Information Fusion 105 (2024) 102217

Contents lists available at ScienceDirect

Information Fusion

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

# A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges

Khaled Bayoudh *

*Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia*
*Laboratoire ImViA, EA 7535, Université de Bourgogne, Dijon, France*

ARTICLE INFO

*Keywords:*
Applications
Computer vision
Multimodal hybrid deep learning
Sensory modalities

ABSTRACT

In recent years, deep learning algorithms have rapidly revolutionized artificial intelligence, particularly machine learning, enabling researchers and practitioners to extend previously hand-crafted feature extraction procedures. In particular, deep learning uses adaptive learning processes to learn more complex and informative patterns from datasets of varying sizes. With the increasing availability of multimodal data streams and recent advances in deep learning algorithms, multimodal deep learning is on the rise. This requires the development of complex models that can process and analyze multimodal information in a consistent manner. However, unstructured data can come in many different forms (also known as modalities). Extracting relevant features from this data remains an ambitious goal for deep learning researchers. According to the literature, most deep learning systems consist of a single architecture (i.e., standalone deep learning). When two or more deep learning architectures are combined over multiple sensory modalities, the result is called a multimodal hybrid deep learning model. Since this research direction has received much attention in the field of deep learning, the purpose of this survey is to provide a broader overview of the topic. In this paper, we provide a comprehensive review of recent advances in multimodal hybrid deep learning, including a thorough analysis of the most commonly developed hybrid architectures. In particular, one of the main challenges in multimodal hybrid analysis is the ability of these architectures to systematically integrate cross-modal features in hybrid designs. Therefore, we propose a generic framework for multimodal hybrid learning that focuses mainly on fusion methods. We also identify trends and challenges in multimodal hybrid learning and provide insights and directions for future research. Our findings show that multimodal hybrid learning can perform well in a variety of challenging computer vision applications and tasks.

## 1. Introduction

Deep learning is an exciting paradigm in the pattern recognition and machine learning communities. In essence, the proliferation of high performance computing and large annotated datasets (big data), coupled with the rapid development of deep neural networks (DNNs), has greatly improved the quality of the data learning process [1]. There has been a rapid shift from shallow learning of hand-crafted features extracted using local descriptors (e.g., HOG, SIFT, etc.) to automatic learning of high-level feature representations derived from raw data, allowing complex features and deep classifiers to be learned together [2]. In general, this new approach requires little hands-on engineering, as the parameter tuning pipeline can be run directly from the raw data with fewer time-consuming data pre-processing steps. Flexible learning mechanisms and the use of deep networks are key factors that facilitate the use of deep learning methods, along with the development of appropriate frameworks. Many vision-based challenges have been effectively solved by integrating the deep learning paradigm instead of hand-crafted methods. In addition, many computer vision tasks, especially image classification, object detection, and semantic segmentation, have been introduced, and the latest competitive solutions are based on deep learning methods [3]. In this paper, we first provide a basic overview of deep learning and explain how deep learning can be used in practice, including a discussion of the most commonly used DNN models and their main properties.

Standalone discriminative and generative DNNs, such as convolutional neural networks (CNNs) [4] and generative adversarial networks (GANs) [5], are popular deep learning techniques that have demonstrated state-of-the-art performance in many real-world applications

* Correspondence to: Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia.
*E-mail address:* khaled.bayoudh@u-bourgogne.fr.

https://doi.org/10.1016/j.inffus.2023.102217
Received 24 August 2023; Received in revised form 1 December 2023; Accepted 26 December 2023
Available online 30 December 2023

and use cases [6]. For example, CNNs are mainly used to process visual cues and have shown better generalization ability than other discriminative networks; RNNs and their variants are another type of DNN that can be trained to extract short- and long-range dependencies between features from continuous data. A model can be classified as a hybrid deep learning model when two deep learning architectures are merged using a bridging network (or fusion method), or when deep learning and machine learning are mixed for prediction purposes. We call this process hybridization. This process, or the act of merging two or more standalone architectures, is often referred to as a hybrid fusion.

The cognitive performance of systems can be enhanced by their ability to integrate complementary and semantic knowledge from multiple data sources and perspectives (in this study referred to as modalities) [7]. The main challenge is to learn how to combine heterogeneous features from different modalities and map them into a common feature space. The specific modality in a multimodal domain depends on the particular format and how the relevant details are embedded in the conceptual design [8]. Modalities such as text, images, and audio typically require the use of specific techniques and fusion processes to systematically encode heterogeneous information, which we refer to in this study as multimodal learning. Meanwhile, multimodal hybrid deep learning is more explicitly defined as the combination of multiple modalities and the use of numerous different architectures (at least two). In particular, a hybrid architecture is designed to integrate homogeneous and adaptive DNNs to improve the discriminative power and versatility of decision systems. In this context, deep learning and machine learning algorithms can be combined in a common framework. Here, deep learning models are used to generate and extract a set of robust visual patterns from distributed datasets, and classical machine learning algorithms are used to generate highly accurate classification models from training datasets. The goal of the hybridization pipeline is to process and combine information from different perspectives to obtain a highly informative and rich representation space from multilevel architectures. Specifically, a deep hybrid architecture can learn the relevant parameters of each of the deep generative and/or discriminative models through a better optimization and regularization process. Hybrid deep learning is an increasingly popular and active research area with great potential. For example, to improve the contextual recognition of dynamic image features, a pair of DNNs can be used to highlight semantic correlations [9].

As the hybridization paradigm is receiving increasing attention from the multimodal learning community, the main goal of this survey is to provide a broader overview of the topic. In particular, this survey will provide a comprehensive review of recent advances in multimodal hybrid deep learning based on systematic hybridization pipelines and their relevance to computer vision-based systems, provide a generic framework for multimodal hybrid analysis, highlight current trends and challenges in multimodal hybrid learning, and identify future research directions.

The remainder of this paper is organized as follows. Section 2 presents the methodology of our study, including our contributions and search strategy. Section 3 summarizes the popular deep learning models and their main characteristics, highlighting the basic concepts relevant to our work. Section 4 introduces the field of multimodal deep learning, including multimodal fusion algorithms. Section 5 describes the theory of hybrid deep learning. Section 6 reviews key application-oriented hybrid architectures based on systematic hybridization of CNNs and other deep networks. Section 7 presents a generic framework for multimodal hybrid learning that integrates standalone networks with both hybrid and multimodal fusion techniques. Section 8 introduces relevant paradigms for multimodal hybrid learning, such as transfer learning and ensemble learning. Section 9 provides an in-depth discussion and critical analysis, focusing on the strengths and limitations of multimodal hybrid deep learning algorithms. Section 10 describes current challenges and future research needs in vision-based multimodal hybrid learning. The last section summarizes the whole paper. The overall structure of the survey is shown in Fig. 1.

## 2. Survey methodology

### 2.1. Contributions

This paper provides a comprehensive review of recent advances in multimodal hybrid deep learning and discusses several related topics, including multimodal representation, multimodal fusion, and hybrid deep learning. In addition to these research areas, this paper focuses on innovative applications of multimodal hybrid learning in computer vision. We also provide a thorough summary of multimodal hybrid techniques, their prospects, trends, and challenges, as well as insights and reinforcements on the main directions of future developments in this area. To the best of our knowledge, there are no recent studies that directly address recent advances in multimodal hybrid deep learning, especially in computer vision. Overall, our survey is similar to the closest studies [7–12], focusing specifically on computer vision applications and discussing recent advances in multimodal and hybrid learning. The main contributions of this study can be summarized as follows:

- In this survey, various algorithms related to multimodal hybrid deep learning for various computer vision tasks and related applications are reviewed. Recent advances in hybrid deep learning with single and multiple sensory modalities are also discussed.
- This paper provides a comprehensive overview of multimodal and hybrid deep learning fusion methods.
- This work proposes a generic framework for multimodal hybrid learning with a focus on the integration of cross-modal information in hybrid designs.
- This study discusses the latest trends, challenges, and multimodal hybrid architectures in computer vision and outlines directions for future research.

### 2.2. Search strategy

The purpose of this narrative survey is to locate relevant research on the development or validation of multimodal hybrid deep learning models using one or more sensory modalities in real-world computer vision applications. The research criteria initially include the terms “hybrid deep learning”, “hybrid multimodal learning”, “hybrid models”, “hybrid machine learning”, and “multimodal hybrid models”, as well as the definitions of “multimodal learning”, “combined models”, and “fusion algorithms”. The focus of this paper is to review recent advances in unimodal and multimodal hybrid deep learning architectures, especially in the area of computer vision. A large number of papers were retrieved using the final search criteria, while a much smaller percentage of papers were relevant to the present survey. In total, 1207 papers were identified using our strategy. The three exclusion criteria were studies not related to artificial intelligence, papers that were not relevant, and papers with insufficient data. After applying the exclusion criteria, 215 references were finally included in the study based on the selection criteria.

### 2.3. Information sources

Most of the papers reviewed in this paper were recently published and presented at high-profile conferences and search platforms such as ICCV, CVPR, ICML, IEEE, Springer, and Elsevier. The main search platforms used in this study are listed in Table 1.

## 3. Fundamentals

This section describes deep learning algorithms in general, from a brief history of deep neural networks to a description of the main categories of features in the literature. These concepts and algorithms are particularly relevant for better understanding the significant impact of hybridization processes in practice.

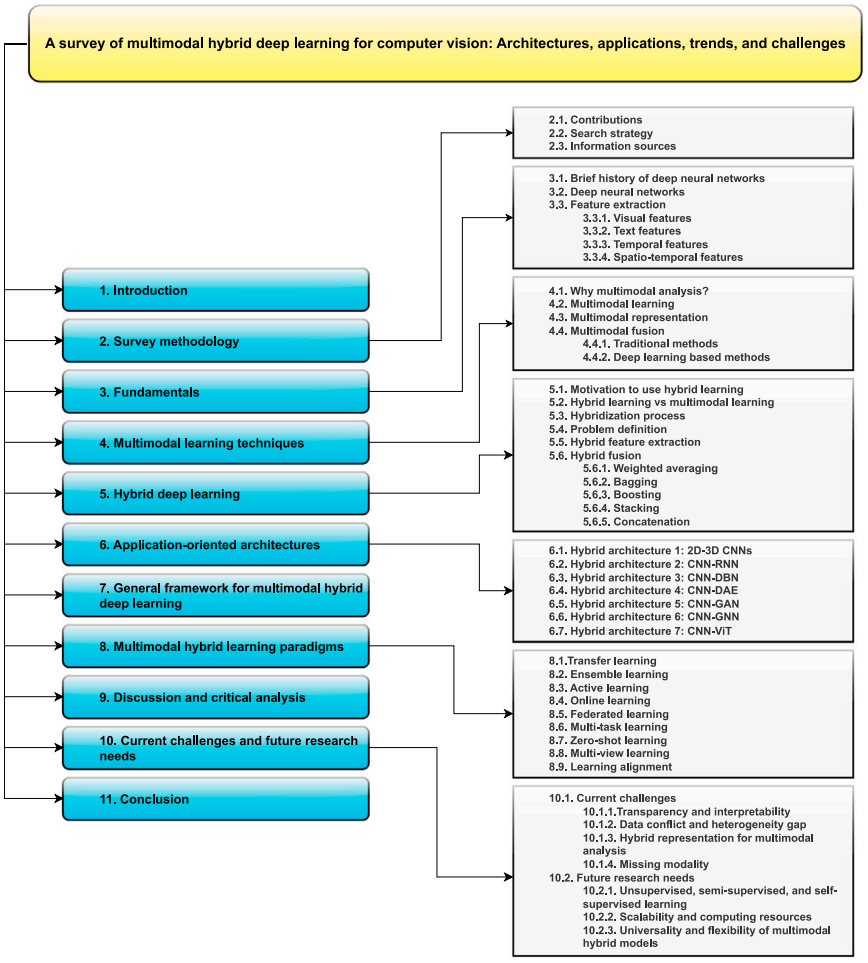


**Fig. 1.** Overall structure of the survey.

**Table 1**
Main search databases used during the review process.

| Search platform | Source |
|---|---|
| Web of Science | http://www.webofknowledge.com/ |
| Springer | http://www.springerlink.com/ |
| Science Direct | http://www.sciencedirect.com/ |
| IEEE explore | http://www.ieeexplore.ieee.org/ |
| Elsevier | http://www.elsevier.com/ |
| arXiv | http://www.arxiv.org/ |
| Wiley online library | http://www.onlinelibrary.wiley.com/ |
| ACM Digital Library | http://dl.acm.org/ |
| PubMed | http://pubmed.ncbi.nlm.nih.gov/ |
| MDPI | http://www.mdpi.com/ |

### 3.1. Brief history of deep neural networks

The concept of neural networks dates back to 1943, after neurophysiologist Warren McCulloch and mathematician Walter Pitts developed an electrical circuit to simulate the workings of formal neurons in the brain. After some advances in hardware technology, Rochester et al. [13] were the first to attempt to modulate neural networks on a computer system. In 1958, Rosenblatt [14] proposed a single-layer network as a simple variant of artificial neural networks, called a perceptron, which showed that a linear stack of artificial neurons could learn from a small set of data samples to perform binary classification tasks.

Later, the community's interest in neural networks was lost for a while after the 1960s, but in the 1980s, Fukushima [15] proposed the first hierarchical model called "Neocognitron", which aimed to recognize global patterns from static data. Specifically, the author used neural networks to show how the pattern recognition mechanism works.

Inspired by the work of Fukushima, a pioneering data scientist, LeCun et al. proposed the first convolutional neural network (CNN) applied to handwritten digit recognition in images [16]. Here, they described how the supervised back-propagation mechanism works and how the generalization capability of the neural network can be improved by stacking multiple convolutional layers within a single architecture called LeNet. Unlike a single-layer perceptron, LeNet is a multi-layer network that uses a gradient-based optimizer to iteratively minimize a function that maps predicted values to real values, called the objective function or loss function. Since then, with the development of hardware infrastructure, the increase in low-cost sensors, and the growth in the volume and velocity of data streams, several approaches have been proposed in recent years to demonstrate that increasing the depth of the network architecture leads to improved performance in feature representation and recognition [2].

### 3.2. Deep neural networks

Historically, shallow networks have been able to learn low-level representations with a limited number of hidden layers. Structurally, a deep model consists of a stack of multiple multi-layer perceptrons [1]. Typically, it contains millions of artificial neurons divided into multiple hidden layers. The input data passes through several layers of processing before producing an intelligent output. The representation of the

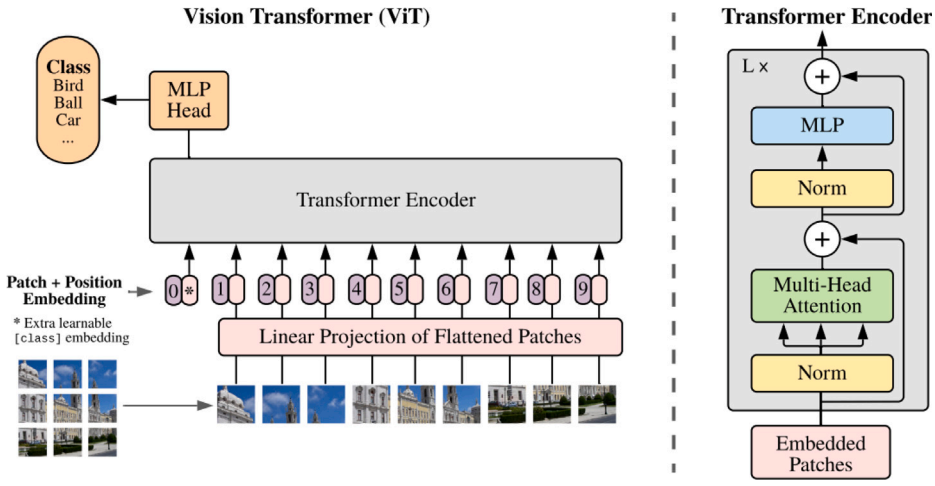


**Fig. 2.** Schematic representation of the ViT architecture [17].

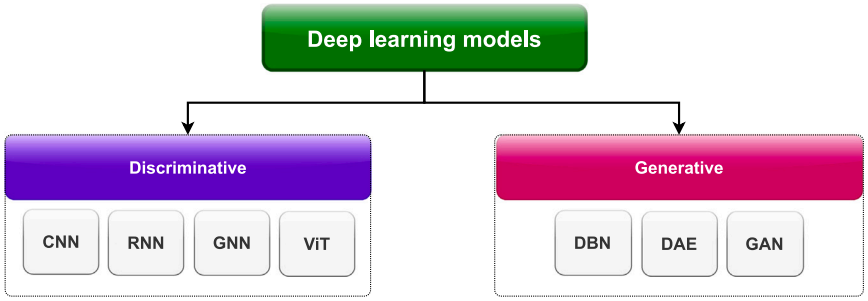


**Fig. 3.** Schematic diagram of deep learning models.

first hidden layer is then used as input to the next processing hierarchy, and so on. In the following, we discuss the most popular DNNs in the recent literature.

- Deep belief networks (DBNs);
- Deep auto-encoders (DAEs);
- Generative adversarial networks (GANs);
- Recurrent neural networks (RNNs);
- Convolutional neural networks (CNNs);
- Graph neural networks (GNNs);
- Vision transformers (ViTs).

Among DNN models, the vision transformer (ViT) [17] has proven to be particularly powerful and popular in recent years. Based entirely on the vanilla transformer [18], ViT is a strong alternative to CNNs, which currently represent the state-of-the-art in computer vision and are used for a variety of image recognition tasks. Structurally, ViT is primarily based on an encoder–decoder structure that allows parallel processing of continuous data streams (see Fig. 2). Unlike conventional CNN architectures, which typically use kernels with local perceptual fields, the ViT model exploits the representation power of attention mechanisms [19] by focusing and integrating salient features in different regions of the input space. To learn more about recent advances in deep learning models in various applications, we strongly encourage readers to read the excellent survey by Dargan et al. [2].

A number of different DNNs and their types, characteristics, and state-of-the-art papers are listed in Table 2. Fig. 3 shows a schematic representation of DNN architectures reported in the literature.

### 3.3. Feature extraction

Deep learning, also known as representation-based learning [1], is a set of modeling techniques that allow a learning system to take raw data as input and identify relevant features for prediction purposes. Unlike machine learning, deep learning techniques are capable of extracting high-quality features across multiple levels of representation and abstraction. Therefore, feature extraction is an important processing step in the overall learning process [43]. In this process, multiple features can be extracted from datasets with different levels of complexity. In the literature, these features can be categorized as visual features, textual features, temporal features, and spatio-temporal features. There are different extraction and representation methods for the different features, which are described in the following subsections.

#### 3.3.1. Visual features

In the field of computer vision and image processing, large-scale visual data such as images and videos can be represented by local and global visual features. Deep learning underlies most current visual feature extraction methods [6]. One of the most common approaches is to use deep CNN models to extract a set of discriminative and abstract spatial features from visual input. Typically, these models are pre-trained on ImageNet-1K (using a transfer learning approach (described in Section 8.1)), and the model parameters can be fine-tuned for a specific task using a particular dataset. However, the final layer of the pre-trained CNN can generate high-dimensional features. The dimensions of these features can be reduced using one of the dimensionality reduction algorithms, principal component analysis (PCA) [44].

#### 3.3.2. Text features

In general, unstructured text data cannot be fed directly into deep learning models. This is because the training algorithms cannot understand the raw text data and its structure without prior processing. Typical text feature extraction involves converting a large amount of text data into numerical values, also called text vectorization in natural language processing (NLP) [45]. Text features are thus simple textual

**Table 2**
DNN models and their associated learning mode and key properties.

| Refs. | DNN | Mode | Key properties |
|---|---|---|---|
| [20] | DBN | Unsupervised learning | Learn the top-down and generative weights<br>Can be applied to labeled data by generating both the label and data distribution |
| [21–24] | DAE | Unsupervised learning | Compress and reconstruct of input representation<br>Output dimension is the same as the input.<br>Reduce the number of training parameters |
| [5,25–27] | GAN | Unsupervised learning<br>Semi-supervised learning<br>Supervised learning | Consists of two networks: generator and discriminator<br>Can enhance image resolution and realism<br>Can handle spatiotemporal streams |
| [28–31] | RNN | Semi-supervised learning | Good for data sequences<br>Have an internal memory (i.e., LSTM, GRU, etc.) |
| [1,32,33] | CNN | Supervised learning | Feed-forward networks<br>Local connection between units<br>Parameters sharing<br>Mainly consists of processing spatial features |
| [34–40] | GNN | Supervised learning<br>Semi-supervised learning<br>Unsupervised learning | Good for irregular structures<br>Can be heterogeneous or homogeneous<br>Can learn inductive models of graphs |
| [17,41,42] | ViT | Self-supervised learning | Can detect and track relationship in sequential data<br>Extremely robust against occlusions, patch perturbations and domain shifts |

descriptions of the input content, containing both implicit and explicit attributes. Therefore, different techniques can be used to extract semantic information from text. For this purpose, pre-trained models (e.g., BERT [46]) are usually developed to capture high-dimensional features.

#### 3.3.3. Temporal features

In deep learning, large amounts of dynamic data such as 3D images, videos, and time series can be used to represent temporal features [47]. Since feature extraction involves temporal visual modalities, RNNs and their variants (i.e., LSTM, GRU, etc.) along with 3D CNNs can be the best deep learning models for this purpose. In recent years, temporal information extraction has become an attractive research area in computer vision and NLP [48]. A major challenge in this field is to identify the various correlations between different signals and temporal representations in multidimensional data.

#### 3.3.4. Spatio-temporal features

Spatio-temporal data acquired from multiple sensors is widely available and can be used for a variety of computer vision tasks [49,50]. Such data can be used for data analysis and prediction to capture both spatial and temporal dimensions. A key factor in improving these models is the ability to extract discriminative features from the spatio-temporal data that contribute significantly to the overall performance of the models. Deep learning models such as CNNs and RNNs have made significant advances in various computer vision tasks due to their powerful hierarchical feature learning capabilities in both spatial and temporal domains. These models are widely used for various data mining tasks such as predictive learning, representation learning, and classification. For example, regarding the potential synergies of combining CNNs and RNNs in the learning process, capturing the local spatial features of CNNs and the temporal features of RNNs are complementary and essential for better performance [51].

## 4. Multimodal learning techniques

### 4.1. Why multimodal analysis?

Recently, the amount of multimedia data has increased dramatically due to the widespread use of low-cost sensors [52]. Still images, video sequences and other visual observations are a large source of data that can be used for multimodal modeling [7]. Therefore, they can be easily used for various multimodal applications such as video synthesis [53] and facial expression recognition [54]. Fig. 4 illustrates a multimodal architecture with three different modalities. By extracting audio-visual attributes from video sequences, common features of the three multimodal learning modalities can be learned [55].

With millions of visual content (videos, images, etc.) collected daily, large databases can be built in multiple formats for efficient learning and rapid analysis. In practice, richer and more informative details can be extracted from multimodal data sources, providing more promising performance than using a single input [7]. Despite the high robustness and accuracy of some unimodal learning systems, many application scenarios still suffer from errors due to noise and missing details [8]. For various reasons, researchers have widely used multimodal learning techniques to improve model performance. Large publicly available multimodal datasets and powerful computers with fast GPUs are some of the reasons.

### 4.2. Multimodal learning

Multimodal learning is one of the hottest research areas in the deep learning community in recent years [7,8]. The goal of multimodal learning is to continuously design and develop robust learning algorithms for data of different modalities, different environments, and different technologies. Multimodal data can be thought of as a combination of different data formats, such as images, text, audio, and time series. More specifically, it is any information received through different perceptual channels, such as sight, hearing, touch, smell, etc. Thus, multimodal learning can be viewed as the simultaneous analysis of heterogeneous data from multiple sources.

Given that different sources of information have different statistical properties, one of the key challenges facing researchers in the deep learning community is discovering meaningful connections across heterogeneous modalities [56]. However, multimodal learning lends itself in many ways as a tool for joint representation across different modalities. In some cases, multimodal learning models can provide a way to compensate for the lack of certain concepts in the observed data [7,8].

### 4.3. Multimodal representation

A multimodal representation can be defined as a complex feature space containing rich information from heterogeneous modalities, or as a set of semantic vectors in a multidimensional space shared by different data sources [7]. In other words, multimodal representation learning takes full advantage of the complementarity and synergy of different modalities, while reducing both inter- and intra-modality redundancy, to learn a high-quality feature representation, called a joint representation. The joint representations of multiple modalities

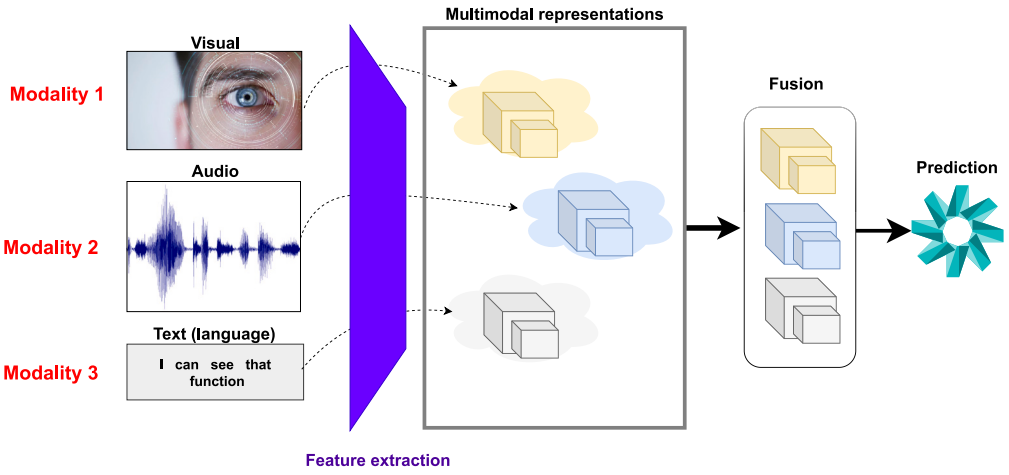


**Fig. 4.** Schematic illustration of a multimodal learning scheme involving three different modalities (visual, acoustic, and textual).

are then fused into a unified multimodal vector space using either traditional or deep learning-based methods.

In multimodal theory, perceptual modalities refer to ways of systematically perceiving, modeling, and interpreting observed data [57]. On the one hand, unimodal data representation can be defined as a way to learn data from a single modality (e.g., images). In supervised learning, the unimodal dataset $D_m = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $(x_n, y_n)$ consists of the input samples $x_i$ and their corresponding ground truth labels $y_i$. Here, $n$ is the number of samples in the dataset. On the other hand, multimodal data representation involves training models from multiple data sources (e.g., images and text) of different modalities (at least two modalities).

### 4.4. Multimodal fusion

Since the ability to represent knowledge at multiple levels of abstraction is one of the most critical challenges in multimodal learning, various fusion mechanisms can be used to fuse sensory stimuli from two or more modalities into a common space [7,10]. This subsection introduces some of the most common multimodal fusion methods, ranging from traditional methods to deep learning-based approaches.

#### 4.4.1. Traditional methods

From an architectural point of view, multimodal fusion is considered an essential step in solving multimodal tasks. Common conventional fusion strategies include early fusion, late fusion, and intermediate fusion, all of which can be applied for this purpose.

- **Early fusion:** fusion of low-level features extracted from each modality before making predictions.
- **Late fusion (also called decision-based fusion):** fusion of features extracted independently from different modalities after predictions are made.
- **Intermediate fusion:** combination of multimodal features from early and late fusion before decision making.

In practice, when dealing with heterogeneous data, early fusion can extract a large amount of information [11]. However, there is often a lack of consistency because the extracted features are very sensitive to variations in modality and visual appearance [7]. A single large feature vector generated during fusion can also be a source of prediction error.

#### 4.4.2. Deep learning based methods

According to the current literature, deep learning-based solutions are the most commonly used for advanced multimodal fusion. In [7], the authors described each method and its advantages and disadvantages in detail. According to [7,10,12], deep belief networks, stacked autoencoders, convolutional neural networks, and recurrent networks are widely used methods with good performance.

However, despite the promising results of deep learning methods in multimodal data fusion, there are still some major drawbacks. First, the computational cost of deep learning models may be too high, especially when dealing with large feature spaces. Second, cross-modality learning may provide limited and noisy information. In addition, multimodal data is usually derived from dynamic environments, requiring flexible models that can quickly adapt to random variations in the data.

## 5. Hybrid deep learning

### 5.1. Motivation to use hybrid learning

In recent years, the artificial intelligence community has increasingly focused on hybrid learning, driven by the development of low-cost sensors and a new generation of artificial neural networks [33]. As the name implies, hybrid learning refers to a coherent design approach that integrates the power of multiple architectures. In general, hybrid information modeling techniques that take advantage of high-speed hardware such as CPUs and GPUs, as well as advanced feature extraction techniques using deep learning, are also leading the way. In the spirit of computer vision, hybrid models using both conventional and advanced fusion techniques are also being developed [58].

Robust algorithms, including generic artificial neural networks (deep and/or shallow networks), can be developed to improve the predictive power of decision systems [59]. In this case, a mixture of deep and/or machine learning methods can be used to build a generic common architecture, called a hybrid architecture. Based on this, at least two basic generic models are then used to extract and learn a set of high-quality patterns from a heterogeneous dataset.

The main reason for the development of hybrid architectures is their ability to extract rich and robust contextual information embedded in multidimensional data volumes, thus providing more optimistic results than single architectures. Although the robustness and effectiveness of most standalone deep learning systems are well documented in the literature, they still suffer from several shortcomings in many use cases, such as the ability to accurately generalize at different levels of abstraction, and inaccuracies due to potential noise and lack of intrinsic detail [60]. In principle, systematic learning using hybrid architectures can facilitate the joint representation of multiple model features and reduce the complexity of the optimization process.

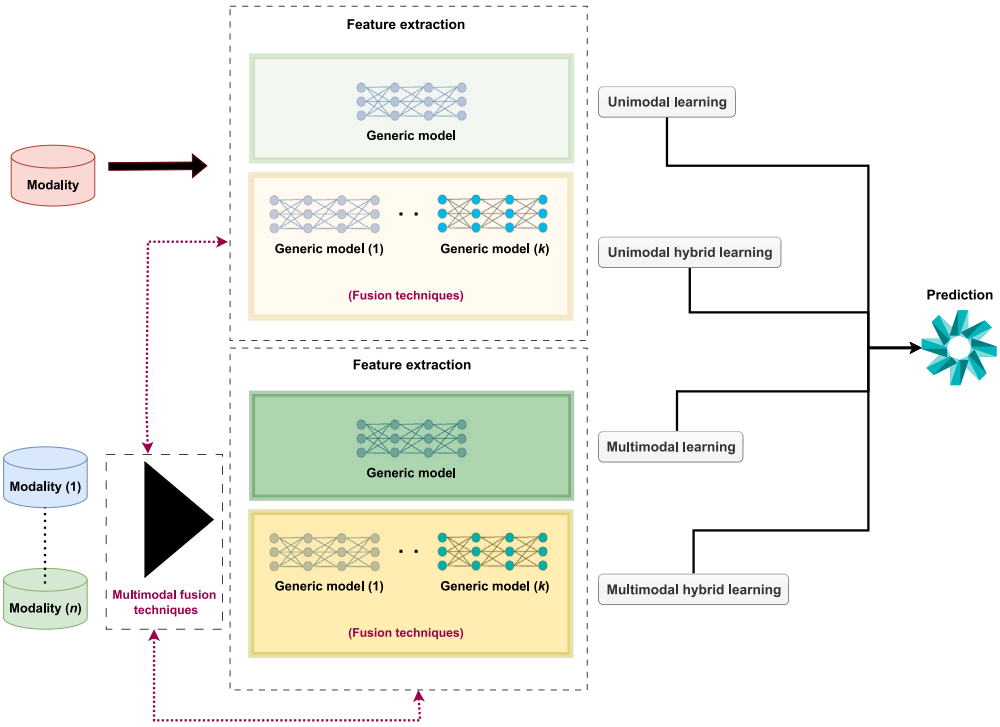


**Fig. 5.** Schematic of the difference between hybrid and multimodal learning pipelines.

### 5.2. Hybrid learning vs multimodal learning

Given the remarkable success of deep learning models in most computer vision tasks, the growth of hybrid modeling theory is not surprising. However, current trends indicate that hybrid models have grown incrementally over the past few years, largely because existing traditional deep learning models could not meet the requirements of the target task [9,61–63]. In fact, hybrid learning modeling has focused on developing new models that provide better predictive performance. Furthermore, there is a clear desire in the deep learning community to not only improve the performance of a standalone deep learning model, but also to improve the performance of common computer vision tasks from different perspectives. In many real-world applications, hybrid deep learning models that use one or more input modalities (data sources) have the ability to perform consistently well [64–67].

When two or more generic neural architectures are combined, it is often referred to as a fusion operation (Fig. 5). More specifically, a fusion learning scheme takes full advantage of multiple neural architectures to handle different types of input signals (e.g., images, video, etc.). Thus, multimodal fusion techniques can be used to combine these architectures into a single unified structure to exploit the representational power of deep networks.

### 5.3. Hybridization process

When two or more generic deep learning architectures are combined (i.e., discriminative and/or generative), often through a fusion mechanism, or when deep learning is cascaded with machine learning for predictive purposes, the resulting models are referred to as hybrid deep learning models. In contrast, standalone deep learning relies on a single deep learning model. Technically, a hybridization process consists of combining two or more different model architectures. More specifically, the hybrid deep learning paradigm studied in this research can be defined as the processing of different types of input data modalities across multiple architectures.

In this subsection, we briefly describe the general steps of the hybridization pipeline. The process can be divided into the following 4 steps:

- **Step 1:** Split the input dataset into two separate subsets: the training and validation set and the test set.
- **Step 2:** Feed the training and validation set into the hybrid model, where two or more deep networks (e.g., CNNs, RNNs, etc.) are combined and unified in a common contextual space.
- **Step 3:** Train the designed hybrid model by extracting highly informative and discriminative features, as well as short- and long-range correlations along with data points.
- **Step 4:** The constructed hybrid model is then validated by testing its effectiveness on the test set. Predictive performance is measured using standard evaluation metrics, and finally, the hybrid model is benchmarked using the results obtained.

### 5.4. Problem definition

In this subsection, we formulate how to combine two or more DNN networks and exploit their semantic synergy to solve core computer vision problems such as object detection, semantic segmentation, and image classification. In this research, DNNs, which are gaining wide popularity in the field of machine learning, are used as the main learning approach. We recall that the hybridization process aims at learning highly informative and rich patterns from multidimensional datasets using both discriminative and/or generative network modules simultaneously.

Let $D$ be an observed dataset of $N$ samples, $X$ be a set of feature vectors, and $y$ be a scalar representing the corresponding class labels, then $D = \{X, y\} = \{(\mathrm{x_n}, \mathrm{y_n})_{n=1}^{N}\}$. Indeed, in most classical applications, deep learning models consider conditional distributions of the form $p(y|x)$. However, in many challenging scenarios, estimating the conditional distribution alone is not sufficient, as the network may miss contextual information or detect artifacts before feeding the conditional distribution into the predictive model for inference. Therefore, it is crucial to adaptively combine the conditional model $p(y|x)$ and the generative model $p(x)$ into a joint model, called a hybrid model. Formally, a joint likelihood distribution of the form $p(y, X)$ can be modeled to generate

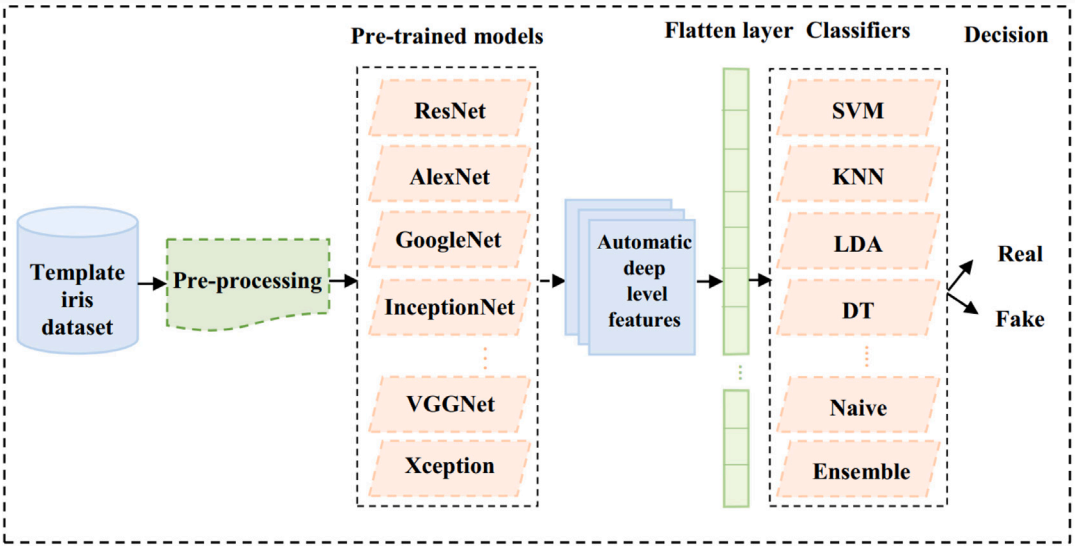


**Fig. 6.** Schematic illustration of a generic hybrid feature extraction scheme for iris spoof detection [70].

more representative and informative features of each samples as:

$$p(y, X; \Theta) = \prod_{n=1}^{N} p(y_n|x_n; \Theta)p(x_n; \Theta), \quad (1)$$

where $\Theta$ refers to the common parameters across models.

### 5.5. Hybrid feature extraction

In recent years, there have been many attempts to represent different network outputs as a common hybrid space using one or more sensory modalities [7,68]. However, the development of robust and efficient feature extraction methods is one of the major challenges faced by deep learning researchers. In particular, the design of efficient and consistent hybrid feature extraction that can achieve high performance in various computer vision tasks remains an open question. With the active development of next-generation fusion technologies, hybrid feature extraction methods are proving to be more effective and powerful than traditional methods. In many cases, hybrid models are developed by combining deep learning with other popular existing methods (e.g., hand-crafted methods, transfer learning, etc.). As a result, hybrid architectures have more powerful feature extraction capabilities than standalone pipelines. For example, it is common to combine CNNs and LSTMs to extract both spatial and temporal features from multidimensional data [69]. The combined features of the two network outputs are called hybrid features.

Fig. 6 shows a schematic illustration of a generic hybrid feature extraction process for iris spoof detection. Here, a set of pre-trained CNN models is combined to extract highly discriminative features, followed by conventional classifiers such as SVM, KNN, and DT.

### 5.6. Hybrid fusion

As mentioned above, in contrast to multimodal fusion, which attempts to merge multiple data modalities, the hybridization process can be viewed as a fusion process in which two or more standalone architectures are combined into a unified common structure. However, choosing the best fusion technique for creating a multimodal hybrid model remains critical for optimal model performance.

This subsection describes how to create a new unified joint model from the outputs of two or more separate models for a single modality. The term "hybrid learning" was first used in the literature by Szegedy et al. [33]. Prior to this, the process of hybridization was only viable when the deep learning paradigm was combined with machine learning. However, since 2016, this research axis has matured due to the huge increase in data and the need to build a new generation of architectures aimed at balancing accuracy and computational complexity [9].

In unimodal learning, ensemble learning (see Section 8.2) is the key process of training and combining a set of individual learners through a number of strategies [71]. Its main purpose is to improve the prediction and classification capabilities of deep learning models. Ensemble learning can also be used to develop new models by mixing multi-level features from different networks. In fact, building new models offers many advantages in terms of accuracy and efficiency (e.g., fewer computational resources, greater capability, etc.) compared to training standalone models from scratch. In ensemble learning, combining the results of the basic classifiers into a single output result is called output fusion. Voting and meta-learning methods can be applied to both parallel and sequential classifiers [72]. In particular, voting methods are often used to improve prediction performance in classification and regression problems. In other words, the goal is to integrate the results of multiple machine learning models by voting. In regression analysis, a voting mechanism calculates the average of several base regressors. In classification, a hard voting approach predicts the class label with the highest number of votes by summing the votes for different class labels provided by other models. In contrast, a soft voting scheme predicts the class label with the highest overall probability by summing the predicted probabilities of the different class labels.

In the following, we will discuss different ensemble-based hybrid fusion methods and their advantages and disadvantages.

#### 5.6.1. Weighted averaging

Weighted averaging is a common machine learning technique that combines predictions from different models and weights them according to the capabilities of each model [73,74]. More specifically, it can be thought of as an extended version of averaging, where multiple predictions are made for a single data point and the average of all predictions is used as the final prediction. In this method, the average of the results from two standalone models is used as the new target model. This is considered the simplest way to merge two different models. The weighting of the different models can be based on the performance of the models or on the amount of training for each model.

However, some high-performing models may contribute more to the integrated predictions, while some lower-performing models may contribute but have little impact on the final predictions. In other words, this strategy relies on each ensemble member contributing equally to the prediction.

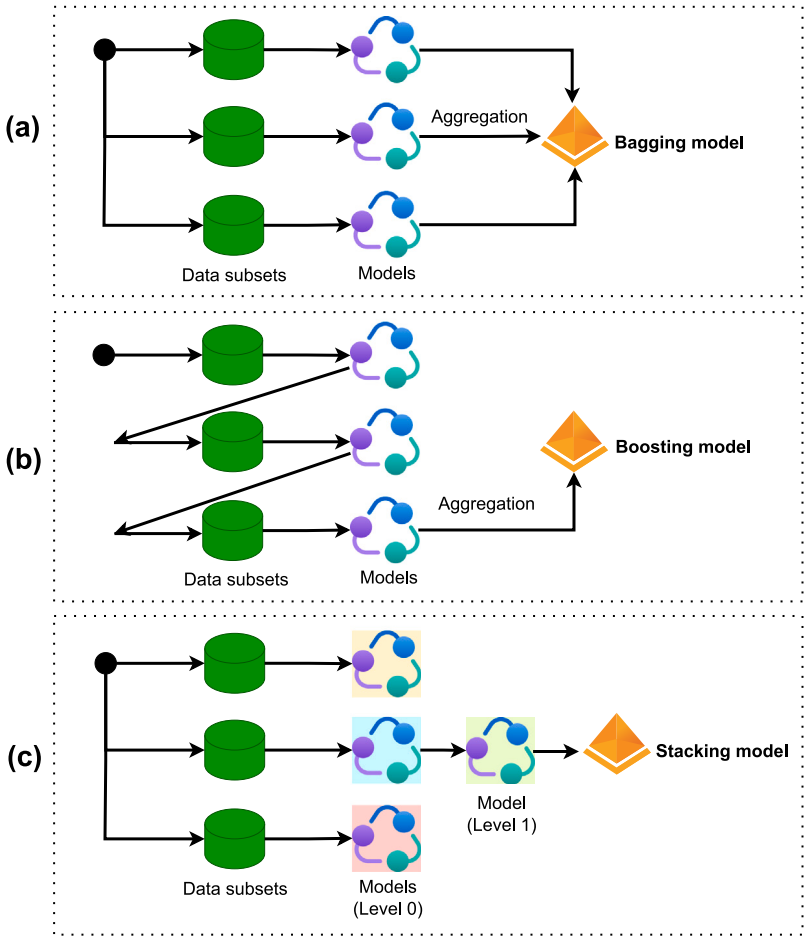


**Fig. 7.** Illustration of bagging (a), boosting (b), and stacking (c) techniques.

#### 5.6.2. Bagging

A bagging algorithm (often referred to as bootstrap aggregation) is a machine learning technique commonly used to reduce the variance of noisy observations in data [75]. Bagging is an algorithm that is completely tailored to a specific dataset. As such, it creates multiple smaller subsets of the original data. In particular, the purpose of bagging algorithms is to create more diverse predictive models by fitting models to trained probability distributions. This process involves multiple iterations of the standalone model (Fig. 7(a)). The different iterations are trained on different subsets with different degrees of improvement.

Although bagging has many advantages (e.g., reducing variance, avoiding overfitting, etc.), it also has some disadvantages. For example, if the model is not well modeled, it can lead to high bias and thus underfitting. Also, since multiple models need to be involved, this method can be very computationally intensive and may not be suitable for a variety of scenarios.

#### 5.6.3. Boosting

A class of machine learning methods known as boosting is based on the idea that new models should attempt to accurately detect previous errors [137]. Many of the weak models (base models) used in boosting methods are highly adaptive, continuously fitting by giving more weight to observations in the dataset that were misinterpreted by the previous model (Fig. 7(b)). In many cases, this approach is effective in minimizing model bias and variability. It also allows for efficient handling of missing data.

While this approach has many advantages, its real-time implementation is inherently complex, which can directly affect the behavior of the training algorithm.

#### 5.6.4. Stacking

Stacking (also known as stacking generalization) is a common machine learning technique that combines relevant knowledge from multiple standalone models to build a new model, called a metamodel [72, 138]. In general, this method considers different weak learners, trains them simultaneously, integrates them, and optimizes the hyperparameters of the generated metamodel (Fig. 7(c)). The idea behind this approach is to combine a set of weak learners from which stronger learners (i.e., metamodels) are created for better performance. In ensemble learning theory, weak learners can be used as building blocks to create more complex models. In some cases, these weak models lack the robustness to perform well on their own due to high levels of bias and variance [139].

While the stacking method has many advantages (e.g., better predictive performance, more accurate data, better generalization, etc.), its main drawback is overfitting. That is, when too many predictors are used to make predictions about the same target, these predictors are subsequently fitted together to produce the final prediction. In addition, multilevel stacking models are expensive.

#### 5.6.5. Concatenation

Concatenation is one of the most widely used feature fusion techniques in machine learning that systematically combines multiple inputs into a single input [140,141]. The core idea of concatenation is to combine two or more output tensors (usually from different network layers) along the channel dimension. More specifically, two or more output tensors from different network layers can be processed with selected multivariate functions to integrate the outputs of multiple models. In this way, multiple data sources can be integrated into a common space.

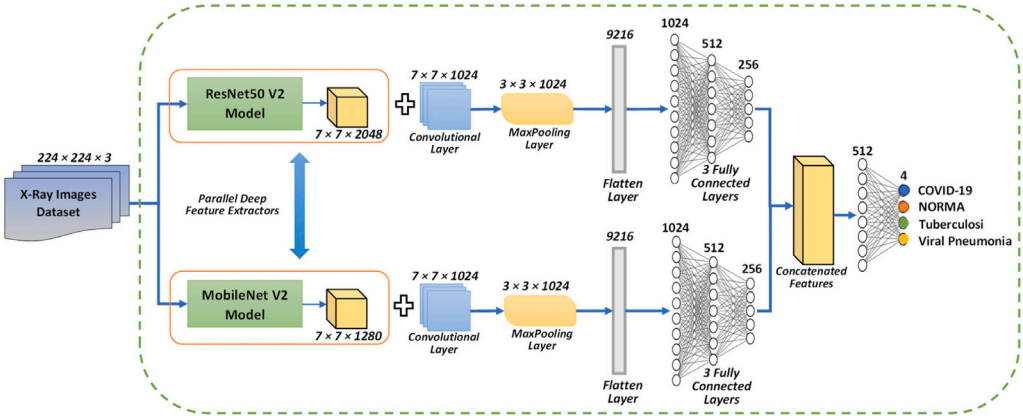


**Fig. 8.** Illustration of a hybrid architecture that combines two pre-trained models using a concatenation process [76].

In practice, concatenation can be thought of as a stacking process. To illustrate, consider a CNN model with 128 channels in each of the two hidden layers. These two layers can be combined channel-wise to obtain an output with 256 channels. The resulting tensor can be subjected to all the basic operations (such as addition, subtraction, etc.) like any other tensor. However, the fact that concatenation expands the feature space by integrating both high- and low-level features can be used to explain why the results of the concatenated output are improved. Subsequent convolution procedures can therefore incorporate additional features based on both high- and low-level patterns.

Fig. 8 shows an example of a robust hybrid architecture for COVID-19 screening, using two pre-trained CNN models as feature extractors, followed by a concatenation process to perform their fusion. Formally, the concatenation of activations (denoted by $\oplus$) obtained from different CNN layers can be formulated as follows:

$$O = X_1 \oplus X_2 \tag{2}$$

where $O$ ($O \in \mathrm{R}^{H\times W\times(C_1+C_2)}$) is the combined features, $X_1$ ($X_1 \in \mathrm{R}^{H\times W\times C_1}$) is the lower-level representation, and $X_2$ ($X_2 \in \mathrm{R}^{H\times W\times C_2}$) is the upper-level representation. Here, $C_i$, $H$, and $W$ represent the number of channels, height, and width, respectively.

While concatenation-based approaches have many advantages (e.g., they can exploit complementary and joint features from different inputs, etc.), they also have some disadvantages. For example, feature fusion of multiple models can result in very high data dimensionality, which can lead to overfitting and may miss some semantic details, affecting the performance of the resulting fused model.

## 6. Application-oriented architectures

In recent years, many researchers have widely applied CNNs as a powerful tool to extract very rich and informative features for many challenging tasks in computer vision [4,142]. To improve the overall performance of prediction systems, several studies have been conducted to adaptively hybridize CNNs with other deep models and unify them into novel hybrid architectures. Typically, multimodal hybrid learning models are application-oriented and attempt to strike a better balance between accuracy and efficiency. In this section, we review recent advances in the hybridization paradigm, with a particular focus on combining CNNs with other deep learning models (e.g., RNNs, DBNs, etc.).

Tables 3 and 4 provide a summary of the reviewed multimodal hybrid models and their associated applications, datasets, and best performance (unimodal- and multimodal-based methods).

### 6.1. Hybrid architecture 1: 2D-3D CNNs

Given the great success of deep learning in recent years, CNN-based methods have shown outstanding performance in various computer vision and video analysis tasks, such as image classification [143], object detection [144], semantic segmentation [145], etc. In video analysis scenarios, both spatial and temporal dimensions should be considered. Typically, a video sequence consists of chronologically ordered frames, each of which represents the spatio-temporal structure of a particular context in which it occurs. However, most existing deep learning methods are based on 2D CNNs, which focus only on spatial patterns and largely ignore the contextual relationships between spatial and temporal cues of objects of interest in dynamic data [4,142,146].

To further improve the overall performance of predictive models, it is essential to leverage both temporal and contextual information in each frame. Recently, pioneering approaches have been developed to successfully model spatio-temporal and contextual information from high-dimensional data space. These methods are based on hierarchical networks, called 3D CNN (a 3D variant of standard CNN), which consists of a series of 3D convolutional operations performed in spatial and temporal feature dimensions [147]. A strong advantage of such models is that they can generate richer and more informative feature maps from 3D data cubes. However, the main disadvantage of these models is that they require more training instances, and training deeper 3D CNN models is very time-consuming in many practical scenarios, especially in real-time tasks.

**Monomodal based.** In recent years, many deep learning researchers have proposed to mix 2D CNN and its 3D counterpart (3D CNN) into a hybrid architecture to exploit their potential synergies and improve performance in terms of accuracy and efficiency [77–81]. For example, the authors in [77,78] proposed robust hybrid CNN architectures (hybrid 2D/3D CNNs) based on transfer learning paradigms to solve some challenging tasks in computer vision, including image classification and semantic segmentation. Similarly, a hybrid 2D-3D CNN model for hyperspectral image classification has been proposed by the authors of [79,81]. Specifically, Roy et al. [79] used 2D and 3D CNNs to learn both abstract spatial and spectral feature representations from remotely sensed images. Yang et al. [81] leveraged the generalization capabilities of 2D and 3D CNNs to simultaneously extract and learn spectral and spatial structures from hyperspectral images. To improve the detection and quantification performance of cardiovascular diseases, Chang et al. [80] proposed a deep learning-based medical diagnostic tool that incorporates 2D and 3D CNNs in an integrated architecture.

**Multimodal based.** Recent research on multimodal computer-aided diagnosis has shown the importance of implementing efficient automatic diagnostic models when using 3D CNNs. To this end, Dai et al. [111]

**Table 3**
A summary of unimodal hybrid models and related applications, datasets, and optimal performance.

| N° | Architecture | Refs. | Applications/ Tasks | Datasets | Performances |
|---|---|---|---|---|---|
| 1 | 2D-3D CNNs | [77] | Image classification | A collected dataset | ACC: 96.91% |
| | | [78] | Image classification | GTSRB | ACC: 99.28% |
| | | [78] | Semantic segmentation | KITTI | MaxF1: 95.57% |
| | | [79] | Image classification | Indian Pines | ACC: 99.81% |
| | | | | Salinas Scene | ACC: 100% |
| | | [80] | Image recognition | Training set: 512,598 images; 901/8.1% hemorrhages<br>Testing set: 23,668 images; 82/12% hemorrhages | ACC: 97% |
| | | [81] | Image classification | Indian Pines Scene | ACC: 97.31% |
| | | | | Botswana Scene | ACC: 99.79% |
| | | | | Kennedy Space Center | ACC: 98.92% |
| | | [82] | Activity recognition | A collected activity dataset | ACC: 83.45% |
| | | [83] | Activity recognition | UCI-WIDSM | ACC: 98.3% |
| | | [84] | Activity recognition | iSPL | ACC: 99% |
| | | | | UCI HAR | ACC: 92% |
| 2 | CNN-RNN | [85] | Sentiment analysis | Kindle | ACC: 93.40% |
| | | | | App | ACC: 92.18% |
| | | | | Movie | ACC: 90.55% |
| | | | | Electronics | ACC: 90.65% |
| | | | | CD | ACC: 88.70% |
| | | | | Sentiment140 | ACC: 81.82% |
| | | | | Airline Twitter | ACC: 92.75% |
| | | [86] | Sentiment analysis | A collected corpus | ACC: 90.75% |
| | | [87] | Sentiment analysis | IMDB Movie Reviews | F-measure: 88% |
| | | | | Amazon Movie Review | F-measure: 86% |
| | | [88] | Image classification | CIFAR-100 | ACC: 83.21% |
| | | | | A subset of ImageNet 2010 | ACC: 90.69% |
| | | [89] | Image classification | A collected dataset | ACC: 90.79% |
| | | [90] | Image classification | A collected dataset | ACC: 99.4% |
| | | [91] | Caption description and generation | COCO Captioning | BLEU: 85.2% |
| | | [92] | Image captioning | MS COCO 2014 | BLEU: 32.6% |
| | | [93] | Fault diagnosis | Rolling bearing datasets | ACC: up to 100% |
| 3 | CNN-DBN | [94] | Fault diagnosis | A high-dimensional fault sample dataset | ACC: 99% |
| | | [95] | Music recognition and classification | A generated dataset | ACC: 98.4% |
| 4 | CNN-DAE | [96] | Fault diagnosis | A collected dataset | ACC: 92.2% |
| | | [97] | Malware detection | A collected dataset | ACC: 99.8% |
| | | | | MIT-BIH arrhythmia | ACC: 98.4% |
| | | [98] | Heartbeats classification | European ST-T | ACC: 94.1% |
| | | | | MIT-BIH ST change | ACC: 94.3% |
| 5 | CNN-GAN | [99] | Image denoising | STL-10 (train set) + SET5 (test set) | PSNR: 27.671 |
| | | [100] | Emotion recognition | DEAP | ACC: 64.81% |
| | | [101] | Network traffic generation | A collected dataset | Success rate: 99% |
| | | [102] | Fault diagnosis | CWRU bearing fault diagnosis | ACC: 100% |
| | | [103] | Gesture recognition | A collected dataset | ACC: 92.7% |
| | | [104] | Design of microstructural materials | A collected dataset | ACC: 90% |
| 6 | CNN-GNN | [105] | Crowd counting | ShanghaiTech Part A | MSE: 94.5 |
| | | | | ShanghaiTech Part B | MSE: 12.7 |
| | | | | UCF_CC_50 | MSE: 270.1 |
| | | | | UCF_QNRF | MSE: 185.3 |
| | | [106] | Image classification | UCM multi-label | F1-Score: 86.39% |
| | | | | AID multi-label | F1-Score: 88.64% |
| | | [107] | Image classification and segmentation | MS COCO 2017 | mAP: 41.6% |
| | | | | ImageNet-1K | ACC: 82.6% |
| | | [108] | Image classification | RSNA | ACC: 98.04% |
| 7 | CNN-ViT | [109] | Image segmentation | A collected dataset (A) | ACC: 94.64% |
| | | | | A collected dataset (B) | ACC: 95.20% |
| | | [110] | Image classification | NWPU-R45 | ACC: 95.85% |
| | | | | BIT-AFGR50 | ACC: 97.27% |

ACC: Accuracy; MaxF1/F-measure/F1-Score: Harmonic mean of precision and recall; BLEU: BiLingual evaluation understudy; PSNR: Peak signal-to-noise ratio; MSE: Mean squared error; mAP: Mean average precision.

proposed a novel multimodal hybrid dimensional network, called MM-Net, for 3D medical image classification. The proposed hybrid model combines a 2D CNN with 3D convolutional layers to generate richer and more informative representations while reducing the complexity of the fusion process. To improve the system's ability to accurately interpret a variety of challenging dynamic scenes, Vynokurova et al. [112] proposed a hybrid multidimensional CNN for human gesture recognition. Their proposed system uses a mid-level feature fusion scheme that incorporates 2D and 3D CNNs of each modality. Recently, Mocanu et al. [113] developed a multimodal emotion recognition model that combines 2D and 3D CNNs in a unified space. In particular, the model aims to capture intra-modal features and effectively detect salient correlations across modalities. To take full advantage of transfer learning techniques, the authors of [114] proposed a hybrid 2D/3D CNN for industrial parts classification. Here, the pre-trained parameters of the 2D CNNs are first transferred to the 3D variant. In particular, the 2D CNNs are used as robust feature extractors and as a starting point for the weight initialization of the 3D side.

**Table 4**
A summary of multimodal hybrid models and related modalities, applications, datasets, and optimal performance.

| N° | Architecture | Refs. | Applications/ Tasks | Modalities/ data sources | Datasets | Performances |
|---|---|---|---|---|---|---|
| 1 | 2D-3D CNNs | [111] | Medical image recognition | MRI modalities (T1 and T2) | PGT | ACC: 90% |
| | | | | | MRNet | ACC: 93.7% |
| | | | | | PROSTATEx | ACC: 82% |
| | | [112] | Gesture recognition | Visual + Acoustic | LRW | ACC: 86% |
| | | [113] | Emotion recognition | Visual + Acoustic | RAVDESS | ACC: 89.25% |
| | | | | | CREMA-D | ACC: 84.57% |
| | | [114] | Industrial parts classification | RGB + D | A generated dataset | ACC: 92.17% |
| | | [115] | Human activity recognition | Multimodal sensors (e.g., accelerometer, gyroscope, etc.) | PAMAP2 | ACC: 81% |
| | | | | | A collected dataset (SCD) | ACC: 80% |
| 2 | CNN-RNN | [116] | Video classification | Visual + Acoustic + Opt. flow | UCF-101 | ACC: 93.1% |
| | | | | | CCV | mAP: 84.5% |
| | | [117] | Emotion recognition | Visual + Acoustic | AFEW 6.0 | ACC: 59.02% |
| | | [118] | Emotion recognition | Visual + Acoustic | MOSI | ACC: 87.56% |
| | | | | | MELD | ACC: 90.06% |
| | | [119] | Emotion recognition | Visual + Acoustic | AFEW 5.0 | ACC: 52.875% |
| | | | | | RML | ACC: 80.36% |
| | | [120] | Emotion recognition | Visual + Acoustic | eNTERFACE05 | ACC: 85.97% |
| | | | | | BAUM-1s | ACC: 54.57% |
| 3 | CNN-DBN | [121] | RNA-binding protein recognition | Multi-resource data (e.g., sequence,structure, etc.) | CLIP-seq datasets | AUC: 91% |
| | | | | | BAUM-1s | ACC: 55.58% |
| | | [122] | Facial expression recognition | RGB + Opt. flow | RML | ACC: 73.73% |
| | | | | | MMI | ACC: 71.43% |
| | | [123] | Emotion recognition | Visual + Acoustic | RECOLA | RMSE: 0.510 |
| 4 | CNN-DAE | [124] | Emotion recognition | Visual + Acoustic + Textual | MOSEI | ACC: 77.2% |
| | | | | | IEMOCAP | ACC: 76.7% |
| | | [125] | Cross-media retrieval and image classification | Visual + Textual | MIRFlickr | mAP: 0.1913 |
| | | | | | NUS-WIDE | mAP: 0.1904 |
| | | [126] | Alzheimer's disease diagnosis | MRI + PET | ADNI | ACC: 98.24% |
| | | [127] | Language understanding | RGB + D + Textual (instruction and context) | A generated dataset (linguistic and visual inputs) | ACC: 87.5% |
| 5 | CNN-GAN | [128] | Semantic segmentation | Hyperspectral + Multispectral | PAVIA | mIoU: 75.92% |
| | | | | | LCZ | mIoU: 51.01% |
| | | [129] | Image captioning | Visual + Textual (captions) | FlickrStyle10K | |
| | | | | | SentiCap | ACC: 97.8% |
| | | | | | COCO | |
| 6 | CNN-GNN | [130] | Micro-video recommendation | Visual + Acoustic + Textual | Tiktok | NDCG: 0.3423 |
| | | | | | Kwai | NDCG: 0.2298 |
| | | | | | MovieLens | NDCG: 0.3062 |
| | | [131] | Visual question answering | Visual + Textual | TextVQA | ACC: 31.21% |
| | | | | | ST-VQA | ACC: 16% |
| | | [132] | Multimedia content classification | Visual + Textual + User | A collected dataset (Stormfront) | ACC: 33.44% |
| 7 | CNN-ViT | [133] | Multimodal image fusion | Visual + Infrared (VI-IR) MRI + PET (Medical) | TNO | PSNR: 56.3492 (Medical) PSNR: 59.1492 (VI-IR) |
| | | | | | VIFB | PSNR: 58.4304 |
| | | [134] | Semantic segmentation | RGB + DSM | Potsdam | mIoU: 72.34 |
| | | [135] | Medical image segmentation | MRI modalities (T1 and T2) | MRBrainS | DSC: 83.47 |
| | | | | | iSEG-2017 | DSC: 87.16 |
| | | [136] | Multimodal medical image fusion | Multimodal brain images (e.g., CT, MR, T1, T2, etc.) | A collected dataset (Harvard Medical School) | VIFF: 0.6226 |

ACC: Accuracy; mIoU: Mean intersection over union; NDCG: Normalized discounted cumulative gain; mAP: Mean average precision; RMSE: Root mean squared error; PSNR: Peak signal-to-noise ratio; DSC: Dice coefficient; VIFF: Visual information fusion fidelity.

### 6.2. Hybrid architecture 2: CNN-RNN

CNN is a class of discriminative models that exploit the spatial structures of visual data streams and have proven effective in most computer vision tasks such as image classification, semantic segmentation, object detection, etc. Compared to CNN, RNN networks are better able to model sequential data and time series, making them more suitable for representing the temporal dependencies of complex tasks such as speech recognition and machine translation. In other words, CNN is a powerful tool for extracting visual features from static data such as images, depth maps, etc.; however, it is not suitable for considering correlations in spatio-temporal information, which is the strong point of RNN. Therefore, attempts have been made to combine the architectures of CNNs and RNNs for different purposes. More specifically, CNN uses convolutional kernels to transform spatial data, while RNN predicts the next state in the observed event sequence based on previous hidden states.

**Monomodal based.** Recently, several approaches have been proposed to exploit CNN and RNN networks and integrate them into a unified framework. In particular, CNNs and RNNs have been effectively integrated and used in several real-world applications, such as activity recognition [82–84], sentiment analysis [85–87], image classification [88–90], caption generation and description [91,92], and so on. For example, Vahora et al. [82] proposed a CNN-RNN hybrid model for group activity recognition. Here, CNNs are used to capture human activity features from a given video sequence, which are then aggregated and fed to an RNN component to capture spatio-temporal flow changes at the group level. In [85], Basiri et al. proposed a sentiment analysis scheme that combines bidirectional LSTM and GRU layers to capture bidirectional temporal information flow and uses CNN

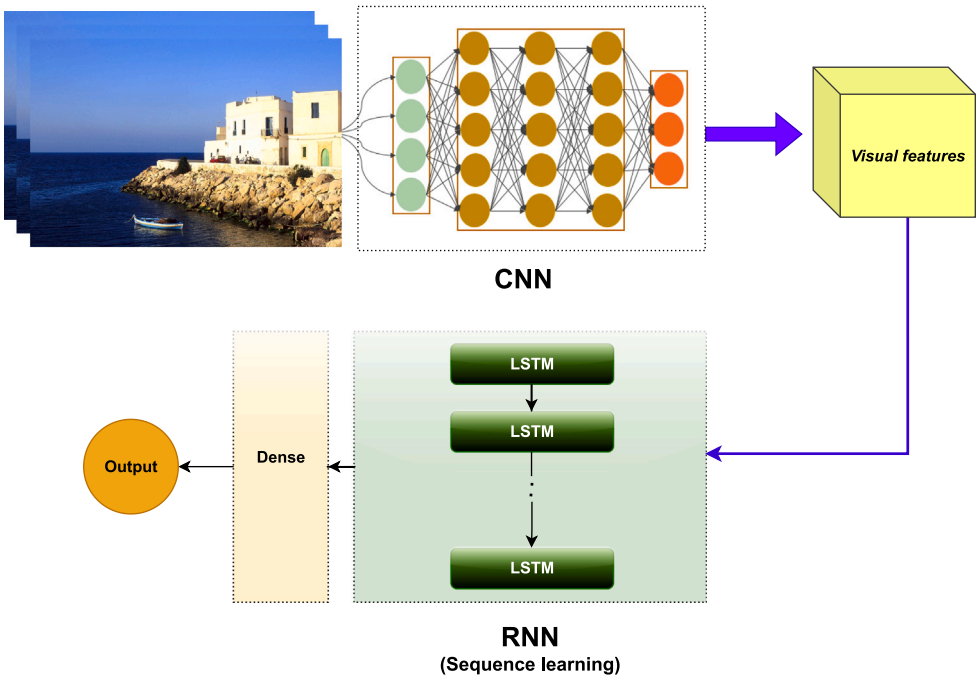


**Fig. 9.** Illustration of a hybrid CNN-RNN architecture for image classification.

blocks to reduce the spatial dimension of local features. Similarly, Guo et al. [88] proposed a hybrid architecture using an end-to-end CNN-RNN. Specifically, the CNN is designed to detect any hidden discriminative features from the input image, while the RNN is used to model the semantic correlation between sequence labels. In [92], Chu et al. found that the combination of CNN (ResNet50) and LSTM units can improve the accuracy of image captioning tasks. They adopted an encoder/decoder architecture in which the encoder (CNN) generates robust feature representations from the input images and embeds them into fixed-length vectors and the decoder (LSTM) takes the embedded vectors as input and predicts the next sentence in the sequence of sentences.

For simplicity, let consider the encoder–decoder structure shown in Fig. 9. The encoder side of the architecture is a deep CNN network that acts as a visual feature extractor that can represent the encoded features as a sequence of $1-K$ items as follows:

$$y = \{y_1, y_2, \ldots, y_c\},\ y_i \in R^k \tag{3}$$

where $K$ is the size of the input dimension and $c$ is the label length. At the decoder side, the LSTM unit is used to generate the correct labels based on the input feature vectors extracted from the CNN output. The LSTM function determines the current input state and decides whether to store the previous state or not. Structurally, an LSTM cell typically consists of three internal gates (i.e., forgetting gate, input gate, and output gate) that modulate the flow of information between the current input state, the previous state, and the output state, respectively. Formally, the input and output states can be expressed as Eqs. (4) and (5), respectively:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{4}$$

and

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0), \tag{5}$$

where $\sigma(.)$ denotes the activation functions (sigmoid and hyperbolic tangent), $h_{t-1}$ is the output vector of the previous unit, $W_i$ and $b_i$ are the network parameters, and $x_t$ is the sequence of input items. For the final classification, the hidden state $h_t(h_t = o_t.\sigma((\sigma((W_c.[h_{t-1}, x_t] + b_c)))))$ of the LSTM is fed into the softmax activation function to estimate the probability distribution $f_t$ over all labels:

$$f_t = Softmax(h_t) \tag{6}$$

In practice, LSTM can be used as a powerful tool to avoid the vanishing and exploding gradient problems.

**Multimodal based.** The authors in [115] proposed a hybrid model for complex activity recognition called HConvRNN. By merging CNN and RNN into a unified framework, the model can be used to jointly model spatial and temporal dependencies in multimodal time series data. Recently, hybrid CNN-RNN architectures have shown excellent performance in video recognition tasks. For example, Jiang et al. [116] proposed a hybrid architecture that incorporates relevant information from multiple modalities, including spatial information, motion patterns, and so on. Architecturally, the LSTM networks are integrated for explicit recursive modeling of sequences, complementing the CNN backbones. In [117], Fan et al. proposed a robust video-based emotion recognition system that combines an audio module, CNNs, and RNNs in a common hybrid architecture to systematically encode visual appearance and motion cues. Specifically, the RNN takes as input appearance features extracted from still video frames by the CNN and subsequently encodes motion information. Similarly, the authors of [118] proposed a multimodal hybrid emotion recognition approach for speech expressions. In this approach, the Inception-ResNet-v2 network and a hybrid CNN/LSTM are combined into a unified framework and used as a robust feature extractor across modalities. In an earlier study, Ebrahimi Kahou et al. [119] proposed the integration of CNNs and RNNs to effectively identify emotions in videos. They found that using their proposed hybrid scheme, the spatio-temporal evolution of features is one of the most powerful cues for emotion recognition.

### 6.3. Hybrid architecture 3: CNN-DBN

DBNs are one of the unsupervised probabilistic graphical models that perform well in learning to stochastically reconstruct their input. Typically, the purpose of a DBN is to learn a hierarchical feature representation of a multidimensional data space and reduce its size to a low-dimensional representation as input for subsequent processing

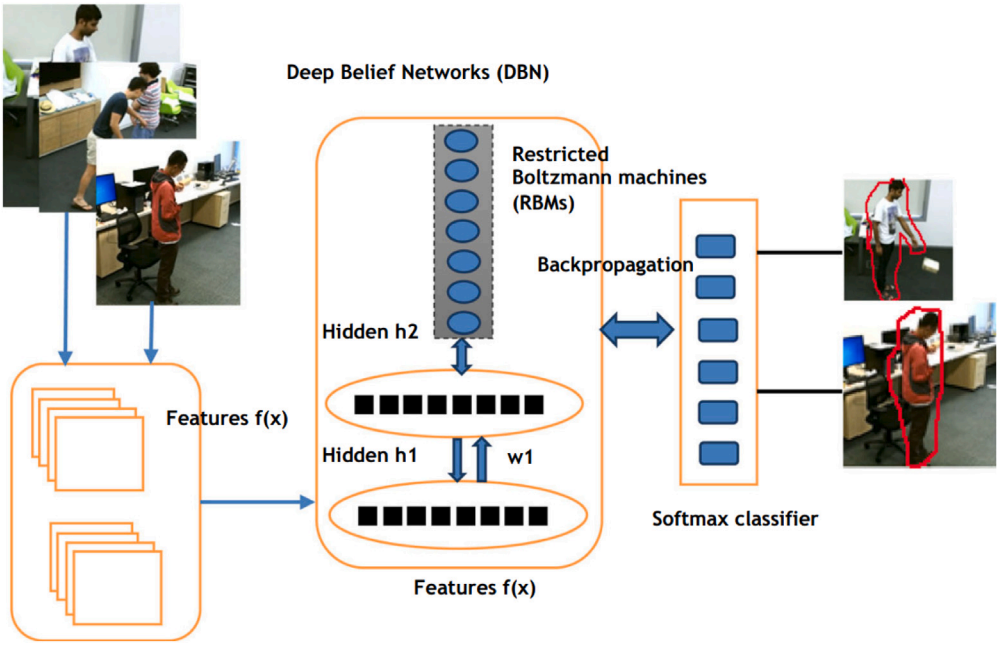


**Fig. 10.** Illustration of a hybrid CNN-DBN architecture for moving objects detection and recognition [149].

stages. In late 2009, Lee et al. [148] proposed a convolutional variant of DBN, called convolutional deep belief network (CDBN), which maps a high-dimensional input image into a low-dimensional space and maintains translation invariance. From a practical point of view, CDBN can perform both discriminative and generative tasks by using the fine-tuning method, which can be optimized using back-propagation or contrastive divergence schemes.

**Monomodal based.** In recent years, the combination of supervised CNNs and unsupervised DBNs has attracted a lot of attention from the deep learning community due to its ability to improve the generalization performance of learning algorithms. For example, the authors of [93,94] proposed unsupervised hybrid models that integrate both CNNs and DBNs into a joint network for mechanical failure diagnosis. Here, the CNN is intentionally used to detect abstract patterns hidden in the low-dimensional representation of the DBN output. With the continuous development of the Internet era, the need for faster, more accurate, more convenient, and more efficient music information retrieval systems is growing. Therefore, a new generation of powerful music information retrieval and recognition models should be developed. For example, Lin et al. [95] proposed a combined CNN-DBN model for music score recognition. The model fully exploits the effectiveness of DBNs and the representation capabilities of CNNs. Experimental results proved the superiority of this method, as the hybrid model was able to efficiently recognize and extract high-level features of the soundtrack.

An example of a hybrid CNN-DBN model architecture is shown in Fig. 10. The architecture is divided into two main sides: a CNN side and a DBN side, followed by a softmax prediction layer. The whole pipeline works as follows. First, the raw input data is fed to the CNN side to extract a set of discriminative feature maps. Then, the DBN side reconstructs the high-dimensional representation vectors into a low-dimensional space. Finally, a final prediction is made using the softmax function.

**Multimodal based.** Recently, Zhang et al. [120] proposed an audio-visual emotion recognition framework based on the design of a 3D CNN-DBN hybrid model. They used 3D CNN to generate audio-visual features from video sequences. The extracted CNN features are then fused and fed into a DBN network to generate fixed-length visual features. In [121], Pan et al. developed a novel hybrid CNN-DBN model for predicting interaction sites and features of RNA-binding proteins from multiple data sources. In addition, the combination of CNNs and DBNs has also been shown to perform well in various video recognition tasks. For example, Zhang et al. [122] proposed a new method for facial expression recognition in video sequences by combining two independent deep CNNs and a DBN model. These CNNs are first integrated to learn high-level spatio-temporal features in segmented video clips. The obtained features are then merged into a deep fusion network constructed in the DBN module.

### 6.4. Hybrid architecture 4: CNN-DAE

DAE is a deep learning model that aims to automatically extract pyramidal features through unsupervised learning to solve the challenge of lack of training data. More specifically, the DAE model can improve the reconstruction performance of complex representations and reduce their dimensionality, which can then be used as input for the subsequent stages. Convolutional autoencoder (CAE) [150] is the convolutional variant of the standard AE architecture, following the encoding and decoding structure. In general, CAE is more feasible than traditional AE for pattern recognition and image processing, as it leverages the feature representation power of CNN to capture the inherent structures of the data [151,152].

**Monomodal based.** From a practical point of view, an adaptive hybrid architecture combining DAE and CNN can provide exceptional performance for decision systems. On the one hand, CNNs can serve as both a dimensionality reducer and a robust feature extractor in high-dimensional domains. On the other hand, the extracted CNN features are then shifted to DAE to detect and extract intrinsic patterns in unsupervised learning mode [153]. Recently, hybrid CNN-DAE-based schemes have been proposed to solve some computer vision problems such as fault diagnosis [96], malware detection [97], heartbeat classification [98], noisy image classification [154], image denoising [99], etc.

Fig. 11 shows an example of using a combination of DAE and CNN for the activity recognition problem [155]. The architecture shown consists of two main parts: the DAE part and the CNN part. The goal of the DAE part (here a denoising autoencoder [22], which is a variant of the standard autoencoder) is to reconstruct typical noisy images from

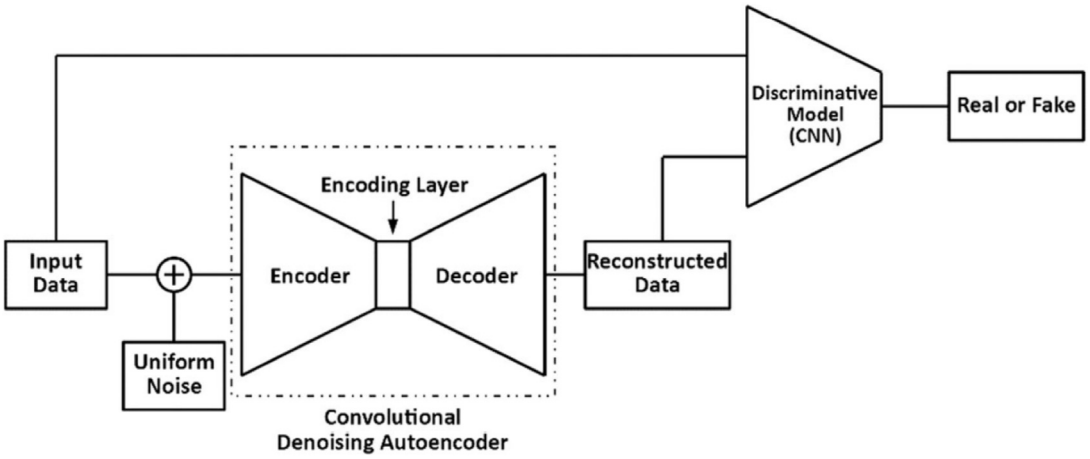


**Fig. 11.** Illustration of a hybrid CNN-DAE architecture for activity classification [155].

unlabeled corrupted data in an unsupervised manner and filter them based on a layer-wise learning strategy for parameter optimization. Then, the CNN part (classifier) is initialized by stacking hierarchical hidden layers such as convolutional layers and pooling layers, followed by dense layers for final prediction (distinguishing between real and fake signals). In summary, by combining the powerful feature extraction capabilities of CNN with the classification performance of DAE, hybrid CNN-DAE models can significantly outperform CNN or DAE alone in terms of accuracy and efficiency.

**Multimodal based.** Recently, Nguyen et al. [123] proposed a deep hybrid architecture for emotion recognition from multimodal data using CAEs to learn heterogeneous feature patterns in the visual and audio domains. The authors of [124] proposed a multimodal emotion recognition system based on CAEs. In this approach, word-level alignment is performed over the feature sequences of each modality. After a suitable standardization process, the convolutional autoencoder then learns unsupervised embeddings by combining multiple datasets. Liu et al. [125] developed a novel multimodal approach for cross-media retrieval and image classification. The proposed approach is designed to learn high-quality features from multiple modalities. To this end, they integrated the convolutional operations of each modality into an autoencoder to learn the joint feature space from image and text content by exploiting the correlation between the hidden features of different modalities. In [126], the authors proposed a novel diagnostic system for Alzheimer's disease and investigated the ability of CAEs to combine information from magnetic resonance imaging (MRI) and positron emission tomography (PET). Specifically, CAEs were used to learn synthetic image features and CNNs were used to learn the original image features.

### 6.5. Hybrid architecture 5: CNN-GAN

To improve the supervised training process of discriminative networks, an unsupervised GAN network can be integrated into a hybrid architecture that includes CNNs and acts as a powerful feature generator. Here, CNNs can be used as a strong backbone to extract a set of feature vectors from the raw data. Typically, the features generated by GAN can represent the spatio-temporal correlations and dependencies between signal channels at different time steps. Recent literature has shown that GANs are robust in generating high-quality and realistic patterns from real-world dynamic datasets [156,157].

**Monomodal based.** In recent years, hybrid architectures combining CNNs and GANs have been successfully developed to accurately and efficiently learn complex structures from high-dimensional data [100–104]. For example, Liang et al. [100] developed a hybrid CNN-GAN

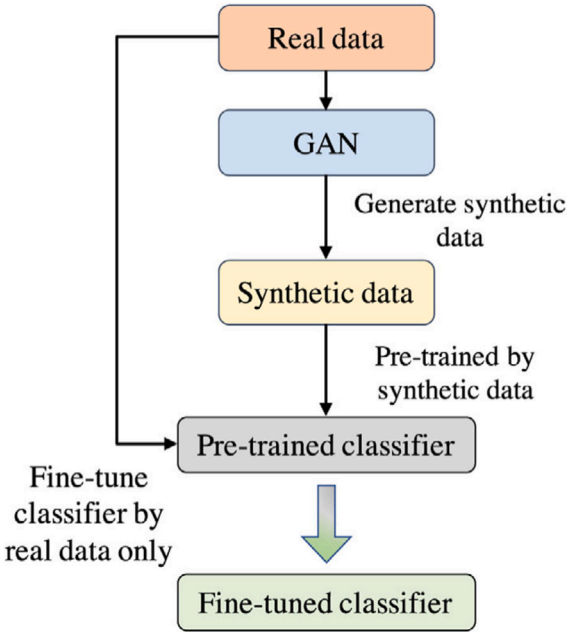


**Fig. 12.** Illustration of a hybrid CNN-GAN architecture for damage estimation from unbalanced data [158].

model to extract emotional features from high-dimensional electroencephalography (EEG) data. They trained the whole architecture in an unsupervised manner following the encoder–decoder design with GAN. In this case, the GAN consists of a generator that reconstructs the EEG signal from the latent representation and a discriminator that detects whether the input EEG signal is a fake signal or a real signal generated by the generator. In [101], Cheng proposed to use CNN-GAN for packet generation of network traffic. In his proposed architecture, a discriminator and a generator are implemented, both of which use CNNs as building blocks. To address the ubiquitous problem of data imbalance, Yin et al. [102] developed a data generation scheme for bearing fault diagnosis based on Wasserstein GAN and CNN models (WG-CNN). They performed adversarial training on a small dataset to generate a high-quality balanced dataset and used CNNs to learn rich features to classify different types of faults. Fang et al. [103] proposed a novel gesture recognition framework incorporating CNN and deep convolution generative adversarial networks (DCGAN). They used DCGAN (an extension of GAN) to generate new, higher quality samples to overcome the overfitting problem and improve the generalization

ability of the training algorithm. Similarly, Tan et al. [104] proposed the combination of DCGAN and CNN models for the purpose of micro-structured materials design. Here, DCGAN is trained in an unsupervised manner to generate micro-structure images of elliptical holes only. A CNN is then trained in a supervised manner, and the generated images are used as input to evaluate the generation performance of the GAN. More recently, Bousmina et al. [159] proposed a hybrid method based on GAN and CNN-LSTM for automatic recognition of human activity in aerial images. Specifically, the authors proposed a hybrid data augmentation technique that combines CNN-LSTM with Wasserstein GAN [160] to learn the spatio-temporal dynamics of the features, and then use the GAN-based technique to generate synthetic CNN-LSTM features based on action categories that have high discriminative power with respect to spatio-temporal variations.

Fig. 12 shows a schematic of a hybrid GAN-CNN architecture, where the GAN is trained to generate synthetic data from an unbalanced dataset to improve the discriminative power and generalization ability of the CNN classifier. The GAN can generate artificial images for each data sample through an adversarial training process. Specifically, the generator and the discriminator play a two-player min–max game. The first network learns from random noise and tries to generate new synthetic samples that are very similar to the original unbalanced dataset. The discriminator then learns from the original samples and tries to maximize the loss function over all samples. In summary, GAN stands out as a powerful generative tool that has made significant advances in the field of data generation and augmentation.

**Multimodal based.** The authors in [127] proposed a new robust multimodal model for language understanding called MMC-GAN. The architecture of this model consists of a CNN, which is part of a feature extraction scheme, and two paragraph vector models. The linguistic input is processed by a distributed memory paragraph vector model (PV-DM) [161], and the visual signal is analyzed by the CNN. In 2020, a novel multimodal semantic segmentation framework for large-scale urban scenes has been proposed by Hong et al. [128]. From an architectural point of view, the dual-SegNet architecture [162] and GANs (self-GAN and mutual-GAN) form the basis for the design of the proposed architecture. For each modality, VGGNet is specifically used as a feature extractor. The self-GANs module receives the output of each modality to accurately represent the random noise. The feature inputs from each modality are then jointly represented before the mutual-GANs module performs their fusion. In [129], Guo et al. introduced a unified model for generating different caption styles. The model learns how to associate images with different caption styles without using style-specific caption pair data. They used CNN as a robust tool to encode images into visual static features. After feature extraction, a GAN is used to generate different captions and accurately describe the image content.

### 6.6. Hybrid architecture 6: CNN-GNN

GNN is a new type of DNN that has proven to be powerful for modeling complex graphs. When training a standard GNN network, the node-level representation of graph structure data can be improved by propagating local information between nodes [163]. However, standard GNN models require large computational resources to learn large and dense graphs, which is where fast graph convolutional network (FastGCN) comes in. FastGCN [164] is a convolutional version of the standard GNN model that is particularly effective for semi-supervised learning. In particular, Chen et al. [164] interpreted the graph convolution as an integral transformation of the embedding function, reducing the need for simultaneous propagation of test data and improving inference time.

The nature of GNNs is beyond the scope of ordinary CNNs [165]. In other words, it is a generalized variant of CNN. On the one hand, a GNN model learns to iteratively aggregate information from a given neighborhood that can be mapped to a high-level representation. On the other hand, CNN models learn to map high-dimensional visual features to low-dimensional space, producing a final abstract, generic representation for inference schemes. As a result, the number of network parameters grows sequentially with the number of hidden layers, requiring an enormous amount of computing power to perform basic operations. In addition, CNN models generally perform well when trained on regularly structured (i.e., Euclidean) data, such as static images. In contrast, GNN models tend to perform well when trained on both regularly and irregularly structured data (i.e., Euclidean and non-Euclidean data), such as images, time series, social networks, state machines, etc. However, GNNs and CNNs share some important characteristics, such as weight sharing, local connectivity between nodes, and the use of multiple processing units.

Fig. 13 shows the semantic relationship between 2D convolution (Fig. 13(a)) and graph convolution (Fig. 13(b)), where an image can be viewed as a special kind of graph where pixels are connected by their neighboring pixels.

**Monomodal based.** Recently, GNNs have received a lot of attention in the deep learning community because they have been successfully applied to many complex tasks that extend the representation capabilities of CNNs. For example, Luo et al. [105] proposed a hybrid CNN-GNN architecture, called HyGnn, to improve the performance of crowd counting tasks. First, a CNN model (VGG-16) is used as a backbone network to extract contextual multi-scale features from the input images. Then, a GNN is used to model multiple kinds of useful associations, where features from different perspectives are drawn as nodes and associations between them are jointly represented as edges. In [106], Li et al. introduced a remote sensing image scene classification framework that integrates CNN and GNN into a unified design called MLRSSC-CNN-GNN. They used CNN to capture high-level visual features in remote sensing image scenes. To optimize the feature learning process, the node neighborhoods are then sampled using a GNN based on the similarity and proximity between superpixel regions using the analogous relationship between CNN features. The authors in [107] proposed a novel hybrid CNN-GNN architecture called MobileViG for mobile vision applications such as image classification, object recognition, and instance segmentation, which exploits the generalization ability of CNNs to learn spatially local representations and GNNs to deal with graph-based structures. According to [107], the proposed model significantly outperforms many existing mobile models in terms of accuracy and efficiency.

**Multimodal based.** With the recent growth of heterogeneous graph data, multimodal methods must be able to account for a variety of inductive biases. However, learning from multimodal data faces fundamental challenges because inductive biases vary across data modalities and the input graphs may not be explicit. To overcome these challenges, multimodal graph-based methods can fuse multiple modalities with the graph and take advantage of cross-modal correlations [166]. For example, Wei et al. [130] developed a new robust framework for multimedia recommendation based on graph convolutional networks (GCNs), called MMGCNs, that exploits information exchange between different contexts in multiple modalities. In [131], the authors proposed a novel architecture for visual question answering (VQA) called multimodal GNN (MM-GNN). The architecture consists of building blocks for effective feature representation and prediction from heterogeneous data. The proposed model represents a multimodal image as a combination of three sub-graphs, each representing a modality. In other words, given an image containing a visual object, a scene text, and a question, the goal is to generate an answer. Prediction is done in three processing steps: extracting the multimodal content of the image and creating a three-layer graph; performing multilevel, cross-modality messaging to refine the representation of the nodes; and predicting the answer based on the graphical representation of the image. The authors in [132] explored the full potential of GCNs in classifying violent online political

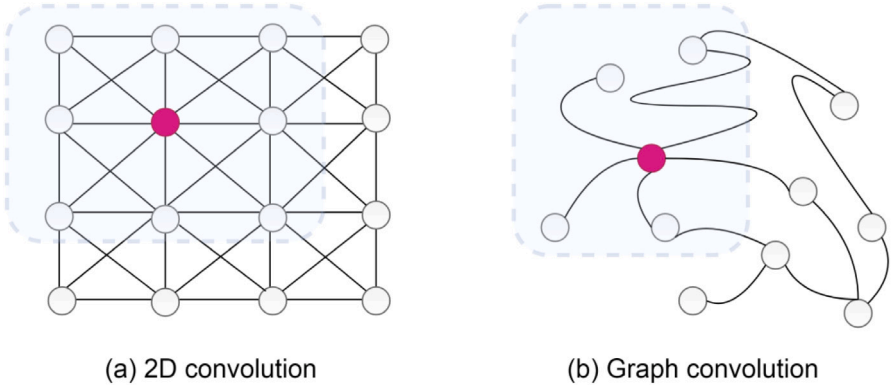


**Fig. 13.** Illustration of 2D CNN (a) and GCN (b).

extremism. They developed a multimodal approach to categorize user posts based on the topic of discussion. Specifically, they introduced GCNs to combine heterogeneous contextual information from posts (e.g., text, visual content, and information about the user's interaction with the online platform).

### *6.7. Hybrid architecture 7: CNN-ViT*

Recently, vision transformers (ViTs) have gained popularity as potential alternatives to standard CNNs in various computer vision applications. Such transformers are very convenient because they can focus on global dependencies in a visual content, but they have a lower generalization ability compared to CNNs [167]. Transformer-based modeling is currently at the forefront of many fields, especially computer vision and NLP [168]. For example, the popular chatbot ChatGPT is a transformer-based language model based on a generative pre-trained transformer known as GPT. GPT uses a self-attention mechanism to model dependencies between terms in a text [169].

More recently, it has become common to combine convolution and attention mechanisms in ViTs. These architectures (also known as hybrid vision transformers or hybrid ViTs) have achieved great performance in vision tasks [170]. As mentioned earlier, the ViT pipeline consists of segmenting the input still image into non-overlapping regions or patches (also called tokens), which are fed to the encoder module via linear layers. Since the number of tokens and the dimension of the tokenized representations are fixed, a vanilla ViT cannot effectively capture fine-grained spatial information at different scales. In contrast, hierarchical transformers such as the convolutional vision transformer (CvT) [171] and the swin transformer (Swin-T) [134] are better suited for large-scale prediction tasks because they can gradually reduce the number of tokens and expand the token feature space. On the one hand, ViT uses a fixed set of input features with variable time delay, while Swin-T uses a weighted sum of input features as output, capturing the temporal dynamics of the input data. On the other hand, CvT, a convolutional variant of ViT, uses convolutional projections to learn spatial and low-level features from the image patches. However, the effectiveness of hierarchical approaches is largely related to the superiority of the transformers rather than the convolutional bias. To this end, Liu et al. in [172] re-examined the internal structure of a pure CNN and explored the limits of what can be achieved. Specifically, they proposed a modernized ResNet architecture, called ConvNeXt, as a backbone network for robust feature extraction that can rival transformers in accuracy and scalability, resulting in state-of-the-art prediction performance.

With the development of lightweight architectures for mobile vision tasks, there is an increasing demand. For example, lightweight CNNs can learn highly informative and complex representations with few parameters. However, such architectures can only learn spatial patterns. In contrast, ViTs have been used to effectively learn global features. However, ViTs are computationally expensive compared to CNNs. To address this problem, Mehta et al. [173] proposed a lightweight version of ViT called MobileViT. MobileViT is a lightweight, low-latency model designed for mobile and embedded vision applications. Specifically, the model combines standard convolution layers with point-wise convolution layers to capture both local and global feature dependencies of the input data with reduced parameters.

Motivated by the rapid development of ViT models, a number of hybrid architectures combining CNNs and ViTs have been proposed in the literature in recent years.

**Monomodal based.** The authors in [108] proposed a new feature generation backbone consisting of CNNs to improve the recently introduced ViT model. Specifically, they applied the developed model to the RSNA intracranial hemorrhage classification problem and showed that n-CNN-ViT can outperform the standard ViT model. In [109], Zhang et al. proposed a novel hybrid architecture for OCT image segmentation called TranSegNet. The internal architecture of TranSegNet consists of a modified U-shaped network for salient feature extraction, i.e. a CNN-ViT encoder. On top of this architecture, a lightweight ViT is introduced to accurately model long-term dependencies in the data. In particular, a multi-head convolutional ViT is integrated throughout the architecture to capture global feature information for accurate retinal layer segmentation. The authors in [110] leveraged the representation power of CNNs and the ability of ViT to more efficiently capture dependencies and relationships between image patches. To this end, they proposed a hybrid CNN-ViT architecture designed to improve the ability to represent global spatial context in remote sensing image classification. Fig. 14 shows the proposed architecture, which includes a patch embedding component, an encoder, and a head classifier. The image is first segmented by the transformer into a set of image patches generated by the encoder module of the ViT, which are then assigned semantic labels for classification. The features extracted by the CNN are then transferred to the CBlock of the first branch (Fig. 14(b)).

**Multimodal based.** To date, the ViT is still a popular learner that has achieved great success in various computer vision tasks. Due to the recent proliferation of multimodal big data, ViT-based multimodal learning has become a hot topic in the deep learning community [174, 175]. However, how to develop a new generation of hybrid models that combine supervised CNNs and transformers with one or more sensory modalities remains an open question in the literature. For example, in [133], the authors designed a hybrid model consisting of a convolutional encoder and a transformer decoder to fuse multiple data inputs. In particular, they proposed a cross-modal attention module in the encoder to systematically capture local and global dependencies of multiple source images. They also developed a branch fusion module to adaptively fuse features from two independent branches. They integrated a swin transformer module [134] in the decoder to improve the reconstruction performance of the proposed network. Recently,

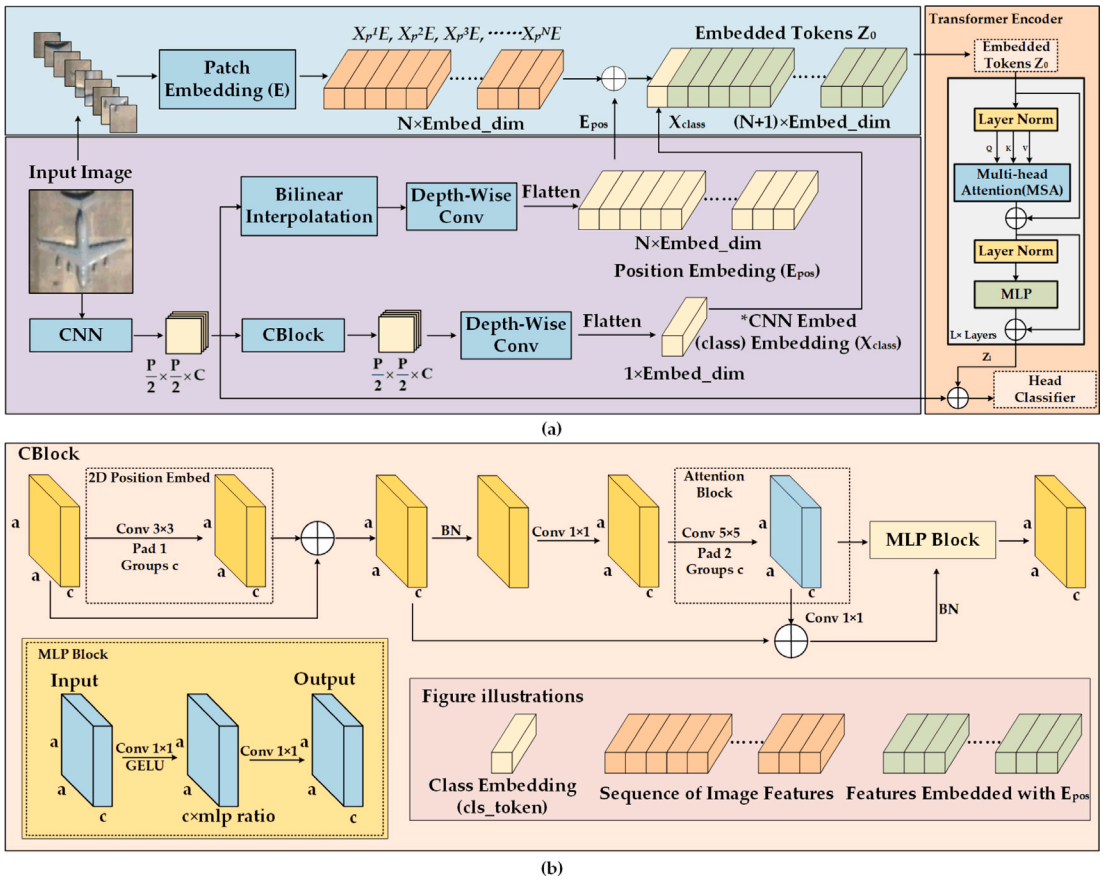


**Fig. 14.** Illustration of a hybrid CNN-ViT architecture for remote sensing image classification [110].

a lightweight model called MFTransNet has been proposed by He et al. [176]. This model integrates CNNs and transformers for semantic segmentation of high spatial resolution (HSR) remote sensing images to reduce model complexity, improve scalability, and increase learning efficiency. To effectively extract features, a tiny convolutional network is first used. These features are then fed into a multi-head module for adaptive feature fusion. Finally, a multi-scale decoder is finaly used to fuse features from multiple scales to perform the prediction process. Similarly, Sun et al. [135] proposed a multimodal hybrid CNN-transformer architecture called HybridCTrm for brain image segmentation. By combining CNNs and transformers, this architecture has been experimentally shown to be able to avoid the problems of overfitting and lack of non-local dependence that arise in CNNs. In [136], Zhou et al. proposed a novel hybrid architecture for effective multimodal medical image fusion. This architecture combines a DHRNet network and a hybrid transformer, where the DHRNet network is particularly suitable for extracting features from the source image. These features are then fed into the fine-grained attention module of the hybrid transformer to generate global features. This module is used to discover correlations between features by examining the long-term dependencies across modalities.

## 7. General framework for multimodal hybrid deep learning

In this section, we present a universal framework for multimodal hybrid learning that systematically incorporates both standalone networks and hybrid and multimodal fusion techniques. As shown in Fig. 15, we can identify several components within this framework that will help develop a better understanding of the key algorithms of multimodal hybrid learning and advance this growing field.

Although multimodal learning-based approaches are well suited to synthesize heterogeneous information from different perspectives, maximizing the use of different data sources to autonomously acquire multimodal data remains a challenge due to its complexity. Before training a model, a unified dataset must be created. This may require a preprocessing and harmonization step. The purpose of this step is to remove redundant and inconsistent information from the multimodal data and to control the complexity of the resulting hybrid network.

At the information level, feature extraction is a critical step in any multimodal hybrid learning scheme. By using the most appropriate feature extraction methods, hybrid models are able to make more accurate and effective predictions. With the advent of a new era of multimodal and hybrid fusion techniques, which we refer to as multimodal hybrid fusion in our proposed framework, the adaptability and robustness of feature extraction methods have continued to improve significantly. From an architectural point of view, multimodal hybrid modeling uses hybrid fusion (e.g., stacking, concatenation, etc.) and multimodal fusion strategies (e.g., traditional and deep learning-based) to merge standalone networks with multiple modalities. Both multimodal and hybrid fusion components contribute significantly to the overall performance of the final hybrid model. However, a lack of multimodal and/or hybrid modeling can lead to inefficient exploitation and capture of cross-modal interactions, which affects the performance of the overall system. In other words, poor multimodal and/or hybrid fusion does not fully exploit inter- and intra-modal correlations, which can affect the stability and effectiveness of the training algorithm. In this study, multimodal hybrid feature extraction pipelines have been investigated using universal CNNs (2D CNN and its 3D counterpart) together with other generative and discriminative algorithms (e.g., RNN, DBN, GAN, etc.). In our proposed framework, two main factors are used to identify the potential domain (i.e., spatial, temporal, or spatio-temporal) of each hybrid architecture, including the sensory input modality and the standalone design. More specifically, these two attributes can be

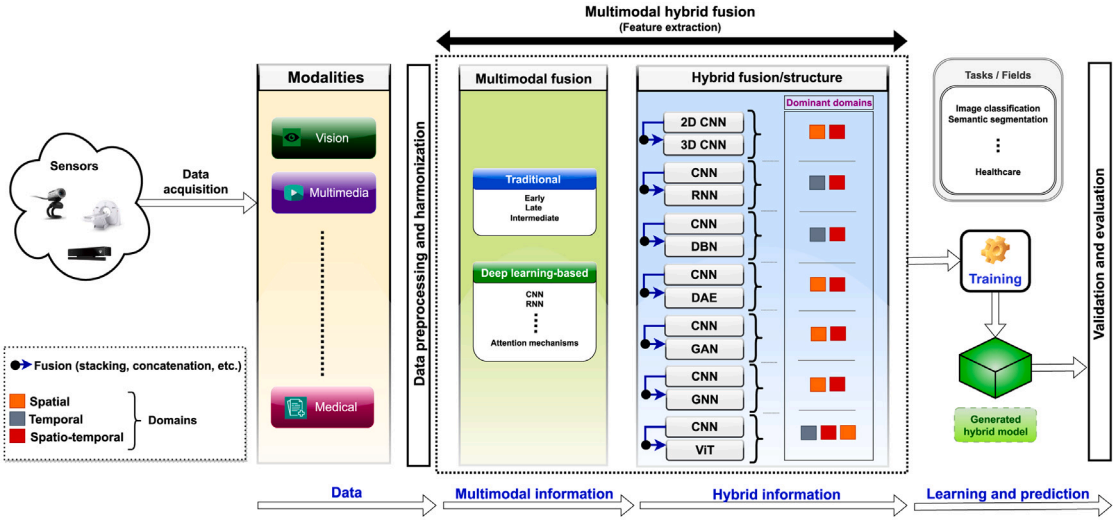


**Fig. 15.** Generic framework for multimodal hybrid deep learning.

considered as a concise solution to identify the dominant domain. For instance, if the designed architecture receives input in the form of a static image, this hybrid model would be assigned to the spatial domain. Another input in the temporal domain could be other forms of temporal and motion information, such as moving images or video content. The spatio-temporal domain encompasses both the temporal and spatial domains.

At the learning and prediction level, as with conventional learning algorithms, it is important to select the appropriate loss function, optimizer, etc., for efficient learning. After learning, evaluation of the results consists in providing relevant information about the parameters and data used and the results obtained in relation to the learning process. Depending on the application domain and task, standard evaluation metrics can be used to interpret the obtained results. For example, image classification models can be evaluated to assess their performance using specific metrics such as accuracy, precision, recall, etc.

The development of integrated multimodal hybrid architectures opens up tremendous opportunities for incremental progress in multimodal data analysis. This will contribute to a better understanding of multimodal hybrid fusion strategies, which in turn will expand the impact of multimodal hybrid learning and its practical applications in many emerging fields, especially in computer vision.

## 8. Multimodal hybrid learning paradigms

### 8.1. Transfer learning

In some real-world scenarios, the amount of training data available to train deep models is generally insufficient, leading to serious problems such as overfitting, inaccuracy, and so on. Transfer learning [177] is an area of research that is receiving increasing attention due to its ability to solve such problems by transferring knowledge gained in one domain to another similar domain using a fine-tuning mechanism. In addition to addressing the above issues, from a practical perspective, by reusing the abstract features of previously learned domains and adapting them to new domains, the ability of the model to generalize and increase the stability of the learning process can be greatly enhanced. In practice, the adaptive use of pre-trained models can be used as a powerful tool to extract abstract features from ImageNet datasets [178] or to initialize the internal parameters of cognitive models for higher performance. For example, Das et al. [179] proposed a two-stage model for COVID-19 screening using chest X-rays based on transfer learning. In the first stage, high-level features are extracted from the input images using a pre-trained VGG19 network. The extracted weights are first fed to the first shallow classifier, which categorizes the samples into infected and non-infected. The samples classified as infected are passed to the second stage. A second shallow classifier is then used to classify the infected patients into pneumonia patients and COVID-19 patients.

In a multimodal context, this approach can speed up the overall optimization process by reducing inference time and computational cost [7]. Furthermore, by maximizing cooperative interactions between modalities, abstract and discriminative feature representations can be effectively learned. It can also improve the quality of the transferred information by filtering out potentially hidden noise patterns. In practice, multimodal hybrid transfer learning has proven to be very effective compared to traditional transfer learning methods [180,181]. However, this learning approach still faces many challenges, such as the need to overcome the heterogeneity between different modalities, the implicit coordination between multiple generic architectures, and how to transfer semantic knowledge from one domain to another when the label sets are shared or disjoint.

### 8.2. Ensemble learning

Ensemble learning [71,182] is one of the key paradigms in machine learning that uses multiple learners to improve the behavior of the training algorithm and increase the discriminative power of models. In practice, ensembles typically contain multiple learning algorithms, called base learners, that help increase the generalization ability of the weak learners and allow them to make very accurate predictions. Specifically, in ensemble learning, a set of features can be extracted from a large amount of data and multiple learners can be used to learn these features to produce weak prediction results. The difference between ensemble learning algorithms lies in the need to make assumptions about how the original multiple learners will evolve.

In multimodal theory, ensemble-based approaches have proven to be very powerful and flexible for modeling multimodal data [183]. However, it is also possible to create hybrid models that incorporate different hybrid predictors [184]. Ensemble learning strategies can be used to systematically create hybrid models that combine multiple modalities to solve a given problem. The key to the success of an ensemble system is the diversity of the ensemble members and their ability to correct some of the prediction errors.

### 8.3. Active learning

While large amounts of unlabeled data are readily available, training enough data to build learning models with supervised learning is challenging due to expensive and time-consuming labeling techniques. By identifying the most informative and representative structures in the instance space, active learning methods can be used to query a small number of data points [185,186].

In multimodal learning, end-to-end learning of multimodal hybrid networks requires a large amount of data compared to unimodal hybrid models due to the large parameter space and the complexity of their interactions [187,188]. As mentioned earlier, active learning is one of the widely used techniques that focuses only on the samples that contribute to the overall performance of the model, thus reducing the cost of data labeling. When dealing with multimodal data, most current active learning algorithms (designed primarily for unimodal tasks) tend to have a bias in selecting important samples. This bias prevents the balanced multimodal learning needed to achieve optimal performance.

### 8.4. Online learning

Offline learning [189], one of the main families of machine learning algorithms, typically attempts to simultaneously learn observations from a static dataset. In contrast, increasingly popular online learning techniques are designed to update predictions as new data enters the system, often requiring a rapid stream of real-time data.

In a multimodal online learning environment, each input dataset is fed into different components of the hybrid model one at a time (rather than in batches), and then its internal parameters are updated after training.

### 8.5. Federated learning

Hospitals and other stakeholders can use a variety of distributed deep learning methods to develop predictive models. These methods are known in machine learning as collaborative or federated learning [190–192]. They are particularly useful when data sharing is prohibited by privacy standards. For example, a hospital can train its model locally and then send the model parameters to a central server. By merging the model parameters for each entity, the central server then generates a final global model. This global model is developed and run iteratively until it converges. In practice, federated learning helps solve the problem of data scarcity and improves the applicability of predictive models [193].

As the amount of multimodal data in advanced real-time embedded systems continues to grow, it has become a key challenge to harness this rich multimodal knowledge without compromising user privacy and data confidentiality. Multimodal federated learning is a privacy-preserving alternative to centralized unimodal deep learning [194]. However, existing multimodal federated learning approaches for multimodal data rely on heterogeneous feature aggregation at the unimodal level, which typically limits the ability of servers and clients to share the same model architecture in each modality [195]. Consequently, this limits both the task variety and the representational capacity of hybrid models.

### 8.6. Multi-task learning

A subfield of machine learning, multi-task learning [196,197], has recently shown that it is possible to learn multiple tasks simultaneously using shared models. Indeed, this approach has a number of advantages, including increased data efficiency, reduced overfitting, and the potential to speed up the optimization process by leveraging auxiliary knowledge. However, selecting which tasks to train together is challenging in itself, and training multiple tasks simultaneously remains a major design and optimization challenge [198].

In recent years, interest in multitasking approaches for learning multimodal representations with hybrid networks has increased for several reasons [199,200]. For example, supervised learning of hybrid deep models typically requires a large amount of task-specific annotations, which are rarely available. Therefore, multitask learning can be an efficient way to use supervised data from many similar tasks. As mentioned earlier, the use of multitask learning can help reduce overfitting to specific tasks and better generalize the learned hybrid representation across tasks.

### 8.7. Zero-shot learning

Most machine learning algorithms rely on classifying data instances into previously learned classes [201]. From a practical standpoint, many advanced applications require the ability to effectively classify unknown instances. Zero-shot learning is a machine learning technique designed to provide the ability to classify unknown objects without requiring the model to specifically learn the classes.

In multimodal zero-shot learning, the learning algorithm typically provides a set of feature vectors for each category [202]. With this multimodal feature space, each category can be recognized, which can improve the representational power of the hybrid model and also provide additional information useful for learning. However, there are still some challenges to overcome before this information can be used effectively. Since each category can have multiple visual appearances, an example of a category will correspond to only one feature vector.

### 8.8. Multi-view learning

To meet the learning requirements, conventional machine learning techniques usually try to combine different views into a single view [203]. However, due to the different statistical properties of each view, such a combination can lead to overfitting problems when the size of the training sample is limited. In contrast to single-view learning, multi-view learning [204–207], an emerging deep learning method, aims to improve the generalization ability of models, including hybrid models.

In the literature, multi-view learning is also referred to as data fusion from multiple feature sets [206]. In recent years, several methods have been developed to learn data from multiple views, taking into account the richness and diversity of the different views. These views can be from different sources or from different subsets of features [208]. In practice, multi-view learning algorithms differ from single-view learning algorithms in that they require multiple views of the same input data.

### 8.9. Learning alignment

In multimodal hybrid deep learning, alignment refers to the process of discovering linear correlations between hybrid patterns from two or more different modalities [7,8,209]. For example, in training pairs consisting of an image and a caption, it is possible to identify the areas of the image that match the words or phrases in the caption [210]. Video activity recognition is another example of the use of alignment [211]. Here, the predictive model analyzes the input video to extract spatio-temporal information and then looks for correlations between objects and actions. Thus, multimodal alignment is required to map the subcomponents of the different modalities and determine the inherent correlations.

## 9. Discussion and critical analysis

In recent years, several types of hybrid architectures have been developed, each targeting a specific computer vision task. Combining different types of standalone architectures has the potential to extract more meaningful and abstract feature information than using fundamental models alone. Choosing which standalone networks to combine is therefore a key step in the hybridization process. In this study, we provided an overview of current multimodal hybrid learning approaches for solving various computer vision problems. In particular, multimodal hybrid approaches aim at combining machine learning and/or deep learning with one or more sensory modalities. The main question of this research is how reliable these methods are and how they overcome some of the problems encountered in complex tasks. Based on Tables 3 and 4, most of the existing hybrid systems perform well on different datasets of different types and scales using standard evaluation metrics. This is due to the expressiveness and flexibility of the hybridization schemes, which can handle different types of input cues through different standalone components. It has also been observed that multimodal hybrid deep learning plays an increasingly important role in most common computer vision tasks, such as image classification and semantic segmentation. Hybrid models have been developed and extended to fuse multidimensional feature information from multiple standalone architectures using one or more modalities. With the increasing need to develop robust decision making systems, relevant information extracted from a single architecture and sensory modality is no longer sufficient to bring the desired performance of deep learning models to an adequate level.

As mentioned above, multimodal hybrid learning algorithms aim to improve system performance by integrating multiple universal architectures and modalities, and are application-oriented. As shown in Fig. 16, image classification and emotion recognition are the most common tasks and application areas covered by hybrid deep learning.

When building hybrid architectures, the goal is not only to strengthen the cooperation between individual components and modalities, but also to capture complementary knowledge from multiple data sources and hierarchical levels. The incremental advantage of using multimodal hybrid approaches is that they produce better results than learning methods based on a single input. Every year, hybrid deep learning algorithms open new horizons in various active areas, especially in computer vision, providing the flexibility to perform complex tasks and build more sophisticated decision tools. By combining different architectures and selecting the best ones through a series of strategies and interactions between independent networks and multiple modalities, hybrid systems can improve the discriminative performance of prediction pipelines. For example, the performance of deep learning algorithms can be dramatically improved by incorporating feature extraction from 2D and 3D CNN networks [77–81]. One of the most attractive aspects of multimodal hybrid learning is the dynamism and strength of the connections between processing modules and modalities, which expands the range of real-world applications. As the number of applications continues to grow, new unimodal and multimodal hybrid architectures are expected to be increasingly supported by improved hardware acceleration.

While multimodal hybrid deep learning has many strengths, it also has drawbacks that need to be recognized and addressed. As mentioned earlier, the cognitive performance of multimodal pipelines can be improved by dynamically merging heterogeneous architectural designs so that they effectively produce consistently high-quality feature representations in a multimodal context. However, this requires a significant amount of multimodal data, and it is typically challenging to obtain a variety of large-scale datasets. In particular, data acquisition is one of the most critical and fundamental steps in the multimodal hybrid learning process. In active domains such as medical imaging (e.g., COVID-19 screening from chest X-rays or CT scans), deep hybrid models with a large number of parameters cannot be trained well with limited data samples, leading to overfitting problems and lack of generalizability, which in some scenarios can also lead to delays in patient treatment. Therefore, multimodal hybrid learning can often lead to better prediction performance, but consideration must be given to the generalizability of hybrid models, the volume and caliber of training data, and the unique challenges that must be addressed in a multimodal environment. In addition, multimodal hybrid learning networks also have certain limitations in terms of interpretability. In other words, there is a need to fully understand and accurately predict the decisions made by the resulting hybrid models. Multimodal hybrid learning also faces the computational burden of model development when integrating multiple data modalities and mixed architectures. In particular, the development of multimodal hybrid models is potentially resource-intensive because they require effective training on a large number of parameters and data of different types and scales. In addition, their complexity can make them more expensive than traditional standalone models. There are also no specific rules for the number of layers and connections within these hybrid models. However, by reducing the number of network parameters and deep processing layers, hybrid models can be lighter and more compact, greatly reducing computational complexity and making them more suitable for edge devices with limited processing power. Also, different hybrid models and architectures can suffer from different levels of noise and bias depending on the hierarchical level of integration.

Given the relative strengths and weaknesses of each hybrid model, no single hybrid system is likely to address all the challenges of multimodal learning. For example, hybrid 2D-3D CNNs can be widely used to extract relevant local and contextual features from static data [77,78]. However, 2D CNNs have no memory and ignore temporal dependencies between correlated features. Hybrid CNN-RNN models, on the other hand, are suitable for problems where temporal dependence plays an important role.

## 10. Current challenges and future research needs

The main purpose of this section is to shed light on the practical directions that multimodal hybrid learning research is currently taking and related challenges that will be of interest to the deep learning community in the near and distant future. Fig. 17 highlights the current challenges and future research needs for hybrid deep learning technologies in the context of multimodal analysis.

### *10.1. Current challenges*

One of the major challenges in the era of multimodal hybrid learning is how to determine which modalities are more relevant and which fusion techniques can be used to combine relevant concepts and dimensions from heterogeneous standalone models across different modalities. It is also critical to reduce noise in the training data, which requires specific model design and fusion techniques. Due to the strong interactions and natural correlations between data modalities and internal model architectures, the use of multimodal hybrid fusion techniques is not an option, but a necessity. The main open question for this study is how reliable these techniques are and how multimodal hybrid learning approaches can overcome some of the current challenges in complex tasks.

Combining multiple sensory modalities and capturing relevant patterns in the data to learn a hybrid model is just one example of how multimodal hybrid learning techniques perform well in many complex scenarios. Therefore, there is a great need for automated algorithms that can automatically infer target information with minimal error. However, most statistical analysis techniques currently in use have limited ability to handle heterogeneous information (i.e., spatial, temporal, and spatio-temporal) and complex architectures [9]. Nevertheless, the scientific community and industry continue to show strong interest in this area. This uncertainty highlights the need to develop innovative

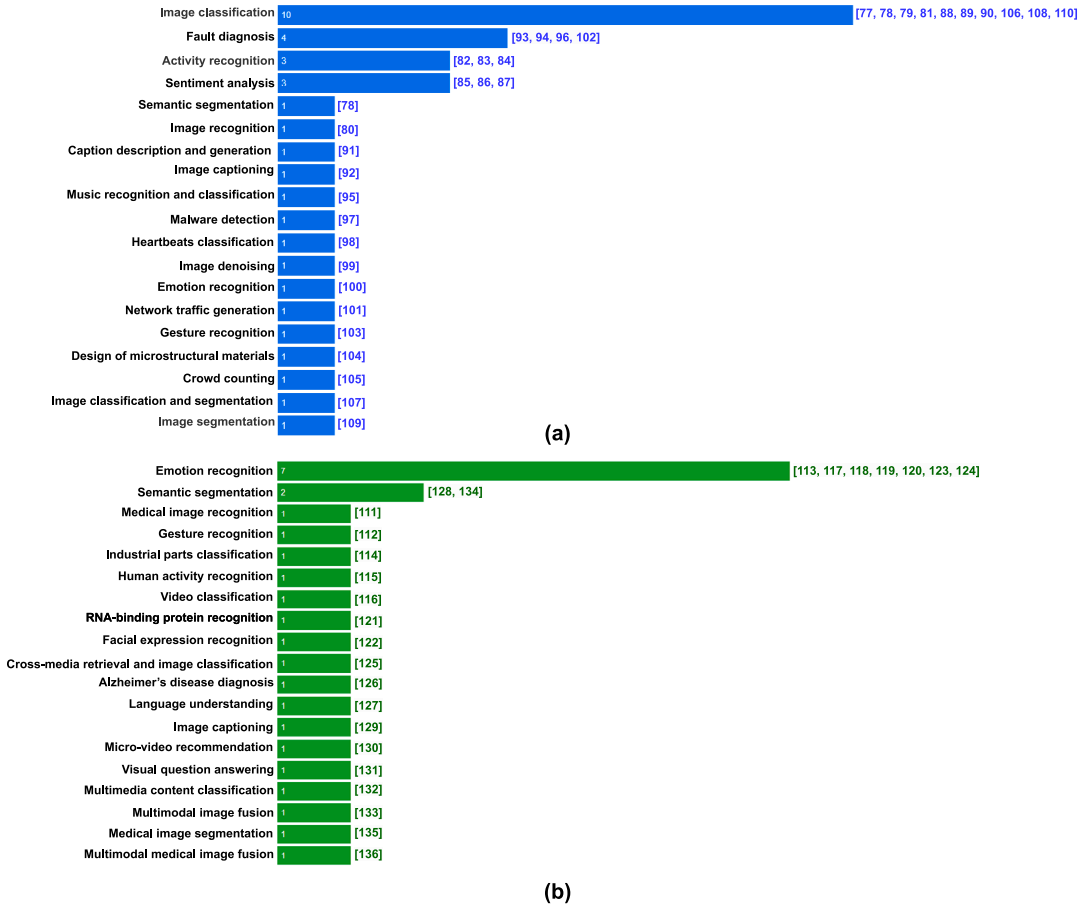


**Fig. 16.** Distribution of unimodal (a) and multimodal (b) hybrid learning studies per application and task.

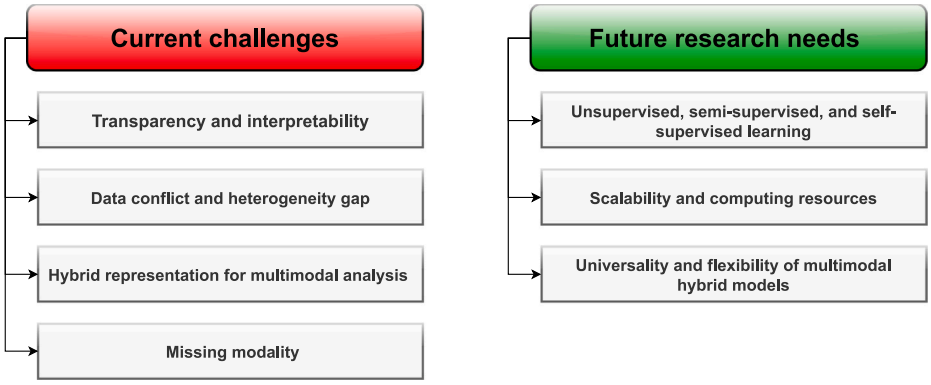


**Fig. 17.** Current challenges and future research needs in multimodal hybrid learning.

practical techniques and theories in environments that are closely related to real-world applications.

In practice, increasing the number of multimodal data streams under different acquisition conditions and ensuring a better trade-off between complex model structures and computational requirements remains one of the key challenges for researchers in the deep learning community. In light of this rapid growth, the current challenges facing multimodal hybrid learning can be summarized in the following subsections.

### *10.1.1. Transparency and interpretability*

Multimodal hybrid learning algorithms are central to the development of artificial intelligence technology and draw on the capabilities of real-world applications. However, these algorithms must ensure the transparency and interpretability of the decision process. In particular, hybrid models are inherently complex due to the numerous hidden layers and connections involved, as well as the challenge of interpreting their results. Therefore, the implications of such models need to be understood by practitioners and deep learning researchers. Conversely, hybrid models are easier to understand when the reasoning behind the decisions and the process by which the system arrived at its results are transparent and understandable.

### *10.1.2. Data conflict and heterogeneity gap*

Multimodal data is inherently redundant and often stored in different types of formats, making it difficult for hybrid models to extract

relevant feature information from heterogeneous data streams. Although the expressiveness of multimodal hybrid learning pipelines can be enhanced by integrating additional information from different data modalities and structures, learning such additional information requires a more sophisticated level of cognitive processing and may be influenced by very different intrinsic properties. For example, color images are composed of 3-RGB pixels, while text is composed of words or sentences of varying length and complexity. In particular, the visual content of images tends to be more comprehensible than text, and text tends to have more contextual and semantic meaning. Furthermore, hybrid models consist of several different and complex connections between heterogeneous components. However, these different components typically have different properties and behaviors, resulting in systems with different learning capabilities. Text embeddings are typically derived from a sequential model (temporal domain), while image features are typically extracted from a hierarchical backbone (spatial domain). This results in a heterogeneity gap, as these features are themselves heterogeneous structures and cannot be directly combined.

#### *10.1.3. Hybrid representation for multimodal analysis*

As mentioned above, multimodal hybrid learning systems can incorporate both intra- and intermodal information, allowing for dynamic interaction between multiple different models that interact in a complementary manner. However, having a universal representation of multiple modalities that can handle both multimodal and hybrid content while preserving intra- and intermodality relationships and hybrid structure remains the most challenging aspect. Therefore, multimodal hybrid learning algorithms must be able to consistently process the different input cues, summarize the extracted hybrid structures, and construct a joint feature map to efficiently learn the designed hybrid architecture. In addition, these algorithms must provide new and powerful tools to standardize data so that computers can learn useful representations in a multimodal setting.

#### *10.1.4. Missing modality*

Due to various constraints, including noise and sensor failure, multimodal hybrid applications often face partial or missing modality problems. In fact, missing modality constraints are extremely critical in multimodal modeling practice, but difficult to solve. Current approaches to address missing modalities in multimodal tasks mainly consider the modality during evaluation or train different hybrid models to handle different instances of the missing modality. For example, an emerging field, co-learning [212], is beneficial to support model learning by sharing or integrating knowledge from complementary modalities. Thus, multimodal hybrid learning algorithms should perform well in situations where multimodal noise is extremely high, or where one sensory modality is unintentionally missing, because they can integrate seamlessly.

### *10.2. Future research needs*

Multimodal hybrid learning holds great promise for improving computer vision tasks due to its ability to generate more accurate predictions. While much progress has been made on this cutting-edge topic in recent years, there are still important areas that have not been fully explored. In this subsection, we discuss some of the most important future research needs from different perspectives, including the scalability and flexibility of multimodal hybrid models.

#### *10.2.1. Unsupervised, semi-supervised, and self-supervised learning*

In general, supervised learning is the primary method for constructing most deep multimodal models, whose main goal is to jointly represent multiple input signals in a methodical way [7]. However, this learning method typically requires a large amount of labeled data, which may be difficult or even infeasible to collect. In cases where only a few labels are available, semi-supervised and unsupervised learning have been used to overcome this obstacle. Multimodal data can be used to enrich information from semi-supervised and unsupervised learning, providing complementary information that reveals fundamental features of high-dimensional data. Such multimodal data can also be used for self-supervised learning, where one input component is learned from another, resulting in a hybrid model.

#### *10.2.2. Scalability and computing resources*

To improve the overall performance of hybrid learning architectures, all aspects of real-time data analysis must be considered. There is a current trend to develop a new generation of hybrid architectures and to build new real-time processing systems that provide improved accuracy/efficiency. In computational neuroscience, for example, spiking neural networks (SNNs) are considered an effective alternative to traditional CNNs because of their ability to represent spiking neurons rather than the continuous-valued features that CNNs extract [213]. Due to their energy efficiency, SNNs are becoming increasingly powerful tools for many challenging tasks. For example, Garain et al. [214] recently developed an efficient SNN-based framework for COVID-19 screening in CT scans, demonstrating better classification performance than many state-of-the-art methods. Therefore, reducing computational resources remains a major challenge for the deep learning community. In particular, resource-intensive technologies include high-performance GPUs and storage systems. For example, edge/cloud computing-based solutions for multimodal analysis provide a straightforward way to process and handle multimodal sensory signals for effective training and deployment of hybrid models.

Currently, multimodal hybrid research uses relatively modest datasets that do not require distributed processing solutions. However, hybrid models typically require more bandwidth than a single CPU or GPU can provide for larger datasets. Therefore, while accelerated systems such as GPU clusters do not currently support multimodal deep learning, they may be beneficial in the future [140]. Nevertheless, multimodal hybrid frameworks with significant processing power will gain popularity in the coming years. Furthermore, significant advances in feature learning for multimodal data have been achieved through quantization and compression techniques [8,10], which can be used to increase the learning efficiency of hybrid models. Thus, a possible line of research could be how to build new compression techniques for multimodal hybrid learning by combining existing compression approaches.

#### *10.2.3. Universality and flexibility of multimodal hybrid models*

In the era of multimodal learning, deep hybrid models can be effectively used to improve the performance of various vision applications. However, the ubiquity and flexibility of these models remains of great interest for further generalization, so that knowledge from one domain can be successfully used to perform tasks in another domain with less data. Given the success of transfer learning, the process of building powerful multimodal hybrid models will be greatly facilitated by the rapid generation of pre-trained models. Ideally, future multimodal hybrid frameworks should be developed using well-defined architectural standards that can be adapted to any type of input modality and task.

With the rapid advances in multimodal and hybrid fusion technologies and the growing need for faster and more sophisticated decision-making systems, multimodal hybrid learning is expected to be widely applied in healthcare. For example, medical imaging is being used extensively to detect the COVID-19 pandemic and other global health challenges. However, there are a number of significant trade-offs, including the ability to understand how hybrid learning models represent the severity of COVID-19 in CT lung scans and other medical modalities. Clearly, there are broad areas of smart healthcare applications that should be explored [215].

## 11. Conclusion

Given the increasing role of multimodal hybrid learning in recent developments in artificial intelligence, this survey provided a comprehensive overview of this topic in the computer vision community. The survey focused on the forward-looking transformation between intra- and intermodal hybrid learning when dealing with hierarchical multidimensional data. In this context, we first briefly reviewed the history of DNNs, then summarized typical fusion algorithms in multimodal hybrid learning, provided a detailed overview of the most common hybrid architectures proposed in the recent literature (e.g., 2D-3D CNNs, CNN-RNN, CNN-DBN, etc.), and showed their potential applicability in computer vision. We also proposed a generic framework focusing on cross-modal information integration using multimodal hybrid fusion techniques, and analyzed the current challenges and future research directions of hybrid modeling in the multimodal domain.

Furthermore, it is important to consider that each multimodal hybrid method requires a specific fusion strategy. Nevertheless, deciding which fusion strategy is most appropriate for different scenarios remains a key challenge in computer vision. In summary, we expect the growth and development of the use of multimodal hybrid theory to be relevant in many active research areas in the future, especially in the fields of computer vision and NLP.

## CRediT authorship contribution statement

**Khaled Bayoudh:** Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

## 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.

## Data availability

No data was used for the research described in the article.

## References


[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[2] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: A new paradigm to machine learning, Arch Comput. Methods Eng. 27 (2020) 1071–1092.
[3] J. Chai, H. Zeng, A. Li, E.W.T. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Mach. Learn. Appl. 6 (2021) 100134.
[4] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 33 (2022) 6999–7019.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144.
[6] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 4037–4058.
[7] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets, Vis. Comput. 38 (2022) 2939–2970.
[8] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 423–443.
[9] B. Jena, S. Saxena, G.K. Nayak, L. Saba, N. Sharma, J.S. Suri, Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review, Comput. Biol. Med. 137 (2021) 104803.
[10] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multimodal data fusion, Neural Comput. 32 (2020) 829–864.
[11] Y. Zhang, D. Sidibé, O. Morel, F. Mériaudeau, Deep multimodal fusion for semantic image segmentation: A survey, Image Vis. Comput. 105 (2021) 104042.
[12] W. Guo, J. Wang, S. Wang, Deep multimodal representation learning: A survey, IEEE Access 7 (2019) 63373–63394.
[13] N. Rochester, J. Holland, L. Haibt, W. Duda, Tests on a cell assembly theory of the action of the brain, using a large digital computer, IRE Trans. Inf. Theory 2 (1956) 80–93.
[14] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408.
[15] K. Fukushima, Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernet. 36 (1980) 193–202.
[16] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021, arXiv preprint arXiv:2010.11929.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[19] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R.R. Martin, M.-M. Cheng, S.-M. Hu, Attention mechanisms in computer vision: A survey, Comp. Visual Media 8 (2022) 331–368.
[20] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[21] A. Makhzani, B. Frey, k-Sparse Autoencoders, 2014, arXiv preprint arXiv:1312.5663.
[22] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, 2008, pp. 1096–1103.
[23] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.
[24] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Machine Learn. Res. 11 (2010) 3371–3408.
[25] J. Zhao, M. Mathieu, Y. LeCun, Energy-Based Generative Adversarial Network, 2017, arXiv preprint arXiv:1609.03126.
[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2016.
[27] T. Che, Y. Li, A.P. Jacob, Y. Bengio, W. Li, Mode Regularized Generative Adversarial Networks, 2017, arXiv preprint arXiv:1612.02136.
[28] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[30] S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (IndRNN): Building a longer and deeper RNN, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5457–5466.
[31] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: 2014 Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734.
[32] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[34] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: 2005 Proceedings IEEE International Joint Conference on Neural Networks, vol. 2, 2005, pp. 729–734.
[35] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (2009) 61–80.
[36] S. Zhang, H. Tong, J. Xu, R. Maciejewski, Graph convolutional networks: a comprehensive review, Comput. Soc. Netw. 6 (2019) 11.
[37] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated Graph Sequence Neural Networks, 2017, arXiv preprint arXiv:1511.05493.
[38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph Attention Networks, 2018, arXiv preprint arXiv:1710.10903.
[39] H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in graph neural networks: A taxonomic survey, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–19.
[40] Y. Xie, Z. Xu, J. Zhang, Z. Wang, S. Ji, Self-supervised learning of graph neural networks: A unified review, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 2412–2429.
[41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, 2021, arXiv preprint arXiv:2012.12877.

[42] Z. Zong, K. Li, G. Song, Y. Wang, Y. Qiao, B. Leng, Y. Liu, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Self-Slimmed Vision Transformer, Springer Nature, Switzerland, Cham, 2022, pp. 432–448.

[43] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, 2014, pp. 372–378.

[44] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), Comput. Geosci. 19 (1993) 303–342.

[45] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimedia Tools Appl. 82 (2023) 3713–3744.

[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understandin, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[47] T. Georgiou, Y. Liu, W. Chen, M. Lew, A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision, Int. J. Multimed. Info. Retr. 9 (2020) 135–170.

[48] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, Z. Wang, A survey of information extraction based on deep learning, Appl. Sci. 12 (2022) 9691.

[49] M. Vasavi, A. Murugan, A survey on spatio-temporal data mining, Mater. Today: Proc. 80 (2023) 2769–2772.

[50] A. Hamdi, K. Shaban, A. Erradi, A. Mohamed, S.K. Rumi, F.D. Salim, Spatiotemporal data mining: a survey on challenges and open problems, Artif. Intell. Rev. 55 (2022) 1441–1488.

[51] A. Zhu, Q. Wu, R. Cui, T. Wang, W. Hang, G. Hua, H. Snoussi, Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN, Neurocomputing 414 (2020) 90–100.

[52] A.O.M. Abuassba, D. Zhang, X. Luo, A. Shaheryar, H. Ali, Improving classification performance through an advanced ensemble based heterogeneous extreme learning machines, Comput. Intell. Neurosci. 2017 (2017) 1–11.

[53] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, K. Olszewski, S. Minaee, D. Metaxas, S. Tulyakov, Show me what and tell me how: Video synthesis via multimodal conditioning, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3605–3615.

[54] N. Ahmed, Z.A. Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, Intell. Syst. Appl. 17 (2023) 200171.

[55] D. Hu, X. Li, X. Lu, Temporal multimodal learning in audiovisual speech recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3574–3582.

[56] X. Pei, K. Zuo, Y. Li, Z. Pang, A review of the application of multi-modal deep learning in medicine: Bibliometrics and future directions, Int. J. Comput. Intell. Syst. 16 (2023) 44.

[57] B. Nanay, Multimodal mental imagery, Cortex 105 (2018) 125–134.

[58] H. Kaur, D. Koundal, V. Kadyan, Image fusion techniques: A survey, Arch. Comput. Methods Eng. 28 (2021) 4425–4447.

[59] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Trans. Radiat. Plasma Med. Sci. 5 (2021) 741–760.

[60] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A.S. Albahri, B.S.N. Al-dabbagh, M.A. Fadhel, M. Manoufali, J. Zhang, A.H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, Y. Gu, A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications, J. Big Data 10 (2023) 46.

[61] J. Shah, D. Vaidya, M. Shah, A comprehensive review on multiple hybrid deep learning approaches for stock prediction, Intell. Syst. Appl. 16 (2022) 200111.

[62] C.N. Dang, M.N. Moreno-García, F. De la Prieta, Hybrid deep learning models for sentiment analysis, Complexity 2021 (2021) e9986920.

[63] Y. Shi, H. Feng, X. Geng, X. Tang, Y. Wang, A survey of hybrid deep learning methods for traffic flow prediction, in: Proceedings of the 2019 3rd International Conference on Advances in Image Processing, Association for Computing Machinery, New York, NY, USA, 2019, pp. 133–138.

[64] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, M. Linden, A comparative analysis of hybrid deep learning models for human activity recognition, Sensors 20 (2020) 5707.

[65] B. Alouffi, A. Alharbi, R. Sahal, H. Saleh, An optimized hybrid deep learning model to detect COVID-19 misleading information, Comput. Intell. Neurosci. 2021 (2021) e9615034.

[66] A. Al-Dulaimi, S. Zabihi, A. Asif, A. Mohammadi, A multimodal and hybrid deep neural network model for remaining useful life estimation, Comput. Ind. 108 (2019) 186–196.

[67] T. Zhou, Q. Li, H. Lu, X. Zhang, Q. Cheng, Hybrid multimodal medical image fusion method based on LatLRR and ED-D2GAN, Appl. Sci. 12 (2022) 12758.

[68] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, A. Raad, Reviewing multimodal machine learning and its use in cardiovascular diseases detection, Electronics 12 (2023) 1558.

[69] X. Qing, A comparison study of convolutional neural network and recurrent neural network on image classification, in: Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City, Association for Computing Machinery, New York, NY, USA, 2022, pp. 112–117.

[70] P. Verma, A. Selwal, D. Sharma, A survey on data-driven iris spoof detectors: state-of-the-art, open issues and future perspectives, Multimedia Tools Appl. 82 (2023) 19745–19792.

[71] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Front. Comput. Sci. 14 (2020) 241–258.

[72] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, J. King Saud Univ. - Comput. Inform. Sci. 35 (2023) 757–774.

[73] T. Iqball, M.A. Wani, Weighted ensemble model for image classification, Int. J. Inf. Tecnol. 15 (2023) 557–564.

[74] M. Shahhosseini, G. Hu, H. Pham, Optimizing ensemble weights and hyperparameters of machine learning models for regression problems, Mach. Learn. Appl. 7 (2022) 100251.

[75] I.D. Mienye, Y. Sun, A survey of ensemble learning: Concepts, algorithms, applications, and prospects, IEEE Access 10 (2022) 99129–99149.

[76] O. El Gannour, S. Hamida, B. Cherradi, M. Al-Sarem, A. Raihani, F. Saeed, M. Hadwan, Concatenation of pre-trained convolutional neural networks for enhanced COVID-19 screening using transfer learning technique, Electronics 11 (2022) 103.

[77] K. Bayoudh, F. Hamdaoui, A. Mtibaa, Hybrid-COVID: a novel hybrid 2D/3D CNN based on cross-domain adaptation approach for COVID-19 screening from chest X-ray images, Phys. Eng. Sci. Med. 43 (2020) 1415–1431.

[78] K. Bayoudh, F. Hamdaoui, A. Mtibaa, Transfer learning based hybrid 2D-3D CNN for traffic sign recognition and semantic road detection applied in advanced driver assistance systems, Appl. Intell. 51 (2021) 124–142.

[79] S.K. Roy, G. Krishna, S.R. Dubey, B.B. Chaudhuri, Hybridsn: Exploring 3D-2d CNN feature hierarchy for hyperspectral image classification, IEEE Geosci. Remote Sensing Lett. 17 (2020) 277–281.

[80] P.D. Chang, E. Kuoy, J. Grinband, B.D. Weinberg, M. Thompson, R. Homo, J. Chen, H. Abcede, M. Shafie, L. Sugrue, C.G. Filippi, M.-Y. Su, W. Yu, C. Hess, D. Chow, Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT, AJNR Am. J. Neuroradiol. 39 (2018) 1609–1616.

[81] X. Yang, X. Zhang, Y. Ye, R.Y.K. Lau, S. Lu, X. Li, X. Huang, Synergistic 2D/3D convolutional neural network for hyperspectral image classification, Remote Sens. 12 (2020) 2033.

[82] S.A. Vahora, N.C. Chauhan, Deep neural network model for group activity recognition using contextual relationship, Eng. Sci. Technol. Int. J. 22 (2019) 47–54.

[83] T.-H. Tan, J.-Y. Shih, S.-H. Liu, M. Alkhaleefah, Y.-L. Chang, M. Gochoo, Using a hybrid neural network and a regularized extreme learning machine for human activity recognition with smartphone and smartwatch, Sensors 23 (2023) 3354.

[84] R. Mutegeki, D.S. Han, A CNN-LSTM approach to human activity recognition, in: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 362–366.

[85] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An attention-based bidirectional CNN-rnn deep model for sentiment analysis, Future Gener. Comput. Syst. 115 (2021) 279–294.

[86] A.H. Ombabi, W. Ouarda, A.M. Alimi, Deep learning CNN–LSTM framework for arabic sentiment analysis using textual information shared in social networks, Soc. Netw. Anal. Min. 10 (2020) 53.

[87] A.U. Rehman, A.K. Malik, B. Raza, W. Ali, A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis, Multimedia Tools Appl. 78 (2019) 26597–26613.

[88] Y. Guo, Y. Liu, E.M. Bakker, Y. Guo, M.S. Lew, CNN-rnn: a large-scale hierarchical image classification framework, Multimedia Tools Appl. 77 (2018) 10251–10271.

[89] G. Liang, H. Hong, W. Xie, L. Zheng, Combining convolutional neural network with recursive neural network for blood cell image classification, IEEE Access 6 (2018) 36188–36197.

[90] Md.Z. Islam, Md.M. Islam, A. Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Inform. Med. Unlocked 20 (2020) 100412.

[91] A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, J.J.P.C. Rodrigues, An integrated hybrid CNN–RNN model for visual description and generation of captions, Circuits Syst. Signal Process 39 (2020) 776–788.

[92] Y. Chu, X. Yue, L. Yu, M. Sergei, Z. Wang, Automatic image captioning based on ResNet50 and LSTM with soft attention, Wirel. Commun. Mob. Comput. 2020 (2020) e8909458.

[93] S. Dong, Z. Zhang, G. Wen, S. Dong, Z. Zhang, G. Wen, Design and application of unsupervised convolutional neural networks integrated with deep belief networks for mechanical fault diagnosis, in: 2017 Prognostics and System Health Management Conference (PHM-Harbin), 2017, pp. 1–7.

[94] Y. Li, L. Zou, L. Jiang, X. Zhou, Fault diagnosis of rotating machinery based on combination of deep belief network and one-dimensional convolutional neural network, IEEE Access 7 (2019) 165710–165723.

[95] Q. Lin, Music score recognition method based on deep learning, Comput. Intell. Neurosci. 2022 (2022) e3022767.

[96] C. Li, D. Zhao, S. Mu, W. Zhang, N. Shi, L. Li, Fault diagnosis for distillation process based on CNN–DAE, Chin. J. Chem. Eng. 27 (2019) 598–604.

[97] W. Wang, M. Zhao, J. Wang, Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network, J. Ambient Intell. Human Comput. 10 (2019) 3035–3043.
[98] J. Jiang, H. Zhang, D. Pi, C. Dai, A novel multi-module neural network system for imbalanced heartbeats classification, Expert Syst. Appl. X 1 (2019) 100003.
[99] K. Bajaj, D.K. Singh, Mohd.A. Ansari, Autoencoders based deep learner for image denoising, Procedia Comput. Sci. 171 (2020) 1535–1541.
[100] Z. Liang, R. Zhou, L. Zhang, L. Li, G. Huang, Z. Zhang, S. Ishii, EegFuseNet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 1913–1925.
[101] A. Cheng, PAC-GAN: Packet generation of network traffic using generative adversarial networks, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0728–0734.
[102] H. Yin, Z. Li, J. Zuo, H. Liu, K. Yang, F. Li, Wasserstein generative adversarial network and convolutional neural network (WG-CNN) for bearing fault diagnosis, Math. Probl. Eng. 2020 (2020) e2604191.
[103] W. Fang, Y. Ding, F. Zhang, J. Sheng, Gesture recognition based on CNN and DCGAN for calculation and text output, IEEE Access 7 (2019) 28230–28237.
[104] R.K. Tan, N.L. Zhang, W. Ye, A deep learning–based method for the design of microstructural materials, Struct. Multidiscip. Optim. 61 (2020) 1417–1438.
[105] A. Luo, F. Yang, X. Li, D. Nie, Z. Jiao, S. Zhou, H. Cheng, Hybrid graph neural networks for crowd counting, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 11693–11700.
[106] Y. Li, R. Chen, Y. Zhang, M. Zhang, L. Chen, Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network, Remote Sens. 12 (2020) 4003.
[107] M. Munir, W. Avery, R. Marculescu, MobileViG: Graph-based sparse attention for mobile vision applications, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2210–2218.
[108] Y. Barhoumi, G. Rasool, Scopeformer: N-CNN-ViT Hybrid Model for Intracranial Hemorrhage Classification, 2021, arXiv preprint arXiv:2107.04575.
[109] Y. Zhang, Z. Li, N. Nan, X. Wang, TranSegNet: Hybrid CNN-vision transformers encoder for retina segmentation of optical coherence tomography, Life (Basel) 13 (2023) 976.
[110] G. Wang, H. Chen, L. Chen, Y. Zhuang, S. Zhang, T. Zhang, H. Dong, P. Gao, P2fevit: Plug-and-play CNN feature embedded hybrid vision transformer for remote sensing image classification, Remote Sens. 15 (2023) 1773.
[111] Y. Dai, Y. Gao, F. Liu, J. Fu, Mutual Attention-Based Hybrid Dimensional Network for Multimodal Imaging Computer-Aided Diagnosis, 2022, arXiv preprint arXiv:2201.09421.
[112] O. Vynokurova, D. Peleshko, Hybrid multidimensional deep convolutional neural network for multimodal fusion, in: 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 2020, pp. 131–135.
[113] B. Mocanu, R. Tapu, T. Zaharia, Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning, Image Vision Comput. 133 (2023) 104676.
[114] I. Merino, J. Azpiazu, A. Remazeilles, B. Sierra, 3D convolutional neural networks initialized from pretrained 2D convolutional neural networks for classification of industrial parts, Sensors 21 (2021) 1078.
[115] M. Lv, W. Xu, T. Chen, A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors, Neurocomputing 362 (2019) 33–40.
[116] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, S.-F. Chang, Modeling multimodal clues in a hybrid deep learning framework for video classification, IEEE Trans. Multimed. 20 (2018) 3137–3147.
[117] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA, 2016, pp. 445–450.
[118] D. Liu, Z. Wang, L. Wang, L. Chen, Multi-modal fusion emotion recognition method of speech expression based on deep learning, Front. Neurorobotics 15 (2021).
[119] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, C. Pal, Recurrent neural networks for emotion recognition in video, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA, 2015, pp. 467–474.
[120] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (2018) 3030–3043.
[121] X. Pan, H.-B. Shen, RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach, BMC Bioinformatics 18 (2017) 136.
[122] S. Zhang, X. Pan, Y. Cui, X. Zhao, L. Liu, Learning affective video features for facial expression recognition via hybrid deep learning, IEEE Access 7 (2019) 32297–32304.
[123] D. Nguyen, D.T. Nguyen, R. Zeng, T.T. Nguyen, S.N. Tran, T. Nguyen, S. Sridharan, C. Fookes, Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition, IEEE Trans. Multimed. 24 (2022) 1313–1324.
[124] P. Koromilas, T. Giannakopoulos, Unsupervised Multimodal Language Representations using Convolutional Autoencoders, 2022, arXiv preprint arXiv:2110.03007.
[125] X. Liu, M. Wang, Z.-J. Zha, R. Hong, Cross-modality feature learning via convolutional autoencoder, ACM Trans. Multimedia Comput. Commun. Appl. 15 (2019) 7:1–7:20.
[126] M. Abdelaziz, T. Wang, A. Elazab, Fusing multimodal and anatomical volumes of interest features using convolutional auto-encoder and convolutional neural networks for alzheimer's disease diagnosis, Front. Aging Neurosci. 14 (2022).
[127] A. Magassouba, K. Sugiura, H. Kawai, A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions, IEEE Robot. Autom. Lett. 3 (2018) 3113–3120.
[128] D. Hong, J. Yao, D. Meng, Z. Xu, J. Chanussot, Multimodal GANs: Toward cross-modal hyperspectral–multispectral image segmentation, IEEE Trans. Geosci. Remote Sens. 59 (2021) 5103–5113.
[129] L. Guo, J. Liu, P. Yao, J. Li, H. Lu, Mscap: multi-style image captioning with unpaired stylized text, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4199–4208.
[130] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1437–1445.
[131] D. Gao, K. Li, R. Wang, S. Shan, X. Chen, Multi-modal graph neural network for joint reasoning on vision and scene text, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12743–12753.
[132] S. Rudinac, I. Gornishka, M. Worring, Multimodal classification of violent online political extremism content with graph convolutional networks, in: Proceedings of the on Thematic Workshops of ACM Multimedia, vol. 2017, Association for Computing Machinery, New York, NY, USA, 2017, pp. 245–252.
[133] Y. Yuan, J. Wu, Z. Jing, H. Leung, H. Pan, Multimodal Image Fusion Based on Hybrid CNN-Transformer and Non-Local Cross-Modal Attention, 2022, arXiv preprint arXiv:2210.09847.
[134] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002.
[135] Q. Sun, N. Fang, Z. Liu, L. Zhao, Y. Wen, H. Lin, HybridCTrm: Bridging CNN and transformer for multimodal brain image segmentation, J. Healthcare Eng. 2021 (2021) e7467261.
[136] Q. Zhou, S. Ye, M. Wen, Z. Huang, M. Ding, X. Zhang, Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer, Neural Comput. Appl. 34 (2022) 21741–21761.
[137] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: Bagging-boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. C 42 (2012) 463–484.
[138] B.A. Ture, A. Akbulut, A.H. Zaim, C. Catal, Stacking-based ensemble learning for remaining useful life estimation, Softw. Comput. (2023).
[139] M. Lu, Q. Hou, S. Qin, L. Zhou, D. Hua, X. Wang, L. Cheng, A stacking ensemble model of various machine learning models for daily runoff forecasting, Water 15 (2023) 1265.
[140] W.C. Sleeman, R. Kapoor, P. Ghosh, Multimodal classification: Current landscape, taxonomy and future directions, ACM Comput. Surv. 55 (2022) 150:1–150:31.
[141] S. Pawar, O. San, P. Vedula, A. Rasheed, T. Kvamsdal, Multi-fidelity information fusion with concatenated neural networks, Sci. Rep. 12 (2022) 5900.
[142] A. Khan, A. Sohail, U. Zahoora, A.S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, Artif. Intell. Rev. 53 (2020) 5455–5516.
[143] T. Deng, A survey of convolutional neural networks for image classification: Models and datasets, in: 2022 International Conference on Big Data, Information and Computer Network (BDICN), 2022, pp. 746–749.
[144] S.S.A. Zaidi, M.S. Ansari, A. Aslam, N. Kanwal, M. Asghar, B. Lee, A survey of modern deep learning based object detection models, Digit. Signal Process. 126 (2022) 103514.
[145] F. Lateef, Y. Ruichek, Survey on semantic segmentation using deep learning techniques, Neurocomputing 338 (2019) 321–348.
[146] H. Song, Y. Wen, A survey of convolutional neural network and its variants, in: Proceedings of the 10th International Conference on Computer and Communications Management, Association for Computing Machinery, New York, NY, USA, 2022, pp. 37–45.
[147] H. Lu, H. Wang, Q. Zhang, S.W. Yoon, D. Won, A 3D convolutional neural network for volumetric image semantic segmentation, Procedia Manuf. 39 (2019) 422–428.
[148] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, 2009, pp. 609–616.
[149] Q. Abbas, M.E.A. Ibrahim, M.A. Jaffar, A comprehensive review of recent advances on deep vision systems, Artif. Intell. Rev. 52 (2019) 39–76.

[150] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in: T. Honkela, W. Duch, M. Girolami, S. Kaski (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2011, Springer, Berlin, Heidelberg, 2011, pp. 52–59.

[151] P. Li, Y. Pei, J. Li, A comprehensive survey on design and application of autoencoder in deep learning, Appl. Soft Comput. 138 (2023) 110176.

[152] X.-J. Mao, C. Shen, Y.-B. Yang, Image Restoration using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections, 2016, arXiv preprint arXiv:1603.09056.

[153] E. Rodríguez, B. Otero, N. Gutiérrez, R. Canal, A survey of deep learning techniques for cybersecurity in mobile networks, IEEE Commun. Surv. Tutor. 23 (2021) 1920–1955.

[154] S.S. Roy, M. Ahmed, M.A.H. Akhand, Classification of massive noisy image using auto-encoders and convolutional neural network, in: 2017 8th International Conference on Information Technology (ICIT), 2017, pp. 971–979.

[155] M.H. Mohd Noor, Feature learning using convolutional denoising autoencoder for activity recognition, Neural Comput. Appl. 33 (2021) 10909–10922.

[156] L. Kumar, D.K. Singh, A comprehensive survey on generative adversarial networks used for synthesizing multimedia content, Multimed. Tools Appl. (2023).

[157] M.R. Pavan Kumar, P. Jayagopal, Generative adversarial networks: a survey on applications and challenges, Int. J. Multimed. Info Retr. 10 (2021) 1–24.

[158] Y. Gao, P. Zhai, K.M. Mosalam, Balanced semisupervised generative adversarial network for damage assessment from low-data imbalanced-class regime, Comput.-Aided Civ. Infrastruct. Eng. 36 (2021) 1094–1113.

[159] A. Bousmina, M. Selmi, M.A. Ben Rhaiem, I.R. Farah, A hybrid approach based on GAN and CNN-LSTM for aerial activity recognition, Remote Sens. 15 (2023) 3626.

[160] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein GANs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc. Red Hook, NY, USA, 2017, pp. 5769–5779.

[161] Q.V. Le, T. Mikolov, Distributed Representations of Sentences and Documents, 2014, arXiv preprint arXiv:1405.4053.

[162] N. Audebert, B. Le Saux, S. Lefèvre, Semantic segmentation of earth observation data using multimodal and multi-scale deep networks, in: S.-H. Lai, V. Lepetit, K. Nishino, Y. Sato (Eds.), Computer Vision – ACCV 2016, Springer International Publishing, Cham, 2016, pp. 180–196.

[163] L. Waikhom, R. Patgiri, A survey of graph neural networks in various learning paradigms: methods, applications, and challenges, Artif. Intell. Rev. 56 (2023) 6295–6364.

[164] J. Chen, T. Ma, C. Xiao, FastGCN: Fast Learning with Graph Convolutional Networks Via Importance Sampling, 2018, arXiv preprint arXiv:1801.10247.

[165] U.A. Bhatti, H. Tang, G. Wu, S. Marjan, A. Hussain, Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence, Int. J. Intell. Syst. 2023 (2023) e8342104.

[166] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, M. Zitnik, Multimodal learning with graphs, Nat. Mach. Intell. 5 (2023) 340–350.

[167] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 87–110.

[168] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. 54 (2022) 200:1–200:41.

[169] P.P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Int. Things Cyber-Phys. Syst. 3 (2023) 121–154.

[170] A. Khan, Z. Rauf, A. Sohail, A. Rehman, H. Asif, A. Asif, U. Farooq, A survey of the Vision Transformers and its CNN-Transformer based Variants, 2023, arXiv preprint arXiv:2305.09880.

[171] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing convolutions to vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 22–31.

[172] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976.

[173] S. Mehta, M. Rastegari, MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer, 2021, arXiv preprint arXiv:2110.02178.

[174] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569.

[175] X. Han, Y.-T. Wang, J.-L. Feng, C. Deng, Z.-H. Chen, Y.-A. Huang, H. Su, L. Hu, P.-W. Hu, A survey of transformer-based multimodal pre-trained modals, Neurocomputing 515 (2023) 89–106.

[176] S. He, H. Yang, X. Zhang, X. Li, MfTransNet: A multi-modal fusion with CNN-transformer network for semantic segmentation of HSR remote sensing images, Mathematics 11 (2023) 722.

[177] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (2021) 43–76.

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

[179] S. Das, S.D. Roy, S. Malakar, J.D. Velásquez, R. Sarkar, Bi-level prediction model for screening COVID-19 patients using chest X-ray images, Big Data Res. 25 (2021) 100233.

[180] M. Zabin, H.-J. Choi, J. Uddin, Hybrid deep transfer learning architecture for industrial fault diagnosis using Hilbert transform and DCNN–LSTM, J. Supercomput. 79 (2023) 5181–5200.

[181] N.A. Samee, A.A. Alhussan, V.F. Ghoneim, G. Atteia, R. Alkanhel, M.A. Al-Antari, Y.M. Kadah, A hybrid deep transfer learning of CNN-based LR-PCA for breast lesion diagnosis via medical breast mammograms, Sensors (Basel) 22 (2022) 4938.

[182] Y. Yang, H. Lv, N. Chen, A survey on ensemble learning under the era of deep learning, Artif. Intell. Rev. 56 (2023) 5545–5589.

[183] M. Zambelli, Y. Demirisy, Online multimodal ensemble learning using self-learned sensorimotor representations, IEEE Trans. Cogn. Dev. Syst. 9 (2017) 113–126.

[184] S. Jain, A. Saha, Improving performance with hybrid feature selection and ensemble machine learning techniques for code smell detection, Sci. Comput. Program. 212 (2021) 102713.

[185] A. Tharwat, W. Schenck, A survey on active learning: State-of-the-art, Pract. Chall. Res. Dir. Math. 11 (2023) 820.

[186] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B.B. Gupta, X. Chen, X. Wang, A survey of deep active learning, ACM Comput. Surv. 54 (2021) 180.

[187] Y.-P. Tang, S.-J. Huang, Active learning for multiple target models, Adv. Neural Inf. Process. Syst. 35 (2022) 38424–38435.

[188] O. Rudovic, M. Zhang, B. Schuller, R. Picard, Multi-modal active learning from human data: A deep reinforcement learning approach, in: 2019 International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA, 2019, pp. 6–15.

[189] S.C.H. Hoi, D. Sahoo, J. Lu, P. Zhao, Online learning: A comprehensive survey, Neurocomputing 459 (2021) 249–289.

[190] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowl.-Based Syst. 216 (2021) 106775.

[191] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, W. Zhang, A survey on federated learning: challenges and applications, Int. J. Mach. Learn. Cyber 14 (2023) 513–535.

[192] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, B. He, A survey on federated learning systems: Vision, hype and reality for data privacy and protection, IEEE Trans. Knowl. Data Eng. 35 (2023) 3347–3366.

[193] A. Brecko, E. Kajati, J. Koziorek, I. Zolotova, Federated learning for edge computing: A survey, Appl. Sci. 12 (2022) 9124.

[194] Y.-M. Lin, Y. Gao, M.-G. Gong, S.-J. Zhang, Y.-Q. Zhang, Z.-Y. Li, Federated learning on multimodal data: A comprehensive survey, Mach. Intell. Res. 20 (2023) 539–553.

[195] Q. Yu, Y. Liu, Y. Wang, K. Xu, J. Liu, Multimodal Federated Learning Via Contrastive Representation Ensemble, 2023, arXiv preprint arXiv:2302.08888.

[196] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (2022) 5586–5609.

[197] S. Sosnin, M. Vashurina, M. Withnall, P. Karpov, M. Fedorov, I.V. Tetko, A survey of multi-task learning methods in chemoinformatics, Mol. Inform. 38 (2019) 1800108.

[198] M. Crawshaw, Multi-Task Learning with Deep Neural Networks: A Survey, 2020, arXiv preprint arXiv:2009.09796.

[199] Y. Jin, T. Zheng, C. Gao, G. Xu, MTMSN: Multi-task and multi-modal sequence network for facial action unit and expression recognition, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3590–3595.

[200] R. Hu, A. Singh, Unit: Multimodal multitask learning with a unified transformer, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1419–1429.

[201] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron Mark. 31 (2021) 685–695.

[202] U. Mall, B. Hariharan, K. Bala, Zero-shot learning using multimodal descriptions, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 3930–3938.

[203] L. Zhou, S. Pan, J. Wang, A.V. Vasilakos, Machine learning on big data: Opportunities and challenges, Neurocomputing 237 (2017) 350–361.

[204] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE Trans. Knowl. Data Eng. 31 (2019) 1863–1883.

[205] S. Sun, A survey of multi-view machine learning, Neural Comput. Appl. 23 (2013) 2031–2038.

[206] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Inf. Fusion 38 (2017) 43–54.

[207] Z. Xie, Y. Yang, Y. Zhang, J. Wang, S. Du, Deep learning on multi-view sequential data: a survey, Artif. Intell. Rev. 56 (2023) 6661–6704.

[208] Y. Gu, J. Yang, G.-Z. Yang, Multi-view multi-modal feature embedding for endomicroscopy mosaic classification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, pp. 1315–1323.

[209] J. Duan, L. Chen, S. Tran, J. Yang, Y. Xu, B. Zeng, T. Chilimbi, Multi-Modal Alignment using Representation Codebook, IEEE Computer Society, 2022, pp. 15630–15639.
[210] L. Huang, W. Wang, Y. Xia, J. Chen, Adaptively aligned image captioning via adaptive attention time, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2019, pp. 8942–8951.
[211] S. Haresh, S. Kumar, H. Coskun, S.N. Syed, A. Konin, M.Z. Zia, Q.-H. Tran, Learning by aligning videos in time, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5544–5554.
[212] A. Rahate, R. Walambe, S. Ramanna, K. Kotecha, Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions, Inf. Fusion 81 (2022) 203–239.
[213] M. Bouvier, A. Valentian, T. Mesquida, F. Rummens, M. Reyboz, E. Vianello, E. Beigne, Spiking neural networks hardware implementations and challenges: A survey, J. Emerg. Technol. Comput. Syst. 15 (2019) 22.
[214] A. Garain, A. Basu, F. Giampaolo, J.D. Velásquez, R. Sarkar, Detection of COVID-19 from CT scan images: A spiking neural network-based approach, Neural Comput. Appl. 33 (2021) 12591–12604.
[215] T. Shaik, X. Tao, L. Li, H. Xie, J.D. Velásquez, A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom, Inf. Fusion 102 (2024) 102040.