Expert Systems With Applications 200 (2022) 117006

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

# An overview of deep learning methods for multimodal medical data mining

Fatemeh Behrad, Mohammad Saniee Abadeh *

*Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran*

ARTICLE INFO

*Keywords:*
Deep learning
Multimodal medical data
Review

ABSTRACT

Deep learning methods have achieved significant results in various fields. Due to the success of these methods, many researchers have used deep learning algorithms in medical analyses. Using multimodal data to achieve more accurate results is a successful strategy because multimodal data provide complementary information. This paper first introduces the most popular modalities, fusion strategies, and deep learning architectures. We also explain learning strategies, including transfer learning, end-to-end learning, and multitask learning. Then, we give an overview of deep learning methods for multimodal medical data analysis. We have focused on articles published over the last four years. We end with a summary of the current state-of-the-art, common problems, and directions for future research.

## 1. Introduction

Medical data analysis using computers has always attracted many researchers, but deep learning methods have revolutionized this field over the past decade. The history of deep learning can be traced back to 1943 when McCulloch and Pitts (1943) created the first mathematical model of a neuron. Although this neuron had no learning mechanism, it laid the foundation for deep learning. Rosenblatt (1957) introduced the perceptron, which is a single-layer neural network with learning capabilities to do binary classification on its own. This invention inspired the revolution in the research of shallow neural networks. After a lot of research in this field, Ivakhnenko (1968) introduced the Group Method of Data Handling (GMDH) for training neural networks in 1968. These networks are widely considered the first deep neural networks of the feedforward multilayer perceptron type. The first convolutional neural network (CNN), called Neocognitron, and recurrent neural network (RNN) were introduced in 1980 and 1982 (Fukushima, 1980; Hopfield, 1982). Implementation of backpropagation in the neural networks by Rumelhart et al. (1986) opened gates for training complex deep neural networks easily. Three years later, LeCun et al. (1989) used backpropagation to train a CNN for handwritten digit recognition. This was a breakthrough moment as it laid the foundation of modern computer vision using deep learning. Hinton et al. (2006) proposed deep belief networks in which the training process is efficient for a large amount of data.

The earliest attempt at using graphics processing units (GPUs) for deep learning was a study by Chellapilla et al. (2006). They implemented a CNN using a GPU, which was four times faster than any equivalent implementation on CPU. However, the winner of the 2012 Imagenet Challenge image classification model, AlexNet (Krizhevsky et al., 2012), proved to be a landmark deep learning model with GPU acceleration. Many complex deep neural networks, such as VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), Inception (Szegedy et al., 2015), and DenseNet (Huang et al., 2017), have been introduced since 2012. All of these networks require high computational resources. Another complex neural network is the generative adversarial network (GAN), created by Goodfellow et al. (2014). The idea of this network is to synthesize realistic data from a random vector. Because of the data scarcity problem, GAN has become popular in medical analyses.

Motivated by the success of deep learning, researchers in medical fields have also attempted to apply deep learning-based approaches to different tasks, such as feature extraction, classification, segmentation in images, prognosis prediction of disease, and overall survival (OS) prediction. Also, some deep learning architectures have been proposed for a specific task in medicine. For example, several architectures have been introduced for medical image segmentation. The most popular one is U-Net, which was first introduced by Ronneberger et al. (2015). The main downside of U-Net is that it can only process 2D images while most medical images used in clinical practice consist of 3D volumes. To solve this problem, Milletari et al. (2016) designed V-Net, which segments volumetric medical images. Architectures such as U-Net++ (Zhou et al., 2018), U-Net 3 + H. Huang et al., 2020), R2U-Net (Alom et al., 2019), and attention U-Net (Oktay et al., 2018) have also been designed to improve the performance of U-Net. In 2021, inspired by the recent

* Corresponding author.
*E-mail address:* saniee@modares.ac.ir (M. Saniee Abadeh).

https://doi.org/10.1016/j.eswa.2022.117006
Received 1 June 2021; Received in revised form 25 December 2021; Accepted 26 March 2022
Available online 4 April 2022

success of transformers in Natural Language Processing (NLP), which leverage self-attention mechanisms and encode long-range dependencies, different transformer-based networks have been designed, such as UNETR (Hatamizadeh et al., 2021), TransUNet (J. Chen et al., 2021), Swin-U-Net (H. Cao et al., 2021), and MedT (Valanarasu et al., 2021). Fig. 1 shows the key developments in deep learning architectures along with state-of-the-art neural networks in medical fields.

In general, medical data divide into three categories: imaging, clinical, and omics data. As every single modality represents various important information, the combination of different modalities provides a more comprehensive view of disease. Therefore, multimodal medical data analysis reduces information uncertainty and improves models' performance. The most widely used imaging modalities are magnetic resonance imaging (MRI), computerized tomography (CT), positron emission tomography (PET), and single-photon emission computed tomography (SPECT). Multiparametric MRI also provides complementary information by combining different MRIs, such as T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and Fluid attenuation inversion recovery (Flair) images. Moreover, clinical data, such as antecedents, age, sex, and medical treatments, help physicians better understand patients' characteristics and disease evolution. Furthermore, genomic data are beneficial in medically-relevant prediction and diagnosis of disease progression (Bell, 2004; Schrodi et al., 2014).

When we use different modalities, we should decide how to integrate their information. Generally, there are three fusion strategies: input-level, layer-level, and decision-level fusion. In input-level fusion, we integrate the information of different modalities before giving them to a single network. In layer-level fusion, one or more modalities are given to the network independently, then their intermediate representations are fused in a layer of the network. In input-level and layer-level fusion, all modalities must be available for each sample in the training set. This is a serious downside because this situation is rarely satisfied in the medical field. However, these two methods find the relationship between different modalities better than decision-level fusion. In decision-level fusion, each modality is used as a single input to train a single neural network. The outputs of individual networks will then be integrated to get the final result. Decision-level fusion is also popular because it does not need all modalities for each sample. Also, individual networks better exploit the unique information of their corresponding modality because the search space in decision-level fusion is much smaller than other fusion methods.

In recent years, the number of articles on multimodal medical data analysis using deep learning has increased steadily. We performed a thorough literature analysis with keywords 'deep learning,' 'medical,' and 'multimodal' on the PubMed database on August 14, 2021. We observed that the number of papers had increased since 2010, which means multimodal medical data analysis using deep learning is obtaining more and more attention (See Fig. 2). We also conducted a similar analysis on the Google Scholar search engine, which showed the same trend.

There are some other reviews on multimodal medical data analysis using deep learning. Some of them gave a detailed review of deep learning applications in medical image analysis. For instance, Zhou et al. (2019) proposed a general pipeline for multimodal medical image segmentation based on deep learning. This pipeline consists of data preparation, network architecture, fusion strategy, and data post-processing. They also compared the results of different deep learning architectures in multimodal medical image segmentation. Xu (2019) introduced a series of studies on deep learning applications in multimodal medical image analysis, emphasizing fusion techniques and feature extraction deep models. Litjens et al. (2017) reviewed major deep learning concepts pertinent to medical image analysis. They summarized deep learning techniques used in medical image analysis and identified the challenges for successful deep learning applications in medical imaging tasks. Zhang et al. (2020) presented an overview of multimodal data fusion in neuroimaging. They also outlined the strengths and limitations of imaging modalities, fundamental fusion rules, fusion quality assessment methods, and current challenges in multimodal fusion. Moreover, they summarised current developments and applications of multimodal neuroimaging in terms of neurological disorders and brain diseases. Ramachandram and Taylor (2017) introduced various applications in which multimodal deep learning had attracted great attention. These applications included human activity recognition, medical applications,

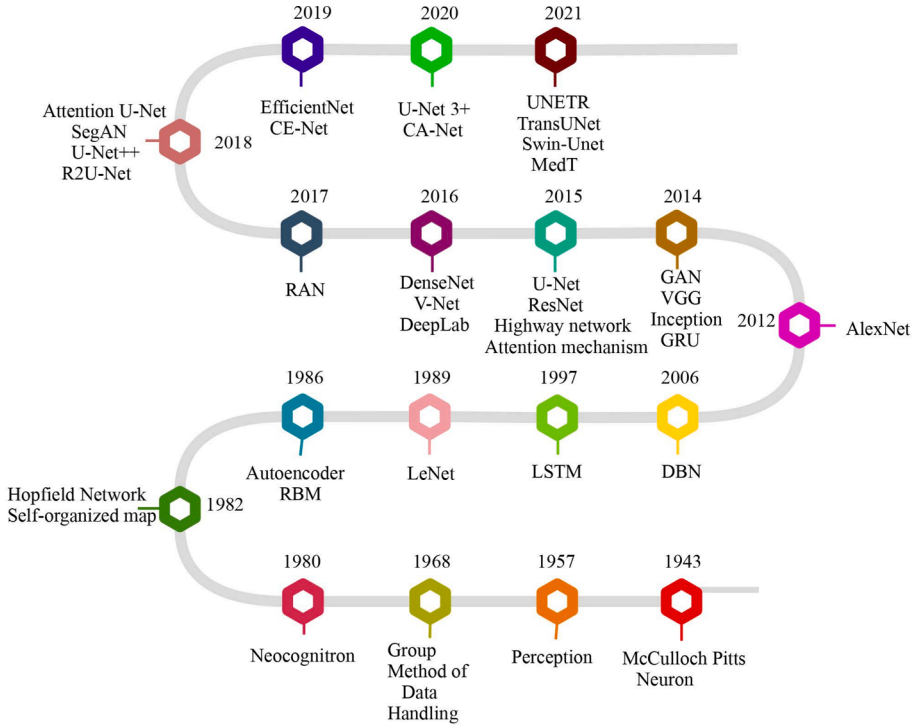


**Fig. 1.** Key developments in deep learning architectures and state-of-the-art neural networks for medical data analysis.

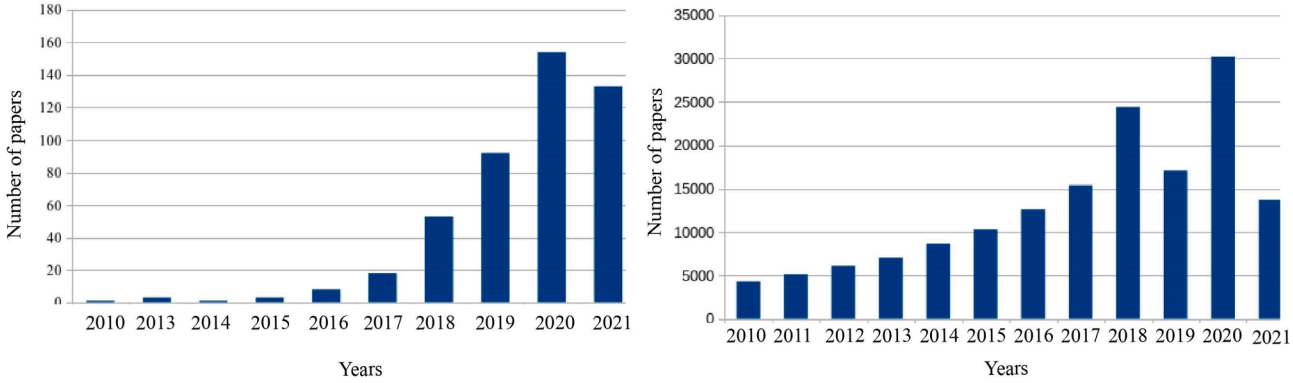


**Fig. 2.** The general trend in multimodal medical data analysis using deep learning obtained from PubMed (left) and Google Scholar (right).

autonomous systems, and multimedia applications. They also reviewed several options, including stochastic regularization, casting architecture optimization, and incremental online reinforcement learning, to learn an optimal architecture.

On the other hand, some review papers focused on omics data and their combinations with other modalities. For example, Antonelli et al. (2019) focused on integrating omics and imaging data. Also, they summarized features extracted by omics imaging data and methods adopted for their analysis and integration. Lo Gullo et al. (2020) explained data related to radiomics and radiogenomics and the differences between them. They reviewed the radiomics and radiogenomics literature in oncology, focusing on breast, brain, gynecological, liver, kidney, prostate, and lung malignancies. Unlike previous review papers, we have not limited our study to one or two modalities. As a result, we cover imaging, clinical, omics, and other types of data simultaneously. To the best of our knowledge, we are the first to review deep learning applications in multimodal medical data analysis without constraints on the data type.

The rest of the paper is structured as follows. In Section 2, we introduce four important decisions on multimodal medical data analysis using deep learning. Section 3 describes the most commonly used modalities in multimodal medical data analysis. When dealing with multimodal data, it is essential to know how to integrate information from different modalities. Consequently, various fusion techniques are explained in Section 4. In Section 5, we present the most popular deep learning architectures, which include CNN, GAN, inception network, U-Net, VGG network, ResNet, RNN, and attention neural network. Section 6 describes different learning strategies, and section 7 reviews papers on deep learning applications in multimodal medical data analysis. We focus on articles published in the last four years. Section 8 introduces some of the most well-known multimodal datasets in the medical field. Finally, we summarize common problems and open challenges in sections 9 and 10.

## 2. Multimodal deep learning process

In the first step of multimodal medical data analysis, researchers should decide on data sources, fusion strategy, learning strategy, and deep learning architecture (as shown in Fig. 3). Choosing the right combination of data sources in multimodal analyses is critical because a wrong combination leads to lower performance. Data sources should provide complementary information to improve results. The next step is to decide how to integrate different modalities. Furthermore, a suitable learning strategy should be chosen. Finally, researchers should choose a network architecture. Knowing different deep learning architectures helps find the most suitable architecture for research. In the following sections, these concepts are explained.

## 3. Data

### 3.1 Imaging data

Image data coming from different imaging technologies are fundamental for medical analysis. Biomedical images reveal a lot of information about human organs' structure and functions (hemodynamic, metabolic, and chemical processes) (Antonelli et al., 2019). Many researchers have used a combination of imaging modalities in recent years. There are two reasons why the integration of imaging modalities is beneficial. First, all individual modalities have their limitations. Second, a disease, disorder, or lesion may manifest itself in different forms, symptoms, or etiology. On the other hand, different diseases may share some common symptoms or appearances. Therefore, an individual image modality may not reveal a complete picture of a disease (Zhang et al., 2020).

As medical images can be 2D or 3D, we should decide on images' dimensions before training a network. The 2D approach takes image slices extracted from the 3D image and feeds them to the network. This approach reduces the computational cost, but it ignores the spatial information of images in the z-direction (T. Zhou et al., 2019). We can also use 3D images and feed them directly to the network. This approach can be expensive, but it does not lose any information. In this section, we introduce some of the most popular imaging modalities and mention their advantages and disadvantages.

#### *3.1.1. Ct*

CT is undoubtedly one of the most important technologies in medical imaging and offers us views inside the human body that are so valuable to physicians (Maier et al., 2018). CT is an imaging procedure in which an x-ray tube rotates around a patient, shooting narrow beams of X-rays to the patient. This procedure produces signals measured by a computer to generate cross-sectional images of the organ under investigation (Antonelli et al., 2019). The primary strength of this modality is that it provides clear anatomical structure information due to its excellent spatial resolution (Lin & Alessio, 2009). In other words, CT images convey details and discriminate between structures located within small

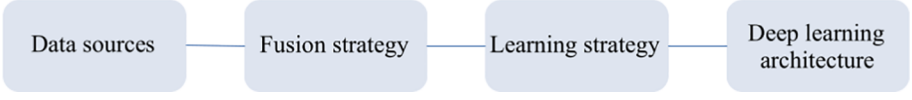


**Fig. 3.** Four decisions on multimodal medical data analysis using deep learning algorithms.

proximity to each other very well. Also, CT images are non-invasive, quick, and painless. On the other hand, CT images cannot identify soft tissues well (Zhang et al., 2020). Furthermore, because of exposure to ionizing radiation, the possibility of developing cancer increases later in life (Kasban et al., 2015). An example of a CT scan is shown in Fig. 4.

#### 3.1.2. Pet

PET provides information about the metabolism of a disease. In PET images, the diseased areas, such as tumor and inflammation, appear as 'hot' areas, reflecting high contrast to the normal surrounding tissues. This high contrast in PET images distinguishes malignant areas from normal tissues easily (Ju et al., 2015). There are some advantages to this technique. First, this technique can check how far cancer has spread and how well the treatment is working. Second, it can provide highly accurate functional information, but this method has some downsides. The main one is that using ionizing radiation makes patients radioactive for a variable period. Also, it has relatively low spatial resolution and high cost (Kasban et al., 2015). This low spatial resolution makes target boundaries blur, so detecting tumors just by PET images is challenging. In the last few years, PET images have been fused with CT or MRI to capture information from different modalities and put this information in one image. Fig. 4 shows an example of PET/CT and PET/MRI (Boss et al., 2010).

#### 3.1.3. SPECt

The goal of SPECT is to determine the three-dimensional radioactivity distribution which is resulted from the radiopharmaceutical (a radioactive-labeled pharmaceutical) uptake inside a patient. SPECT and PET are very similar. For example, both of them use radioactive tracer and detect γ-rays to reflect functional information about a disease. However, unlike PET, the radioisotopes used for SPECT emit only a single γ-ray during decay. Also, SPECT scans produce lower resolution images than PET. SPECT scans are significantly less expensive than PET scans because the nuclides used in SPECT have a longer half-life and are more easily obtained than PET (Zhang et al., 2020). Fusion of SPECT and other modalities provide more information than SPECT images alone. For example, hybrid SPECT/CT adds clinical value over SPECT imaging due to more precise anatomical lesion localization ("Mathematics and Physics of Emerging Biomedical Imaging" 1996). Fig. 5 shows an example of SPECT/CT (Jacene et al., 2008).

#### 3.1.4. MRi

Modern MRI systems allow physicians to look inside the body without ionizing radiation. They provide excellent soft-tissue contrast and high spatial resolution for morphological imaging as well as a range of possibilities for functional imaging (Maier et al., 2018). MRI is also non-invasive and painless. However, relatively low sensitivity, long scan, long post-processing time, and being expensive are its downsides. Moreover, this technique cannot detect intraluminal abnormalities (Kasban et al., 2015). Multiparametric MRI provides complementary information due to its dependence on variable acquisition parameters, which improves models' performance. MRI modalities include T1, T1c, T2, and Flair images, as shown in Fig. 6. For further details about different imaging techniques, see Table 2 in Kasban et al. (2015).

### 3.2 Omics data

Omics data characterize the behaviors of cells, tissues, and organs at a molecular level and provide a comprehensive understanding of the etiology of human diseases (Raja et al., 2017). Omics data are high-dimensional, but only a small subset of them has important implications (Kristensen et al., 2014). As a result, feature selection and feature extraction strategies are helpful when we use omics data. Genomics, transcriptomics, proteomics, and epigenomics are the four main categories of omics data. Although there are some other categories, such as metabolomics and lipidomics, we only describe four main groups.

Genomics is the study of the genomes of organisms. The genome is

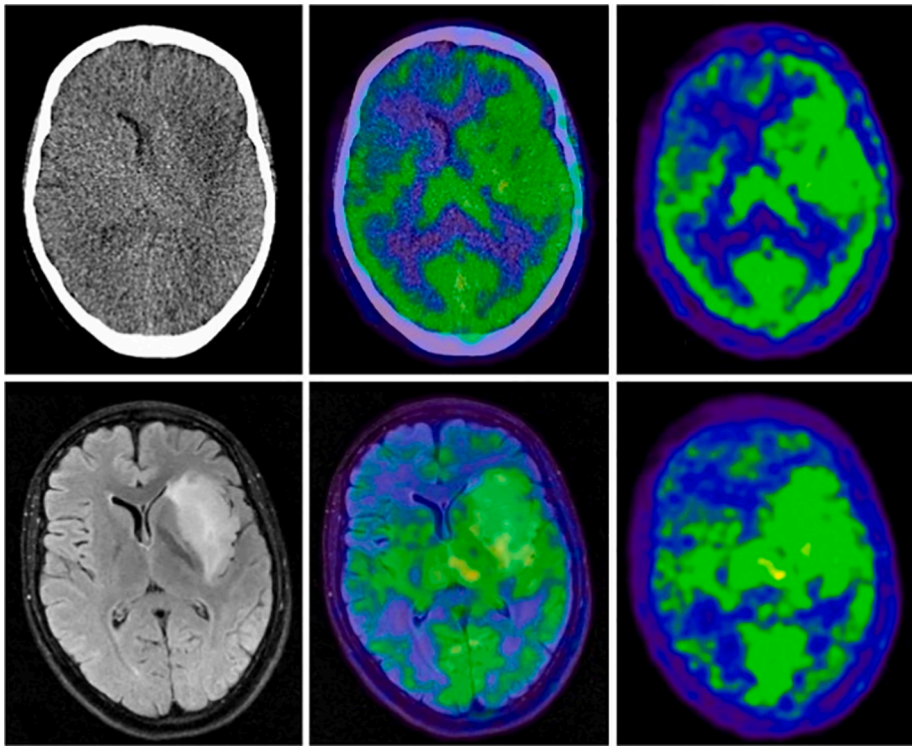

**Fig. 4.** PET/MRI and PET/CT images of a 30-year-old patient with low-grade glioma. (Top) low-dose non–contrast-enhanced CT image (left), PET/CT (center), and PET images (right). (Bottom) T2-weighted Flair image (left), PET/MR (center), and PET image (right).

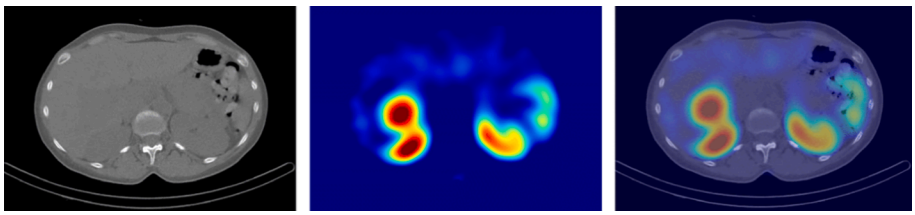

**Fig. 5.** CT-image (left), SPECT-image (center), and hyprid SPECT/CT-image (right).

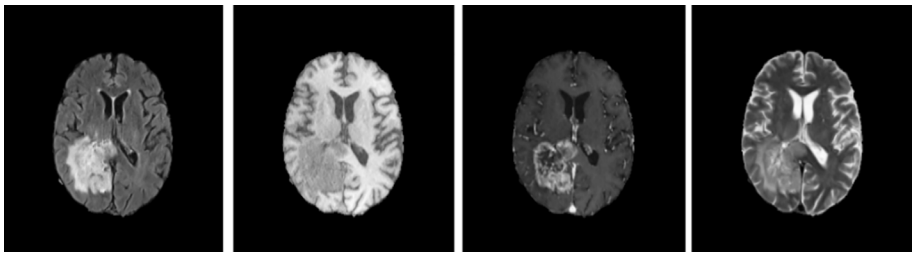

**Fig. 6.** Different MRI modalities in the BraTS dataset. From left to right: Flair, T1, T1ce, and T2 image.

the complete sequence of DNA in a cell or organism. Complete or partial DNA sequences can be assayed using various experimental platforms, such as single-nucleotide polymorphisms (SNP). Genomic analysis can detect insertions, deletions, and copy number variation (CNV), referring to the loss of or amplification of the expected two copies of each gene (one from the mother and one from the father at each gene locus). Gene sequences and regulatory motifs are other types of genomic data. The transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue. The transcriptome includes ribosomal RNA (rRNA), messenger RNA (mRNA), transfer RNA (tRNA), micro RNA (miRNA), and other non-coding RNA (ncRNA; Götz, 2019). Proteomics is concerned with the structure, function, and modification of proteins expressed in a biological system. Proteomics data include protein expression and protein structure. Epigenomics characterizes the epigenetic modifications of the genome and aims to understand the regulations of the gene expression (Raja et al., 2017).

There is a correlation between different types of omics data. For example, Fig. 7 illustrates a two-step process by which a sequence of nucleotides from DNA is converted into a sequence of amino acids to build the desired protein (Betts et al., 2013). These two steps are called transcription and translation.

### 3.3 Clinical data

Different types of clinical data can be used in medical analyses. The results of physiological measurements (lab results, vital signs), demographic information (gender, location, age, and marital status), payment and insurance information (which is indirectly related to the disease), and clinical notes are different groups of clinical data. Clinical notes include descriptions of lab test results, physician diagnoses, drugs, and treatments. Clinical notes may also contain other information such as the chief complaint, family history, medical history, and allergies (Yu et al., 2019). Usually, NLP techniques are used to process clinical notes. Some researchers have integrated clinical data with other data types to improve performance.

### 3.4 Others

Depending on the type of disease, other data types are also available. For example, molecular aberrations in Alzheimer's disease (AD) are reflected in the cerebrospinal fluid (CSF), so many studies on AD have used the combination of CSF with other modalities (Kim & Lee, 2018; Lee, Kang, Nho, Sohn, & Kim, 2019; Lee, Nho, Kang, Sohn, & Kim, 2019;

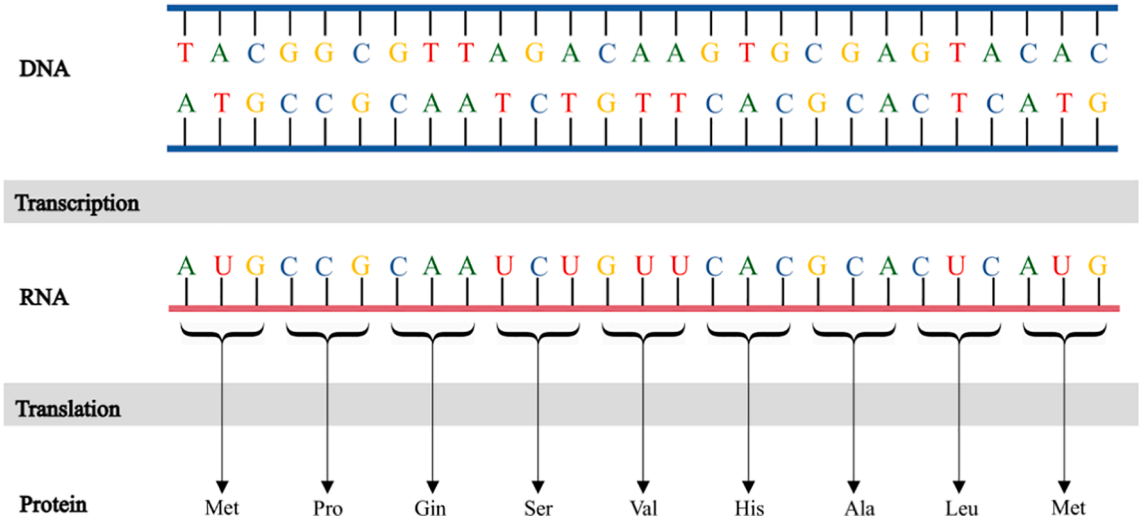


**Fig. 7.** The flow of information from DNA to proteins.

Lin et al., 2020). Another modality is electroencephalography (EEG) which is a monitoring method to record the brain's electrical activity and helps physicians be more accurate in diagnosing some diseases such as epilepsy (Hosseini et al., 2020), sleep disorders, depth of anesthesia, coma, encephalopathies, AD, and brain death (Shikalgar & Sonavane, 2020). Another modality is fluorescein angiography (FA) image which has been used in several multimodal eye analyses using deep learning (Hervella et al., 2020, 2019; Li et al., 2020; Vaghefi et al., 2020). FA is a medical procedure in which a fluorescent dye is injected into the bloodstream. The dye highlights the blood vessels in the back of the eye, which helps photograph vessels.

### 3.5 Selection of appropriate modalities

Although multimodality was not very popular in machine learning before, it has recently gained a lot of interest. The main reason why few studies used multimodal data in the past is that it used to be very difficult to access a multimodal dataset. Also, most of the papers and challenges only focused on one modality. Consequently, models were optimized for a single data type and a specific task. Several multimodal data repositories, which contain hundreds of matched patient samples for imaging, genomic, and clinical data, have been created in recent years. These repositories let researchers benefit from multimodality and adopt a more comprehensive multimodal data approach to tackle more challenging and global tasks. We introduce some of these data repositories in Section 8.

Multimodal data help researchers have the most comprehensive understanding of patients, their background, and their disease evolution because each modality provides different important information. For example, clinical data help understand patients' characteristics and biological differences in their disease evolution. Genomic data provide prognostic signals, which are not accessible through other modalities, so they are beneficial in medically-relevant prediction and diagnosis of disease progression (Micheel et al., 2012). Similarly, each imaging modality provides different types of biological information about a disease. For instance, CT images diagnose muscle and bone disorders, such as bone tumors and fractures, while MRI offers a good soft-tissue contrast without radiation. Functional images, such as PET and SPECT, lack anatomical characterization but provide quantitative metabolic and functional information about diseases (Bhatnagar et al., 2015). Histology slides help clinicians understand the structure of the problem at a cellular scale. Additionally, analyzing the lymphocyte infiltration on a histology slide is an excellent indicator of organism resistance. In the case of brain tumor segmentation, various imaging modalities can be used to map brain tumor-induced tissue changes. For instance, T2 and Flair MRI highlight differences in tissue water relaxational properties and detect the tumor with peritumoral edema. However, T1 and T1c detect the tumor core without peritumoral edema. MRSI shows relative concentrations of selected metabolites, while post gadolinium T1 MRI shows pathological intratumoral take-up of contrast agents. Perfusion and diffusion MRI show local water diffusion and blood flow (Menze et al., 2015; Zhou et al., 2019).

Choosing the best combination of modalities is crucial because a wrong combination leads to a bad performance. Knowledge of the limitations and strengths of available modalities and an understanding of disease biology help determine which modalities are optimal for a task. Also, reading related articles helps choose the right combination of modalities. For example, for disease assessment in Hodgkin lymphoma, Kanoun et al. (2018) noted that FDG-PET plus CT increased the sensitivity from 10% to 20% compared to conventional CT. Many researchers have integrated genomic data with other modalities to make more precise predictions about a disease (Bell, 2004; Schrodi et al., 2014). For example, Chen et al. (2019a) demonstrated that combining clinical data and gene expression data improves the accuracy of prognostic prediction for breast cancer patients. Lai et al. (2019) integrated gene expression and clinical data using a deep model to predict the 5-year survival of non-small cell lung cancer patients. In short, reading related papers guides new researchers in multimodal medical data analysis a lot. Table 1 summarizes a list of papers that employed the most popular combinations of modalities.

## 4. Fusion structures

When we work with multimodal data, we need to decide how to integrate them. There are three ways to fuse different modalities: input-level fusion, layer-level fusion, and decision-level fusion. In this section, we describe these methods.

### 4.1 Input-level fusion

This method is also known as early integration, feature-based integration, and data-based integration. In this method, different modalities are fused before conducting an analysis, as illustrated in Fig. 8(a). One benefit of input-level fusion is that it finds the relationship between different modalities. However, to find this relationship, all modalities must be available for each sample in the training set, which is hardly satisfied in practice. Another disadvantage of this method is that it leads to a very large feature vector, which causes a high computational cost.

### 4.2 Layer-level fusion

This method is also known as intermediate integration and transformation-based integration. In this method, one or more modalities are given to a network independently, then their intermediate representations are fused in a layer of the network. Like input-level fusion, this method finds the relationship between different modalities and requires all modalities for each sample in the training set. Fig. 8 (b) illustrates this method.

### 4.3 Decision-level fusion

Late integration and model-based integration are the other names of this method. In this method, each modality is used as a single input to train a neural network, then the outputs of models are fused to make the final decision (see Fig. 8(c)). This method is prevalent because it does not require all modalities for each sample. Unlike other fusion techniques, this technique cannot find the relationship between different modalities. However, models which use decision-level fusion can

**Table 1**
Different combinations of modalities used in the reviewed articles.

| Combination | Articles |
|---|---|
| Multiparametric MRI | Abrol et al. (2019); Ge et al. (2020); Isensee et al. (2019); Jiang et al. (2020); Jiang et al. (2020); Liang et al. (2018); McKinley et al. (2019, 2020); Milecki et al. (2021); Myronenko (2019); Nie et al. (2019); Saba et al. (2020); Soltaninejad et al. (2019); Taleb et al. (2021); Tang et al. (2020); Varghese et al. (2016); Wang et al. (2018); Wang et al. (2020); Zhao et al. (2020) |
| PET / CT | Guo, et al. (2019); Kirienko et al. (2018); Li et al. (2019); Peng et al. (2019); Qin et al. (2020); Rubinstein et al. (2019); Shi et al.(2018); Zhao et al. (2018); Zhao et al. (2020); Zhou et al. (2018) |
| MRI / PET | Feng et al. (2019a); Liu et al. (2018); Lu et al. (2018); Shi et al. (2018); Suk et al. (2014); Vu et al. (2018); Zhang and Shi (2020a) |
| CT / X-ray | El Asnaoui and Chawki (2020); Kassani et al. (2020); Maghdid et al. (2020); Mukherjee et al. (2020); Rehman et al. (2020); Zhang et al. (2021) |
| Omics / Clinical | Chen et al. (2019a); Hooshmand et al. (2020); Lai et al. (2019); Sun et al. (2019) |
| CT / Clinical | Bai et al. (2020); Lassau et al. (2020); Xu et al. (2020) |
| MRI / CT | Cao et al. (2020); Ma et al. (2018); Xu et al.(2021) |
| MRI / Clinical | Huang and Chung (2020); Liu et al. (2019) |

**Table 2**
The most common deep learning architectures in multimodal medical data analysis.

| Architecture | Articles |
|---|---|
| CNN | Cheerla and Gevaert (2019); El-Sappagh et al. (2020); Feng et al. (2019b); Ge et al. (2020); Li, et al. (2019); Hosseini et al. (2020); Huang and Chung (2020); Kirienko et al. (2018); Li et al. (2019); Liu et al. (2018); Liu et al. (2019); Ma et al. (2018); Milecki et al. (2021); Nie et al. (2019); Peng et al. (2019); Qiu et al. (2020); Shi et al. (2018); Shikalgar and Sonavane (2020); Tang et al. (2020); Wang et al. (2018); Zhang et al. (2019); Zhang et al. (2021); Zhou et al. (2018) |
| Inception | El Asnaoui and Chawki (2020); Kassani et al. (2020); Vaghefi et al. (2020) |
| U-Net | Hervella et al. (2020, 2019); Isensee et al. (2019); Jiang et al. (2020); Wang et al. (2020); Zhao et al. (2020); Zhao et al. (2020) |
| VGGNet | El Asnaoui and Chawki (2020); Jiang et al. (2020); Kassani et al. (2020); Liu et al. (2018); Saba et al. (2020); Yan et al. (2019) |
| ResNet | Abrol et al. (2019); El Asnaoui and Chawki (2020); Lassau et al. (2020); Rehman et al. (2020); Vaghefi et al. (2020); van Sonsbeek and Worring (2020); Yap et al. (2018); Zhang and Shi (2020b) |
| DenseNet | El Asnaoui & Chawki (2020); Guo, et al. (2019); Kassani et al. (2020); Liang et al. (2018); Qin et al. (2020); Wang et al. (2020) |
| RNN | Lee, Kang, et al. (2019); Lee, Nho, et al. (2019); Shukla and Marlin (2020) |
| LSTM | Bagheri et al. (2020); Bai et al. (2020); El-Sappagh et al. (2020); Feng et al. (2019b); Hosseini et al. (2020); van Sonsbeek and Worring (2020); Zhang et al. (2020) |
| FCN | Ali et al. (2020); Bagheri et al. (2020); Cheerla and Gevaert (2019); Chen et al. (2019b); Hung et al. (2019); Lai et al. (2019); Soltaninejad et al. (2019); Sun et al. (2019) |
| Auto-encoder | Khamparia et al. (2019); Myronenko (2019); Rubinstein et al. (2019); Varghese et al. (2016) |
| GAN | B. Cao et al. (2020); Ge et al. (2020); Li et al. (2020) |
| RBM | Hooshmand et al. (2020); Suk et al. (2014) |
| Attention-based | Chen et al. (2019a); Zhang and Shi (2020b); Zhang et al. (2021) |

achieve high performance because the search space is smaller than other fusion techniques. In this technique, information is independently learned from different modalities, so the likelihood of overfitting is lower than other methods. There are a variety of techniques for decision-level fusion (Rokach, 2010). The most popular technique is based on majority voting, in which after training each model separately, the final prediction is chosen based on the majority of the predictions of the individual networks (Shikalgar & Sonavane, 2020).

## 5. Deep learning

### 5.1 Introduction to deep neural network

Neural networks are built of many neurons with specific activation functions. Each set of neurons belongs to a hidden layer, many of which create a deep neural network. Deep learning methods are a kind of representation learning algorithms that allow a machine to automatically discover the needed representation from the raw input data (LeCun et al., 2015).

### 5.2 Important architectures in deep learning

This section introduces some of the most popular neural network architectures, which greatly impact medical analyses. Fully connected networks (FCNs), CNNs, and RNNs are mainly used for supervised tasks. On the other hand, unsupervised neural networks, including GANs, restricted Boltzmann machines (RBMs), and auto-encoders can be employed in the absence of labeled data. Moreover, a combination of these techniques can be used in semi-supervised tasks. The detailed list of studies using each deep learning architecture in multimodal medical data analysis is shown in Table 2.

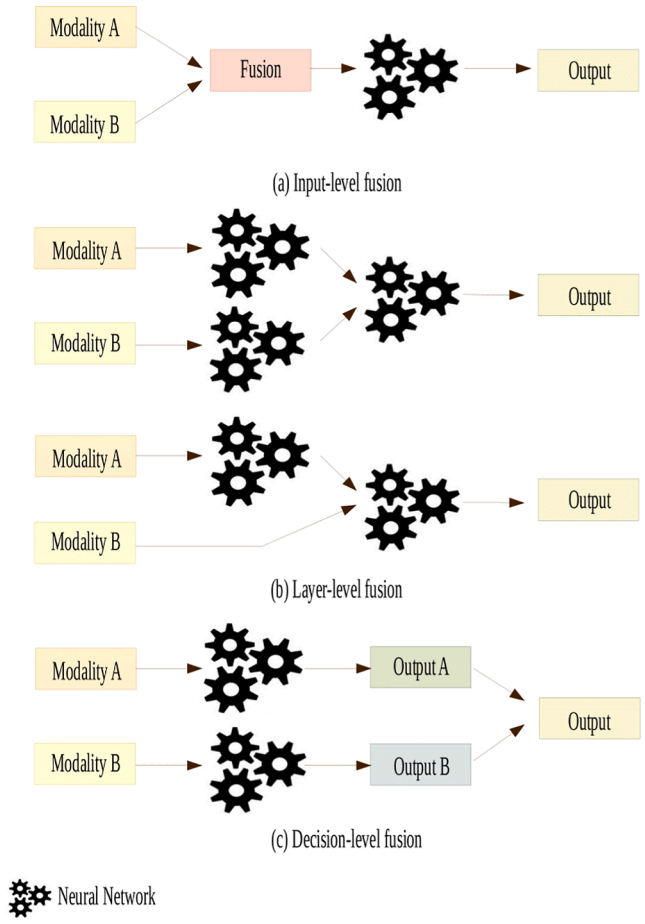


**Fig. 8.** The illustration of various fusion strategies for multimodal learning. (a) Input-level fusion, (b) Layer-level fusion, and (c) Decision-level fusion.

#### 5.2.1. CNn

CNN architecture comprises three basic layers: convolutional layer, pooling layer, and fully connected layer. The convolutional layer uses multiple filters for extracting high-level features. The pooling layer decreases the spatial size of feature maps obtained from the convolutional layer, which leads to a decrease in the computational power required to process image data. It also causes translation invariance, which is the ability to ignore translations of the target in the input. Usually, after convolutional and pooling layers, there are some fully connected layers. Various modifications, such as structural reformulation, regularization, and parameter optimizations, have been made in CNN architecture from 1989 until today (Kanoun et al., 2018). Fig. 9 illustrates an example of a basic CNN architecture. Although CNNs are mainly used for image analysis, some studies have used them for sequential data analysis (Zhang & Wallace, 2015). We explain the most popular CNN architectures in this section.

##### 5.2.1.1. Inception network

Inception network, also known as GoogleNet, achieved perfect results for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14) (Szegedy et al., 2015). Finding the optimal value for filter size, which is one of the hyperparameters of neural networks, is challenging. The main idea of the inception network is to apply various filter sizes and max-pooling simultaneously. An inception network consists of inception modules. In an inception module, $1 \times 1$, $3 \times 3$, $5 \times 5$ filters are applied to the output of a layer. Then, their outputs are concatenated to make a large vector, which is the input of the next layer. However, this concatenation would inevitably increase

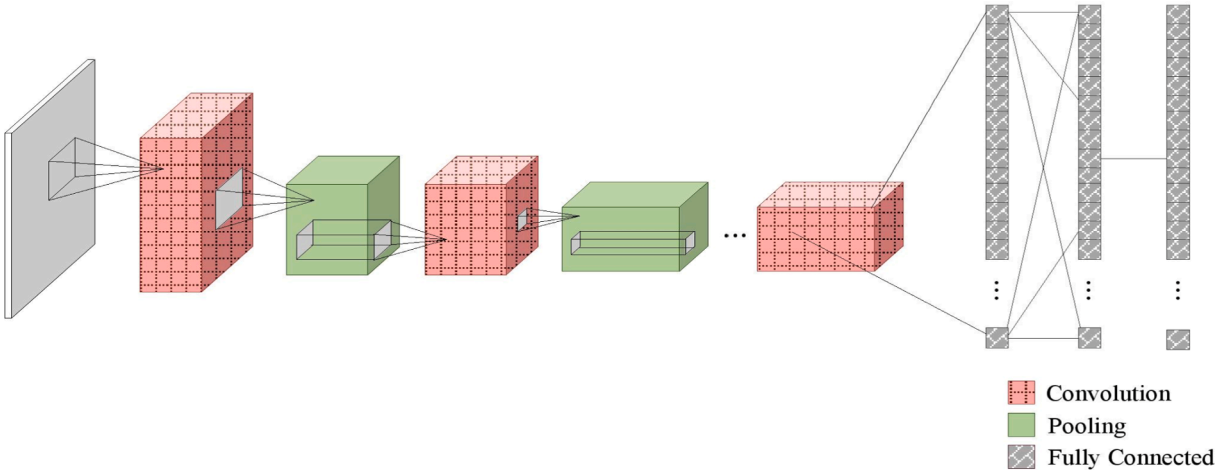


**Fig. 9.** An example of CNN.

the number of outputs from stage to stage. To solve this problem, a 1 × 1 filter is placed before 3 × 3 and 5 × 5 filters (Fig. 10).

*5.2.1.2. U-Net*

The U-net architecture, the winner of ISBI Cell Tracking Challenge 2015, has achieved good performance on different biomedical segmentation applications (Ronneberger et al., 2015). This network only needs a few annotated images. As shown in Fig. 11, the architecture consists of a contracting path to capture context and an expansive path that enables precise localization. The contracting path comprises the repeated application of two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. Finally, a 1x1 convolution maps each feature vector to the desired number of classes.

*5.2.1.3. ResNet*

ResNet, which won first place in the classification task of ILSVRC 2015, was proposed to ease the training of very deep neural networks (He et al., 2016). Increasing the network depth may cause the vanishing gradient problem. ResNet attempts to solve this problem by the idea of skip connection or shortcut, which adds the input of one layer to the output of the linear function of the next layer. This structure is called a residual block, and ResNet is built by stacking these blocks. Fig. 12 shows the residual block.

*5.2.1.4. VGG network*

The input of this architecture is a fixed-sized 224 × 224 × 3 image. The image is passed through a stack of convolutional layers, where all filters are 3 × 3 with stride one and same padding. Some of the convolutional layers are followed by max-pooling, which is performed over a 2 × 2 pixel window with stride 2. Convolutional layers are followed by three fully connected layers: the first two have 4096 channels, and the third one, which is a soft-max layer, contains 1000 channels (Simonyan & Zisserman, 2014). What is explained here is VGG16 (Fig. 13). Another version of the VGG network is VGG19 that is even deeper than VGG16 but does almost as well as VGG16. The main downside of this network is its large number of parameters (i.e., around 138 million). However, the uniformity of this architecture has attracted many researchers.

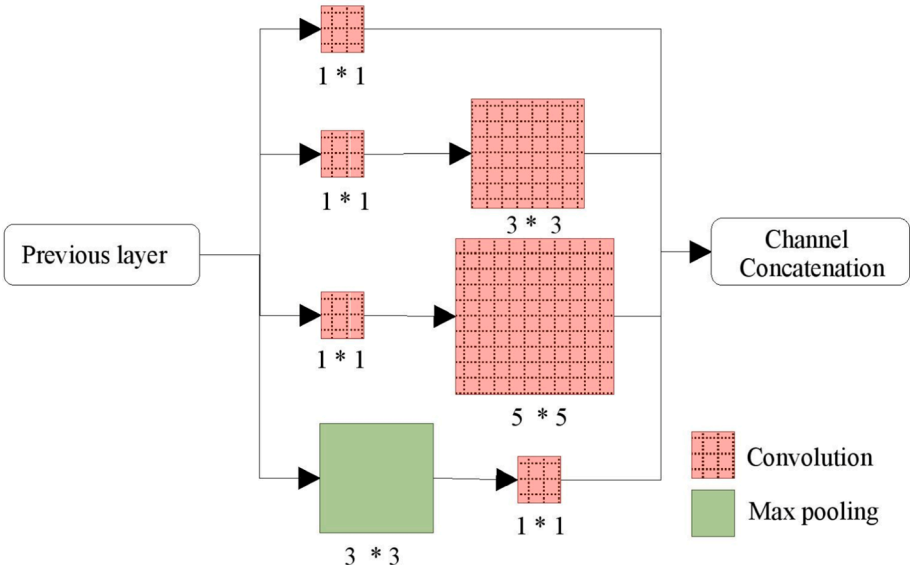


**Fig. 10.** The inception module.

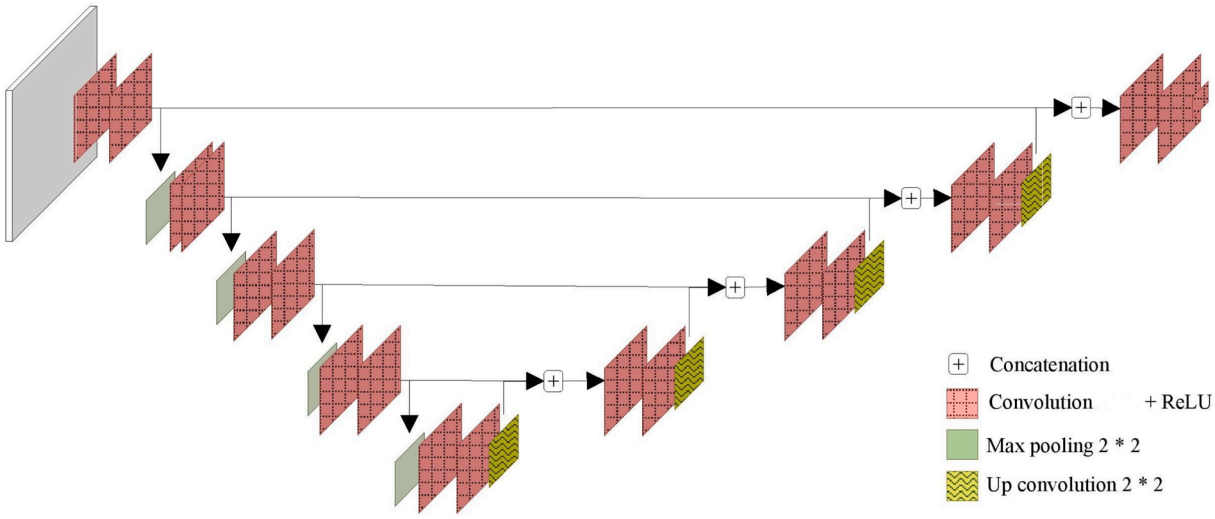


**Fig. 11.** U-net which consists of a contracting path (left side) and an expansive path (right side).

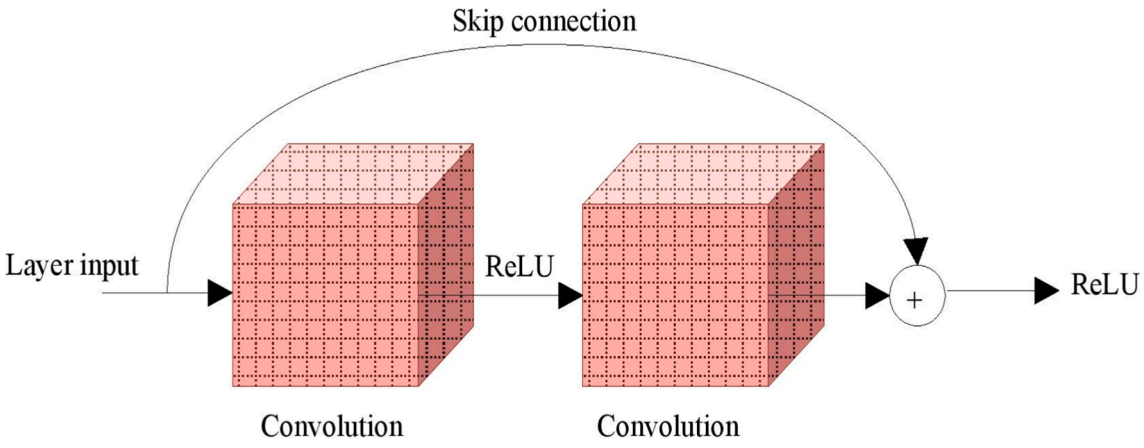


**Fig. 12.** The residual block.

#### 5.2.2. RNn

RNNs are a class of neural networks that successfully model sequence data (Mandic & Chambers, 2001). Sequential data are ordered data in which related things follow each other, such as a DNA sequence. The basic premise of a traditional RNN is to parse every item in an input series, one after the other, and keep updating its "hidden state" vector every step of the way. At the end of every step, this hidden vector learns to represent the context of all prior inputs. Therefore, when it makes a decision, it considers the current input and what it has learned from previous inputs. This is an important advantage of RNN because a sequence of data contains important information about what is coming next. Another advantage of RNN is its ability to process variable-length sequence input. Also, input size does not affect the RNN size. An example of a traditional RNN is shown in Fig. 14.

On the other hand, the basic RNN is not very good at capturing long dependencies. Another disadvantage of the basic RNN is the vanishing gradient problem. Gated recurrent unit (GRU) is an extension of RNN, which makes the basic RNN better in facing these two problems (Chung et al., 2014). Long-short-term memory (LSTM) is another extension of RNN, which is more powerful than GRU (Hochreiter & Schmidhuber, 1997). In LSTM, memory is extended, so it is well suited to learn from meaningful experiences that have very long time lags in between.

#### 5.2.3. Attention neural network

Attention models have recently become very popular within the artificial intelligence community as an essential component of neural architecture (Alzubaidi et al., 2021). The intuition behind attention models can be explained using human biological systems. Our visual processing system tends to focus selectively on some parts of an image while ignoring other irrelevant information in a manner that can assist in perception (Chaudhari et al., 2019). Similarly, in several problems such as language and speech, some parts of the input are more important than others. The attention mechanism allows a model to dynamically pay attention to only certain parts of the input that help perform the task effectively. Important parts of the data are chosen based on the context and learned through training procedure by gradient descent. One of the reasons why attention models have become so popular is that they improve the interpretability of neural networks, which are otherwise considered black-box models. This is a great benefit mainly because of the growing interest in the fairness, accountability, and transparency of machine learning models, especially in applications that influence human lives. Furthermore, the attention mechanism helps overcome some challenges with RNNs, such as performance degradation with an increase in the input size and the computational inefficiencies resulting from sequential processing of the input (Xu et al., 2015).

The first attention model, proposed by Bahdanau et al. (2014), is shown in Fig. 15. This architecture consists of two RNNs, one of which is

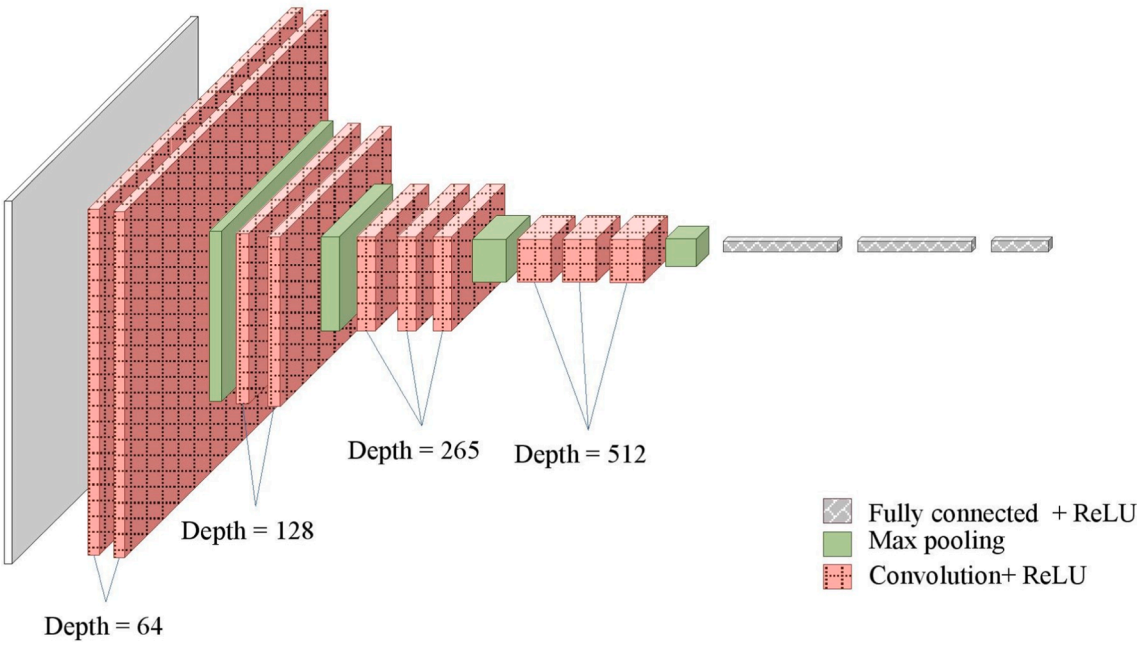


**Fig. 13.** VGG16.

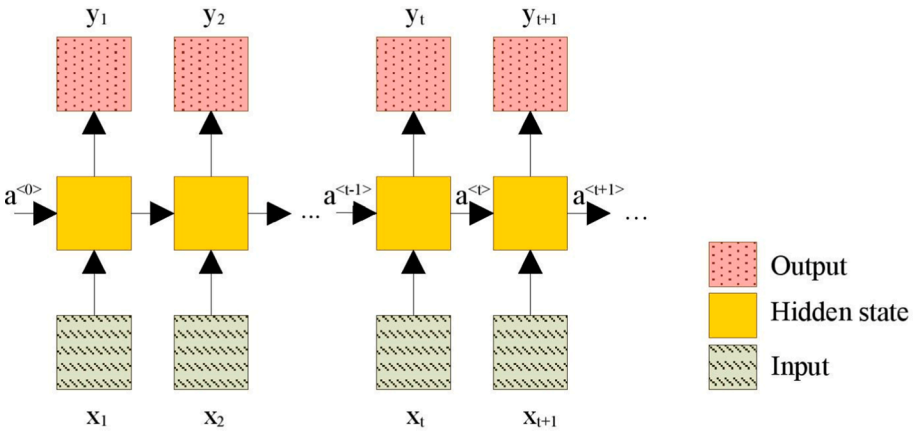


**Fig. 14.** A traditional RNN. $a^{<t>}$ denotes the activation function.

called an encoder and reads a variable-length sequence input. The other RNN is called a decoder and produces a sequence output. A context vector is utilized to preserve information from all hidden states of the encoder and align them with the current target output. Attention weights α are assigned to the input sequence to prioritize the set of positions where relevant information is present. The weighted sum of all hidden states of the encoder and their corresponding attention weights creates the context vector c. By doing so, the model can attend to a certain part of the source input and learn the complex relationship between the source and target better.

#### *5.2.4. GAn*

GANs are unsupervised learning algorithms that use a supervised cost function as part of their training process. GANs comprise generator and discriminator networks, as shown in Fig. 16 (Goodfellow et al., 2014). The generator network uses a noise vector as its input to generate an image, then this generated image is given to the discriminator network. The discriminator network tries to distinguish between the real and the generated image. In other words, the discriminator is a classifier that determines which image is real and which one is fake. In the training process, the generator aims to create realistic images that confuse the discriminator. However, it is important to keep these networks at the same level and make them improve together. In recent years, GANs have become popular in medical analyses and have been used for different tasks, such as data augmentation and image translation. One of the applications of GAN to image translation is CycleGAN (Zhu et al., 2017), which allows converting an image from one domain to another domain.

## 6. Different learning strategies

### 6.1 Transfer learning

The main idea of transfer learning is to use the knowledge gained from one problem and apply it to a different problem. These problems are usually related to each other. When a large dataset is available for the first problem, while there is not enough data available for the target problem, transfer learning can be helpful. Transfer learning includes two steps: pre-training and fine-tuning. A model is first trained on a large

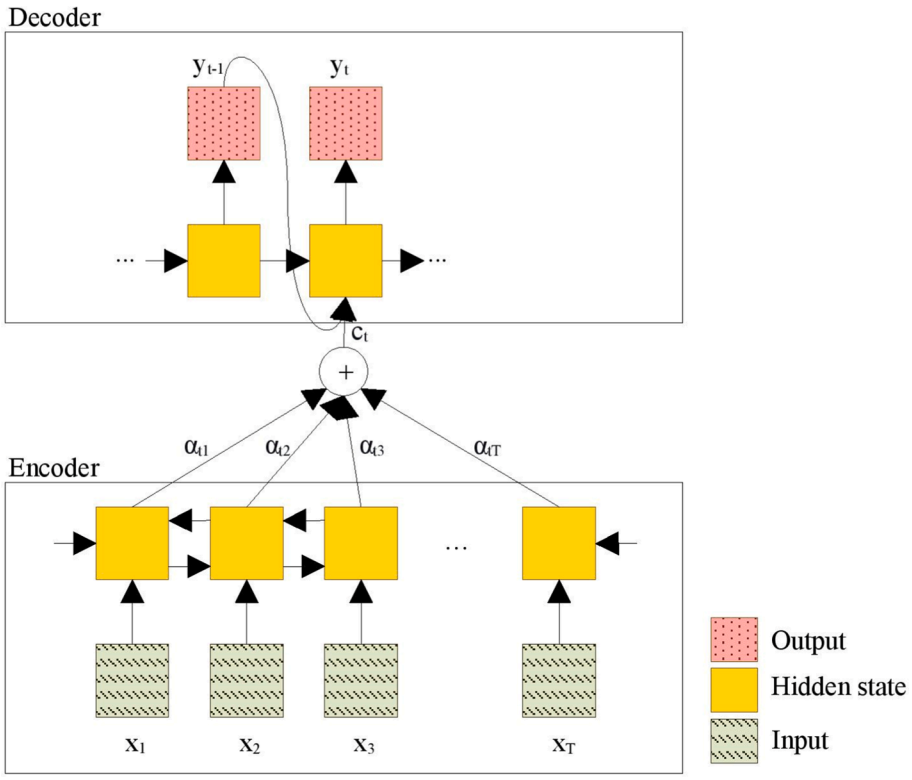


**Fig. 15.** The first attention neural network.

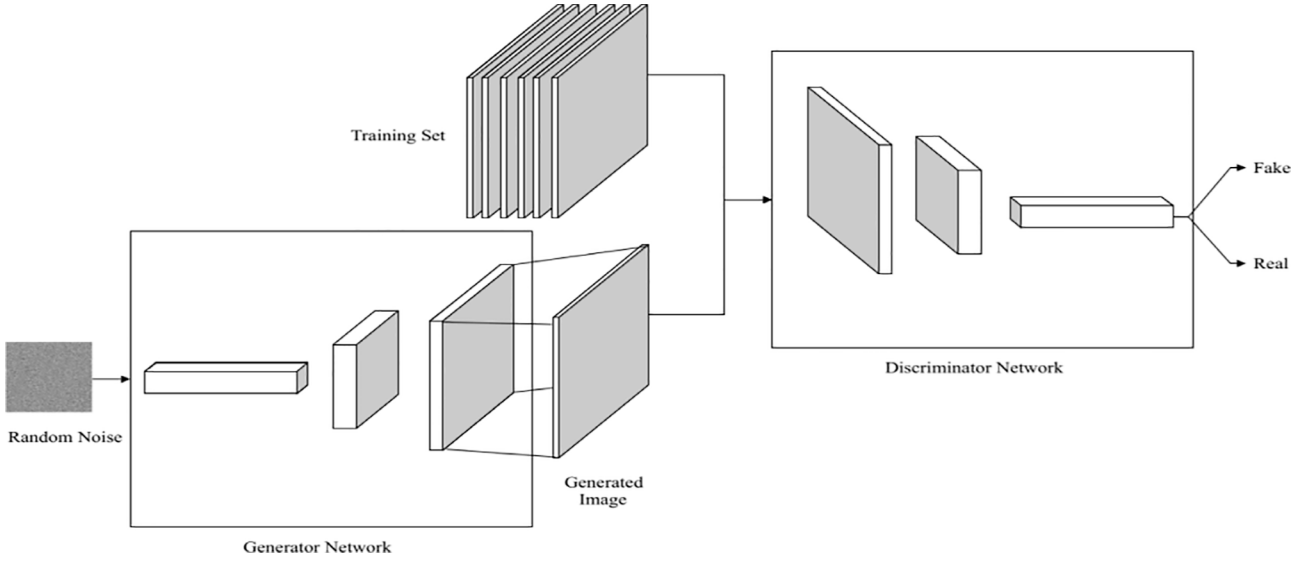


**Fig. 16.** GAN structure.

dataset to learn all network parameters. Next, this pre-trained network is fine-tuned on a new dataset. In fine-tuning, a few last layers of the network are randomly initialized, then the entire network is retrained using the new dataset. When the new dataset is very small, only the last layer should be randomly initialized to avoid the overfitting problem. Moreover, If the second task is not related to the first task, only initial layers of the pre-trained network, which capture generic features, should be used for the second task. The most widely used dataset for pre-training a network is the ImageNet dataset (Ng, 2018; Pratt et al., 1991).

### 6.2 Multitask learning

Unlike transfer learning, which is a sequential process, multitask learning aims to do multiple tasks simultaneously by a shared model. In this technique, a big enough neural network is needed to do several tasks with similar low-level features and an equal amount of data in the training set. (Caruana, 1997; Ng, 2018).

### 6.3 End-to-end learning

Some systems need multiple stages of processing. In end-to-end learning, a single neural network replaces all these stages. End-to-end learning simplifies a system and reduces the need to design components manually. However, this learning strategy has some downsides.

For example, this method needs a lot of data to work well and excludes potentially useful hand-designed components. On the other hand, if there is enough data to train a big neural network, end-to-end learning can find the most appropriate mapping function between inputs and outputs (Ng, 2018). Table 3 shows different learning strategies.

## 7. Applications of deep learning in multimodal medical data analysis

This section provides an overview of different deep learning applications in multimodal medical data analysis. Generally, deep learning techniques are classified into four major categories: unsupervised, semi-supervised, self-supervised, and supervised. Fig. 17 shows the hierarchy of papers considered in this article. "Alzheimer multimodal deep learning," "cancer multimodal deep learning," "multimodal deep learning in medical analysis," and "COVID-19 multimodal deep learning" are queries that we used to search and filter papers in Google scholar. The main focus of this survey is on COVID-19, cancer, and Alzheimer's.

### 7.1 Supervised methods

Supervised learning indicates learning methods that use data with human-annotated labels to train networks (Jing & Tian, 2020). The objective of a supervised learning model is to predict a correct label for new input data based on prior training data.

#### *7.1.1. Classification*

Many studies have used multimodal deep learning algorithms to improve classification performance in the medical field. Using multimodal data with deep learning helps achieve superior results, provided that the right combination of modalities is used. Some combinations of modalities are prevalent in a disease analysis, such as MRI and PET in Alzheimer's. Zhang et al. (2019) employed two independent CNNs to diagnose AD based on MRI and PET images. They used the Pearson correlation coefficient to judge the consistency of CNNs' predictions. Also, they proposed a formula to combine the neuroimaging diagnoses with clinical neuropsychological diagnoses (MMSE and CDR) based on the Pearson correlation coefficient. When the result of neuroimaging diagnoses is consistent, the algorithm only takes the results of the neuroimaging diagnoses. Otherwise, the algorithm only focuses on the diagnoses of clinical neuropsychology. Zhang and Shi (2020b) proposed a deep learning model based on the attention mechanism for AD diagnosis using MRI and PET images. In this model, the fusion ratio of each modality is assigned automatically according to its importance. The final output of their model determines whether the input sample has Alzheimer's or not.

Also, the combination of PET and CT images is prevalent for multimodal cancer analysis. For example, H. Shi et al. (2018) used CT, PET, and PET/CT images for lung tumor detection. They trained three CNNs separately on each modality and integrated the outputs of these CNNs to make the final decision. Similarly, Qin et al. (2020) used a CNN architecture to combine the fine-grained features from PET and CT images for lung cancer detection. After the outbreak of COVID-19, many researchers have used deep learning models to analyze this disease. Since chest CT and chest X-ray provide complementary information, using their combination is very popular among researchers. For instance, Zhang et al. (2021) proposed a deep convolutional attention network for COVID-19 diagnosis based on chest CT and chest X-ray. This network has two branches, one of each receives 3D CT images, and the other one receives 2D X-ray images. After five convolutional block attention modules (Woo et al., 2018) in each branch, the extracted deep CT features, and deep X-ray features are flattened. Next, the concatenation of these feature vectors is given to fully connected layers to diagnose COVID-19. This method achieved a high accuracy of $98.02 \pm 1.35\%$ on a private dataset collected from local hospitals.

In some diseases, modalities should be chosen based on the purpose of the research. For instance, Vasquez-Correa et al. (2018) used speech, handwriting, and gait signals to detect patients with Parkinson's disease. They trained three individual CNNs on each modality to create feature maps. Then, they averaged these feature maps across different tasks and transitions of a given subject. Finally, they concatenated embeddings from three bio-signals and fed them to a radial basis support vector machine (SVM) to detect Parkinson's disease patients. Vaghefi et al. (2020) investigated the role of combining different image modalities in diagnosing intermediate dry age-related macular degeneration. They trained a network based on Inception-ResNet-v2 using optical coherence tomography (OCT), OCT angiography (OCT-A), and color fundus photographs. Ali et al. (2020) used deep learning to diagnose heart disease from sensor data and electronic medical records. They combined extracted features from both data modalities and used the information gain technique for feature selection. With this technique, they decreased the computational burden and enhanced the system performance. They also employed the conditional probability approach to identify the significance of features. Next, they trained an ensemble deep learning model for heart disease prediction. Finally, they recommended ontology-based dietary plans and activities based on each patient's health condition. Khamparia et al. (2019) proposed a multimodal deep learning model for chronic kidney disease classification. Their model is constructed using stacked autoencoders with one softmax classifier. The model proposed by Shikalgar and Sonavane (2020) classifies AD using MRI images and EEG signals. The key objective of this method is to enhance the learning procedure in which the weight factor of the deep belief network (DBN) is incorporated with CNN for dealing with multimodal heterogeneous information. First, the median filter on MRI images and the Gaussian filter on EGG signals are applied to minimize noise. After extracting texture properties of images by gray level co-occurrence matrix (GLCM), features from both modalities are concatenated. Finally, a hybrid CNN-DBN model classifies final features.

On the other hand, some articles focus on side issues instead of

**Table 3**
Different learning strategies.

| Learning strategy | Articles |
|---|---|
| Transfer learning | El Asnaoui and Chawki (2020); Jiang et al. (2020); Kassani et al. (2020); Maghdid et al. (2020); Rehman et al.(2020); Saba et al. (2020); van Sonsbeek and Worring (2020); Vu et al. (2018); Wang et al.(2020); Yap et al. (2018) |
| End-to-End learning | Bai et al. (2020); Cao et al. (2020); Cheerla and Gevaert (2019); Chen et al. (2019b); Feng et al. (2019b); Ge et al. (2020); Guo, et al. (2019); Li, et al. (2019); Hervella et al. (2019, 2020); Hooshmand et al. (2020); Huang and Chung (2020); Hung et al. (2019); Isensee et al. (2019); Jiang et al. (2020); Khamparia et al. (2019); Kirienko et al. (2018); Lai et al. (2019); Lee, Nho, et al. (2019); Li et al. (2020); Liang et al.(2018); Lin et al. (2020); Liu et al.(2018); Ma et al. (2018); McKinley et al. (2019); Milecki et al.(2021); Mukherjee et al. (2020); Myronenko (2019); Peng et al. (2019); Qin et al. (2020); Shi et al. (2018); Shi et al. (2018); Shikalgar and Sonavane (2020); Suk et al. (2014); Sun et al. (2019);Taleb et al. (2021); Vaghefi et al. (2020); Wang et al. (2018); Wang et al. (2020); Xu et al. (2021); Zhang et al. (2020); Zhang et al. (2019); Zhang and Shi (2020b); Zhang et al. (2021); Zhao et al. (2018); Y. Zhao et al. (2020) |
| Multitask learning | El-Sappagh et al.(2020); Liu et al. (2019); McKinley et al. (2020); Tang et al. (2020);Zhao et al. (2020) |
| Hybrid[1] | Abrol et al. (2019); Ali et al. (2020); Bagheri et al. (2020); Hosseini et al. (2020); Kim and Lee (2018); Lassau et al. (2020); Lee, Kang, et al. (2019); Li et al. (2019); Liu and Hu (2019); Lu et al. (2018); Nie et al. (2019); Qiu et al. (2020); Rubinstein et al. (2019); Shukla and Marlin (2020); Soltaninejad et al.(2019); Vasquez-Correa et al. (2018); Xu et al. (2020); Yan et al. (2019); Zhou et al.(2018) |

[1] Articles that used techniques such as machine learning or image processing for some stages of processing.

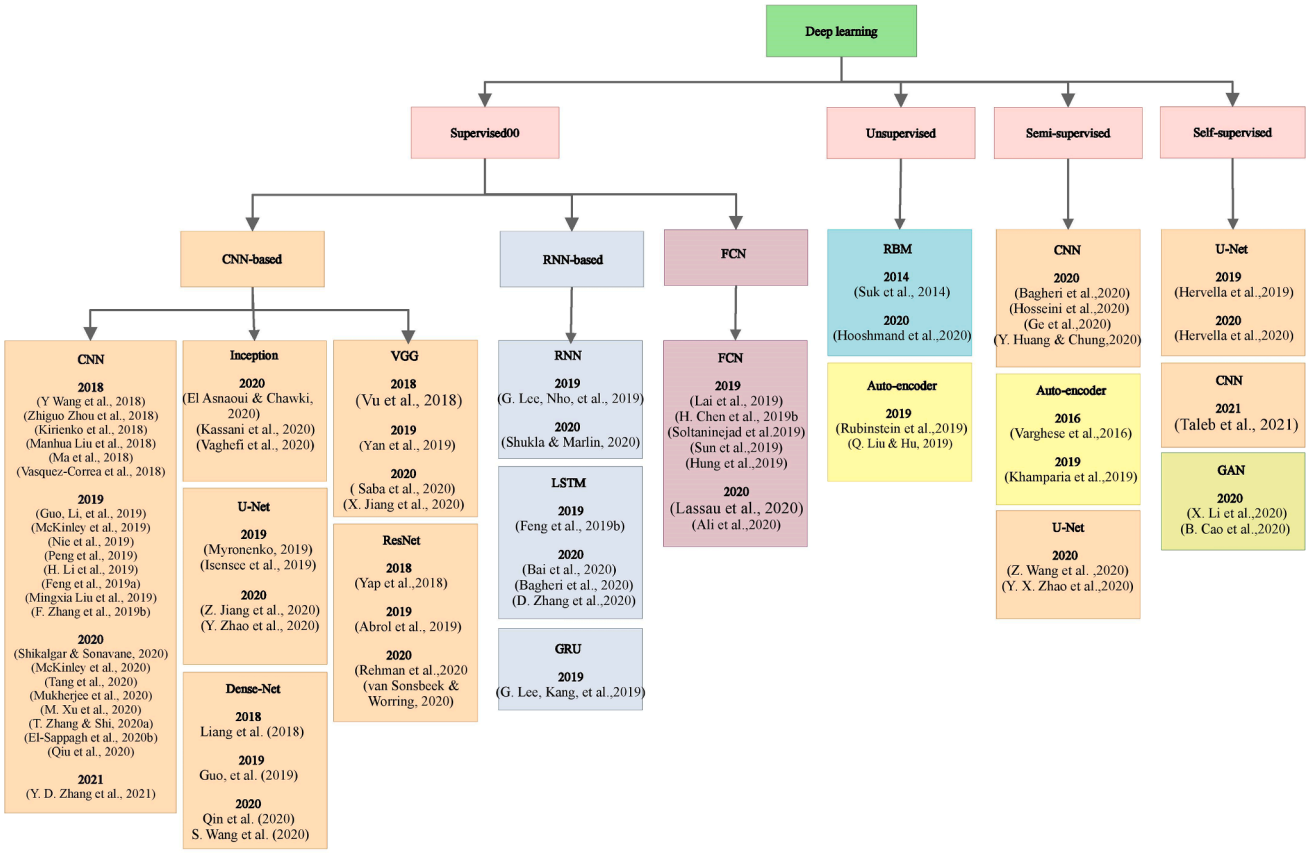


**Fig. 17.** The tree of papers related to deep learning applications in multimodal medical data analysis.

classification. For example, most existing methods for AD analysis need image preprocessing, such as registration and segmentation, but Manhua Liu et al. (2018) proposed a cascaded CNN, which requires no image segmentation and rigid registration in the preprocessing of brain images. This model learns multimodal features from MRI and PET images for AD classification, saves computation costs, and achieves more robustness to some variations of translation and rotation in images. Many CNN-based methods use a flattening layer after convolutional layers because a fully connected layer only processes 1D information. The feature maps of 3D-CNN are always in 3D, so using a flattening layer leads to the loss of 3D spatial information in feature maps. To solve this problem, Feng et al. (2019a) used a fully stacked bidirectional LSTM to get rich spatial and semantic information from feature maps. They employed their method for the diagnosis of AD using MRI and PET.

Another issue on which many researchers are working is the data scarcity problem, especially in COVID-19 analysis. Since large COVID-19 datasets have not been publicly available yet, data augmentation techniques and transfer learning have been used in many articles to prevent the overfitting problem. For example, Kassani et al. (2020) compared different pre-trained CNN models for COVID-19 detection based on MobileNet, DenseNet, Xception, ResNet, InceptionV3, Inception-ResNet-v2, VGGNet, and NASNet. They used pre-trained CNNs for feature extraction from X-ray and CT scans and fed the extracted features into several machine learning classifiers to identify whether a subject has COVID-19 or not. The DenseNet121 feature extractor with bagging tree classifier achieved the best performance with 99% classification accuracy. El Asnaoui and Chawki (2020) also employed different pre-trained models (VGG16, VGG19, DenseNet201, Inception-ResNet-v2, Inception_V3, Resnet50, and MobileNet_V2) for feature extraction from X-ray and CT scans. Then, they flattened these features and passed them to a multilayer perceptron to classify each image. Their results concluded that Inception Resnet V2 is the best feature extractor, which achieved 92.18% accuracy. Similarly, Maghdid et al. (2020) and Rehman et al. (2020) used transfer learning along with the combination of X-ray and CT scans.

Some articles address several problems simultaneously by using a multitask model. For instance, El-Sappagh et al. (2020) proposed a multitask multimodal deep learning model for AD classification and four critical cognitive scores (ADAS, MMSE, FAQ, and CDRSB) regression. They extracted a set of static baseline features and temporal features from five heterogeneous data sources, including MRI, PET, cognitive scores, neuropsychological data, and assessment data. In this method, each time-series source is separately learned using a pipeline of stacked CNN-Bidirectional LSTM blocks. The CNN automatically extracts local features from each time series, and LSTM extracts temporal features and temporal relationships among them. Then, a decision fusion by a dense layer is used to get more abstract deep features from time-series sources and baseline data. The final step is the task of specific learning to produce final results. Mingxia Liu et al. (2019) proposed a multitask deep learning model for simultaneous AD classification and clinical score regression using MRI and demographic information (age, education, and gender). After extracting multiple image patches from MRI images, they fed them into a CNN. Next, they concatenated demographic information and the feature vector of each MRI patch and used these feature vectors for AD classification and clinical score regression.

### 7.1.2. Prediction

This section explains some articles that used deep learning algorithms for predicting a medical event in the future. Various predictive tasks can be performed based on disease type, such as metastasis prediction in cancer. Peng et al. (2019) used PET and CT images to predict distant metastases of soft-tissue sarcoma. The CNN proposed in this study has two branches for processing PET and CT images

simultaneously. After feature extraction in each branch, the extracted features are concatenated. After another convolutional layer, these features are concatenated with texture features, which are obtained by applying grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM), grey-level size zone matrix (GLSZM), and neighborhood grey-tone difference matrix (NGTDM) on PET images and feature selection. Finally, the new feature vector is given to fully connected layers to predict whether the input PET-CT image will develop metastasis. Zhou et al. (2018) proposed a hybrid model that predicts lymph node metastasis in head and neck cancer. This method includes two models, one of which extracts intensity, texture, and geometric features from PET and CT images. Then, these features are fed into an SVM to predict the three classes of lymph nodes, including normal, suspicious, and involved. The second model uses 3D-CNN to predict lymph node classes based on CT and PET images. Finally, the outputs of these two models are fused through the evidential reasoning (ER) approach.

Other common predictive tasks in cancer are prognosis and survival prediction. In a study by Li et al. (2019), PET and CT images are fed into a CNN model to predict survival risk in patients with rectal cancer. The CNN learns imaging features by optimizing the partial likelihood of a proportional hazards model. Several studies have combined genomic data with other modalities to achieve high performance. For instance, Chen et al. (2019a) proposed an attention-based neural network that fuses patients' gene expression and clinical data for the breast cancer prognosis. They focused on the efficient fusion of multiple feature extraction algorithms. In this model, five non-negative matrix factorization (NMF) algorithms extract features from gene expression and obtain five feature matrices. Next, the attention mechanism calculates the weight of each NMF algorithm according to the clinical data of each patient. The weighted sum of five eigenvectors of feature matrices is then concatenated with clinical data to generate the final representation, which is fed into a deep neural network for prediction. Lai et al. (2019) combined gene expression and clinical data to predict the survival status of patients within five years. First, they removed patients with incomplete clinical data. Then, they identified eight novel survival-related genes based on seven previously well-known NSCLC biomarkers. The combined 15 biomarkers and clinical data are fed into an FCN with two separate branches for gene markers and clinical data. After four hidden layers in each branch, these branches are stacked for the final prediction. Genomic data is beneficial for OS prediction, but tumor genotype is not available pre-operatively. Tang et al. (2020) proposed a multitask CNN to simultaneously predict the genotype and OS of glioblastoma patients from pre-operative multiparametric MRI. After two convolutional blocks, their model is split into five branches, four branches for genotype prediction (MGMT, IDH, 1p/19q, TERT), and one for OS prediction. High-level features learned from four genotype prediction branches are fed into the fully connected layer of the OS prediction branch to provide it with tumor genomic features. Furthermore, patients' age, gender, tumor size, and tumor location are fed to the fully connected layers of all prediction tasks. This study achieves a remarkable result for OS prediction. Another study that focused on predicting IDH genotype from multiparametric MRI is Liang et al. (2018) which developed a multimodal 3D deep learning model based on the 3D DenseNet framework. This model takes four MRI sequences (T1, T2, T1Gd, FLAIR) as input and gives IDH genotype as output.

A popular predictive task in AD is to predict conversion from mild cognitive impairment (MCI) to AD. For example, Lin et al. (2020) recommended an extreme learning machine (ELM) based (Huang et al., 2012) grading method to fuse multimodal data and predict MCI-to-AD conversion efficiently. All modalities, including MRI, PET, CSF, and gene data, are first individually graded using the ELM method. Then, these grading scores calculated from different modalities are fed into an ELM classifier to discriminate subjects with progressive MCI from those with stable MCI. One of the major problems while dealing with longitudinal data, such as longitudinal CSF and longitudinal cognitive performance, in Alzheimer's analysis is handling variable-length of sequential data and missing values. Usually, a feature extraction phase is required to produce a fixed-size feature representation. However, Lee, Nho, et al. (2019) proposed an RNN that uses any irregular length of data as input without preprocessing to predict conversion from MCI to AD using demographic information (age, sex, years of education, and APOE $\varepsilon 4$ status), neuroimaging phenotypes measured by MRI, cognitive performance, and CSF measurements. For extracting features, a single GRU is trained separately on each modality. These GRU components make an encoding process in which longitudinal data are transformed into a vector. Next, four extracted feature vectors from each modality are concatenated and given to an l1-regularized logistic regression model for the final prediction.

After the outbreak of COVID-19, another predictive task has attracted many researchers. In this task, researchers find patients who easily deteriorate into critical cases because these patients have a higher priority to receive medical treatment and special care. As there are not enough medical resources in epidemic areas, this task becomes more important. Some articles have addressed this problem by predicting disease severity. Lassau et al. (2020) used clinical characteristics, lab tests, and CT images to predict the severity of hospitalized COVID-19 patients. They first trained a deep learning model to extract features from CT images. Then, they concatenated these features with lab tests and clinical characteristics and gave them to a logistic regression model for predicting the severity score of patients. Similarly, Wang et al. (2020) proposed a method to identify the potential high-risk COVID-19 patients who are more likely to become severe. In this method, DenseNet121-FPN first finds lung masks in CT images. Some non-lung tissues such as the spine and heart inside the lung mask may still exist, so a non-lung area suppression operation suppresses the intensities of non-lung areas inside the lung mask. Then, the standardized lung mask is sent to another DenseNet-based model, which is pre-trained using CT images and gene information. This model generates deep features and predicts the probability that a patient has COVID-19. After feature selection from the combination of deep features and clinical features (age, sex, and comorbidity), a multivariate Cox proportional hazard model is used to identify high-risk COVID-19 patients. In a similar study, Fang et al. (2021) used a 3D ResNet to extract features from CT images and used a multilayer perceptron to obtain features from clinical laboratory data and personal information. Then, they concatenated these feature vectors and gave them to an LSTM to determine which patients deteriorate into severe cases.

#### 7.1.3. Segmentation

Image segmentation has greatly benefited from recent developments in multimodal deep learning. In image segmentation, we determine the outline of an organ or anatomical structure as accurately as possible (Maier et al., 2019) Due to the variable size, shape, and location of the target tissue, medical image segmentation is one of the most challenging tasks in medical image analysis (Zhou et al., 2019) U-Net and V-Net are the most successful architectures for medical image segmentation, so many researchers have employed them. Zhao et al. (2020) developed a model to automatically detect local prostate tumors, bone lesions, and lymph node metastasis based on PET/CT images. Their model consists of three 2.5D U-Nets, any of which is separately applied to one of the different planes (axial, coronal, and sagittal) to make predictions about each voxel. Then, the majority voting strategy combines the outputs of U-Nets to segment lesions. Zhao et al. (2018) developed two individual V-Nets for extracting features from CT and PET images. Then, they fused extracted features and gave them to a softmax layer to find the tumor mask in lung cancer. Other architectures can also segment tissues and achieve good results. For instance, Guo, et al. (2019) proposed a 3D DenseNet model for gross tumor volume segmentation in head and neck cancer. In this study, after standardization and normalization, PET and CT images are fed into the model to adopt the tumor volume contour.

Since 2012 Brain Tumor Image Segmentation Benchmark (BraTS)

challenge has been organized in conjunction with the international conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) to assess and compare state-of-the-art methods in automated brain tumor segmentation. This challenge has encouraged many researchers to focus on brain tumor segmentation because it provides a multiparametric dataset of MRI, which contains T1, T1c, T2, and Flair images (see Section 8). Before this challenge, virtually all studies were validated on relatively small private datasets with varying metrics for performance quantification, making objective comparisons between methods highly challenging (Menze et al., 2015) Brain tumor segmentation is the main task in the BraTS challenge, but other tasks have also been added over the years. For example, the OS prediction task was added to this challenge in 2017. For OS prediction based on multiparametric MRI in the BraTS dataset, machine learning techniques with hand-crafted features are more popular than deep learning techniques. As this topic is beyond the scope of this paper, we do not cover it. For further details about datasets provided by this challenge, see Table 1 in Ghaffari et al. (2020).

An example of studies on the BraTS dataset is Myronenko (2019) In this study, after normalization and data augmentation, input images are given to an encoder-decoder model in which the encoder extracts features from inputs, and the decoder predicts segmentation masks. Furthermore, a variational auto-encoder (VAE) branch is added to the encoder endpoint to reconstruct the original image. Also, the VAE loss is considered in the overall loss function to regularize the shared encoder. This study won first place in the segmentation task of the BraTS 2018 challenge. Jiang et al. (2020) the winner of the segmentation task in the BraTS 2019 challenge, took a variant of (Myronenko, 2019) as the basic segmentation architecture and proposed a two-stage cascaded U-Net. In the first stage, U-Net predicts coarse segmentation maps, which are used to calculate the first loss. Then, these maps and raw images are fed into the second stage U-Net, to provide more accurate segmentation maps. The second stage U-net has two decoders with the same structure, except that one decoder uses deconvolution and the other uses trilinear interpolation. The outputs of these decoders are used to calculate the second and third losses, which are added to the first loss to create the final loss. The interpolation decoder is only used during training for regularizing the shared encoder.

### 7.2 Unsupervised methods

Unsupervised learning refers to learning methods without using human-annotated labels (Jing & Tian, 2020) In unsupervised learning, we group an unlabeled dataset based on underlying hidden features. Auto-encoders, GANs, deep Boltzmann machines (DBMs), RBMs, and DBNs are popular deep learning models in unsupervised tasks (Raza & Singh, 2021) Labeling data in the medical field is difficult and time-consuming; also, it needs a lot of knowledge and experience. As a result, unsupervised, self-supervised, and semi-supervised methods have recently received considerable attention in medical analyses due to their potential for reducing the effort to label data. However, a few studies have used these techniques in multimodal medical data analysis. For instance, Hooshmand et al. (2020) addressed the problem of drug repurposing in COVID-19 in an unsupervised manner. Drug repurposing is a way to discover new applications of existing drugs for treating other diseases (Pushpakom et al., 2018). They employed RBM to categorize two types of drug data, including differentially expressed genes and chemical structures, into 12 clusters. Then, they chose clusters that consisted of drugs used for curing COVID-19 to discover medications that may be useful in treating this disease. Xu et al. (2021) proposed an unsupervised deep learning method for multimodal image registration. This framework has two branches, both of which use encoder-decoder architecture. The image registration branch estimates the primary deformation field for a moving CT and a fixed MRI, while the gradient map registration branch takes the corresponding gradient maps of CT and MRI as inputs to produce another deformation field. They also designed a gated dual-branch fusion mechanism to adaptively fuse the estimated deformation fields. With the help of auxiliary gradient-space guidance, their network concentrates more on the spatial relationship of the organ boundary.

In another study, Rubinstein et al. (2019) extracted three feature classes, including statistical, kinetic biological, and deep features, from PET/CT images of patients with prostate tumors. They trained a stacked convolutional auto-encoder and considered its reconstruction errors in different training epochs as deep features. Then, they used all features to compute an anomaly score for each voxel. Finally, they used density estimation to detect anomalies, which are classified as tumors, in the feature space. Milecki et al. (2021) employed an unsupervised deep learning method to segment kidney grafts in T2 and Dynamic Contrast-Enhanced (DCE) MRI. They applied thresholding techniques and morphological operators to detect the area of interest. Then, an unsupervised CNN model, based on differentiable feature clustering, was used for the pixel-wise segmentation of the kidney graft. Suk et al. (2014) devised an unsupervised deep learning method for learning high-level latent and shared features from MRI and PET images. In this method, paired patches of MRI and PET are given to a Gaussian RBM, which is used as a preprocessor to transform the real-valued observations into binary vectors. Then, these binary vectors are given to a DBM, which finds a shared feature representation from the paired patches. As DBM is an undirected graphical model, bidirectional information flows from one modality to another one and vice versa. Therefore, feature representations are distributed over different layers in the path between modalities, and thus a shared representation is discovered. To validate their model, authors trained multiple SVM classifiers with this shared representation for AD/MCI diagnosis.

### 7.3 Semi-supervised methods

Semi-supervised learning is a branch of machine learning that combines supervised and unsupervised learning (Chapelle et al., 2006). Usually, in semi-supervised learning, a small amount of labeled data is used in conjunction with a large amount of unlabeled data (Jing & Tian, 2020). In recent years, semi-supervised learning has attracted many researchers, especially in the medical field, where small annotated datasets are often available.

Some studies use a small labeled dataset to estimate labels for unlabeled data. For instance, Ge et al. (2020) trained a multi-stream 2D CNN using only labeled data in the training dataset. They also devised a graph-based semi-supervised method to estimate labels for unlabeled data. Then, they trained a GAN using training data from labeled and unlabeled sets to generate synthetic MRIs for data augmentation. Finally, they passed the labeled training dataset, the unlabeled training dataset with estimated labels, and augmented data to the pretrained multi-stream CNN to classify glioma. In the study by Huang and Chung (2020), a graph convolutional neural network is designed to determine whether a subject is healthy or diseased. Their model accepts subjects' imaging and non-imaging data and represents them as a population graph (partially labeled). Features extracted from imaging data of subjects are considered nodes in the population graph. A trainable module, called the edge adapter, encodes the non-imaging data, such as phenotypic information (age, gender, and site), into the population connectivity. Then, the population graph is given to a graph convolutional neural network, which allows the edge adapter to learn the pairwise associations between subjects during the training of the network. While their goal is to predict the disease state of unlabeled subjects under the supervision of the labeled ones in the population graph, the learned graph can be easily applied to clustering analyses by thresholding.

In another study, Zhao et al. (2020) combined different tricks to achieve better accuracy for 3D brain tumor segmentation. To cope with the problem of data imbalance in segmentation, they employed heuristic sampling and hard sample mining. They also constructed a training batch pool with batches of different patch sizes. For each iteration

during training, they randomly selected a batch from the pool to update the model, so they took advantage of both large patches and the large batch size. Moreover, they used a multi-space semi-supervised method to tackle the lack of annotated data. In this method, they trained different models on the training set under different conditions, such as different subsets of the training set or different subspaces of features, at each iteration. Then, they combined all of these models and used them to label the unlabeled dataset. After each iteration, they merged the manually labeled dataset and model labeled dataset as the new training set. Furthermore, they developed a self-ensemble U-Net, which makes predictions at different scales and joins them to obtain segmentation masks. They also applied various optimization techniques, including gradual warming up learning rate (Goyal et al., 2018) and multitask learning.

On the other hand, some researchers use both labeled and unlabeled data to improve the performance of their model without labeling unlabeled data. For example, Wang et al. (2020) proposed a semi-supervised method to synthesize high-quality pairs of Apparent Diffusion Coefficient (ADC) and T2 images containing clinically significant (CS) prostate cancer. Their model comprises an encoder to obtain latent vectors from real ADC maps, a decoder to derive low-dimensional ADC maps from latent vectors, a StitchLayer to convert low-dimensional ADC maps to full-size ADC images, and a U-Net to convert the full-size ADC images to T2 images. By training the synthesizer in a supervised manner, the model enforces the correct paired relationship between synthesized ADC and T2 images of a pair. To increase the diversity of the generated data and avoid overfitting, they also trained the synthesizer in an unsupervised manner by providing various random latent vectors. Furthermore, they minimized the Wasserstein distances between the marginal distributions of synthesized and real images of two modalities to ensure high visual similarity between real ADC/T2 images and fake ones. Finally, to enforce the synthetic images to contain distinguishable CS prostate cancer lesions, they maximized the distance of Jensen-Shannon divergence between CS and non-CS images. In the study by Hosseini et al. (2020), an unsupervised CNN extracts high-level features and wavelet & spatial group ICA extract time & frequency features from preprocessed EEG. Then, these features are given to a nonlinear SVM to classify interictal epileptiform discharge (IED) and non-IED time intervals. Based on the SVM outputs, a differential connectivity graph is built. Also, a reclustering method is proposed to identify brain networks from preprocessed rs-fMRI data. Finally, for seizure focus localization in epilepsy, an LSTM is used to merge the estimation of brain network and connectivity found by rs-fMRI analysis and the differential connectivity graph found by EEG analysis. The authors claimed their method achieves better results than other methods for IED detection and localization of epileptogenicity.

In another study, Varghese et al. (2016) designed a new post-processing technique to improve glioma segmentation using multi-parametric MRI. In this study, two stacked denoising auto-encoders (SDAE) were pretrained using many unlabeled High-Grade Glioma (HGG) patches. One of the SDAEs is fine-tuned using labeled HGG patches to segment HGG images, and another one is fine-tuned using labeled Low-Grade Glioma (LGG) patches to segment LGG images. Furthermore, a one-layer denoising auto-encoder (DAE), called Novelty detector (ND), is trained to create reconstruction error maps by assigning every voxel the mean reconstruction error of the patch centered at that voxel. This leads to a heat map-like image with large error regions corresponding to the location of the Glioma. Error maps are then binarized using Otsu's thresholding. After applying connected component analysis on images predicted by two SDAEs, connected components that have an empty intersection with their corresponding binary error mask are discarded. Their results demonstrate that the novelty detector causes a reduction in false-positive voxels and improves glioma segmentation. Cheerla and Gevaert (2019) developed an unsupervised encoder to compress clinical data, gene expression, miRNA, and histopathology whole slide images (WSIs) into a single feature vector for each patient. Then, they used this vector for predicting pan-cancer OS. They tailored encoding methods to each data type using deep highway networks (Srivastava et al., 2015) to extract features from clinical and genomic data, and SqueezeNet (Iandola et al., 2016) to extract features from WSIs.

### 7.4. Self-supervised methods

Self-supervised learning algorithms solve a series of handcrafted auxiliary tasks (so-called pretext tasks) in which supervision signals are acquired from the data itself, without the need for manual annotation (Liu et al., 2021). Pretext tasks result in a model or representation that can be used to solve the original modeling problem (Kolesnikov et al., 2019). With the help of well-designed pretext tasks, self-supervised learning enables models to learn more informative representations from unlabeled data to achieve better performance, generalization, and robustness on various downstream tasks (Liu et al., 2021). Tasks such as image inpainting, colorizing grayscale images, jigsaw puzzles, and super-resolution have proven effective for learning good representations (Jaiswal et al., 2020). For example, Taleb et al. (2021) introduced a self-supervised jigsaw puzzle-solving task for learning semantic representations that facilitate downstream tasks in the multimodal medical imaging context. In this model, all modalities are cut into puzzle pieces or patches and are shuffled randomly according to a specific permutation. These shuffled image pieces are then assembled and create a set of patches called P. In other words, P is a restored image in which each element is drawn from a different modality. Then, a neural network processes each element independently to produce a single output feature vector for each element in P. The matrix created by the concatenation of feature vectors is then passed to the Sinkhorn operator to obtain the soft permutation matrix. This soft permutation matrix is applied to the scrambled input P to reconstruct it. The network aims to minimize the mean squared error between the sorted ground-truth P and the reconstructed version of the scrambled input P. In this way, the network learns different tissue structures across given modalities. After the training process, the network parameters can be used in downstream tasks by fine-tuning on target domains.

However, some researchers devise new pretext tasks to learn valuable representations. For instance, Li et al. (2020) proposed a representation learning method that exploits FA and color fundus images for retinal disease diagnosis. First, they reconstructed FA images from corresponding color fundus images using a CycleGAN to learn the mapping function between these modalities. Each patient data consists of a color fundus image, a transformed fundus image obtained from a random data augmentation technique, and the corresponding synthesized FA. Next, they randomly sampled n patients and fed the batch of data into the ResNet18 to get high-level representations. Then, they classified high-level representations into n classes, where each class represents a patient, and used the classification error for optimizing the model. Finally, they trained a KNN classifier using high-level representations to solve the downstream task, which was retinal disease diagnosis. Hervella et al. (2020) proposed a self-supervised method for the optic disc and cup segmentation using unlabeled pairs of retinography and FA images. In the first phase, a U-Net reconstructs FA images from their corresponding retinographies. This multimodal reconstruction is a self-supervised task that aims at learning domain-specific patterns. This trained network is then fine-tuned using the annotated data for the downstream task (i.e., the optic disc and cup segmentation). This technique causes a significant improvement in the segmentation performance. When multimodal data is used, some imaging modalities may be unavailable due to clinical and practical restrictions. To impute missing data with adequate clinical accuracy, Cao et al. (2020) designed a self-supervised collaborative learning framework to synthesize a missing modality using other available imaging modalities. In the training phase, all modalities are available for each sample. In this phase, the encoder of a translation network encodes the input images from different sources into a common latent feature space. The latent features are then concatenated and given

to the translation network's decoder to construct the target image. The self-representation network is an auto-encoder that attempts to model the distribution of target images by reconstructing them. Once well-trained, feature maps extracted from the decoder of the self-representation network are used to guide the optimization of the translation network's decoder. The pseudo images generated by translation and self-representation networks are utilized for training a discriminator network with ground-truth images. The discriminator tries to classify whether each patch in the input image is real or fake. In the testing phase, self-representation and discriminator networks are removed, and only the translation network is used to translate images from multiple source domains to the target domain.

In another form of self-supervised learning, pseudo labels are generated for an unlabeled dataset according to the structure or characteristics of the data itself. Then, these pseudo labels are used to train a model in a supervised manner (Yuan et al., 2021). For instance, Hervella et al. (2019) proposed a self-supervised strategy for retinal vessel segmentation. They used multiscale laplacian operation, an edge detection filter in image processing, to obtain vessel maps for each angiography. They aligned these generated maps with corresponding retinographies and used them as pseudo labels. Then, they used these pseudo labels to train a U-Net with standard pixel-wise metrics. They chose these modalities because, unlike retinography, the vasculature is already highlighted in angiography. As a result, angiography provides complementary information that allows the network to segment retinal vessels in retinography without manually annotated labels.

## 8. Data sources

In Table 4, some of the most well-known multimodal datasets are introduced. Some of them only focus on a specific disease such as ANDI (Alzheimer), BCDR (breast cancer), DDSM (breast cancer), and BraTS (brain tumor), while others such as TCIA and Grand Challenge provide data for different diseases.

## 9. Common problems

This section describes common problems of applying deep learning algorithms to medical analyses.

### 9.1 Lack of data

Labeling medical data is time-consuming and needs a lot of knowledge and expertise. Researchers have made a tremendous effort to create several medical datasets; however, most of them are limited in size. Small medical datasets usually cause the overfitting problem because training a neural network needs a lot of data. When data is insufficient, the network only memorizes samples in the training set. Consequently, it performs very well on the training set but not on the test set.

There are different strategies to overcome this problem. One of them is decreasing the network complexity by regularizing the network or controlling the number of layers and parameters. Another common technique is increasing the training set size by data augmentation techniques. Traditional data augmentation methods include mirroring, random crop, rotation, shearing, and local wrapping. These techniques improve the results slightly, but they cannot gain much additional information. A new sophisticated data augmentation technique is to synthesize high-quality examples of the data using generative models, such as GAN (Frid Adar et al., 2018). For example, Frid Adar et al. (2018) trained a GAN for generating synthetic liver lesions. They used these synthetic images to increase the size of the training set and fed this larger training set into a CNN for liver lesion classification. Their results show that using synthetic images generated by GAN improves the accuracy of their model. Waheed et al. (2020) used a GAN-based network to

**Table 4**
Different multimodal medical datasets.

| Dataset | Description | Data types | Website |
|---|---|---|---|
| ADNI | The Alzheimer's disease dataset designed for the early detection and tracking of AD | MRI, PET images, genetics, cognitive tests, CSF, and blood biomarkers | http://adni.loni.usc.edu/ |
| BCDR | A digital repository for breast cancer analysis | Mammography, related ultrasound images, and clinical reports | https://bcdr.eu/ |
| OASIS | Normal Aging and Alzheimer's disease dataset | Multiparametric MRI (T1, T2, FLAIR, ASL, SWI, DTI) and PET images | https://www.oasis-brains.org/ |
| TCIA | An archive which includes imaging data for different cancer types.<br>Also, some COVID-19 datasets have been added to this archive recently. | Based on the disease type, different image modalities are available in this archive, such as MRI, CT, and digital histopathology.<br>Supporting data, including patient outcomes, treatment details, genomics, and image analyses, are also provided. | https://www.cancerimagingarchive.net/ |
| IDA | An archive consisted of medical images for autism, brain mapping, brain aging, and Parkinson's progression | Based on the project, different image modalities, including MRI, fMRI, DTI, PET, and SPECT, are available. | https://ida.loni.usc.edu/services/Menu/IdaData.jsp?project= |
| DDSM | A digital database for mammographic image analysis in breast cancer | Mammography and patient information such as age, ACR breast density rating, subtlety rating for abnormalities, and ACR keyword description of abnormalities | http://www.eng.usf.edu/cvprg/Mammography/Database.html |
| Grand Challenge | A repository for different challenges in medical image analysis.<br>For each challenge, a dataset is provided by organizers. | Based on the challenge, different image modalities are available. | https://grand-challenge.org/ |
| BraTS | A challenge with three tasks: brain tumor segmentation, survival time prediction, and the evaluation of the uncertainty in tumor segmentation | Multiparametric MRI, including T1, T1c, T2, and Flair, and age of patients | https://www.med.upenn.edu/cbica/brats2020/ |
| MIMIC | A dataset comprising de-identified health-related data from ~ 40,000 critical care patients for computational physiology | Patients' demographics, lab tests, and textual patient notes | https://mimic.physionet.org/about/mimic/ |
| EMBL-EBI's COVID-19 Data Portal | COVID-19 datasets provided by The European Bioinformatics Institute | Biological images, genomics, protein expression data, and chemical structure data | https://www.covid19dataportal.org/about |
| ISLES | A dataset for the segmentation of stroke lesions in brain images | Multiparametric MRI including T1, T1c, Flair, and DWI sequences | http://www.isles-challenge.org/ |
| ISEG 2017 | A dataset for the segmentation of 6-month infant brain tissues | Multiparametric MRI including T1and T2 | http://iseg2017.web.unc.edu/ |
| MRBrainS | A dataset for the segmentation of grey matter, white matter, and cerebrospinal fluid of the brain | Multiparametric MRI including T1, T1-weighted inversion recovery, and Flair | https://mrbrains13.isi.uu.nl/ |

generate synthetic chest X-ray images. They added these generated images to the training set and gave this new training set to a CNN model for COVID-19 detection. With this technique, the accuracy of their model increased from 85% to 95%. Another prevalent technique to overcome the overfitting problem is transfer learning, which is explained in Section 6. Moreover, unsupervised, semi-supervised, and self-supervised techniques are beneficial when the training set is small.

### 9.2 Data preprocessing

Different modalities provide different aspects of the same problem; however, they have been acquired differently. Therefore, standardization and normalization techniques are required to make them comparable (Antonelli et al., 2019). Furthermore, some modalities are high-dimensional, sparse, irregular, biased, or multi-scale, so it is critical to preprocess them before an analysis (Kwak & Hui, 2019). For example, genomic data are high-dimensional, so it is difficult to match a large amount of data from whole-genome sequencing with other kinds of data. As a result, the dimensionality of genomic data should be reduced to match with other modalities (Xu, 2019).

### 9.3 Class imbalance

A significant difference between the number of negative and positive samples in the training set leads to the class imbalance problem. This problem causes models to over-classify the majority group due to the higher prior probability of this group. As a result, the instances belonging to the minority group are misclassified more often than those belonging to the majority group (Johnson & Khoshgoftaar, 2019). There are three main approaches to deal with this problem: data-level methods, algorithm-level methods, and hybrid methods. Data-level methods modify the training set to make it suitable for a standard learning algorithm. These methods include generating new objects for the minority group (oversampling), removing examples from the majority group (undersampling), and random selection of target samples for preprocessing. Algorithm-level methods modify the model to alleviate their bias towards the majority group. The most popular technique is the cost-sensitive approach, which assigns a higher penalty to the minority class and a lower penalty to the majority class. Hybrid methods combine the advantages of previous groups (Krawczyk, 2016).

### 9.4 Model interpretability and reliability

Accuracy is a critical factor for convincing users to use methods proposed in medical studies because these studies are related to human lives (Kwak & Hui, 2019). Using multimodal data improves models' accuracy, which leads more people to trust models. Another important factor that persuades the medical community to use deep learning models is interpretability. Many studies have used deep learning in the medical field, and some of them have achieved perfect results. However, the medical community has not accepted these methods yet. The main reason is that they cannot trust the predictions of deep learning models, which are considered uninterpretable and black-box models. In recent years, several studies have focused on improving the interpretability of deep neural networks. Different visualization techniques are also developed to understand deep learning models. These methods include weight histograms, saliency maps (Simonyan et al., 2014), occlusion maps, class maximization, and activation maximization.

### 9.5 Missing data

Handling missing data is challenging in medical data analyses, especially when multimodal data are used. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Sterne et al., 2009). MCAR means the probability of missing is the same for all cases, and missing data are unrelated to any observed or unobserved elements. MAR means missing data are related to some observed values; therefore, they can be predicted by other features in the dataset. MNAR means missing data are related to both observed and unobserved values.

To handle missing data, we should determine their type first. With MCAR, any missing data can be dropped without biasing models. This method is not prevalent as medical datasets are usually small. In MAR and MNAR, dropping data may bias models. To handle MNAR, we should find the causes of missingness. Also, we should perform what-if analyses to see how sensitive the results are under various scenarios (van Buuren, 2012). Finally, data imputation techniques are practical to address MAR. In data imputation, we replace missing data with some estimated ones. Regression imputation, k-nearest neighbor (KNN) imputation (Cover & Hart, 1967), and multiple imputation (Rubin, 1987) are well-known data imputation methods.

### 9.6 Privacy concerns about data

Deep learning algorithms need large centralized training datasets. Gathering data from different repositories is a possible strategy to increase the size of training sets and gain a better understanding of data. However, institutions cannot legally share their medical data with other institutions because medical data are the most sensitive information associated with an individual. As data protection is crucial in the medical field, different strategies, such as federated learning (Konečný et al., 2015), differential privacy (Dwork, 2006), and multiparty computation (Lindell & Pinkas, 2009), have been suggested to deal with data privacy problems.

### 9.7 Generalization

Deep learning models should generalize well to unseen data. Most existing deep learning models perform well until the test set distribution is similar to the training set distribution; otherwise, their performance might degrade significantly. Domain adaptation and domain generalization techniques reduce differences between training and test distributions by learning domain invariant features (Khandelwal & Yushkevich, 2020). Also, several studies have focused on the generalization of deep learning models. For instance, Lee et al. (2020) trained a CycleGAN to learn the training set intensity distribution. Then, they used this network to adapt the arbitrary intensity distribution to the specific intensity distribution of the training set. They confirmed that their method creates images similar to the training set domain without significant feature loss.

### 9.8 High-performance computational resources

Deep learning algorithms usually demand high-performance computational resources. The high computational cost of medical data analysis makes this problem worse. Recently, many hardware accelerators have been developed to provide the required computational power for deep learning projects (Talib et al., 2020). However, the training of deep learning algorithms is still time-consuming, especially with medical data. Parallel processing techniques also alleviate this problem.

## 10. Open challenges and future works

The integration of multiple modalities helps a model perform much better because the model takes advantage of additional information provided by different data types. As a result, multimodal data have attracted many researchers recently, especially in the medical field. This section describes open challenges and future scopes of this topic.

- The most challenging problem of training a deep neural network using medical data is data scarcity. A state-of-the-art approach to deal with the lack of data is few-shot learning (Fei Fei et al., 2006;

Fink, 2004). As the name implies, few-shot learning is the practice of training a model with a limited amount of training data, which also leads to less computational costs. Few-shot learning involves identifying the key features of each class to distinguish between the classes (Kotia et al., 2021). Few-shot learning with one training sample in each class is called one-shot learning. Another type of few-shot learning is zero-shot learning in which no training sample is available (Wang et al., 2020). Although few-shot learning is an active research area, a few studies have used it in multimodal medical data analyses. As a result, further investigation is required.

- Unsupervised, self-supervised, and semi-supervised methods have recently received considerable attention in medical analyses due to their potential for reducing the effort of labeling data. However, a few studies in multimodal medical data analysis have used these techniques.
- GAN has become wildly popular in the medical field recently. One of its current applications is data augmentation; however, further research is required to make GAN's output more reliable. A novel target in medical analyses is differential privacy-based GAN, which creates new fake data without memorizing individual characteristics of samples in the training set. Consequently, these models make large amounts of new fake medical data that can be released publicly. On the other hand, GAN requires an enormous computational cost and GPU memory. Some researchers have focused on this problem, but it is still an open area for further research. Finally, GAN can be used for unsupervised learning and generalization.
- Large neural networks generally address complex problems better, but they require more computational power and memory. One popular approach to tackle this problem is neural network pruning, which aims to eliminate a significant number of parameters from a neural network without affecting its accuracy. Neural networks can be pruned before, during, or after training the model (Cun et al., 1990; Han et al., 2015; Frankle et al., 2020). Furthermore, many studies have concentrated on reducing overall memory by compression, but only a few have aimed at speeding up layers (Maji & Mullins, 2018).
- Some combinations of modalities are commonly used in multimodal medical analysis. The main reason is that they have proved effective in improving models' performance. However, it may be beneficial to consider other combinations because they may reveal more information. As a result, future studies should try new combinations.

## 11. Conclusion

Deep learning is a powerful technique to analyze complex medical data. Recently, deep learning methods have become very popular in medical analyses because they have achieved outstanding results in this field. Multimodal data improve neural networks' performance as they provide complementary information. This paper presents a comprehensive overview of the latest studies on multimodal medical data analysis using deep learning algorithms. We divided related articles into four main categories, including supervised, semi-supervised, self-supervised, and unsupervised methods. We observed that many articles on COVID-19 had used transfer learning because they did not access large datasets. We conclude transfer learning methods are invaluable in situations such as pandemics when not enough data is available. Different modalities, deep learning architectures, and fusion strategies are also introduced in this paper. Furthermore, we provided links to access some of the most well-known multimodal datasets and identified common problems and open challenges in this field. We believe that deep learning methods in multimodal medical data analysis will remain an active research area in the coming years.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


Abrol, A., Fu, Z., Du, Y., & Calhoun, V. D. (2019). Multimodal Data Fusion of Deep Learning and Dynamic Functional Connectivity Features to Predict Alzheimer's Disease Progression *. *In IEEE Xplore*. https://doi.org/10.1109/EMBC.2019.8856500

Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K.-S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion, 63*. https://doi.org/10.1016/j.inffus.2020.06.008

Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., & Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging, 6*(01). https://doi.org/10.1117/1.jmi.6.1.014006

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*(1). https://doi.org/10.1186/s40537-021-00444-8

Antonelli, L., Guarracino, M. R., Maddalena, L., & Sangiovanni, M. (2019). Integrating imaging and omics data: A review. *Biomedical Signal Processing and Control, 52*. https://doi.org/10.1016/j.bspc.2019.04.032

Bagheri, A., Groenhof, T. K. J., Veldhuis, W. B., de Jong, P. A., Asselbergs, F. W., & Oberski, D. L. (2020). Multimodal Learning for Cardiovascular Risk Prediction using EHR Data. In *arXiv:2008.11979 [cs, eess, stat]*. https://arxiv.org/abs/2008.11979.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.

Bai, X., Fang, C., Zhou, Y., Bai, S., Liu, Z., Chen, Q., Xu, Y., Xia, T., Gong, S., Xie, X., Song, D., Du, R., Zhou, C., Chen, C., Nie, D., Tu, D., Zhang, C., Liu, X., Qin, L., & Chen, W. (2020). Predicting COVID-19 malignant progression with AI techniques. *SSRN*. https://doi.org/10.1101/2020.03.20.20037325

Bell, J. (2004). Predicting disease using genomics. *Nature, 429*(6990). https://doi.org/10.1038/nature02624

Betts, J.G., Young, K.A., Wise, J.A., Johnson, E., Poe, B., Kruse, D.H., Korol, O., Johnson, E.J., Womble, M., & DeSaix, P. (2013). *Anatomy and Physiology*. OpenStax.

Bhatnagar, G., Wu, Q. M. J., & Liu, Z. (2015). A new contrast based multimodal medical image fusion framework. *Neurocomputing, 157*. https://doi.org/10.1016/j.neucom.2015.01.025

Boss, A., Bisdas, S., Kolb, A., Hofmann, M., Ernemann, U., Claussen, C. D., Pfannenberg, C., Pichler, B. J., Reimold, M., & Stegger, L. (2010). Hybrid PET/MRI of Intracranial Masses: Initial Experiences and Comparison to PET/CT. *Journal of Nuclear Medicine, 51*(8). https://doi.org/10.2967/jnumed.110.074773

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. CRC Press: In Google Books.

Cao, B., Zhang, H., Wang, N., Gao, X., & Shen, D. (2020). Auto-GAN: Self-supervised collaborative learning for medical image synthesis. In *AAAI 2020–34th AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i07.6619

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Qi, T., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. ArXiv Preprint ArXiv:2105.05537.

Caruana, R. (1997). Multitask Learning. *Machine Learning, 28*(1). https://doi.org/10.1023/a:1007379606734

Chapelle, O., Chi, M., & Zien, A. (2006). A continuation method for semi-supervised SVMs. *ACM International Conference Proceeding Series, 148*. https://doi.org/10.1145/1143844.1143868

Chaudhari, S., Mithal, V., Polatkan, G., & Ramanath, R. (2019). An attentive survey of attention models. ArXiv Preprint ArXiv:1904.02874.

Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics, 35*(14). https://doi.org/10.1093/bioinformatics/btz342

Chellapilla, K., Puri, S., & Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing BT – Tenth International Workshop on Frontiers in Handwriting Recognition. *Tenth International Workshop on Frontiers in Handwriting Recognition*.

Chen, H., Gao, M., Zhang, Y., Liang, W., & Zou, X. (2019). Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model. *BioMed Research International, 2019*. https://doi.org/10.1155/2019/9523719

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Le Lu, Yuille, A. L., & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. ArXiv Preprint ArXiv:2102.04306.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. https://arxiv.org/abs/1412.3555.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1). https://doi.org/10.1109/tit.1967.1053964

Cun, L., Le Cun, Y., Denker, J., & Solla, S. (1990). Optimal Brain Damage. *Advances in Neural Information Processing Systems, 2*.

Dwork, C. (2006). Differential Privacy. *Automata, Languages and Programming, 4052*. https://doi.org/10.1007/11787006_1

El-Sappagh, S., Abuhmed, T., Riazul Islam, S. M., & Kwak, K. S. (2020). Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing, 412*. https://doi.org/10.1016/j.neucom.2020.05.087

El Asnaoui, K., & Chawki, Y. (2020). Using X-ray images and deep learning for automated detection of coronavirus disease. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2020.1767212

Fang, C., Bai, S., Chen, Q., Zhou, Y., Xia, L., Qin, L., Gong, S., Xie, X., Zhou, C., Tu, D., Zhang, C., Liu, X., Chen, W., Bai, X., & Torr, P. H. S. (2021). Deep learning for predicting COVID-19 malignant progression. *Medical Image Analysis, 72*. https://doi.org/10.1016/j.media.2021.102096

Fei Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4). https://doi.org/10.1109/tpami.2006.79

Feng, C., Elazab, A., Yang, P., Wang, T., Zhou, F., Hu, H., Xiao, X., & Lei, B. (2019). Deep Learning Framework for Alzheimer's Disease Diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access, 7*. https://doi.org/10.1109/access.2019.2913847

Fink, M. (2004). Object classification from a single example utilizing class relevance metrics. *Advances in Neural Information Processing Systems, 17*.

Frankle, J., Dziugaite, G. K., Roy, D. M., & Carbin, M. (2020). Pruning Neural Networks at Initialization: Why are We Missing the Mark? https://arxiv.org/abs/2009.08576.

Frid Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing, 321*. https://doi.org/10.1016/j.neucom.2018.09.013

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4). https://doi.org/10.1007/BF00344251

Ge, C., Gu, I. Y. H., Jakola, A. S., & Yang, J. (2020). Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging, 20*(1). https://doi.org/10.1186/s12880-020-00485-0

Ghaffari, M., Sowmya, A., & Oliver, R. (2020). Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012–2018 Challenges. *IEEE Reviews in Biomedical Engineering, 13*. https://doi.org/10.1109/RBME.2019.2946868

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1406.2661.

Götz, T. I. (2019). Technical report: time-activity-curve integration in Lu-177 therapies in nuclear medicine. ArXiv Preprint ArXiv:1907.06617.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2018). *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. https://arxiv.org/abs/1706.02677.

Guo, Z., Li, X., Huang, H., Guo, N., & Li, Q. (2019). Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences, 3*(2). https://doi.org/10.1109/TRPMS.2018.2890359

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems*, *2015-January*.

Hatamizadeh, A., Yang, D., Roth, H., & Xu, D. (2021). UNETR: Transformers for 3D Medical Image Segmentation. *ArXiv Preprint ArXiv:2103.10504*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hervella, A. S., Ramos, L., Rouco, J., Novo, J., & Ortega, M. (2020). Multi-Modal Self-Supervised Pre-Training for Joint Optic Disc and Cup Segmentation in Eye Fundus Images. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, *2020-May*. https://doi.org/10.1109/ICASSP40776.2020.9053551.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2019). Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data. *Proceedings of the International Joint Conference on Neural Networks*, *2019-July*. https://doi.org/10.1109/IJCNN.2019.8851844.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7). https://doi.org/10.1162/neco.2006.18.7.1527

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8). https://doi.org/10.1162/neco.1997.9.8.1735

Hooshmand, S. A., Zarei Ghobadi, M., Hooshmand, S. E., Azimzadeh Jamalkandi, S., Alavi, S. M., & Masoudi-Nejad, A. (2020). A multimodal deep learning-based drug repurposing approach for treatment of COVID-19. *Molecular Diversity*. https://doi.org/10.1007/s11030-020-10144-9

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8). https://doi.org/10.1073/pnas.79.8.2554.

Hosseini, M. P., Tran, T. X., Pompili, D., Elisevich, K., & Soltanian-Zadeh, H. (2020). Multimodal data analysis of epileptic EEG and rs-fMRI via deep learning and edge computing. *Artificial Intelligence in Medicine, 104*. https://doi.org/10.1016/j.artmed.2020.101813

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(2). https://doi.org/10.1109/tsmcb.2011.2168604

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.243

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W., & Wu, J. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, *2020-May*. https://doi.org/10.1109/ICASSP40776.2020.9053405.

Huang, Y., & Chung, A. C. S. (2020). Semi-Supervised Multimodality Learning with Graph Convolutional Neural Networks for Disease Diagnosis. *Proceedings – International Conference on Image Processing, ICIP*, *2020-October*. https://doi.org/10.1109/ICIP40778.2020.9191172.

Hung, C. Y., Lin, C. H., Chang, C. S., Li, J. L., & Lee, C. C. (2019). Predicting Gastrointestinal Bleeding Events from Multimodal In-Hospital Electronic Health Records Using Deep Fusion Networks. In *IEEE Xplore*. https://doi.org/10.1109/EMBC.2019.8857244.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. https://arxiv.org/abs/1602.07360.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2019). No New-Net. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 11384*. https://doi.org/10.1007/978-3-030-11726-9_21

Ivakhnenko, A. G. (1968). The group method of data of handling; a rival of the method of stochastic approximation. *Soviet Automatic Control, 1*(3), 43–55.

Jacene, H. A., Goetze, S., Patel, H., Wahl, R. L., & Ziessman, H. A. (2008). Advantages of Hybrid SPECT/CT vs SPECT Alone. *The Open Medical Imaging Journal, 2*(1). https://doi.org/10.2174/1874347100802010067

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A Survey on Contrastive Self-Supervised Learning. *Technologies, 9*(1). https://doi.org/10.3390/technologies9010002

Jiang, X., Li, J., Kan, Y., Yu, T., Chang, S., Sha, X., Zheng, H., Luo, Y., & Wang, S. (2020). MRI Based Radiomics Approach with Deep Learning for Prediction of Vessel Invasion in Early-Stage Cervical Cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2019.2963867

Jing, L., & Tian, Y. (2020). Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/tpami.2020.2992393

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1). https://doi.org/10.1186/s40537-019-0192-5

Ju, W., Xiang, D., Zhang, B., Wang, L., Kopriva, I., & Chen, X. (2015). Random Walk and Graph Cut for Co-Segmentation of Lung Tumor on PET-CT Images. *IEEE Transactions on Image Processing, 24*(12). https://doi.org/10.1109/TIP.2015.2488902

Kanoun, S., Rossi, C., & Casasnovas, O. (2018). [18F]FDG-PET/CT in hodgkin lymphoma: Current usefulness and perspectives. In *Cancers* (Vol. 10, Issue 5). https://doi.org/10.3390/cancers10050145.

Kasban, H., El-Bendary, M. A. M., & Salama, D. H. (2015). A Comparative Study of Medical Imaging Techniques. International Journal of Information Science and Intelligent System, *4*(2).

Kassani, S. H., Kassasni, P. H., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2020). *Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning-Based Approach*. https://arxiv.org/abs/2004.10641.

Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. (2019). KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. *Multimedia Tools and Applications, 79*(47–48). https://doi.org/10.1007/s11042-019-07839-z

Khandelwal, P., & Yushkevich, P. (2020). Domain Generalizer: A Few-Shot Meta Learning Framework for Domain Generalization in Medical Imaging. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, 12444*. https://doi.org/10.1007/978-3-030-60548-3_8

Kim, J., & Lee, B. (2018). Identification of Alzheimer's disease and mild cognitive impairment using multimodal sparse hierarchical extreme learning machine. *Human Brain Mapping, 39*(9). https://doi.org/10.1002/hbm.24207

Kirienko, M., Sollini, M., Silvestri, G., Mognetti, S., Voulaz, E., Antunovic, L., Rossi, A., Antiga, L., & Chiti, A. (2018). Convolutional Neural Networks Promising in Lung Cancer T-Parameter Assessment on Baseline FDG-PET/CT. *Contrast Media & Molecular Imaging*. https://www.hindawi.com/journals/cmmi/2018/1382309/abs/.

Kolesnikov, A., Zhai, X., & Beyer, L. (2019). Revisiting self-supervised visual representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June*. https://doi.org/10.1109/CVPR.2019.00202.

Konečný, J., McMahan, B., & Ramage, D. (2015). *Federated Optimization: Distributed Optimization Beyond the Datacenter*. https://arxiv.org/abs/1511.03575.

Kotia, J., Kotwal, A., Bharti, R., & Mangrulkar, R. (2021). Few Shot Learning for Medical Imaging. *Machine Learning Algorithms for Industrial Applications. Studies Computational Intelligence, 907*. https://doi.org/10.1007/978-3-030-50641-4_7

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress Artificial Intelligence, 5*(4). https://doi.org/10.1007/s13748-016-0094-0

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., & Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer, 14*(5). https://doi.org/10.1038/nrc3721

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 2*.

Kwak, G. H., & Hui, P. (2019). *DeepHealth: Review and challenges of artificial intelligence in health informatics*. https://arxiv.org/abs/1909.00384.

Lai, Y. H., Chen, W. N., Hsu, T. C., Lin, C., Tsao, Y., & Wu, S. (2019). Predicting the Prognosis of Non-Small Cell Lung Cancer by Integrating Microarray and Clinical Data with Deep Learning. *bioRxiv*. https://doi.org/10.1101/656140

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., … Blum, M. G. B. (2020). Integration of clinical characteristics, lab tests and a deep learning CT scan analysis to predict severity of hospitalized COVID-19 patients. *MedRxiv*. https://doi.org/10.1101/2020.05.14.20101972

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation, 1*(4). https://doi.org/10.1162/neco.1989.1.4.541

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553). https://doi.org/10.1038/nature14539

Lee, D. H., Li, Y., & Shin, B.-S. (2020). Generalization of intensity distribution of medical images using GANs. *Human-Centric Computing and Information Sciences, 10*(1). https://doi.org/10.1186/s13673-020-00220-2

Lee, G., Kang, B., Nho, K., Sohn, K.-A., & Kim, D. (2019). MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework. *Frontiers in Genetics, 10*. https://doi.org/10.3389/fgene.2019.00617

Lee, G., Nho, K., Kang, B., Sohn, K.-A., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-018-37769-z

Li, H., Boimel, P., Janopaul-Naylor, J., Zhong, H., Xiao, Y., Ben-Josef, E., & Fan, Y. (2019). Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. *Proceedings. IEEE International Symposium on Biomedical Imaging (ISBI 2019)*. https://doi.org/10.1109/ISBI.2019.8759301.

Li, X., Jia, M., Islam, M. T., Yu, L., & Xing, L. (2020). Self-Supervised Feature Learning via Exploiting Multi-Modal Data for Retinal Disease Diagnosis. *IEEE Transactions on Medical Imaging, 39*(12). https://doi.org/10.1109/TMI.2020.3008871

Liang, S., Zhang, R., Liang, D., Song, T., Ai, T., Xia, C., Xia, L., & Wang, Y. (2018). Multimodal 3D DenseNet for IDH Genotype Prediction in Gliomas. *Genes, 9*(8). https://doi.org/10.3390/genes9080382

Lin, E., & Alessio, A. (2009). What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT? *Journal of Cardiovascular Computed Tomography, 3*(6). https://doi.org/10.1016/j.jcct.2009.07.003

Lin, W., Gao, Q., Yuan, J., Chen, Z., Feng, C., Chen, W., Du, M., & Tong, T. (2020). Predicting Alzheimer's Disease Conversion From Mild Cognitive Impairment Using an Extreme Learning Machine-Based Grading Method With Multimodal Data. *Frontiers in Aging Neuroscience, 12*. https://doi.org/10.3389/fnagi.2020.00077

Lindell, Y., & Pinkas, B. (2009). Secure Multiparty Computation for Privacy-Preserving Data Mining. *Journal of Privacy and Confidentiality, 1*(1). https://doi.org/10.29012/jpc.v1i1.566.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*. https://doi.org/10.1016/j.media.2017.07.005

Liu, M., Cheng, D., Wang, K., & Wang, Y. (2018). Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Neuroinformatics, 16*(3–4). https://doi.org/10.1007/s12021-018-9370-4

Liu, M., Zhang, J., Adeli, E., & Shen, D. (2019). Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer's Disease Diagnosis. *IEEE Transactions on Biomedical Engineering, 66*(5). https://doi.org/10.1109/TBME.2018.2869989

Liu, Q., & Hu, P. (2019). Association Analysis of Deep Genomic Features Extracted by Denoising Autoencoders in Breast Cancer. *Cancers, 11*(4). https://doi.org/10.3390/cancers11040494

Liu, Y., Pan, S., Jin, M., Zhou, C., Xia, F., & Yu, P. S. (2021). Graph Self-Supervised Learning: A Survey. *CoRR, abs/2103.00111*. https://arxiv.org/abs/2103.00111.

Lo Gullo, R., Daimiel, I., Morris, E. A., & Pinker, K. (2020). Combining molecular and imaging metrics in cancer: Radiogenomics. *Insights into Imaging, 11*(1). https://doi.org/10.1186/s13244-019-0795-6

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., & Beg, M. F. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-22871-z

Ma, Z., Wu, X., Sun, S., Xia, C., Yang, Z., Li, S., & Zhou, J. (2018). A discriminative learning based approach for automated nasopharyngeal carcinoma segmentation leveraging multi-modality similarity metric learning. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. https://doi.org/10.1109/ISBI.2018.8363696.

Maghdid, H. S., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., & Khan, M. K. (2020). *Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms*. https://arxiv.org/abs/2004.00038.

Maier, A., Steidl, S., Christlein, V., & Hornegger, J. (Eds.). (2018). *Medical Imaging Systems* (Vol. 11111). Springer International Publishing. https://doi.org/10.1007/978-3-319-96520-8.

Maier, A., Syben, C., Lasser, T., & Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift fur Medizinische Physik, 29*(2). https://doi.org/10.1016/j.zemedi.2018.12.003

Maji, P., & Mullins, R. (2018). On the Reduction of Computational Complexity of Deep Convolutional Neural Networks. *Entropy, 20*(4). https://doi.org/10.3390/e20040305

Mandic, D. P., & Chambers, J. A. (2001). Recurrent Neural Networks for Prediction. *Wiley Series in Adaptive and Learning Systems for Signal Processing, Communications, and Control*. https://doi.org/10.1002/047084535x

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4). https://doi.org/10.1007/BF02478259

McKinley, R., Meier, R., & Wiest, R. (2019). Ensembles of Densely-Connected CNNs with Label-Uncertainty for Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 11384*. https://doi.org/10.1007/978-3-030-11726-9_40

McKinley, R., Rebsamen, M., Meier, R., & Wiest, R. (2020). Triplanar Ensemble of 3D-to-2D CNNs with Label-Uncertainty for Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 11992*. https://doi.org/10.1007/978-3-030-46640-4_36

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., … Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10). https://doi.org/10.1109/TMI.2014.2377694

Micheel, C. M., Nass, S. J., Omenn, G. S., & Policy, H. S. (2012). Evolution of Translational Omics Lessons Learned and the Path Forward. *Evolution*.

Milecki, L., Bodard, S., Correas, J. M., Timsit, M. O., & Vakalopoulou, M. (2021). 3D unsupervised kidney graft segmentation based on deep learning and multi-sequence MRI. *Proceedings – International Symposium on Biomedical Imaging, 2021-April*. https://doi.org/10.1109/ISBI48211.2021.9433854.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*. https://doi.org/10.1109/3dv.2016.79

Mukherjee, H., Ghosh, S., Dhar, A., Obaidullah, S. M., Santosh, K. C., & Roy, K. (2020). Deep neural network to detect COVID-19: One architecture for both CT Scans and Chest X-rays. *Applied Intelligence*. https://doi.org/10.1007/s10489-020-01943-6

Myronenko, A. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11384 LNCS*. https://doi.org/10.1007/978-3-030-11726-9_28.

Ng, A. (2018). Structuring Machine Learning Projects. *Coursera*. https://www.coursera.org/learn/machine-learning-projects?specialization=deep-learning.

Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., Liu, L., Wang, Q., Wu, J., & Shen, D. (2019). Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-018-37387-9

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M. C. H., Heinrich, M. P., Misawa, K., Mori, K., McDonagh, S. G., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning Where to Look for the Pancreas. *CoRR, abs/1804.03999*. http://arxiv.org/abs/1804.03999.

Peng, Y., Bi, L., Guo, Y., Feng, D., Fulham, M., & Kim, J. (2019). Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://doi.org/10.1109/EMBC.2019.8857666

Pratt, L. Y., Mostow, J., & Kamm, C. A. (1991). Direct Transfer of Learned Information Among Neural Networks. In *Proceedings of the ninth National conference on Artificial intelligence (AAAI 91)* (Vol. 2). https://www.aaai.org/Library/AAAI/1991/aaai91-091.php.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., & Pirmohamed, M. (2018). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery, 18*(1). https://doi.org/10.1038/nrd.2018.168

Qin, R., Wang, Z., Jiang, L., Qiao, K., Hai, J., Chen, J., Xu, J., Shi, D., & Yan, B. (2020). Fine-Grained Lung Cancer Classification from PET and CT Images Based on Multidimensional Attention Mechanism. *Complexity, 2020*. https://doi.org/10.1155/2020/6153657

Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., … Kolachalama, V. B. (2020). Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain, 143*(6). https://doi.org/10.1093/brain/awaa137

Raja, K., Patrick, M., Gao, Y., Madu, D., Yang, Y., & Tsoi, L. C. (2017). A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. *International Journal of Genomics, 2017*. https://doi.org/10.1155/2017/6213474

Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine, 34*(6). https://doi.org/10.1109/msp.2017.2738401

Raza, K., & Singh, N. K. (2021). A Tour of Unsupervised Deep Learning for Medical Image Analysis. *Current Medical Imaging Formerly Current Medical Imaging Reviews, 17*. https://doi.org/10.2174/1573405617666210127154257

Rehman, A., Naz, S., Khan, A., Zaib, A., & Razzak, I. (2020). *Improving Coronavirus (COVID-19) Diagnosis using Deep Transfer Learning*. https://doi.org/10.1101/2020.04.11.20054643.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2). https://doi.org/10.1007/s10462-009-9124-7

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science, 9351*. https://doi.org/10.1007/978-3-319-24574-4_28

Rosenblatt, F. (1957). The Perceptron, A Perceiving and Recognizing Automaton Project Para. *Cornell Aeronautical Laboratory*.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. In *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc.. https://doi.org/10.1002/9780470316696

Rubinstein, E., Salhov, M., Nidam-Leshem, M., White, V., Golan, S., Baniel, J., Bernstine, H., Groshar, D., & Averbuch, A. (2019). Unsupervised tumor detection in Dynamic PET/CT imaging of the prostate. *Medical Image Analysis, 55*. https://doi.org/10.1016/j.media.2019.04.001

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088). https://doi.org/10.1038/323533a0

Saba, T., Sameh Mohamed, A., El-Affendi, M., Amin, J., & Sharif, M. (2020). Brain tumor detection using fusion of hand crafted and deep learning features. *Cognitive Systems Research, 59*. https://doi.org/10.1016/j.cogsys.2019.09.007

Schrodi, S. J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J. J., Callear, A. P., Carter, T. C., Ye, Z., Haines, J. L., Brilliant, M. H., Crane, P. K., Smelser, D. T., Elston, R. C., & Weeks, D. E. (2014). Genetic-based prediction of disease traits:

Prediction is very difficult, especially about the future. *Frontiers in Genetics, 5*, 162. https://doi.org/10.3389/fgene.2014.00162

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2018). Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE Journal of Biomedical and Health Informatics, 22*(1). https://doi.org/10.1109/JBHI.2017.2655720

Shikalgar, A., & Sonavane, S. (2020). Hybrid Deep Learning Approach for Classifying Alzheimer Disease Based on Multimodal Data. *Advances in Intelligent Systems and Computing, 1025*. https://doi.org/10.1007/978-981-32-9515-5_49

Shukla, S. N., & Marlin, B. M. (2020). Integrating Physiological Time Series and Clinical Notes with Deep Learning for Improved ICU Mortality Prediction. In *arXiv: 2003.11059 [cs, stat]*. https://arxiv.org/abs/2003.11059.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. http://arxiv.org/abs/1312.6034.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556.

Soltaninejad, M., Zhang, L., Lambrou, T., Yang, G., Allinson, N., & Ye, X. (2019). MRI Brain Tumor Segmentation using Random Forests and Fully Convolutional Networks. In *arXiv:1909.06337 [cs]*. https://arxiv.org/abs/1909.06337.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). *Highway Networks*. https://arxiv.org/abs/1505.00387.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ, 338*, 1. https://doi.org/10.1136/bmj.b2393

Suk, H. I., Lee, S. W., & Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage, 101*. https://doi.org/10.1016/j.neuroimage.2014.06.077

Sun, D., Wang, M., & Li, A. (2019). A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16*(3). https://doi.org/10.1109/TCBB.2018.2806438

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper With Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taleb, A., Lippert, C., Klein, T., & Nabi, M. (2021). *Multimodal Self-supervised Learning for Medical Image Analysis*. https://doi.org/10.1007/978-3-030-78191-0_51.

Talib, M. A., Majzoub, S., Nasir, Q., & Jamal, D. (2020). A systematic literature review on hardware implementation of artificial intelligence algorithms. *The Journal of Supercomputing*. https://doi.org/10.1007/s11227-020-03325-8

Tang, Z., Xu, Y., Jin, L., Aibaidula, A., Lu, J., Jiao, Z., Wu, J., Zhang, H., & Shen, D. (2020). Deep Learning of Imaging Phenotype and Genotype for Predicting Overall Survival Time of Glioblastoma Patients. *IEEE Transactions on Medical Imaging, 39*(6). https://doi.org/10.1109/TMI.2020.2964310

Vaghefi, E., Hill, S., Kersten, H. M., & Squirrell, D. (2020). Multimodal Retinal Image Analysis via Deep Learning for the Diagnosis of Intermediate Dry Age-Related Macular Degeneration: A Feasibility Study. In. *Journal of Ophthalmology*. https://www.hindawi.com/journals/joph/2020/7493419/.

Valanarasu, J., Jose, M., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. *ArXiv Preprint ArXiv:2102.10662*.

van Sonsbeek, T., & Worring, M. (2020). Towards Automated Diagnosis with Attentive Multi-modal Learning Using Electronic Health Records and Chest X-Rays. *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures, 12445*. https://doi.org/10.1007/978-3-030-60946-7_11

Varghese, A., Vaidhya, K., Thirunavukkarasu, S., Kesavdas, C., & Krishnamurthi, G. (2016). Semi-supervised Learning using Denoising Autoencoders for Brain LesionDetection and Segmentation. *CoRR, abs/1611.08664*. http://arxiv.org/abs/1611.08664.

Vasquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J., & Noth, E. (2018). Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach. *IEEE Journal of Biomedical and Health Informatics, 23*(4). https://doi.org/10.1109/jbhi.2018.2866873

Vu, T.-D., Ho, N.-H., Yang, H.-J., Kim, J., & Song, H.-C. (2018). Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection. *Soft Computing, 22*(20). https://doi.org/10.1007/s00500-018-3421-5

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). CovidGAN: Data Augmentation using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access, 8*. https://doi.org/10.1109/ACCESS.2020.2994762

Wang, S., Zha, Y., Li, W., Wu, Q., Li, X., Niu, M., Wang, M., Qiu, X., Li, H., Yu, H., Gong, W., Bai, Y., Li, L., Zhu, Y., Wang, L., & Tian, J. (2020). A Fully Automatic Deep Learning System for COVID-19 Diagnostic and Prognostic Analysis. *European Respiratory Journal*. https://doi.org/10.1183/13993003.00775-2020

Wang, Y., Yang, Y., Guo, X., Ye, C., Gao, N., Fang, Y., & Ma, H. T. (2018). A Novel Multimodal MRI Analysis for Alzheimer's Disease Based on Convolutional Neural Network. *IEEE Xplore*. https://doi.org/10.1109/EMBC.2018.8512372

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11211 LNCS*. https://doi.org/10.1007/978-3-030-01234-2_1.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, 2048–2057.

Xu, M., Ouyang, L., Gao, Y., Chen, Y., Yu, T., Li, Q., Sun, K., Bao, F. S., Safarnejad, L., Wen, J., Jiang, C., Chen, T., Han, L., Zhang, H., Gao, Y., Yu, Z., Liu, X., Yan, T., Li, H., … Chen, S. (2020). *Accurately Differentiating COVID-19, Other Viral Infection, and Healthy Individuals Using Multimodal Features via Late Fusion Learning*. https://doi.org/10.1101/2020.08.18.20176776.

Xu, Y. (2019). Deep Learning in Multimodal Medical Image Analysis. *Health Information Science, 11837*. https://doi.org/10.1007/978-3-030-32962-4_18

Xu, Z., Yan, J., Luo, J., Li, X., & Jagadeesan, J. (2021). *Unsupervised Multimodal Image Registration with Adaptative Gradient Guidance*. https://doi.org/10.1109/icassp39728.2021.9414320.

Yan, R., Ren, F., Rao, X., Shi, B., Xiang, T., Zhang, L., Liu, Y., Liang, J., Zheng, C., & Zhang, F. (2019). Integration of Multimodal Data for Breast Cancer Classification Using a Hybrid Deep Learning Method. *Intelligent Computing Theories and Application, 11643*. https://doi.org/10.1007/978-3-030-26763-6_44

Yap, J., Yolland, W., & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Experimental Dermatology, 27*(11). https://doi.org/10.1111/exd.13777

Yu, Y., Li, M., Liu, L., Li, Y., & Wang, J. (2019). Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics, 2*(4). https://doi.org/10.26599/bdma.2019.9020007.

Yuan, Y., Borrmann, D., Hou, J., Ma, Y., Nüchter, A., & Schwertfeger, S. (2021). Self-supervised point set local descriptors for point cloud registration. *Sensors (Switzerland), 21*(2). https://doi.org/10.3390/s21020486

Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., & Zhang, X. (2019). Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. *Neurocomputing, 361*. https://doi.org/10.1016/j.neucom.2019.04.093

Zhang, T., & Shi, M. (2020). Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease. *Journal of Neuroscience Methods, 341*. https://doi.org/10.1016/j.jneumeth.2020.108795

Zhang, Y. D., Dong, Z., Wang, S. H., Yu, X., Yao, X., Zhou, Q., Hu, H., Li, M., Jiménez-Mesa, C., Ramirez, J., Martinez, F. J., & Gorriz, J. M. (2020). Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion, 64*. https://doi.org/10.1016/j.inffus.2020.07.006

Zhang, Y. D., Zhang, Z., Zhang, X., & Wang, S. H. (2021). MIDCAN: A multiple input deep convolutional attention network for Covid-19 diagnosis based on chest CT and chest X-ray. *Pattern Recognition Letters, 150*. https://doi.org/10.1016/j.patrec.2021.06.021

Zhang, Y., & Wallace, B. C. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) ConvolutionalNeural Networks for Sentence Classification. *CoRR, abs/1510.03820*. http://arxiv.org/abs/1510.03820.

Zhao, X., Li, L., Lu, W., & Tan, S. (2018). Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Physics in Medicine & Biology, 64*(1). https://doi.org/10.1088/1361-6560/aaf44b

Zhao, Y., Gafita, A., Vollnberg, B., Tetteh, G., Haupt, F., Afshar-Oromieh, A., Menze, B., Eiber, M., Rominger, A., & Shi, K. (2020). Deep neural network for automatic characterization of lesions on 68Ga-PSMA-11 PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging, 47*(3). https://doi.org/10.1007/s00259-019-04606-y

Zhou, T., Ruan, S., & Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array, 3–4*. https://doi.org/10.1016/j.array.2019.100004

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11045 LNCS*. https://doi.org/10.1007/978-3-030-00889-5_1

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-To-Image Translation Using.