoma diagnosis.

It is difficult to determine a unified solution for a wide variety of multimodal fusion medical image fusion tasks. In spite of this, we can draw some preliminary conclusions from the above analysis. For medical modalities with similar structures, modal registration is easier, so input fusion, single-level fusion, and hierarchical fusion are all network structures worth investigating. Generally, single-level fusion and hierarchical fusion fuse deeper features, which will improve the classification performance. When data have a wide range of structures or dimensions, single-level fusion and attention-based fusion are preferable solutions, as they are capable of handling a wide range of modal feature fusion scenarios. Lastly, if we have a large number of unimodal datasets for each modality in multimodal data, output fusion will perform well.

In addition to using a single multimodal fusion method, multiple fusion methods can be combined [66,143,144,190]. Tang et al. [66] achieved the classification of skin lesions using a combination of single-level fusion and output fusion. For the classification of Diabetic Retinopathy, Li et al. [190] utilized different configurations of OCT Angiography data. Their approach combined hierarchical fusion for registered modalities with late fusion for unregistered modalities. In order to improve the diagnosis of breast cancer, Hu et al. [144] fused multi-parametric MRI data at three levels: input, feature (intermediate), and decision (output). Combining different fusion methods can cumulate their advantages, allowing data from various perspectives to be fused and improving classification performance to some extent. It is one of the promising strategies that can be used when performing multimodal medical classification.

5.2. How to find the best architecture?

During the investigation of multimodal approaches, we have found that researchers need not only many tests to compare various fusion methods but also a large number of hyperparametric tests to determine the best network architecture for each fusion method. It takes a great deal of time and labor to conduct these extensive tests. There have been many recent studies that have applied Neural Architecture Search (NAS) to multimodal networks that can integrate various fusion techniques to determine the best architecture for a given dataset. The use of these methods is widespread in the fields of diagnosing dementia [191], gliomas segmentation [192], multimodal action recognition [193], visual question answering [194], multimodal damage identification [195], multimodal gesture recognition [196], etc. We believe that this approach has the potential to be explored in the future for multimodal classification in medicine.

5.3. How to manage incomplete multimodal data?

The problem of modality incompleteness is one of the most pressing challenges in multimodality medical research. The high cost and potentially harmful effects of medical images may lead many patients to refuse being scanned with multiple imaging modalities for clinical diagnosis [197]. In the ADNI dataset, all subjects had MRI data; however, only about half of the subjects had PET scans [197]. The most common approach to solving the modality incomplete problem is to discard the modality incomplete subjects [59,82,122,142], but this approach reduces the number of trainable subjects for the deep learning model, resulting in reduced classification performance. There is also the option of estimating the features of the missing subjects [198,199] based on, for example, the mean or median of the subjects with complete modal data, although this requires some prior knowledge and may introduce errors.

Generative Adversarial Networks (GAN) [200] is a type of generative model used to produce data of a modality from another modality [201]. With the development of GAN, more and more fields are using this technology to generate images. The modal incompleteness problem has recently been solved through the use of GAN in many studies [2,61,146,161,169,197,202,203]. The GAN is used to generate the missing data and then the generated data is used for multimodal classification. It provides a significant increase in the number of subjects in the dataset, improves the model's classification performance, and is an effective solution when dealing with multimodal incompleteness. GAN-based solutions are currently the most promising.

Recent research endeavors have sought to confront the difficulties associated with addressing unbalanced datasets and incomplete modalities through the implementation of a distinctive fusion network design and training strategy. This approach aims to optimize the utilization of available dataset information while mitigating bias introduced by independently generating missing modalities. Gravina et al. [62] introduced a Multi-Input–Multi-Output 3D CNN designed for the assessment of dementia severity, specifically tailored for scenarios involving incomplete multimodal brain MRI and PET data. In alignment with our hierarchical fusion architecture, they incorporated a fusion branch named PAIRED-NET during feature extraction, employing distinct CNN branches for MRI and PET modalities, each capable of producing independent outputs. Simultaneous parameter updates for all three branches occurred during training when the patient sample encompassed the full modality set. In instances where a modality was missing, only the parameters of the branch corresponding to the available modality were updated. This methodology facilitates comprehensive network training using the entire dataset, enabling classification of instances with a single missing modality during testing. Liu et al. [176] introduced a cascaded Multi-Modal Mixing Transformer (3MT) designed for the classification of Alzheimer's Disease with incomplete data. The architecture of 3MT comprises a sequence of Cascaded Modality Transformers (CMTs), each incorporating features from a specific modality. At the conclusion of this sequence, a more informed class prediction is obtained by aggregating the extracted multi-modal features. In scenarios involving missing data, the CMTs corresponding to the absent modalities receive zero embeddings, indicating "not available" to the model. This training approach equips the model with prior knowledge for handling diverse missing data scenarios.

5.4. Does multimodal fusion always performs better?

Multimodal data not only contain complementary information but may also contain a great deal of redundant information. One study found that multimodal fusion did not enhance classification performance [204]. In one sense, this relates to the design of the network, and in another sense, multimodal fusion may not improve the performance of classification if the information in the modalities is relatively similar or if a particular modality does not accurately define the target class. Before starting a multimodal fusion project and gathering multimodal data, it is advisable to conduct a redundancy analysis [205].

Alternately, Narazani et al. [206] questioned the multimodal diagnostic objectives. Clinical studies primarily aim to determine the type of dementia, whereas studies on DL focused on only one type, AD. multimodal fusion studies have performed well in terms of classifying NC versus AD, but the clinical goal is the classification of AD at multiple levels. As a result of their tests, multimodal fusion networks did not improve the multi-level classification. Therefore, multimodal fusion classification studies should be conducted in conjunction with clinical needs.

5.5. Can we advantageously combine multimodal image classification with other tasks?

We should point out that input fusion [207–211], intermediate fusion [212–216], and output fusion [217,218] methods can also be applied to medical image segmentation, medical image fusion, and similar tasks. A notable trend is to incorporate the fusion network into

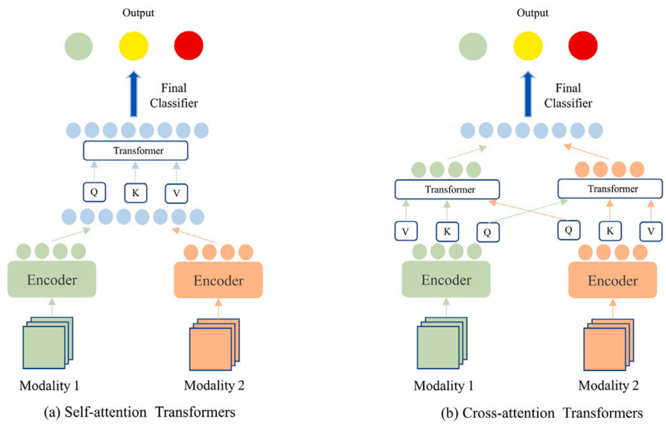

Fig. 23. Two architectures of Transformer-based fusion.

the feature extraction process, thereby enabling the creation of multi-task multimodal networks. Cheng et al. [219] proposed an end-to-end multi-task learning network for simultaneous glioma segmentation and IDH genotyping based on the sharing of spatial and global feature representations extracted from the hybrid CNN-Transformer encoder. The performance of both classification and segmentation can be enhanced through the use of a joint network.

5.6. Can we advantageously combine images with contextual data in a classification pipeline?

In this survey, our main focus was on integrating data from various imaging modalities. However, out of 114 reviewed papers, 27 incorporate non-image contextual information in various forms, often structured demographic and clinical metadata extracted from electronic health records [70,71,135,220,221]. Clinical metadata includes the result of physical tests, like visual acuity or refraction for chronic central serous chorioretinopathy diagnosis [11], the result of chemical tests, like Pap or HPV for cervical dysplasia diagnosis [151], and the result of cognitive tests, like executive functioning (ADNI-EF) and memory (ADNI-MEM) test for Alzheimer's disease diagnosis [222]. Sometimes, images are combined with more complex contextual data like voice signals [185], free-form text [175], or genomic data [71,154,160,174]. Image and contextual metadata have no geometrical relationship, so input and hierarchical fusion are not relevant in this scenario. Earlier solutions thus relied on single-level intermediate fusion or output fusion. Recent studies tend to use attention-based intermediate fusion [174,175,180]. One challenge with contextual data is that they are often incomplete: the reader is referred to Section 5.3 for this challenge. Nevertheless, all these studies report increased classification performance when using contextual data.

5.7. Trends for the future

As the network structure evolves and hardware devices become increasingly available, there has been a growing interest in multimodality research. From Table 5, it is evident that the rapid development of multimodal fusion has led to significant improvements in classification results. In fact, multimodal networks based on deep learning exhibit greater potential for development than unimodal networks.

Transformer is one of the most popular network architectures, and multimodal fusion based on Transformer has developed rapidly in the past two years. In particular, for visual-language tasks, Transformer can handle the fusion of images, languages, and text very effectively. Based on the research conducted in different fields, we classify Transformer-based multimodal fusion networks into self-attention Transformers [223–230] and cross-attention Transformers [231,232,232–238], as shown in Fig. 23. Following the extraction of features using encoders, self-attention Transformers concatenate features from different modalities and compute the attention relationship between the fused features using Transformer blocks. Alternatively, cross-attentional Transformers compute the attentional relationships among different modalities in order to achieve information fusion. Nowadays, these two architectures are the most popular multimodal fusion networks.

Compared to CNNs, Transformers have the advantage of efficiently identifying long-range relationships between sequences. In medical images, most visual representations are ordered due to the similarity of human organs. Medical images contain more information regarding sequence relationships than natural images [173]. This indicates that Transformer-based multimodal medical image fusion is a promising approach, and the above two network architectures are worth exploring. While recent medical research has employed analogous structures [61,179,181], it is imperative to undertake broader investigations and validations to extend the applicability of these findings.

In addition to these developments, the field is also witnessing an emerging trend in the exploration of representation learning for multimodal data using techniques such as pretext tasks or contrastive learning. These methods aim to learn robust and transferable representations that can be applied to downstream tasks of a classification nature or directly in parallel with the classification task [259–265]. This field is new and a lot of research questions are emerging. How to learn an aligned representation across modalities? Can we learn one representation space for the different modalities? The research community is actively moving in this direction to address challenges associated with representation learning.

In response to the challenge posed by the limited scale of medical datasets and the laborious nature of manual labeling, certain investigations have resorted to harnessing information from image-text pairs available on the web. This approach facilitates the construction of transfer learning multimodal models, which have demonstrated notable efficacy in subsequent fine-tuning tasks and zero-shot classification endeavors. Zhang et al. [266] introduced ConVIRT, an unsupervised methodology for acquiring medical visual representations through the analysis of naturally paired images and text. Their approach involves evaluating image representations in conjunction with paired text representations through a bidirectional objective, surpassing alternative methods in performance across various downstream medical classification tasks. Drawing inspiration from the Contrastive Language–Image Pre-training (CLIP) approach [267], Huang et al. [268] undertook fine-tuning specifically tailored for medical applications, resulting in the formulation of Pathology Language–Image Pretraining (PLIP). Exploiting numerous de-identified images and abundant textual data disseminated by clinicians on public platforms like Twitter, PLIP introduces a multimodal, unsupervised, Transformer-based transfer learning model. This model effectively classifies new pathological images across four external datasets, exhibiting state-of-the-art performance. These initiatives highlight the immense potential of publicly shared medical information as a valuable resource for developing multimodal medical AI systems, thereby enhancing the landscape of medical diagnosis.

In support to these advancements and needs, the TorchMultimodal library [1] has been created as a framework for training state-of-the-art multimodal multi-task models at scale by Meta Research using PyTorch framework. This library is the result of concerted community efforts, reflecting the growing focus on multimodal tasks and methodologies, as well as representation learning for multimodal data. Furthermore, in alignment with the increasing enthusiasm for multimodal learning, libraries such as Transformers[2] by Hugging Face offer robust resources for constructing and training such models. These Transformers present pretrained models that can handle various modalities, such as text

1 https://github.com/facebookresearch/multimodal
2 https://huggingface.co/docs/transformers/index

Table 6
List of terms.

Term	Description
ADC	Apparent Diffusion Coefficient
AD	Alzheimer's Disease
AdaBoost	Adaptive Boosting
ADHD	Attention Deficit Hyperactivity Disorder
AE	Auto-Encoder
AGDAF	Attention Guided Discriminative and Adaptive Fusion
ASD	Autism Spectrum Disorder
CatBoost	Categorical Boosting
CFP	Color Fundus Photographs
CLIMAT	Clinically-Inspired Multi-Agent Transformers
CMR	Cardiac Magnetic Resonance
CNN	Convolutional Neural Network
CSF	CerebroSpinal Fluid
CT	Computed Tomography
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCE	Dynamic Contrast-Enhanced
DL	Deep Learning
DNN	Deep Neural Network
Dsc	Dermatoscopic Image
DWI	Diffusion-Weighted Imaging
DXA	Dual-energy X-ray Absorptiometry
EHR	Electronic Health Records
EMR	Electronic Medical Record
FA	Fractional Anisotropy
FC Layer	Fully Connected Layer
FCNN	Fully Connected Neural Network
FHR	Fetal Heart Rate
FLAIR	Fluid Attenuated Inversion Recovery
FFN	Feed-Forward Network
fMRI	functional Magnetic Resonance Imaging
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
HoFN	High-order Factorization Network
HPV	Human Papillomavirus
KELM	Kernel-based Extreme Learning Machine
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSTM	Long-Short Term Memory
MD	Mean Diffusivity
mDSNet	multimodality Disease-Image-Specific Network
MGF	Multiheaded Gating Fusion
MLP	Multilayer Perceptron
MMNet	Multimodal 3D medical image classification Network
MM-SDPN	Multimodal Stacked Deep Polynomial Networks
MRI	Magnetic Resonance Imaging
NBI	Narrow Band Imaging
PCA	Principal Component Analysis
PEI	Positive Enhancement Integral
PET	Positron Emission Tomography
RBM	Restricted Boltzmann Machine
ROI	Regions-Of-Interest
SAE	Stacked Auto-Encoder
SBM	Swapping Bi-Attention Mechanism
sELM-AE	stacked sparse Extreme Learning Machine Auto-Encoder
sMRI	structural Magnetic Resonance Imaging
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
T1	T1-Weighted
T1C	T1-Contrast
T1-Gd	T1-weighted Gadolinium Contrasted
T2	T2-Weighted
T2WI	T2-Weighted MP-MRI Image
TSI	Tumor Shape Image
TWIST	Time-resolved Angiography with Interleaved Stochastic Trajectories
OCT	Optical Coherence Tomography
UC	Uterine Contraction
US	Ultrasound
VF	Visual Field
ViT	Vision Transformer
WSI	Whole Slide Image
XGBoost	Extreme Gradient Boosting

(*continued on next page*)

Table 6 (*continued*).

and images, simplifying tasks like image categorization or answering questions with multimodal data. This emphasis on multimodal functionalities underscores the growing demand for models that possess the ability to comprehend and analyze diverse types of data. Although Transformers are not exclusively designed for the medical domain, they can serve as a beneficial resource for scholars aiming to create and distribute code for tasks related to multimodal classification of medical images.

As illustrated in Table 7, very few multimodal image classification papers in the medical field are associated with public code. We hope TorchMultimodal, Transformers or similar libraries will facilitate code sharing.

6. Conclusion

In this paper, we conducted a comprehensive review of the development of deep learning-based multimodal medical classification tasks over the past few years. We examined the complementary relationships among several common clinical modalities and delved into five types of architectures for deep learning multimodal classification networks: input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion. Our study covered a wide range of multimodal fusion scenarios in medical classification and the application domains for which different network architectures are most suitable.

Additionally, we discussed emerging trends and challenges in the field, including the exploration of representation learning techniques and the development of dedicated frameworks like TorchMultimodal. These advancements provide efficient tools for training state-of-the-art multimodal multi-task models at scale. In particular, we highlighted the advantages of using Transformer-based multimodal fusion

Table 7
List of publications for different fusion networks. IF: Input Fusion; SLF: Single-level Fusion; HF: Hierarchical Fusion; ABF: Attention-based Fusion; OF: Output Fusion.

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Li et al. [124]	MRI, PET, CSF	IF	Concatenation (Inputs)	RBM	SVM	Brain	ADNI	202	2	N/A
Liu et al. [59]	MRI, PET	IF	Concatenation (Inputs)	Manual	Softmax	Brain	ADNI	331	2,4	N/A
Akhavan Aghdam et al. [79]	sMRI, fMRI	IF	Concatenation (Inputs)	DBN	FC Layer	Brain	ABIDE I, ABIDE II	185	2	N/A
Aldoj et al. [145]	MRI (ADC, DWI, T2)	IF	Concatenation (Inputs)	CNN	FC Layer	Prostate	TCIA (PROSTATEx)	200	2	N/A
Qian et al. [4]	US (US B-mode, US color Doppler)	IF	Concatenation (Inputs)	CNN	FC Layer	Breast	Private Data	59 959	4	N/A
Zong et al. [147]	MRI (ADC, DWI, T2)	IF	Concatenation (Inputs)	CNN	FC Layer	Prostate	TCIA (PROSTATEx)	201	2	N/A
Sanford et al. [239]	MRI (T2, ADC, High-b)	IF	Merge (Inputs)	Resnet	FC Layer	Prostate	Private Data, TCIA (PROSTATEx)	687	2	N/A
Lin et al. [146]	MRI, PET	IF	Concatenation (Inputs)	CNN	FC Layer	Brain	ADNI	1086	2	N/A
Song et al. [149]	MRI, PET	IF	Merge (Inputs)	CNN	FC Layer	Brain	ADNI	381	2,3	N/A
Decuyper et al. [54]	MRI (T1, T1C, T2M FLAIR)	IF	Merge (Inputs)	CNN	FC Layer	Brain	TCIA (GBM, LGG), BraTS 2019	628	2	N/A
Kong et al. [1]	MRI, PET	IF	Merge (Inputs)	CNN	FC Layer	Brain	ADNI	370	2,3	N/A
Zhang et al. [2]	MRI, PET	IF	Concatenation (Inputs)	Resnet	FC Layer	Brain	ADNI	873	2,3	N/A
Zhou et al. [148]	MRI (PEI, DWI)	IF	Concatenation (Inputs)	CNN	FC Layer	Breast	Private Data	210	2	N/A
Rallabandi and Seetharaman [150]	MRI, PET	IF	Merge (Inputs)	CNN	SoftMax	Brain	OASIS-3	1098	2	N/A
Odusami et al. [240]	MRI, PET	IF	Concatenation (Inputs)	ResNet	FC Layer	Brain	ADNI	412	2	N/A
Suk and Shen [133]	MRI, PET, CSF	SLF	Concatenation (Classic)	SAE	SVM	Brain	ADNI	202	2	N/A
Suk et al. [142]	MRI, PET	SLF	Concatenation (Classic)	DBM	SVM	Brain	ADNI	398	2	N/A
Suk et al. [134]	MRI, PET	SLF	Concatenation (Classic)	SAE	SVM	Brain	ADNI	202	2	N/A
Xu et al. [151]	Photograph of the cervix, Pap tests, HPV tests	SLF	Concatenation (Classic)	CNN	FC Layer	Cervix	TCIA (Guanacaste Project)	690	2	N/A
Mehrtash et al. [9]	MRI (ADC, DWI, DCE)	SLF	Concatenation (Classic)	CNN	FC Layer	Prostate	TCIA (PROSTATEx)	201	2	N/A

(continued on next page)

Table 7 (*continued*).

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Vu et al. [241]	MRI, PET	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	317	2	N/A
Zou et al. [121]	sMRI, fMRI	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADHD-200	730	2	N/A
Yang et al. [8]	MRI (ADC, T2)	SLF	Merge (Classic)	CNN	FC Layer	Prostate	Private Data	160	2	[a]
Shi et al. [122]	MRI, PET	SLF	Concatenation (Network)	MM-SDPN	FC Layer	Brain	ADNI	202	2,4	N/A
Zhou et al. [158]	MRI, PET	SLF	Concatenation (Network)	DNN	Score Merge	Brain	ADNI	805	3,4	N/A
Ye et al. [55]	MRI (TI, T2, T1C, FLAIR)	SLF	Merge (Classic)	CNN	FC Layer	Brain	BraTS 2015	274	2	N/A
Cheng and Liu [159]	MRI, PET	SLF	Concatenation (Network)	CNN	FC Layer	Brain	ADNI	193	2	N/A
Le et al. [7]	MRI (ADC, T2WI)	SLF	Merge (Classic)	CNN	SVM	Prostate	TCIA, Private Data	364	2	N/A
Pan et al. [202]	MRI, PET	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	1457	2	N/A
Ge et al. [242]	MRI (TI, T2, FLAIR)	SLF	Merge (Classic)	CNN	FC Layer	Brain	BraTS 2017, TCIA (LGG)	444	2	N/A
Liu et al. [3]	MRI, PET	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	397	2	N/A
Wang et al. [120]	MRI (ADC, T2)	SLF	Concatenation (Classic)	CNN	Softmax	Prostate	Private Data, TCIA (PROSTATEx)	360	2	N/A
Kim and Lee [123]	MRI, PET, CSF	SLF	Concatenation (Network)	sELM-AE	KELM	Brain	ADNI	202	2	N/A
Kawahara et al. [67]	Dsc, Clinical Image, Metadata	SLF	Concatenation (Classic)	CNN	FC Layer	Skin	SPC	1011	2	N/A
Lu et al. [243]	MRI, PET	SLF	Concatenation (Network)	DNN	Score Merge	Brain	ADNI	626	2	N/A
Zhou et al. [164]	MRI, PET, SNP	SLF	Concatenation (Network)	DNN	Score Merge	Brain	ADNI	360	2	N/A
Feng et al. [244]	MRI, PET	SLF	Concatenation (Classic)	CNN, LSTM	Softmax	Brain	ADNI	397	2	N/A
Dalmis et al. [5]	MRI (ADC, T2, TWIST)	SLF	Concatenation (Classic)	CNN	Random Forest	Breast	Private Data	576	2	N/A
Huang et al. [84]	MRI, PET	SLF	Concatenation (Classic)	VGG	FC Layer	Brain	ADNI	1512	2	N/A
Lee et al. [222]	MRI, CSF, Demographic Information, Cognitive Performance	SLF	Concatenation (Classic)	GRU	LR	Brain	ADNI	1618	2	N/A
Punjabi et al. [153]	MRI, PET	SLF	Merge (Classic)	CNN	FC Layer	Brain	ADNI	723	2	N/A

(*continued on next page*)

Table 7 (*continued*).

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Zhang et al. [188]	MRI, PET	SLF	Concatenation (Classic)	CNN	Softmax	Brain	ADNI	392	2	N/A
Qin et al. [63]	PET, CT	SLF	Merge (Classic)	CNN	FC Layer	Lung	Private Data	397	3	N/A
Pan et al. [197]	MRI, PET	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	2355	2	N/A
El-Sappagh et al. [125]	MRI, PET, Clinical Data	SLF	Concatenation (Classic)	CNN, LSTM	FC Layer	Brain	ADNI	1536	4	N/A
Venugopalan et al. [71]	MRI, EHR, SNP	SLF	Concatenation (Classic)	CNN, SAE	FC Layer	Brain	ADNI	808	3	N/A
Qian et al. [6]	US (US B-mode, US color Doppler, US elastography images)	SLF	Concatenation (Classic)	CNN	FC Layer	Breast	Private Data	775	2	N/A
Massalimova and Varol [245]	MRI (T1, FA, MD)	SLF	Concatenation (Classic)	ResNet	FC Layer	Brain	OASIS-3	1098	3	N/A
Abdelaziz et al. [246]	MRI, PET, SNP	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	805	2	N/A
Zhou et al. [221]	CT, EMR	SLF	Concatenation (Classic)	CNN, HoFN	FC Layer	Lung	Private Data	733	4	N/A
Pan et al. [203]	MRI, PET	SLF	Concatenation (Classic)	mDSNet	FC Layer	Brain	ADNI	1455	2	N/A
Yan et al. [135]	EMR, Pathological images	SLF	Concatenation (Classic)	VGG, AE	FC Layer	Breast	Private Data	185	2	N/A
Zhang et al. [247]	MRI, PET	SLF	Concatenation (Network)	CNN	Score Merge	Brain	ADNI	875	2	N/A
Joo et al. [152]	MRI (T1, T2), Clinical Information	SLF	Concatenation (Classic)	ResNet	FC Layer	Breast	Private Data	536	2	N/A
Rahaman et al. [154]	sMRI, fMRI, SNP	SLF	Concatenation (Classic)	DNN	FC Layer	Brain	COBRE	437	2	N/A
Abdelaziz et al. [189]	MRI, PET	SLF	Concatenation (Network)	CNN	FC Layer	Brain	ADNI	959	2	N/A
Puyol-Antón et al. [248]	Echocardiography, CMR	SLF	Merge (Network)	CNN	SVM	Heart	Private Data	50	2	N/A
Al-Absi et al. [249]	DXA, CFP	SLF	Concatenation (Network)	CNN	FC Layer	Heart	Private Data	483	2	N/A
Rahaman et al. [160]	sMRI, fMRI, Genomic Sequence	SLF	Merge (Network)	AE, MLP, LSTM	Softmax	Brain	COBRE	437	2	N/A
Wu et al. [64]	CFP, OCT	SLF	Concatenation (Classic)	CNN	FC Layer	Eye	GAMMA	300	3	N/A

(continued on next page)

Table 7 (*continued*).

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Jin et al. [161]	MRI, PET	SLF	Concatenation (Network)	ResNet	FC Layer	Brain	ADNI	360	2	N/A
Xiong et al. [155]	VF, OCT	SLF	Merge (Classic)	CNN	FC Layer	Eye	Private Data	1083	2	N/A
Huang et al. [12]	VF, CFP	SLF	Concatenation (Classic)	CNN	FC Layer	Eye	Private Data	1027	2	N/A
Xing et al. [136]	PET-AV45, PET-FDG	SLF	Concatenation (Classic)	ViT	FC Layer	Brain	ADNI	381	2	N/A
Dolci et al. [250]	sMRI, fMRI, SNP	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	788	2	N/A
Tu et al. [251]	MRI, Metadata	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	1461	2	N/A
Leng et al. [162]	MRI, PET	SLF	Merge (Network)	CNN	FC Layer	Brain	ADNI	536	2	N/A
Liu et al. [72]	MRI, Metadata	SLF	Concatenation (Classic)	VGG	CatBoost	Brain	ADNI	242	2	N/A
Kollias et al. [56]	MRI (T1, T2, T1C, FLAIR)	SLF	Concatenation (Classic)	CNN, RNN	FC Layer	Brain	BraTS 2021	585	2	N/A
Kadri et al. [156]	PET, MRI, CT	SLF	Concatenation (Classic)	ViT, CNN	FC Layer	Brain	OASIS-1, OASIS-3, ADNI	N/A	2,3,4	N/A
Abbas et al. [252]	sMRI, fMRI	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	ABIDE I	855	2	N/A
Lu et al. [163]	MRI, SNP, Clinical data	SLF	Merge (Network)	CNN	FC Layer	Brain	ADNI	577	2	N/A
Saponaro et al. [157]	sMRI, fMRI	SLF	Concatenation (Classic)	DNN	FC Layer	Brain	ABIDE I, ABIDE II	1383	2	N/A
Gravina et al. [62]	MRI, PET	SLF	Concatenation (Classic)	CNN	FC Layer	Brain	OASIS-3	1025	2,3	N/A
Mahmood et al. [165]	White Light RGB, NBI	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Digestive tract	ISIT-UMR	76	3	N/A
Zhang and Shi [166]	MRI, PET	HF	Merge (Network), Merge (Classic)	CNN	Softmax	Brain	ADNI	500	2,4	N/A
Zhou et al. [126]	MRI, PET	HF	Merge (Network),Merge (Classic)	RBM	Softmax	Brain	ADNI	340	2	N/A
He et al. [141]	MRI (T1C, FLAIR)	HF	Merge (Network), Merge (Classic)	CNN	FC Layer	Brain	TCIA, BraTS 2017	499	2	N/A
Gao et al. [171]	MRI, PET	HF	Concatenation (Network), Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	977	2	N/A
Li et al. [10]	CFP, OCT	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Eye	GAMMA, Private Data	264	2,3	N/A
Liu et al. [80]	sMRI, fMRI	HF	Merge (Network), Concatenation (Classic)	MGF	FC Layer	Brain	ABIDE, ADHD-200, COBRE	2120	2	N/A

(*continued on next page*)

Table 7 (*continued*).

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Xu et al. [168]	MRI, PET	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	579	2	N/A
Wu et al. [58]	MRI (T1, T2, T1C, FLAIR)	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Brain	BraTS 2018, BraTS 2019	326	2	N/A
Omeroglu et al. [167]	Dsc, Clinical Image, Metadata	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Skin	SPC	1011	2,3,5	N/A
Miao et al. [170]	MRI, PET	HF	Merge (Network), Merge (Classic)	CNN, ViT	FC Layer	Brain	ADNI	720	2	N/A
Xu et al. [57]	MRI (T1, T2, T1C, FLAIR)	HF	Concatenation (Network), Concatenation (Classic)	CNN	SoftMax	Brain	BraTS 2018, BraTS 2019	620	2	N/A
Tu et al. [169]	MRI, PET	HF	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Brain	ADNI	821	2	N/A
Dai et al. [173]	MRI (T1,T2)	ABF	Attention Fusion	TransMed	FC Layer	Parotid, Knee	MRNet	344	2	N/A
Hoang Nguyen et al. [253]	Visual Data, Clinical Data	ABF	Attention Fusion	CLIMAT	FFN	Brain, Knee	ADNI, OAI [254]	10 260	5	[b]
Liu et al. [176]	Neuro-imaging Data, Clinical Data	ABF	Attention Fusion	3MT	FC Layer	Brain	ADNI	816	2	N/A
Zhang et al. [175]	Image, Text	ABF	Attention Fusion	MMIF	FC Layer	Uterus	CTU-UHB	160	2	N/A
Li et al. [177]	Multi-parametric MRI	ABF	Attention Fusion	AGDAF	FC Layer	Liver	Private Data	112	2	N/A
Dai et al. [178]	MRI (T1,T2)	ABF	Attention Fusion	MMNet	FC Layer	Parotid, Prostate	MRNet, TCIA (PROSTATEx)	2089	2	N/A
Qiu et al. [174]	Genomic Data, Pathology Data	ABF	Attention Fusion	AHM-Fusion	FC Layer	Brain, Lung	TCGA (LUAD/LUSC, GBM/LGG)	1670	2,3	N/A
Zuo et al. [179]	sMRI, fMRI	ABF	Attention Fusion	SBM	FC Layer	Brain	ADNI	268	2	N/A
Chen et al. [180]	MRI, PET, Metadata	ABF	Attention Fusion	CBAM	FC Layer	Brain	ADNI	227	2,3	N/A
Gao et al. [61]	MRI, PET	ABF	Attention Fusion, Concatenation (Classic)	CNN, ViT	FC Layer	Brain	ADNI, OASIS-3	1700	2	N/A
Bi et al. [181]	sMRI, fMRI	ABF	Attention Fusion	ViT, CNN	FC Layer	Brain	COBRE	827	2,3,5	N/A

(*continued on next page*)

Table 7 (*continued*).

Research work	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset	Patients	Class	Code
Wei and Ji [68]	Dsc, Clinical Image	ABF	Attention Fusion, Concatenation (Classic)	CNN	FC Layer	Skin	SPC	1011	2,3,5	N/A
Abdol-maleki and Abadeh [184]	sMRI, fMRI	OF	Merge (Outputs)	CNN	SVM, KNN, LDA	Brain	ADHD-200	730	2	N/A
Xi et al. [255]	MRI (T1C, T2), Clinical Data	OF	Merge (Outputs)	ResNet	Bagging	Kidney	Private Data	1162	2	N/A
Moon et al. [182]	US (ROI, Tumor image, TSI, Fused image)	OF	Merge (Outputs)	CNN	Score Merge	Breast	BUSI [256]	697	2	N/A
Fang et al. [60]	MRI, PET	OF	Merge (Outputs)	CNN	AdaBoost	Brain	ADNI	398	2	N/A
Huang et al. [220]	CT, EMR	OF	Merge (Outputs)	DNN, CNN	Score Merge	Lung	Private Data	1794	2	N/A
Ying et al. [257]	MRI, SNP	OF	Merge (Outputs)	CNN, MLP	Ensemble Gate	Brain	ADNI	100	2	N/A
Yoo et al. [11]	CFP, Clinical Data	OF	Merge (Outputs)	ResNet	XGBoost	Eye	Private Data	166	2	N/A
Prabhu et al. [70]	MRI, EHR	OF	Merge (Outputs)	CNN, AE	Score Merge	Brain	ADNI	3996	2,3	N/A
Guo et al. [183]	MRI (TI, T2, T1C, FLAIR)	OF	Merge (Outputs)	CNN	Score Merge	Brain	CPM-RadPath 2020	221	3	N/A
Kwon et al. [185]	Laryngeal image, Voice	OF	Merge (Outputs)	CNN	Decision Tree	Larynx	Private Data	431	2	N/A
Qiu et al. [186]	MRI, Metadata	OF	Merge (Outputs)	CNN	CatBoost	Brain	OASIS-3	666	2	[c]
Hu et al. [144]	MRI (DCE, T2)	IF, SLF, OF	Merge (Input), Merge (Classic), Merge (Outputs)	VGG	Score Merge	Breast	Private Data	616	2	N/A
Wang et al. [143]	WSI, MRI (T1, T1-Gd, T2, FLAIR)	IF, OF	Concatenation (Input), Merge (Outputs)	CNN	Score Merge	Brain	CPM-RadPath 2020	378	3	N/A
Tang et al. [66]	Dsc, Clinical Image, Metadata	SLF, OF	Concatenation (Classic), Merge (Outputs)	CNN	Score Merge	Skin	SPC	1011	2,3,5	N/A
Mustafa and Luo [258]	CT, MRI	IF, OF	Merge (Input), Merge (Outputs)	CNN	Score Merge	Brain	OASIS-3	1377	4	N/A

[a] https://github.com/Andysis/co-trained-CADx
[b] https://github.com/oulu-imeds/climatv2
[c] https://github.com/vkola-lab/ncomms2022

architectures, particularly in medical imaging applications, where sequence relationships are more prevalent. This demonstrates the potential of these architectures in advancing the field of multimodal medical classification tasks.

Looking forward, we encourage the research community to continue investigating novel fusion techniques, optimization methods, and network architectures to further enhance the performance of multimodal classification tasks. Developing interpretable models, addressing data imbalance and scarcity, and exploring unsupervised and semi-supervised learning approaches are other areas worth investigating. Additionally, we recommend future research focus on the application of multimodal fusion in emerging areas such as genomics, proteomics, and patient-centered care, where the integration of diverse data types can potentially lead to significant improvements in diagnostic and therapeutic outcomes.

CRediT authorship contribution statement

Yihao Li: Conceptualization, Data curation, Formal analysis, Software, Visualization, Writing – original draft, Investigation, Methodology. **Mostafa El Habib Daho:** Conceptualization, Investigation, Supervision, Validation, Writing – review & editing, Formal analysis. **Pierre-Henri Conze:** Supervision, Validation, Writing – review & editing. **Rachid Zeghlache:** Visualization, Writing – review & editing. **Hugo Le Boité:** Writing – review & editing. **Ramin Tadayoni:** Funding acquisition, Project administration. **Béatrice Cochener:** Data curation, Resources. **Mathieu Lamard:** Supervision, Writing – review & editing. **Gwenolé Quellec:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ramin Tadayoni, reports financial support was provided by French National Research Agency. Gwenolé Quellec reports financial support was provided by French National Research Agency.

Acknowledgments

The work takes place in the framework of the ANR RHU project Evired. This work benefits from State aid managed by the French National Research Agency under the "Investissement d'Avenir" program bearing the reference ANR-18-RHUS-0008. It was also funded in part by the ANR, France under the LabCom program (ANR-19-LCV2-0005 - ADMIRE project).

References

[1] Z. Kong, M. Zhang, W. Zhu, Y. Yi, T. Wang, B. Zhang, Multi-modal data Alzheimer's disease detection based on 3D convolution, Biomed. Signal Process. Control 75 (2022) 103565.
[2] J. Zhang, X. He, L. Qing, F. Gao, B. Wang, BPGAN: brain PET synthesis from MRI using generative adversarial network for multi-modal Alzheimer's disease diagnosis, Comput. Methods Programs Biomed. 217 (2022) 106676.
[3] M. Liu, D. Cheng, K. Wang, Y. Wang, ADNI, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics 16 (2018) 295–308.
[4] X. Qian, B. Zhang, S. Liu, Y. Wang, X. Chen, J. Liu, Y. Yang, X. Chen, Y. Wei, Q. Xiao, et al., A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network, Eur. Radiol. 30 (2020) 3023–3033.
[5] M.U. Dalmis, A. Gubern-Mérida, S. Vreemann, P. Bult, N. Karssemeijer, R. Mann, J. Teuwen, Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI, Invest. Radiol. 54 (6) (2019) 325–332.
[6] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng, et al., Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning, Nat. Biomed. Eng. 5 (6) (2021) 522–532.
[7] M.H. Le, J. Chen, L. Wang, Z. Wang, W. Liu, K.-T.T. Cheng, X. Yang, Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks, Phys. Med. Biol. 62 (16) (2017) 6497.
[8] X. Yang, C. Liu, Z. Wang, J. Yang, H. Le Min, L. Wang, K.-T.T. Cheng, Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI, Med. Image Anal. 42 (2017) 212–227.
[9] A. Mehrtash, A. Sedghi, M. Ghafoorian, M. Taghipour, C.M. Tempany, W.M. Wells III, T. Kapur, P. Mousavi, P. Abolmaesumi, A. Fedorov, Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks, in: Medical Imaging 2017: Computer-Aided Diagnosis, Vol. 10134, SPIE, 2017, pp. 589–592.
[10] Y. Li, M. El Habib Daho, P.-H. Conze, H. Al Hajj, S. Bonnin, H. Ren, N. Manivannan, S. Magazzeni, R. Tadayoni, B. Cochener, et al., Multimodal information fusion for glaucoma and diabetic retinopathy classification, in: Ophthalmic Medical Image Analysis: 9th International Workshop, OMIA 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings, Springer, 2022, pp. 53–62.
[11] T.K. Yoo, S.H. Kim, M. Kim, C.S. Lee, S.H. Byeon, S.S. Kim, J. Yeo, E.Y. Choi, DeepPDT-Net: predicting the outcome of photodynamic therapy for chronic central serous chorioretinopathy using two-stage multimodal transfer learning, Sci. Rep. 12 (1) (2022) 18689.
[12] X. Huang, J. Sun, K. Gupta, G. Montesano, D.P. Crabb, D.F. Garway-Heath, P. Brusini, P. Lanzetta, F. Oddone, A. Turpin, et al., Detecting glaucoma from multi-modal data using probabilistic deep learning, Front. Med. 9 (2022).
[13] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375.
[14] M.A. Azam, K.B. Khan, S. Salahuddin, E. Rehman, S.A. Khan, M.A. Khan, S. Kadry, A.H. Gandomi, A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, Comput. Biol. Med. 144 (2022) 105253.
[15] H. Hermessi, O. Mourali, E. Zagrouba, Multimodal medical image fusion review: Theoretical background and recent advances, Signal Process. 183 (2021) 108036.
[16] D. Ramachandram, G.W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108.
[17] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[18] G. Xie, J. Wang, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, Y. Jin, Cross-modality neuroimage synthesis: A survey, 2022, http://dx.doi.org/10.48550/ARXIV.2202.06997, arXiv. URL: https://arxiv.org/abs/2202.06997.
[19] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array 3 (2019) 100004.
[20] F.E.-Z.A. El-Gamal, M. Elmogy, A. Atwan, Current trends in medical image registration and fusion, Egyptian Inf. J. 17 (1) (2016) 99–124.
[21] A. Shoeibi, M. Khodatars, M. Jafari, N. Ghassemi, P. Moridian, R. Alizadesani, S.H. Ling, A. Khosravi, H. Alinejad-Rokny, H. Lam, et al., Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review, Inf. Fusion (2022).
[22] J. Lipkova, R.J. Chen, B. Chen, M.Y. Lu, M. Barbieri, D. Shao, A.J. Vaidya, C. Chen, L. Zhuang, D.F. Williamson, et al., Artificial intelligence for multimodal data integration in oncology, Cancer Cell 40 (10) (2022) 1095–1110.
[23] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: A scoping review, Npj Digit. Med. 5 (1) (2022) 171.
[24] R. Stokking, K.J. Zuiderveld, M.A. Viergever, Integrated volume visualization of functional image data and anatomical surfaces using normal fusion, Hum. Brain Mapp. 12 (4) (2001) 203–218.
[25] G. Bhatnagar, Q.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimedia 15 (5) (2013) 1014–1024.
[26] C. He, Q. Liu, H. Li, H. Wang, Multimodal medical image fusion based on IHS and PCA, Procedia Eng. 7 (2010) 280–285.
[27] R. Bashir, R. Junejo, N.N. Qadri, M. Fleury, M.Y. Qadri, SWT and PCA image fusion methods for multi-modal imagery, Multimedia Tools Appl. 78 (2019) 1235–1263.
[28] M.R. Princess, V.S. Kumar, M.R. Begum, Comprehensive and comparative study of different image fusion techniques, Int. J. Adv. Res. Electr. Electron. Instrum. Eng (2014) 11800–11806.
[29] K. Parmar, R. Kher, A comparative analysis of multimodality medical image fusion methods, in: 2012 Sixth Asia Modelling Symposium, IEEE, 2012, pp. 93–97.
[30] F. Sadjadi, Comparative image fusion analysais, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, IEEE, 2005, p. 8.
[31] S. Das, M.K. Kundu, A neuro-fuzzy approach for medical image fusion, IEEE Trans. Biomed. Eng. 60 (12) (2013) 3347–3353.
[32] Y. Liu, J. Yang, J. Sun, PET/CT medical image fusion algorithm based on multiwavelet transform, in: 2010 2nd International Conference on Advanced Computer Control, Vol. 2, IEEE, 2010, pp. 264–268.
[33] X.-X. Xi, X.-Q. Luo, Z.-C. Zhang, Q.-J. You, X. Wu, Multimodal medical volumetric image fusion based on multi-feature in 3-D shearlet transform, in: 2017 International Smart Cities Conference, ISC2, IEEE, 2017, pp. 1–6.
[34] Q. Zhang, Y. Liu, R.S. Blum, J. Han, D. Tao, Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review, Inf. Fusion 40 (2018) 57–75.
[35] Z. Zhu, M. Zheng, G. Qi, D. Wang, Y. Xiang, A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain, IEEE Access 7 (2019) 20811–20824.
[36] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[37] D. Mishra, B. Palkar, Image fusion techniques: a review, Int. J. Comput. Appl. 130 (9) (2015) 7–13.
[38] S. Bhat, D. Koundal, Multi-focus image fusion techniques: a survey, Artif. Intell. Rev. 54 (2021) 5735–5787.
[39] A.M. Sharma, A. Dogra, B. Goyal, R. Vig, S. Agrawal, From pyramids to state-of-the-art: a study and comprehensive comparison of visible–infrared image fusion techniques, IET Image Process. 14 (9) (2020) 1671–1689.
[40] J. Lee, I. Mawla, J. Kim, M.L. Loggia, A. Ortiz, C. Jung, S.-T. Chan, J. Gerber, V.J. Schmithorst, R.R. Edwards, et al., Machine learning-based prediction of clinical pain using multimodal neuroimaging and autonomic metrics, Pain 160 (3) (2019) 550.
[41] X. Tang, X. Xu, Z. Han, G. Bai, H. Wang, Y. Liu, P. Du, Z. Liang, J. Zhang, H. Lu, et al., Elaboration of a multimodal MRI-based radiomics signature for the preoperative prediction of the histological subtype in patients with non-small-cell lung cancer, Biomed. Eng. Online 19 (2020) 1–17.
[42] G. Quellec, M. Lamard, G. Cazuguel, C. Roux, B. Cochener, Case retrieval in medical databases by fusing heterogeneous information, IEEE Trans. Med. Imaging 30 (1) (2010) 108–118.
[43] P.A. Lalousis, S.J. Wood, L. Schmaal, K. Chisholm, S.L. Griffiths, R.L. Reniers, A. Bertolino, S. Borgwardt, P. Brambilla, J. Kambeitz, et al., Heterogeneity and classification of recent onset psychosis and depression: a multimodal machine learning approach, Schizophrenia Bull. 47 (4) (2021) 1130–1140.
[44] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111.

[45] S.Y. Boulahia, A. Amamra, M.R. Madi, S. Daikh, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, Mach. Vis. Appl. 32 (6) (2021) 121.
[46] B. Singh, R. Gautam, S. Kumar, S. Umapathy, Application of vibrational microspectroscopy to biology and medicine, Curr. Sci. (2012).
[47] D.B. Plewes, W. Kucharczyk, Physics of MRI: a primer, J. Magn. Resonance Imag. 35 (5) (2012) 1038–1054.
[48] D.L. Bailey, M.N. Maisey, D.W. Townsend, P.E. Valk, Positron Emission Tomography, Vol. 2, Springer, 2005.
[49] T.M. Buzug, Computed Tomography, Springer, 2011.
[50] T.G. Leighton, What is ultrasound? Progr. Biophys. Mol. Biol. 93 (1–3) (2007) 3–83.
[51] D. Huang, E.A. Swanson, C.P. Lin, J.S. Schuman, W.G. Stinson, W. Chang, M.R. Hee, T. Flotte, K. Gregory, C.A. Puliafito, et al., Optical coherence tomography, Science 254 (5035) (1991) 1178–1181.
[52] R. MacKie, C. Fleming, A. McMahon, P. Jarrett, The use of the dermatoscope to identify early melanoma using the three-colour test, Br. J. Dermatol. 146 (3) (2002) 481–484.
[53] R. Besenczi, J. Tóth, A. Hajdu, A review on automatic analysis techniques for color fundus photographs, Comput. Struct. Biotechnol. J. 14 (2016) 371–384.
[54] M. Decuyper, S. Bonte, K. Deblaere, R. Van Holen, Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma, Comput. Med. Imaging Graph. 88 (2021) 101831.
[55] F. Ye, J. Pu, J. Wang, Y. Li, H. Zha, Glioma grading based on 3D multimodal convolutional neural network and privileged learning, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2017, pp. 759–763.
[56] D. Kollias, K. Vendal, P. Gadhavi, S. Russom, BTDNet: A multi-modal approach for brain tumor radiogenomic classification, Appl. Sci. 13 (21) (2023) 11984.
[57] D. Xu, X. Wang, J. Cai, P.-A. Heng, Cross-modality guidance-aided multi-modal learning with dual attention for MRI brain tumor grading, 2024, arXiv preprint arXiv:2401.09029.
[58] P. Wu, Z. Wang, B. Zheng, H. Li, F.E. Alsaadi, N. Zeng, AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, Comput. Biol. Med. 152 (2023) 106457.
[59] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M.J. Fulham, et al., Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, IEEE Trans. Biomed. Eng. 62 (4) (2014) 1132–1140.
[60] X. Fang, Z. Liu, M. Xu, Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis, IET Image Process. 14 (2) (2020) 318–326.
[61] X. Gao, F. Shi, D. Shen, M. Liu, Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease, Comput. Med. Imaging Graph. 110 (2023) 102303.
[62] M. Gravina, A. García-Pedrero, C. Gonzalo-Martín, C. Sansone, P. Soda, Multi input–Multi output 3D CNN for dementia severity assessment with incomplete multimodal data, Artif. Intell. Med. 149 (2024) 102774.
[63] R. Qin, Z. Wang, L. Jiang, K. Qiao, J. Hai, J. Chen, J. Xu, D. Shi, B. Yan, Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism, Complexity 2020 (2020) 1–12.
[64] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, L. Huang, Q. Yu, S. Song, X. Xu, et al., Gamma challenge: glaucoma grading from multi-modality images, 2022, arXiv preprint arXiv:2202.06511.
[65] M. El Habib Daho, Y. Li, R. Zeghlache, Y.C. Atse, H. Le Boité, S. Bonnin, D. Cosette, P. Deman, L. Borderie, C. Lepicard, R. Tadayoni, B. Cochener, P.-H. Conze, M. Lamard, G. Quellec, Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of UWF-CFP and OCTA images, in: B. Antony, H. Chen, H. Fang, H. Fu, C.S. Lee, Y. Zheng (Eds.), Ophthalmic Medical Image Analysis, Springer Nature Switzerland, Cham, 2023, pp. 11–20.
[66] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, T. Lasser, FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification, Med. Image Anal. 76 (2022) 102307.
[67] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE J. Biomed. Health Inf. 23 (2) (2018) 538–546.
[68] Y. Wei, L. Ji, Multi-modal bilinear fusion with hybrid attention mechanism for multi-label skin lesion classification, Multimedia Tools Appl. (2024) 1–27.
[69] J. Yap, W. Yolland, P. Tschandl, Multimodal skin lesion classification using deep learning, Exp. Dermatol. 27 (11) (2018) 1261–1267.
[70] S.S. Prabhu, J.A. Berkebile, N. Rajagopalan, R. Yao, W. Shi, F. Giuste, Y. Zhong, J. Sun, M.D. Wang, Multi-modal deep learning models for Alzheimer's disease prediction using MRI and EHR, in: 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2022, pp. 168–173.
[71] J. Venugopalan, L. Tong, H.R. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of Alzheimer's disease stage, Sci. Rep. 11 (1) (2021) 3254.
[72] S. Liu, Y. Zheng, H. Li, M. Pan, Z. Fang, M. Liu, Y. Qiao, N. Pan, W. Jia, X. Ge, Improving Alzheimer diagnoses with an interpretable deep learning framework: Including neuropsychiatric symptoms, Neuroscience 531 (2023) 86–98.
[73] P.P. Pai, P.K. Mandal, K. Punjabi, D. Shukla, A. Goel, S. Joon, S. Roy, K. Sandal, R. Mishra, R. Lahoti, BRAHMA: Population specific t1, t2, and FLAIR weighted brain templates and their impact in structural and functional imaging studies, Magn. Resonance Imag. 70 (2020) 5–21.
[74] T. Lindig, R. Kotikalapudi, D. Schweikardt, P. Martin, F. Bender, U. Klose, U. Ernemann, N.K. Focke, B. Bender, Evaluation of multimodal segmentation based on 3D T1-, T2-and FLAIR-weighted images–the difficulty of choosing, Neuroimage 170 (2018) 210–221.
[75] M. Hecht, F. Fellner, C. Fellner, M. Hilz, D. Heuss, B. Neundörfer, MRI-FLAIR images of the head show corticospinal tract alterations in ALS patients more frequently than T2-, T1-and proton-density-weighted images, J. Neurol. Sci. 186 (1–2) (2001) 37–44.
[76] D.A. Kuban, H.D. Thames, L.B. Levy, E.M. Horwitz, P.A. Kupelian, A.A. Martinez, J.M. Michalski, T.M. Pisansky, H.M. Sandler, W.U. Shipley, et al., Long-term multi-institutional analysis of stage T1–T2 prostate cancer treated with radiotherapy in the PSA era, Int. J. Radiat. Oncol.* Biol.* Phys. 57 (4) (2003) 915–928.
[77] D.C. Preston, Magnetic resonance imaging (mri) of the brain and spine: Basics, MRI Basics, Case Med. 30 (2006).
[78] J.-M. Shen, X.-W. Xia, W.-G. Kang, J.-J. Yuan, L. Sheng, The use of MRI apparent diffusion coefficient (ADC) in monitoring the development of brain infarction, BMC Med. Imag. 11 (1) (2011) 1–4.
[79] M. Akhavan Aghdam, A. Sharifi, M.M. Pedram, Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network, J. Digit. Imag. 31 (2018) 895–903.
[80] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, K.C. Tan, Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[81] P.A. Bandettini, Twenty years of functional MRI: the science and the stories, Neuroimage 62 (2) (2012) 575–588.
[82] V.D. Calhoun, J. Sui, Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness, Biol. Psychiatry: Cognit. Neurosci. Neuroimaging 1 (3) (2016) 230–244.
[83] M. Liu, Y. Gao, P.-T. Yap, D. Shen, Multi-hypergraph learning for incomplete multimodality data, IEEE J. Biomed. Health Inf. 22 (4) (2017) 1197–1208.
[84] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, ADNI, Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, Front. Neurosci. 13 (2019) 509.
[85] H. Xu, Y. Li, W. Zhao, G. Quellec, L. Lu, M. Hatt, Joint nnU-Net and radiomics approaches for segmentation and prognosis of head and neck cancers with PET/CT images, 2022, arXiv preprint arXiv:2211.10138.
[86] V. Andrearczyk, V. Oreiller, S. Boughdad, C.C.L. Rest, H. Elhalawani, M. Jreige, J.O. Prior, M. Vallières, D. Visvikis, M. Hatt, et al., Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images, in: Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, Springer, 2022, pp. 1–37.
[87] G. Muehllehner, J.S. Karp, Positron emission tomography, Phys. Med. Biol. 51 (13) (2006) R117.
[88] W.J. Zwiebel, J.S. Pellerito, Introduction to vascular ultrasonography, Elsevier Saunders, Philadelphia, 2005.
[89] E.A. Abdelgawad, M.F. Abu-samra, N.M. Abdelhay, H.M. Abdel-Azeem, B-mode ultrasound, color Doppler, and sonoelastography in differentiation between benign and malignant cervical lymph nodes with special emphasis on sonoelastography, Egyptian J. Radiol. Nucl. Med. 51 (1) (2020) 1–10.
[90] M.-H. Lu, X.-Y. Pu, X. Gao, X.-F. Zhou, J.-G. Qiu, J. Si-Tu, A comparative study of clinical value of single B-mode ultrasound guidance and B-mode combined with color doppler ultrasound guidance in mini-invasive percutaneous nephrolithotomy to decrease hemorrhagic complications, Urology 76 (4) (2010) 815–820.
[91] M. Schelling, M. Braun, W. Kuhn, G. Bogner, R. Gruber, J. Gnirs, K.T. Schneider, K. Ulm, S. Rutke, A. Staudach, Combined transvaginal B-mode and color Doppler sonography for differential diagnosis of ovarian tumors: results of a multivariate logistic regression analysis, Gynecologic Oncol. 77 (1) (2000) 78–86.
[92] M. Schelling, J. Gnirs, M. Braun, R. Busch, S. Maurer, W. Kuhn, K. Schneider, H. Graeff, Optimized differential diagnosis of breast lesions by combined B-mode and color Doppler sonography, Ultrasound Obstetr. Gynecol.: Official J. Int. Soc. Ultrasound Obstetr. Gynecol. 10 (1) (1997) 48–53.
[93] L. Li, X. Zhou, X. Zhao, S. Hao, J. Yao, W. Zhong, H. Zhi, B-mode ultrasound combined with color Doppler and strain elastography in the diagnosis of non-mass breast lesions: A prospective study, Ultrasound Med. Biol. 43 (11) (2017) 2582–2590.
[94] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, R. Garnavi, Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images, in: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III. Vol. 20, Springer, 2017, pp. 250–258.

[95] R.C. Petersen, P.S. Aisen, L.A. Beckett, M.C. Donohue, A.C. Gamst, D.J. Harvey, C.R. Jack, W.J. Jagust, L.M. Shaw, A.W. Toga, et al., Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization, Neurology 74 (3) (2010) 201–209.

[96] L.A. Beckett, M.C. Donohue, C. Wang, P. Aisen, D.J. Harvey, N. Saito, A.D.N. Initiative, et al., The Alzheimer's Disease Neuroimaging Initiative phase 2: Increasing the length, breadth, and depth of our understanding, Alzheimer's Dementia 11 (7) (2015) 823–831.

[97] M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack Jr., W. Jagust, J.C. Morris, et al., The Alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement, Alzheimer's Dementia 13 (5) (2017) 561–571.

[98] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[99] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, J. Digit. Imag. 26 (2013) 1045–1057.

[100] D.S. Marcus, A.F. Fotenos, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults, J. Cogn. Neurosci. 22 (12) (2010) 2677–2684.

[101] P.J. LaMontagne, T.L. Benzinger, J.C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A.G. Vlassenko, et al., OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease, Cold Spring Harbor Laboratory Press, 2019, MedRxiv. 2019-2012.

[102] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120.

[103] K. Tomczak, P. Czerwińska, M. Wiznerowicz, Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol./Współczesna Onkologia 2015 (1) (2015) 68–77.

[104] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (6) (2014) 659–667.

[105] A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J.S. Anderson, M. Assaf, J.H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, et al., Enhancing studies of the connectome in autism using the autism brain imaging data exchange II, Sci. Data 4 (1) (2017) 1–15.

[106] A.-. consortium, The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience, Front. Syst. Neurosci. 6 (2012) 62.

[107] V.D. Calhoun, J. Sui, K. Kiehl, J. Turner, E. Allen, G. Pearlson, Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder, Front. Psychiatry 2 (2012) 75.

[108] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, Y. Huang, Q. Yu, S. Song, X. Xu, et al., Gamma challenge: glaucoma grading from multi-modality images, Med. Image Anal. 90 (2023) 102938.

[109] W.-W. Hsu, J.-M. Guo, L. Pei, L.-A. Chiang, Y.-F. Li, J.-C. Hsiao, R. Colen, P. Liu, A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs, Sci. Rep. 12 (1) (2022) 6111.

[110] T. Kurc, S. Bakas, X. Ren, A. Bagari, A. Momeni, Y. Huang, L. Zhang, A. Kumar, M. Thibault, Q. Qi, et al., Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches, Front. Neurosci. 14 (2020) 27.

[111] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, A. Bartoli, Computer-aided classification of gastrointestinal lesions in regular colonoscopy, IEEE Trans. Med. Imaging 35 (9) (2016) 2051–2063.

[112] N. Bien, P. Rajpurkar, R.L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B.N. Patel, K.W. Yeom, K. Shpanskaya, et al., Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet, PLoS Med. 15 (11) (2018) e1002699.

[113] V. Chudacek, J. Spilka, M. Bursa, P. Janku, L. Hruban, M. Huptych, L. Lhotska, Open access intrapartum CTG database, BMC Pregnancy Childbirth 14 (2014) 1–12.

[114] W.C. Sleeman, R. Kapoor, P. Ghosh, Multimodal classification: Current landscape, taxonomy and future directions, ACM Comput. Surv. 55 (7) (2022) 1–31.

[115] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, ADNI, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56 (2) (2011) 766–781.

[116] C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification, Neurobiol. Aging 32 (12) (2011) 2322–e19.

[117] O. Kohannim, X. Hua, D.P. Hibar, S. Lee, Y.-Y. Chou, A.W. Toga, C.R. Jack Jr., M.W. Weiner, P.M. Thompson, ADNI, et al., Boosting power for clinical trials using classifiers based on multiple biomarkers, Neurobiol. Aging 31 (8) (2010) 1429–1442.

[118] M. Liu, D. Zhang, D. Shen, ADNI, et al., Ensemble sparse classification of Alzheimer's disease, NeuroImage 60 (2) (2012) 1106–1116.

[119] M. Liu, D. Zhang, D. Shen, ADNI, Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis, Hum. Brain Mapp. 35 (4) (2014) 1305–1319.

[120] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, K.-T. Cheng, Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network, IEEE Trans. Med. Imaging 37 (5) (2018) 1127–1139.

[121] L. Zou, J. Zheng, C. Miao, M.J. Mckeown, Z.J. Wang, 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI, IEEE Access 5 (2017) 23626–23636.

[122] J. Shi, X. Zheng, Y. Li, Q. Zhang, S. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease, IEEE J. Biomed. Health Inf. 22 (1) (2017) 173–183.

[123] J. Kim, B. Lee, Identification of Alzheimer's disease and mild cognitive impairment using multimodal sparse hierarchical extreme learning machine, Hum. Brain Mapp. 39 (9) (2018) 3728–3741.

[124] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, J. Li, A robust deep model for improved classification of AD/MCI patients, IEEE J. Biomed. Health Inf. 19 (5) (2015) 1610–1616.

[125] S. El-Sappagh, T. Abuhmed, S.R. Islam, K.S. Kwak, Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data, Neurocomputing 412 (2020) 197–215.

[126] P. Zhou, S. Jiang, L. Yu, Y. Feng, C. Chen, F. Li, Y. Liu, Z. Huang, Use of a sparse-response deep belief network and extreme learning machine to discriminate alzheimer's disease, mild cognitive impairment, and normal controls based on amyloid PET/MRI images, Front. Med. 7 (2021) 621204.

[127] M.A. Azam, K.B. Khan, M. Ahmad, M. Mazzara, Multimodal medical image registration and fusion for quality enhancement, Comput., Mater. Continua 68 (1) (2021) 821–840.

[128] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[129] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[130] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[131] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, http://dx.doi.org/10.48550/ARXIV.1512.03385, arXiv. URL: https://arxiv.org/abs/1512.03385.

[132] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, 2016, http://dx.doi.org/10.48550/ARXIV.1608.06993, arXiv. URL: https://arxiv.org/abs/1608.06993.

[133] H.-I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II. Vol. 16, Springer, 2013, pp. 583–590.

[134] H.-I. Suk, S.-W. Lee, D. Shen, ADNI, Latent feature representation with stacked auto-encoder for AD/MCI diagnosis, Brain Struct. Funct. 220 (2015) 841–859.

[135] R. Yan, F. Zhang, X. Rao, Z. Lv, J. Li, L. Zhang, S. Liang, Y. Li, F. Ren, C. Zheng, et al., Richer fusion network for breast cancer classification based on multimodal data, BMC Med. Inf. Decis. Mak. 21 (1) (2021) 1–15.

[136] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, N. Jacobs, Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, IEEE, 2022, pp. 1–4.

[137] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[138] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[139] D.H. Ballard, Modular learning in neural networks, in: AAAI, Vol. 647, 1987, pp. 279–284.

[140] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/ARXIV.1706.03762, arXiv. URL: https://arxiv.org/abs/1706.03762.

[141] M. He, K. Han, Y. Zhang, W. Chen, Hierarchical-order multimodal interaction fusion network for grading gliomas, Phys. Med. Biol. 66 (21) (2021) 215016.

[142] H.-I. Suk, S.-W. Lee, D. Shen, ADNI, et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage 101 (2014) 569–582.

[143] X. Wang, R. Wang, S. Yang, J. Zhang, M. Wang, D. Zhong, J. Zhang, X. Han, Combining radiology and pathology for automatic glioma classification, Front. Bioeng. Biotechnol. 10 (2022).

[144] Q. Hu, H.M. Whitney, M.L. Giger, A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI, Sci. Rep. 10 (1) (2020) 10536.

[145] N. Aldoj, S. Lukas, M. Dewey, T. Penzkofer, Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network, Eur. Radiol. 30 (2) (2020) 1243–1253.

[146] W. Lin, W. Lin, G. Chen, H. Zhang, Q. Gao, Y. Huang, T. Tong, M. Du, ADNI, Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, Front. Neurosci. 15 (2021) 646013.
[147] W. Zong, J.K. Lee, C. Liu, E.N. Carver, A.M. Feldman, B. Janic, M.A. Elshaikh, M.V. Pantelic, D. Hearshen, I.J. Chetty, et al., A deep dive into understanding tumor foci classification using multiparametric MRI based on convolutional neural network, Med. Phys. 47 (9) (2020) 4077–4086.
[148] Z. Zhou, B.E. Adrada, R.P. Candelaria, N.A. Elshafeey, M. Boge, R.M. Mohamed, S. Pashapoor, J. Sun, Z. Xu, B. Panthi, et al., Prediction of pathologic complete response to neoadjuvant systemic therapy in triple negative breast cancer using deep learning on multiparametric MRI, Sci. Rep. 13 (1) (2023) 1171.
[149] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, P. Shen, An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis, Front. Digit. Health 3 (2021) 637386.
[150] V.S. Rallabandi, K. Seetharaman, Deep learning-based classification of healthy aging controls, mild cognitive impairment and Alzheimer's disease using fusion of MRI-PET imaging, Biomed. Signal Process. Control 80 (2023) 104312.
[151] T. Xu, H. Zhang, X. Huang, S. Zhang, D.N. Metaxas, Multimodal deep learning for cervical dysplasia diagnosis, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II. Vol. 19, Springer, 2016, pp. 115–123.
[152] S. Joo, E.S. Ko, S. Kwon, E. Jeon, H. Jung, J.-Y. Kim, M.J. Chung, Y.-H. Im, Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer, Sci. Rep. 11 (1) (2021) 18800.
[153] A. Punjabi, A. Martersteck, Y. Wang, T.B. Parrish, A.K. Katsaggelos, ADNI, Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks, PLoS One 14 (12) (2019) e0225759.
[154] M.A. Rahaman, J. Chen, Z. Fu, N. Lewis, A. Iraji, V.D. Calhoun, Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2021, pp. 3267–3272.
[155] J. Xiong, F. Li, D. Song, G. Tang, J. He, K. Gao, H. Zhang, W. Cheng, Y. Song, F. Lin, et al., Multimodal machine learning using visual fields and peripapillary circular OCT scans in detection of glaucomatous optic neuropathy, Ophthalmology 129 (2) (2022) 171–180.
[156] R. Kadri, B. Bouaziz, M. Tmar, F. Gargouri, Efficient multimodel method based on transformers and CoAtNet for Alzheimer's diagnosis, Digit. Signal Process. 143 (2023) 104229.
[157] S. Saponaro, F. Lizzi, G. Serra, F. Mainas, P. Oliva, A. Giuliano, S. Calderoni, A. Retico, Deep learning based joint fusion approach to exploit anatomical and functional brain information in autism spectrum disorders, Brain Inform. 11 (1) (2024) 2.
[158] T. Zhou, K.-H. Thung, X. Zhu, D. Shen, Feature learning and fusion of multi-modality neuroimaging and genetic data for multi-status dementia diagnosis, in: Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings. Vol. 8, Springer, 2017, pp. 132–140.
[159] D. Cheng, M. Liu, CNNs based multi-modality classification for AD diagnosis, in: 2017 10th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics, CISP-BMEI, IEEE, 2017, pp. 1–5.
[160] M.A. Rahaman, Y. Garg, A. Iraj, Z. Fu, J. Chen, V. Calhoun, Two-dimensional attentive fusion for multi-modal learning of neuroimaging and genomics data, in: 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing, MLSP, IEEE, 2022, pp. 1–6.
[161] L. Jin, K. Zhao, Y. Zhao, T. Che, S. Li, A hybrid deep learning method for early and late mild cognitive impairment diagnosis with incomplete multimodal data, Front. Neuroinform. 16 (2022).
[162] Y. Leng, W. Cui, Y. Peng, C. Yan, Y. Cao, Z. Yan, S. Chen, X. Jiang, J. Zheng, A.D.N. Initiative, et al., Multimodal cross enhanced fusion network for diagnosis of Alzheimer's disease and subjective memory complaints, Comput. Biol. Med. 157 (2023) 106788.
[163] P. Lu, L. Hu, A. Mitelpunkt, S. Bhatnagar, L. Lu, H. Liang, A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease, Biomed. Signal Process. Control 88 (2024) 105669.
[164] T. Zhou, K.-H. Thung, X. Zhu, D. Shen, Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis, Hum. Brain Mapp. 40 (3) (2019) 1001–1016.
[165] F. Mahmood, Z. Yang, T. Ashley, N.J. Durr, Multimodal densenet, 2018, arXiv preprint arXiv:1811.07407.
[166] T. Zhang, M. Shi, Multi-modal neuroimaging feature fusion for diagnosis of alzheimer's disease, J. Neurosci. Methods 341 (2020) 108795.
[167] A.N. Omeroglu, H.M. Mohammed, E.A. Oral, S. Aydin, A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification, Eng. Appl. Artif. Intell. 120 (2023) 105897.
[168] H. Xu, S. Zhong, Y. Zhang, Multi-level fusion network for mild cognitive impairment identification using multi-modal neuroimages, Phys. Med. Biol. 68 (9) (2023) 095018.
[169] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, Z. Wang, D. Wang, Multimodal fusion diagnosis of Alzheimer's disease based on FDG-PET generation, Biomed. Signal Process. Control 89 (2024) 105709.
[170] S. Miao, Q. Xu, W. Li, C. Yang, B. Sheng, F. Liu, T.T. Bezabih, X. Yu, MMTFN: Multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis, Int. J. Imaging Syst. Technol. 34 (1) (2024) e22970.
[171] X. Gao, F. Shi, D. Shen, M. Liu, Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease, IEEE J. Biomed. Health Inf. 26 (1) (2021) 36–43.
[172] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
[173] Y. Dai, Y. Gao, F. Liu, Transmed: Transformers advance multi-modal medical image classification, Diagnostics 11 (8) (2021) 1384.
[174] L. Qiu, L. Zhao, R. Hou, W. Zhao, S. Zhang, Z. Lin, H. Teng, J. Zhao, Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features, Comput. Med. Imaging Graph. (2023) 102176.
[175] Y. Zhang, Y. Deng, Z. Zhou, X. Zhang, P. Jiao, Z. Zhao, Multimodal learning for fetal distress diagnosis using a multimodal medical information fusion framework, Front. Physiol. (2022) 2362.
[176] L. Liu, S. Liu, L. Zhang, X.V. To, F. Nasrallah, S.S. Chandra, Cascaded multi-modal mixing transformers for Alzheimer's disease classification with incomplete data, NeuroImage 277 (2023) 120267.
[177] S. Li, Y. Xie, G. Wang, L. Zhang, W. Zhou, Attention guided discriminative feature learning and adaptive fusion for grading hepatocellular carcinoma with Contrast-enhanced MR, Comput. Med. Imaging Graph. 97 (2022) 102050.
[178] Y. Dai, Y. Gao, F. Liu, J. Fu, Mutual attention-based hybrid dimensional network for multimodal imaging computer-aided diagnosis, 2022, arXiv preprint arXiv:2201.09421.
[179] Q. Zuo, Y. Shen, N. Zhong, C.P. Chen, B. Lei, S. Wang, Alzheimer's disease prediction via brain structural-functional deep fusing network, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 4601–4612.
[180] H. Chen, H. Guo, L. Xing, D. Chen, T. Yuan, Y. Zhang, X. Zhang, Multimodal predictive classification of Alzheimer's disease based on attention-combined fusion network: Integrated neuroimaging modalities and medical examination data, IET Image Process. 17 (11) (2023) 3153–3164.
[181] Y. Bi, A. Abrol, Z. Fu, V. Calhoun, A Multimodal Vision Transformer for Interpretable Fusion of Functional and Structural Neuroimaging Data, Cold Spring Harbor Laboratory, 2023, bioRxiv. 2023-2007.
[182] W.K. Moon, Y.-W. Lee, H.-H. Ke, S.H. Lee, C.-S. Huang, R.-F. Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, Comput. Methods Programs Biomed. 190 (2020) 105361.
[183] S. Guo, L. Wang, Q. Chen, L. Wang, J. Zhang, Y. Zhu, Multimodal MRI image decision fusion-based network for glioma classification, Front. Oncol. 12 (2022).
[184] S. Abdolmaleki, M.S. Abadeh, Brain MR image classification for ADHD diagnosis using deep neural networks, in: 2020 International Conference on Machine Vision and Image Processing, MVIP, IEEE, 2020, pp. 1–5.
[185] I. Kwon, S.-G. Wang, S.-C. Shin, Y.-I. Cheon, B.-J. Lee, J.-C. Lee, D.-W. Lim, C. Jo, Y. Cho, B.-J. Shin, Diagnosis of early glottic cancer using laryngeal image and voice based on ensemble learning of convolutional neural network classifiers, J. Voice (2022).
[186] S. Qiu, M.I. Miller, P.S. Joshi, J.C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P.H. Hwang, J.A. Cramer, et al., Multimodal deep learning for Alzheimer's disease dementia assessment, Nat. Commun. 13 (1) (2022) 3404.
[187] B. Dubois, H.H. Feldman, C. Jacova, S.T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, et al., Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria, Lancet Neurol. 6 (8) (2007) 734–746.
[188] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, X. Zhang, Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease, Neurocomputing 361 (2019) 185–195.
[189] M. Abdelaziz, T. Wang, A. Elazab, Fusing multimodal and anatomical volumes of interest features using convolutional auto-encoder and convolutional neural networks for Alzheimer's disease diagnosis, Front. Aging Neurosci. 14 (2022).
[190] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, S. Bonnin, D. Cosette, S. Magazzeni, B. Lay, A. Le Guilcher, R. Tadayoni, B. Cochener, M. Lamard, G. Quellec, Hybrid fusion of high-resolution and ultra-widefield OCTA acquisitions for the automatic diagnosis of diabetic retinopathy, Diagnostics 13 (17) (2023) http://dx.doi.org/10.3390/diagnostics13172770, URL: https://www.mdpi.com/2075-4418/13/17/2770.
[191] M. Chatzianastasis, L. Ilias, D. Askounis, M. Vazirgiannis, Neural architecture search with multimodal fusion methods for diagnosing dementia, 2023, arXiv preprint arXiv:2302.05894.
[192] F. Wang, Neural architecture search for gliomas segmentation on multimodal magnetic resonance imaging, 2020, arXiv preprint arXiv:2005.06338.
[193] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, F. Jurie, MFAS: Multi-modal fusion architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6966–6975.

[194] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, Q. Tian, Deep multimodal neural architecture search, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3743–3752.
[195] A. Singh, H. Nair, A neural architecture search for automated multimodal learning, Expert Syst. Appl. 207 (2022) 118051.
[196] Y. Yin, S. Huang, X. Zhang, Bm-nas: Bilevel multimodal neural architecture search, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, (8) 2022, pp. 8901–8909.
[197] Y. Pan, M. Liu, C. Lian, Y. Xia, D. Shen, Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages, IEEE Trans. Med. Imaging 39 (9) (2020) 2965–2975.
[198] A.R.T. Donders, G.J. Van Der Heijden, T. Stijnen, K.G. Moons, A gentle introduction to imputation of missing values, J. Clin. Epidemiol. 59 (10) (2006) 1087–1091.
[199] J.A. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, J.R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, BMJ 338 (2009).
[200] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[201] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, http://dx.doi.org/10.48550/ARXIV.1406.2661, arXiv. URL: https://arxiv.org/abs/1406.2661.
[202] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, D. Shen, Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III. Vol. 11, Springer, 2018, pp. 455–463.
[203] Y. Pan, M. Liu, Y. Xia, D. Shen, Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 6839–6853.
[204] B. Khagi, G.-R. Kwon, 3D CNN design for the classification of Alzheimer's disease using brain MRI and PET, IEEE Access 8 (2020) 217830–217847.
[205] R. Salvador, E. Canales-Rodríguez, A. Guerrero-Pedraza, S. Sarró, D. Tordesillas-Gutiérrez, T. Maristany, B. Crespo-Facorro, P. McKenna, E. Pomarol-Clotet, Multimodal integration of brain images for MRI-based diagnosis in schizophrenia, Front. Neurosci. 13 (2019) 1203.
[206] M. Narazani, I. Sarasua, S. Pölsterl, A. Lizarraga, I. Yakushev, C. Wachinger, Is a PET all you need? A multi-modal study for Alzheimer's disease using 3D CNNs, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I, Springer, 2022, pp. 66–76.
[207] S. Pereira, A. Pinto, V. Alves, C.A. Silva, Brain tumor segmentation using convolutional neural networks in MRI images, IEEE Trans. Med. Imaging 35 (5) (2016) 1240–1251.
[208] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers. Vol. 3, Springer, 2018, pp. 287–297.
[209] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, No new-net, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II. Vol. 4, Springer, 2019, pp. 234–244.
[210] S. Cui, L. Mao, J. Jiang, C. Liu, S. Xiong, Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network, J. Healthc. Eng. 2018 (2018).
[211] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.
[212] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I.B. Ayed, HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation, IEEE Trans. Med. Imaging 38 (5) (2018) 1116–1126.
[213] J. Dolz, C. Desrosiers, I. Ben Ayed, IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet, in: Computational Methods and Clinical Applications for Spine Imaging: 5th International Workshop and Challenge, CSI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Springer, 2019, pp. 130–143.
[214] L. Chen, Y. Wu, A.M. DSouza, A.Z. Abidin, A. Wismüller, C. Xu, MRI tumor segmentation with densely connected 3D CNN, in: Medical Imaging 2018: Image Processing, Vol. 10574, SPIE, 2018, pp. 357–364.
[215] G. Andrade-Miranda, V. Jaouen, V. Bourbonne, F. Lucia, D. Visvikis, P.-H. Conze, Pure versus hybrid transformers for multi-modal brain tumor segmentation: a comparative study, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 1336–1340.
[216] J. Li, C. Bu, C. Qian, A cross-attention based image fusion network for prediction of mild cognitive impairment, in: Journal of Physics: Conference Series, Vol. 2284, (1) IOP Publishing, 2022, 012002.
[217] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, et al., Ensembles of multiple models and architectures for robust brain tumour segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers. Vol. 3, Springer, 2018, pp. 450–462.
[218] M. Aygün, Y.H. Şahin, G. Ünal, Multi modal convolutional neural networks for brain tumor segmentation, 2018, arXiv preprint arXiv:1809.06191.
[219] J. Cheng, J. Liu, H. Kuang, J. Wang, A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping, IEEE Trans. Med. Imaging 41 (6) (2022) 1520–1532.
[220] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, M.P. Lungren, Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection, Sci. Rep. 10 (1) (2020) 1–9.
[221] J. Zhou, X. Zhang, Z. Zhu, X. Lan, L. Fu, H. Wang, H. Wen, Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2021) 2535–2549.
[222] G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, Predicting Alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (1) (2019) 1952.
[223] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, B. Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, Adv. Neural Inf. Process. Syst. 34 (2021) 24206–24221.
[224] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, Attention bottlenecks for multimodal fusion, Adv. Neural Inf. Process. Syst. 34 (2021) 14200–14213.
[225] B. Shi, W.-N. Hsu, K. Lakhotia, A. Mohamed, Learning audio-visual speech representation by masked multimodal cluster prediction, 2022, arXiv preprint arXiv:2201.02184.
[226] R. Li, S. Yang, D.A. Ross, A. Kanazawa, Ai choreographer: Music conditioned 3d dance generation with AIST++, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13401–13412.
[227] A. Pashevich, C. Schmid, C. Sun, Episodic transformer for vision-and-language navigation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15942–15952.
[228] S. Appalaraju, B. Jasani, B.U. Kota, Y. Xie, R. Manmatha, Docformer: End-to-end transformer for document understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 993–1003.
[229] J.-M.O. Steitz, J. Pfeiffer, I. Gurevych, S. Roth, TxT: Crossmodal end-to-end learning with transformers, in: Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings, Springer, 2022, pp. 405–420.
[230] P.Y. Wu, W.R. Mebane Jr., MARMOT: A deep learning framework for constructing multimodal representations for vision-and-language tasks, Comput. Commun. Res. 4 (1) (2022).
[231] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, vol. 32, 2019.
[232] R.J. Chen, M.Y. Lu, W.-H. Weng, T.Y. Chen, D.F. Williamson, T. Manz, M. Shady, F. Mahmood, Multimodal co-attention transformer for survival prediction in gigapixel whole slide images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4015–4025.
[233] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, 2019, arXiv preprint arXiv:1908.07490.
[234] L. Zhu, Y. Yang, Actbert: Learning global-local video-text representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8746–8755.
[235] K. Ramesh, C. Xing, W. Wang, D. Wang, X. Chen, Vset: A multimodal transformer for visual speech enhancement, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6658–6662.
[236] T. Rahman, M. Yang, L. Sigal, Tribert: Full-body human-centric audio-visual representation learning for visual sound separation, 2021, arXiv preprint arXiv:2110.13412.
[237] S. Chen, P.-L. Guhur, C. Schmid, I. Laptev, History aware multimodal transformer for vision-and-language navigation, in: Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 5834–5847.
[238] Y. Li, A.W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q.V. Le, et al., Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17182–17191.
[239] T. Sanford, S.A. Harmon, E.B. Turkbey, D. Kesani, S. Tuncer, M. Madariaga, C. Yang, J. Sackett, S. Mehralivand, P. Yan, et al., Deep-learning-based artificial intelligence for PI-RADS classification to assist multiparametric prostate MRI interpretation: A development study, J. Magn. Resonance Imag. 52 (5) (2020) 1499–1507.
[240] M. Odusami, R. Maskeliūnas, R. Damaševičius, S. Misra, Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images, J. Med. Biol. Eng. (2023) 1–12.

[241] T.D. Vu, H.-J. Yang, V.Q. Nguyen, A.-R. Oh, M.-S. Kim, Multimodal learning using convolution neural network and sparse autoencoder, in: 2017 IEEE International Conference on Big Data and Smart Computing, BigComp, IEEE, 2017, pp. 309–312.

[242] C. Ge, I.Y.-H. Gu, A.S. Jakola, J. Yang, Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2018, pp. 5894–5897.

[243] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images, Sci. Rep. 8 (1) (2018) 5697.

[244] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, B. Lei, Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM, IEEE Access 7 (2019) 63605–63618.

[245] A. Massalimova, H.A. Varol, Input agnostic deep learning for Alzheimer's disease classification using multimodal MRI images, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2021, pp. 2875–2878.

[246] M. Abdelaziz, T. Wang, A. Elazab, Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks, J. Biomed. Inform. 121 (2021) 103863.

[247] X. Zhang, W. Lin, M. Xiao, H. Ji, Multimodal 2.5 D convolutional neural network for diagnosis of Alzheimer's disease with magnetic resonance imaging and positron emission tomography, Prog. Electromagn. Res. 171 (2021).

[248] E. Puyol-Antón, B.S. Sidhu, J. Gould, B. Porter, M.K. Elliott, V. Mehta, C.A. Rinaldi, A.P. King, A multimodal deep learning model for cardiac resynchronisation therapy response prediction, Med. Image Anal. 79 (2022) 102465.

[249] H.R. Al-Absi, M.T. Islam, M.A. Refaee, M.E. Chowdhury, T. Alam, Cardiovascular disease diagnosis from DXA scan and retinal images using deep learning, Sensors 22 (12) (2022) 4310.

[250] G. Dolci, M.A. Rahaman, J. Chen, K. Duan, Z. Fu, A. Abrol, G. Menegaz, V.D. Calhoun, A deep generative multimodal imaging genomics framework for Alzheimer's disease prediction, in: 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2022, pp. 41–44.

[251] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, P. Zhang, Alzheimer's disease diagnosis via multimodal feature fusion, Comput. Biol. Med. 148 (2022) 105901.

[252] S.Q. Abbas, L. Chi, Y.-P.P. Chen, DeepMNF: Deep multimodal neuroimaging framework for diagnosing autism spectrum disorder, Artif. Intell. Med. 136 (2023) 102475.

[253] H. Hoang Nguyen, M.B. Blaschko, S. Saarakkala, A. Tiulpin, Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data, 2022, arXiv e-prints, arXiv–2210.

[254] M. Nevitt, D. Felson, G. Lester, The osteoarthritis initiative, in: Protocol for the Cohort Study, Vol. 1, 2006.

[255] I.L. Xi, Y. Zhao, R. Wang, M. Chang, S. Purkayastha, K. Chang, R.Y. Huang, A.C. Silva, M. Vallières, P. Habibollahi, et al., Deep learning to distinguish benign from malignant renal lesions based on routine MR ImagingDeep learning for characterization of renal lesions, Clin. Cancer Res. 26 (8) (2020) 1944–1952.

[256] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2019) 104863, (2020).

[257] Q. Ying, X. Xing, L. Liu, A.-L. Lin, N. Jacobs, G. Liang, Multi-modal data analysis for alzheimer's disease diagnosis: An ensemble model using imagery and genetic features, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2021, pp. 3586–3591.

[258] Y. Mustafa, T. Luo, Diagnosing Alzheimer's disease using early-late multimodal data fusion with Jacobian maps, 2023, arXiv preprint arXiv:2310.16936.

[259] Y. Wei, X. Chen, L. Zhu, L. Zhang, C.-B. Schönlieb, S.J. Price, C. Li, Multi-modal learning for predicting the genotype of glioma, 2022, arXiv:2203.10852.

[260] M. Prabhushankar, K. Kokilepersaud, Y.-Y. Logan, S.T. Corona, G. AlRegib, C. Wykoff, OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics, 2022, http://dx.doi.org/10.5281/ZENODO.7105232, Zenodo. URL: https://zenodo.org/record/7105232.

[261] Z. Cai, L. Lin, H. He, X. Tang, Corolla: An efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, 2022, pp. 1–4, http://dx.doi.org/10.1109/ISBI52829.2022.9761712.

[262] Y. Gutiérrez, J. Arevalo, F. Martánez, Multimodal contrastive supervised learning to classify clinical significance MRI regions on prostate cancer, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2022, pp. 1682–1685, http://dx.doi.org/10.1109/EMBC48229.2022.9871243.

[263] X. Xing, Z. Chen, M. Zhu, Y. Hou, Z. Gao, Y. Yuan, Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading, in: L. Wang, Q. Dou, P.T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 636–646.

[264] A. Taleb, M. Kirchler, R. Monti, C. Lippert, ContIG: Self-supervised multimodal contrastive learning for medical imaging with genetics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 20908–20921.

[265] P. Hager, M.J. Menten, D. Rueckert, Best of both worlds: Multimodal contrastive learning with tabular and imaging data, 2023, arXiv:2303.14080.

[266] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 2–25.

[267] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[268] Z. Huang, F. Bianchi, M. Yuksekgonul, T.J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nat. Med. 29 (9) (2023) 2307–2316.