Check for updates

# Multimodal Classification: Current Landscape, Taxonomy and Future Directions

WILLIAM C. SLEEMAN IV, RISHABH KAPOOR, and PREETAM GHOSH,
Virginia Commonwealth University, USA

Multimodal classification research has been gaining popularity with new datasets in domains such as satellite imagery, biometrics, and medicine. Prior research has shown the benefits of combining data from multiple sources compared to traditional unimodal data that has led to the development of many novel multimodal architectures. However, the lack of consistent terminologies and architectural descriptions makes it difficult to compare different solutions. We address these challenges by proposing a new taxonomy for describing multimodal classification models based on trends found in recent publications. Examples of how this taxonomy could be applied to existing models are presented as well as a checklist to aid in the clear and complete presentation of future models. Many of the most difficult aspects of unimodal classification have not yet been fully addressed for multimodal datasets, including big data, class imbalance, and instance-level difficulty. We also provide a discussion of these challenges and future directions of research.

CCS Concepts: • **Computing methodologies** → **Classification and regression trees**; **Neural networks;**

Additional Key Words and Phrases: Multimodal learning, neural networks, machine learning, classification

150

**ACM Reference format:**
William C. Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. 2022. Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *ACM Comput. Surv.* 55, 7, Article 150 (December 2022), 31 pages.
https://doi.org/10.1145/3543848

## 1 INTRODUCTION

In the past decade, there has been an increased focus on combining data from multiple modalities to further improve machine learning–based classification models. Data is becoming more important to every sector of business and research, which has led to the creation of larger and more diverse datasets. The ever-increasing rate of data collection and the reported benefits of multimodal data for machine learning modals has fueled interest in this field. By using information from several representations of the same subject, a more complete picture of the problem at hand can be constructed. Multiple data modalities are naturally present in many problem domains such as medicine [27, 35, 40, 49], hyperspatial imagery [7, 30], sentiment analysis [14, 83, 89], and many others.

Authors' address: W. C. Sleeman IV, R. Kapoor, and P. Ghosh, Virginia Commonwealth University, 601 West Main Street, Richmond, Virginia, 23284-3068; emails: wcsleeman@vcu.edu, rishabh.kapoor@vcuhealth.org, pghosh@vcu.edu.

0360-0300/2022/12-ART150 $15.00
https://doi.org/10.1145/3543848

The majority of existing classification algorithms were designed for unimodal datasets that represent a single data source for each specific problem. These datasets typically use only one data type such as tabular, images, or text but many real-world scenarios include data of mixed types. Price prediction datasets may include both text from news articles and tabular data from financial reports, or medical records may have both signal-based heart monitoring data and diagnostic imaging. Combining data modalities can be challenging when each individual representation is significantly different with combinations like image-text, audio-video, or multiple sensors that are not time-synchronized. These challenges have led to solutions that utilize unimodal algorithms to solve multimodal problems.

The increasing interest in multimodal learning has led to a number of recent survey papers covering entire domains [9, 59, 122, 124], with many of these surveys focusing on deep learning [32, 77, 115], domain-specific solutions [7, 27, 29], or non-classification methods [35, 49, 78]. While these works cover the breadth of multimodal learning, there are no surveys that specifically investigate classification problems and their unique properties. This article has been guided by the following motivations:

**(1) The lack of a specific multimodal classification taxonomy**
Although several taxonomies have been previously presented, they are directed at multimodal learning in whole instead of classification. For example, the work by Baltrušaitis et al. [9] addresses many types of learning such as image captioning, video descriptions, text-to-image conversions, co-training, transfer learning, and zero-shot learning. While that taxonomy can be applied to almost any multimodal problem, it is not specific enough to fully describe the recent multimodal classification architectures.

**(2) Identifying recent trends in model architectures**
Since the prior surveys have focused on the high-level aspects of multimodal learning or domain-specific problems, there has not been a review of recent classification models and their architectures. A comparison of the architectures is needed to identify current trends and how they could be described with a common taxonomy.

**(3) Providing a way to describe multimodal classification architectures**
While reviewing multimodal classification papers, we observed that a significant amount of work was required to decompose many of the model architectures. Each model used its own set of terms and method of presentation, so reviewing each paper was a new experience that made the process more difficult. Paired with the taxonomy, a common descriptive framework is needed to make model architecture depictions and comparisons easier.

**(4) Discussion of future challenges**
Issues related to big data, distributed computing, and difficult datasets have been well studied with unimodal problems, but limited work has been done for multimodal learning. A discussion on how these challenges could affect multimodal classification is needed.

The rest of this article is organized as follows: Section 2 provides a brief overview of multimodal learning and examples of existing domain-specific solutions. Section 3 gives descriptions of multimodal classification architectures and our proposed taxonomy, and Section 4 reviews recent multimodal classification research using these common terms, which addresses the motivations 1 and 2 above. Section 5 gives examples of how to apply the taxonomy on both existing and future models for motivation 3. Section 6 discusses the challenges with classification that have not been addressed for multimodal problems as mentioned in motivation 4. Finally, Section 7 provides our concluding remarks. In Table 1, we introduce the list of abbreviations that will be frequently used in the rest of this article.

Table 1. A List of Terms Used to Describe Algorithms or Methods Used in Multimodal Classification Architectures

| List of Terms | |
|---|---|
| **Term** | **Description** |
| AE | Autoencoder |
| AN | Attention Network |
| CCA | Canonical Correlation Analysis |
| CNN | Convolutional Neural Network |
| DBN | Deep Belief Network |
| DL | Deep Learning |
| DNN | Deep Neural Network |
| DT | Decision Tree |
| ELM | Extreme Learning Machine |
| FCNN | Fully Connected Neural Network |
| GAN | Generative Adversarial Networks |
| GBT | Gradient Boosted Trees |
| GRU | Gated Recurrent Unit |
| *k*NN | *k*-Nearest Neighbors |
| LR | Linear Regression |
| LSTM | Long Short-Term Memory |
| Manual | Feature extraction performed by hand or with a manually chosen method |
| MKL | Multiple Kernel Learning |
| ML | Machine Learning |
| MLP | Multilayer Perceptron |
| Multi-SVNN | Multi Support Vector Neural Network |
| RF | Random Forest |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |
| RNN | Recurrent Neural Network |
| WE | Word Embedding |
| XGBoost | Extreme Gradient Boosting |

## 2 PREVIOUS MULTIMODAL RESEARCH

The potential benefit of utilizing information from multiple data sources has led to many recent papers focusing on multimodal learning. In this section, we review core concepts of multimodal learning, taxonomies, and domain-specific research.

However, we must first discuss the terms *multimodal* and *multi-view,* as both are commonly used throughout the literature. These terms are often used interchangeably as learning systems that incorporate information from multiple sources, usually to differentiate from traditional problems using a single (unimodal) data source. Although these terms are used to describe learning models that explicitly combine multiple data sources, *multi-view* appears to be more commonly associated with different kinds of algorithms, such as co-training [92, 124], **Canonical Correlation Analysis (CCA)**-based cross-modality retrieval [59, 124], semi-supervised [124], clustering [15, 50], unsupervised feature selection [99, 100], and subspace learning [110].

For the sake of consistency, we have chosen to use the term *multimodal* to describe learning algorithms that include information from multiple data sources for the purpose of improving predictive performance. Other use cases, such as transfer learning or co-training, could then be described as *multi-view* to provide some distinction between these learning approaches. While we believe these definitions would be useful to the research community, further discussion will be required to form a consensus.

### 2.1 Multimodal Concepts

Several surveys have identified co-training [92, 110, 124] and co-regularization [92, 124] as major categories of multimodal learning. Co-training is used in semi-supervised problems where a mix of labeled and unlabeled data is present where knowledge from one modality can be used to support a model trained on another modality. Potential applications of co-training include zero-shot learning, transfer learning, and the annotation of unlabeled examples. Co-regularization transforms each modality to ensure that they are compatible. This can be achieved with techniques such as CCA or regularizing modality-specific classifiers.

A more comprehensive work was later presented by Baltrušaitis et al. [9] that identified five challenges with multimodal learning: representation [9, 122], translation [9], alignment [9, 59, 92], fusion [9, 59, 92, 122], and co-learning [9, 124]. A number of other multimodal learning tasks have also been investigated including semi-supervised learning [92, 110, 124], encoding [25], clustering [92, 124], and multi-task learning [55, 124]. Although many of these surveys address aspects of classification problems, such as alignment and fusion, they tend to cover a much wider range of topics.

### 2.2 Taxonomies

Several review papers have presented taxonomies, often targeting specific problem domains. Di Mitri et al. [19] investigated the use of multimodal data for learning human behavior with sensors. A feedback system called **Multimodal Learning Analysis Model (MLeAM)** used multimodal sensor data, manual annotation, and machine learning to guide behavioral change. The authors also provided a tree structure taxonomy that described different data sources that could be acquired with sensors. Garcia-Ceja et al. [27] also reviewed sensor-based systems addressing mental health issues. Their taxonomy addressed the study type, study duration, and the sensing types.

Yan et al. [115] reviewed multimodal methods in the context of deep learning. Their taxonomy separated algorithms depending on if they were native to deep learning (i.e., CNN, GAN) or traditional machine learning techniques adapted to deep learning (i.e., CCA, spectral clustering). This review also discussed different network fusion strategies, bimodal auto-encoders, and GAN-based methods.

Jiang et al. [49] provided a comprehensive review of multimodal image matching and registration techniques that correlated the same concept across multiple images. They also identified two major classes of solutions: area-based and feature-based image registration. Using intensity information of the entire image to find matching regions, area-based methods produce a transformation that can be used for the image registration. Feature-based registration methods include feature detection (corner, blob, and learnable features), feature description (float, binary, and learnable descriptors), and feature matching (graph matching, point set registration, indirect methods).

Ramachandram and Taylor [77] provided a taxonomy with a review of deep learning multimodal research. In their taxonomy, models were described by their input modalities, problem space, fusion method, model type, and architecture. Input modalities included audio, video, images, text, and others specific to the field of medicine. The problem space covered domains such as action recognition, medical diagnosis, and robot grasping. The fusion method described how data from each modality was combined and used the terms *early*, *intermediate*, and *late*. The model type was either *generative* or *discriminate*, with the architecture defined as the actual model used, such as CNN, RNN, or LSTM.

We next provide a brief overview of the multimodal taxonomy proposed by Baltrušaitis et al. [9] that included the learning concep ts of representation, translation, alignment, fusion, and co-learning.

**Representation** is described as how data from each modality is presented as feature vectors. Since data can be text, image, or video, the potential heterogeneity may introduce additional complexity to learning models. The representation challenges were grouped as *Joint* or *Coordinated*.

*Joint* representations combine data from multiple modalities to create a single representation. This can be achieved with neural networks by concatenating modality-specific layers to create a new hidden layer. Probabilistic graphical models like deep Boltzmann machines can be used to create representations from latent space, also allowing for the generation of missing data from one of the modalities. Sequential representations are used for handling variable length data, such as sentences or audio clips, often with RNNs.

*Coordinated* representations are learned using similarities between modalities and enforced constraints. Similarity models can force representations of each modality to be close to each other, such as the word "car" should be closer to a picture of car than that of a boat. Structured coordinate spaces use hashing-based compression, which constrains the placement of embedded modality data.

**Translation** is used to map one modality to another, such as generating text captions from image or video data. These tasks were categorized as *Example-based* or *Generative*.

*Example-based* translations use dictionary look-ups to find a matching value in another modality. In addition to dictionaries, *k*-nearest neighbor searches have been used to perform consensus-based retrievals. Both approaches are restricted by the specific modality data they have at training.

*Generative* translations can create new translated values from the source modality instead of simple retrieval. Grammar-based solutions can create text for the target modality using high-level concepts in the source modality but only within the predefined grammatical rules. Encoder-decoder networks encode source modality data that can then be decoded into examples in the target modality. Tasks like speech-to-text can utilize continuous generation modals by sampling a latent space common to both modalities.

**Alignment** finds corresponding sub-components between each modality. This is often done for multimedia retrieval such as syncing audio with video frames or marking images that include a specific individual. The alignment techniques were identified as *Explicit* or *Implicit*.

*Explicit* alignment is used when the goal is to align multiple modalities bases on related components. Unsupervised alignments do not use labels but rely on similarity metrics, such as matching gene sequences. Supervised learning methods can also be used if the modality alignments are labeled.

*Implicit* alignment is used when the specific alignment is not known and has been used with tasks such as speech recognition and translation. One approach has been to use graphical models where the structure of language relationships is mapped to audio data. Neural networks using encoder-decoder and attention-based models have been used to align audio-video data using a latent space.

**Fusion** is the method for combining data from multiple modalities before applying a learning algorithm. Data fusion is a core concept of all multimodal approaches and was grouped into *Model-agnostic* and *Model-based* solutions.

*Model-agnostic* approaches use unimodal classifiers with early, late, and hybrid fusion techniques, which has also been discussed by Di Mitri et al. [19] and Simonetta et al. [86]. Early fusion combines modality data before classification, late fusion performs modality-specific learning before the results are combined, and hybrid fusion uses a combination of both.

*Model-based* approaches are designed to address modality fusion more directly than the Model-agnostic methods, which do not take in account inter-modality relationships. Multiple kernel learning, deep belief networks, and neural network models have all been used for multimodal fusion while considering all of the modalities.

**Co-learning** uses knowledge from one modality to support the learning of a different modality. This approach can address missing or low-quality data in one modality by using high-quality data from another. Co-learning methods were identified as *Parallel*, *Non-parallel*, and *Hybrid*.

*Parallel* methods use data from examples shared across multiple modalities at the same time. Co-training uses information from a well-labeled modality to generate missing labels for another modality. Transfer learning can use data from one parallel model to perform new but similar tasks.

*Non-parallel* methods do not need shared modality examples but only shared concepts. Like with Parallel models, transfer learning can be utilized as well as zero-shot learning that can identify unseen classes. Conceptual grounding is another technique that learns semantic meaning from multiple modalities within a common latent space.

*Hybrid* methods use two non-parallel modalities that are bridged using a common modality or dataset. This has been used for multilingual image captioning where image data is shared between different language-based models.

### 2.3 Domain-specific Solutions

One of the common applications for multimodal learning is remote sensing with hyperspectral satellite imagery. This method collects image data from the target area using multiple light wavelengths such as standard RGB, infrared, or imaging technologies like LiDAR. In one review paper [30], multiple kernel learning approaches were investigated for image classification. These kernel methods map input data to a new feature space that then can be used by SVM-based classifiers, resulting in something similar to late fusion architectures, as later discussed in Section 3.3. Focusing on deep learning methods, Audebert et al. [7] covered different networks designed for hyperspectral classification. The authors observed that 2-D approaches work well on data with spatial relationships and 3-D approaches for hyperspectral data where the third dimension represents image modalities. It was also suggested that Gaussian mixture models and GANs can be used to augment training data by approximating the embedded space. Results of the 2017 IEEE Geoscience and Remote Sensing Society Data Fusion Contest showed that the top teams all utilized data from multiple sources and used ensemble methods [118].

In a similar fashion, multimodal learning has also been applied to medical imaging. Today, it is common for multiple image modalities such as **computed tomography (CT)**, **magnetic resonance imaging (MRI)**, or **positron emission tomography (PET)** to be fused together to provide additional information for determining a diagnosis or the best treatment procedure. A number of image fusion architectures were reviewed by Huang et al. [40], and they observed a current limitation that few existing fusion methods utilized more than two image modalities at the same time. A survey by Haskins et al. [35] also covered medical image fusion while comparing both rigid and deformable registration techniques. To address unrealistic image deformations used for non-linear image registrations, GANs were proposed, as they often can learn to generate plausible synthetic images. In both papers, it was mentioned that the lack of standard evaluation metrics makes accurate assessment of image fusion methods difficult. The field of neuroimaging has also utilized multimodal imagery to improve scientific understanding and diagnostic performance [18], including two surveys [68, 80] that focused on Alzheimer's disease.

Another problem space well adapted for multimodal learning is human activity tracking. With the reduced cost and size, wearable sensors including microphones, accelerometers, and GPS are now practical to use. In one work [76], deep learning techniques for activity and context recognition were investigated. Several neural network architectures with the traditional early and late fusion methods were evaluated as well as different feature extraction and data modality combination methods. The authors also mentioned a challenge with feature extraction, as it was not clear if signal data should be treated as time domain points or be further processed with methods like a **Fast Fourier Transform (FFT)**. In another survey [27], wearable sensors were used to monitor mental health conditions with traditional machine learning algorithms. Based on modalities used in prior driver stress detection systems, Rastgoo et al. [79] proposed a multimodal framework using various types of sensors.

Other domain-specific surveys have also covered biometrics, 3-D image classification, and music information processing. The field of biometrics uses human features, such as images of the face, ear, or fingerprints, to identify individuals. Oloyede and Hancke [70] reviewed different multimodal architectures and fusion methods for that domain. Griffiths and Boehm [29] also reviewed works on 3-D object classification using multimodal inputs. By using different representations of objects, such as with 2-D images taken from multiple angles or **RGB plus depth (RGB-D)** images, multimodal models were used for identification. In addition to the presentation of many different network architectures, it was observed that multimodal 2-D models can perform well on a 3-D task, especially since pre-trained 2-D networks were more mature than 3-D networks. Zhang et al. [123] also performed a review of research using multimodal image data such as RGB-D for image segmentation. In the context of music processing, Simonetta et al. [86] explored preprocessing steps such as modality synchronization, feature extraction methods, and the conversion of multiple modalities to a common feature space.

## 3 MULTIMODAL CLASSIFICATION TAXONOMY

A current challenge with multimodal learning research is the wide mix of terms describing different aspects of the learning process. Many of the previously discussed papers used terms such as *early*, *late*, *intermediate*, or *hybrid* to describe such architectures, but their definitions are not always the same. Today, practitioners of deep learning methods are immediately familiar with networks described in terms of CNNs, GANs, or fully connected neural networks, but those kinds of portrayals are not present for multimodal classification architectures. Prior taxonomies tend to be either domain-specific or general for all multimodal learning systems, so their applicability for arbitrary multimodal classification problems is limited.

The taxonomy presented by Ramachandram and Taylor [77] was most closely aligned to classification, however, it only focused on deep learning, earlier research (2011–2016), and may not be specific enough to fully describe current multimodal pipelines. To address some of these challenges, we propose a new multimodal classification taxonomy that provides a descriptive, high-level set of terms that can be used to more completely describe existing or future model architectures. Table 2 provides a list of the five main stages used by this taxonomy and Table 3 includes other topics that are important to consider when describing multimodal architectures.

From the reviewed works in Tables 4 and 5 and the surveys dating since 2017, we propose a taxonomy with five major stages used for building multimodal classification models: *Preprocessing*, *Feature Extraction*, *Data Fusion*, *Primary Learner*, and *Final Classifier*. In Section 4, we present recent research using these terms, discuss some of the scenarios where multiple architectural concepts are employed in a single multimodal architecture, and where the exact architectural description is more subjective.

### 3.1 Preprocessing

Although not always used, many classification models require some preprocessing, whether it will be addressing missing data values, cropping images, filtering noise, or class balancing. Here, we describe preprocessing as a data-cleaning step, done with some level of domain expertise that may

Table 2. A List of the Five Stages Used by This Taxonomy to Describe Multimodal Classification Architectures

| | **A Taxonomy for Describing Multimodal Classification Models** | |
|---|---|---|
| | **Stage** | **Description** |
| 1 | Preprocessing | This is the initial step of data modification, which may include tasks such as removing noise, class balancing, or augmentation. |
| 2 | Feature Extraction | Higher-level features are extracted from the raw input data before being used for the direct model training. This stage can include methods such as manual feature engineering, text encoding, or CNN-generated filters. |
| 3 | Data Fusion | This stage combines raw features, extracted features, or class prediction vectors from multiple modalities to create a single data vector. Multimodal models can be further defined by their architecture and data fusion techniques. |
| | – Fusion Architecture | A descriptor of how and when modality-specific portions of the multimodal architecture is being combined. These styles can include: *early fusion*, *late fusion*, and *cross-modality fusion*. |
| | – Data Fusion Technique | A descriptor of how data from each modality is fused, including *concatenation* and *merge*. These fusion methods could be applied to traditional features or neural network node activations. |
| 4 | Primary Learner | The bulk of the overall learning process is performed in this stage and may be done independently for each modality or shared with the *Feature Extraction* or *Final Classifier* stages. |
| 5 | Final Classifier | This stage produces the final results, such as predicted labels or class likelihood scores. The classification stage could include anything from a shallow neural network or decision tree to a complicated ensemble model. |

Table 3. Other Topics beyond the Taxonomy Stages that Should Be Discussed when Describing Multimodal Architectures

| **Other Considerations when Describing Multimodal Modals** | |
|---|---|
| | **Description** |
| Shared Stages | In some cases, the same instance of a model or algorithm is shared between multiple stages such as *Feature Extraction - Primary Learner* or *Primary Learner - Final Classifier*. These shared resources are denoted with a * marker in Tables 4 and 5. |
| Cross-modality Architectures | These types of multimodal models can be quite complicated and may not fit well in any specific pattern. The tasks described in Table 2 are still applicable, but the specifics of such architectures must be described in detail. |

be difficult to generalize in this proposed taxonomy. While there are many ways to clean unimodal data, even more options are possible with multimodal datasets if each modality is processed independently. For example, if CT and MRI data is used, then different strategies for cropping, scaling, and noise reduction may be required. These images may also need to be registered, or aligned, using a rigid or deformable transformation. However, all preprocessing could be skipped for one or both modalities to use the raw data instead. Because this work is very user- and domain-dependent, the preprocessing step is not further discussed in detail but should be considered in practice on a case-by-case basis.

### 3.2 Feature Selection

Each multimodal classification model uses feature selection in some capacity, which may include manual feature engineering, deep learning methods, or be an inherent part of a classifier algorithm. The feature selection process can be performed independently for each modality or be part of multiple steps in the overall model architecture. Deep learning methods like CNNs are often used for feature extraction but the same network may also perform the primary learning step, thus doing two tasks at once. Although classifiers like Random Forests can perform the feature selection process by identifying the most useful cut points during the creation of decision trees, this operation could be performed at one explicit step, such as with a CNN for image feature extraction followed by an FCNN classifier. Dimensionality reduction using **Principal Component Analysis (PCA)** or **Linear Discriminant Analysis (LDA)** can be used as part of the feature selection process, most commonly used in traditional machine learning models that can struggle with very high dimensional data.

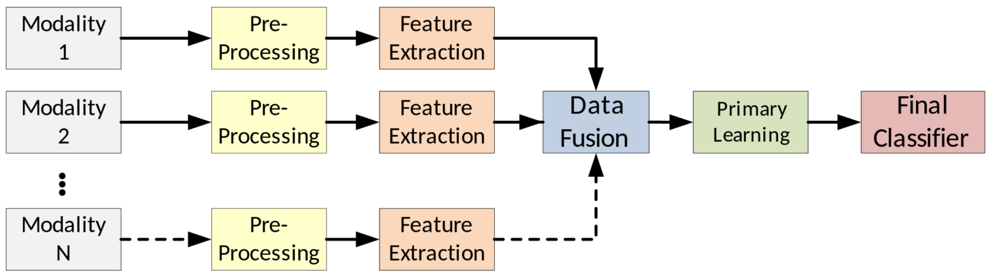


Fig. 1. An example architecture for multimodal classification with early fusion.

### 3.3 Data Fusion

Data fusion is a unique aspect of multimodal learning where information from different data sources need to be combined when building a joint model. This process may happen right after the input data is presented, right before the final classification, or multiple times in the middle. Commonly used terms for these architectures include early or late fusion, but these terms may not be enough to fully describe a multimodal model. Since previous works presented these fusion approaches differently, we propose a series of definitions that will be used throughout this article.

*3.3.1 Fusion Architecture.* **Early Fusion** occurs when all of the multimodal data is merged before the primary learning model has been performed. As shown in Figure 1, data from two or more modalities are joined and then passed to a learning algorithm. One of the most common ways to achieve this kind of fusion is to simply concatenate the incoming modality data, which can include traditional feature vectors or output nodes from pre-trained neural networks. Each modality could also represent a different channel in a CNN model, and this early fusion method may be most appropriate when there is a strong association between each data source. For example, radiotherapy datasets with imaging (CT) and planning dose volumes can be stacked as CNN channels, since each modality usually has a one-to-one voxel (3-D pixel) relationship. Satellite imagery using different light wavelengths could also be fused in a similar manner if each modality is representing the same ground area.

**Late Fusion** performs the feature extraction independently for each modality before the final classification, as shown in Figure 2. Output from the fusion stage can include low-level learned features with deep networks or class probabilities from full classifier algorithms. In both cases, the learned results are combined for the final classification. This architecture benefits from the ability to train each modality with a specific algorithm and may make it easier to add or exchange different modalities in the future. One downside is the lack of cross-modality data sharing, which could hinder learning the relationships between modalities.

**Cross-modality Fusion** allows for the sharing of modality-specific data before or during the primary learning stage. Unlike early or late fusion, this approach provides a way that each modality can use the context of each other to improve the predictive power of the overall model. This data sharing can be represented in many ways including at different parts of the learning process, the type or amount of data shared, or which modalities participate in the sharing. Figure 3 shows an architecture sharing data between each modality once before fusion, and Figure 4 shows data sharing occurring multiple times during training. A number of the presented papers showed that this kind of data sharing can outperform the traditional early or late fusion approaches, suggesting

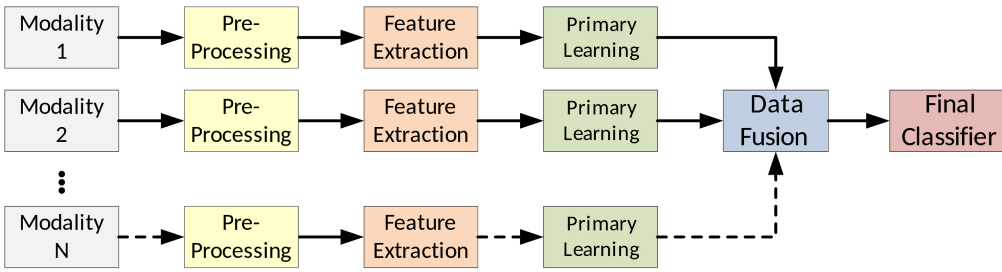


Fig. 2. An example architecture for multimodal classification with late fusion.

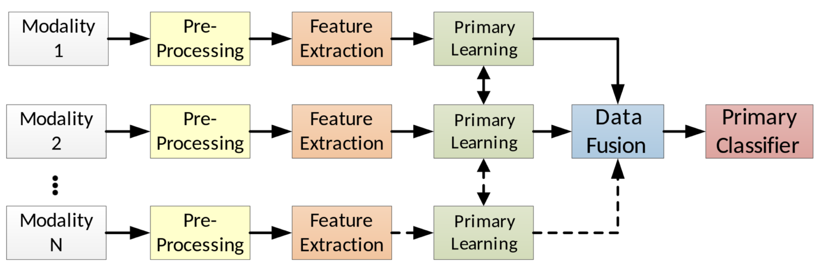


Fig. 3. An example cross-modality architecture with a single data-sharing operation.

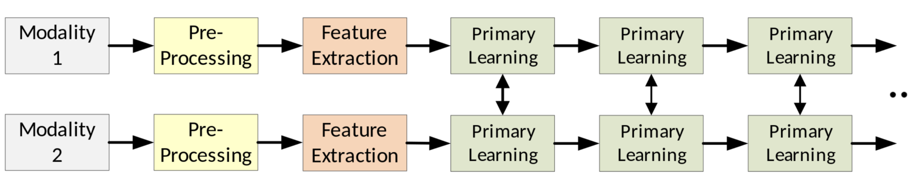


Fig. 4. The first part of a sample cross-modality architecture where data between modalities is shared multiple times during the learning process.

a promising direction for solving multimodal problems. In one work classifying satellite images [39], each modality was partially trained using a CNN, and the results were merged with the original data from the other modality. Another set of CNNs was used to continue the learning of those combined features, and results were merged again before performing the final classification. Similar to the sharing style shown in Figure 4, the model presented by Gao et al. [26] shared partially learned features between parallel CNN networks multiple times for Alzheimer's disease classification. The results from the last stage of each modality-specific network was concatenated before making predictions. This cross-modality fusion architecture is often also used with **deep belief network (DBN)** or autoencoder-style networks [25, 32, 59].

*3.3.2 Data Fusion Technique.* Based on our review of previous works, we have separated data fusion techniques into concatenation and merge fusion techniques. Figure 5 shows visual representations of these commonly used styles.

**Concatenation**: This data fusion method simply concatenates modality data to form a single feature vector. When using this technique, the input data can be raw features, class likelihood vectors, or neural network nodes. Examples of concatenation are shown in Figure 5(a) for traditional machine learning features and in Figure 5(b) for neural network nodes.

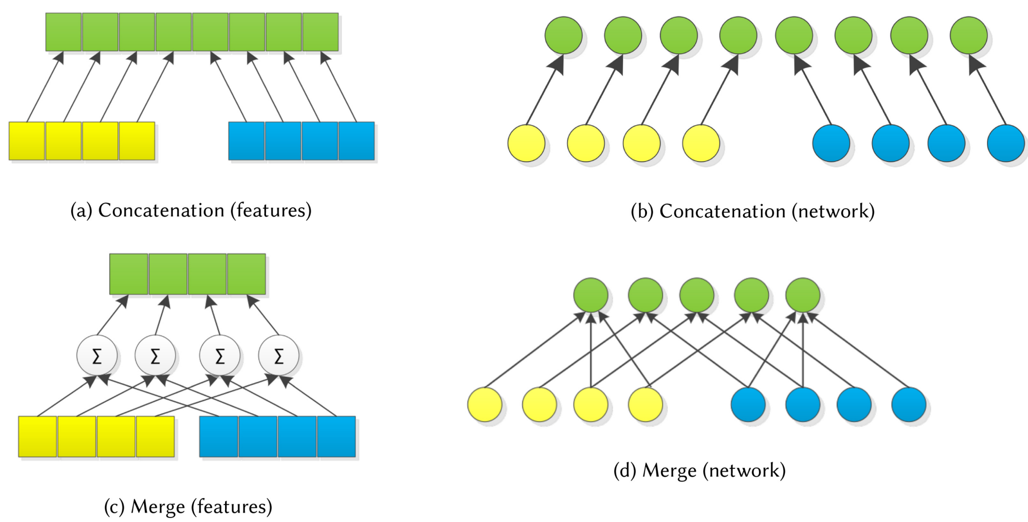


Fig. 5. A depiction of the main data fusion techniques (concatenation, merge) with two modalities (yellow, blue) and the final result (green). Traditional machine learning features are depicted as squares, neural network nodes as circles, and feature merging operators as Σ.

**Merge**: This approach combines modality data with business logic more complicated than simple concatenation. The merging process is often performed for traditional features with an arithmetic operator (represented as Σ) that transforms the input values into a new feature vector, as shown in Figure 5(c). Neural network merging connects modality-specific nodes to an output merged layer that utilizes the network weights and biases to combine features, as shown in Figure 5(d). Although the merge technique usually produces fewer output features or nodes as with an encoder, it could also be used as part of a decoding operation.

### 3.4 Primary Learner

Each traditional machine learning or deep learning system is designed to extract knowledge from the training data, often as network weights, decision boundaries, or splitting criteria. Multimodal pipelines can perform this learning process several ways, which requires clear explanation to support future advancements and the replication of experiments. For example, early fusion models produce a single joined data source so the learning process can happen all at once for all modalities. However, late fusion performs independent learning for each modality, and cross-modality models can perform the learning process multiple times. The work performed by the *Primary Learner* stage can also be shared with either the *Feature Extraction* or *Final Classifier* stage, which is further discussed in Section 3.6.

### 3.5 Final Classifier

Unlike the *Primary Learner* stage, the *Final Classifier* is used to produce the end result of the multimodal pipeline, usually predicted labels or class likelihood vectors. The algorithm used for this stage can be the same as the one used for *Primary Learner*, a completely different algorithm, or the work could be shared between the two stages. The algorithms used at this stage can range from a single softmax layer to entire ensemble models. We believe that explicitly defining this stage makes it easier to describe the overall multimodal architecture of a new model that can be implemented by future researchers.

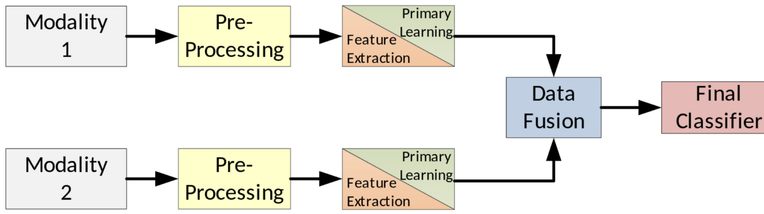


Fig. 6. An example multimodal architecture using late fusion with the feature extraction and learning tasks performed with a shared model.

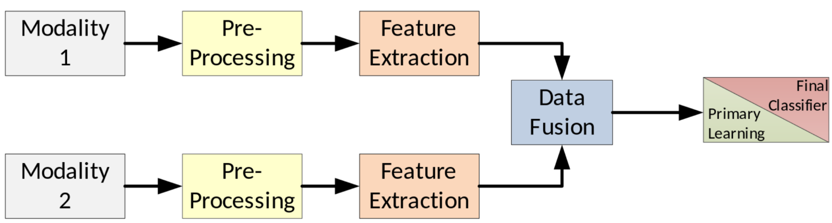


Fig. 7. An example multimodal architecture using early fusion with the learning and classification tasks performed with a shared model.

### 3.6 Stage Sharing

There are many real-world scenarios where individual stages in our taxonomy use the same model. Figure 6 shows an example of a late fusion architecture where each modality performs the feature extraction and learning with the same model. This is often done with CNNs, as later shown in Table 5. Figure 7 shows an early fusion architecture where the learning and classification is performed with a single model. This is done in the majority of traditional machine learning models as extracted features are concatenated and passed to a single classifier.

## 4 REVIEW OF RECENT MULTIMODAL CLASSIFICATION RESEARCH

In this section, we review recent works on multimodal classification. Tables 4 and 5 provide a selection of models and their architectural design using the previously defined taxonomy. This review process was used to discover patterns and commonalities between different multimodal classification approaches so the taxonomy could be created. We have separated the traditional machine learning and deep learning works, as they tended to use different types of architectures. Section 4.3 discusses these differences and other observations in more detail.

Only papers published since 2017 were considered to keep this proposed taxonomy focused on the state-of-the-art works, especially relevant for the quickly evolving field of deep learning. We considered over 400 papers during our literature search and eventually selected 121 papers, which included surveys, multimodal classification models, and topics related to future challenges. Published models that lacked the required details for decomposition were excluded. In cases where a paper presented multiple models, we chose the one that gave the best overall results, and for clarity some model types like CNNs were given their generic name instead of their specific implementation or the pre-trained model used. Determining the exact model configuration of these models may be open to different interpretations, but we chose descriptions that best fit our proposed definitions. In these tables (Tables 4 and 5), the $^*$ marker is used to identify cases where two of the stages are shared by the same model.

Table 4. An Overview of Publications Using Traditional Machine Learning Methods for Multimodal Classification

| Multimodal Classification with Traditional Machine Learning Techniques | | | | | | |
|---|---|---|---|---|---|---|
| **Reference** | **Feature Extraction** | **Fusion Architecture** | **Data Fusion Technique** | **Primary Learner** | **Final Classifier** | **Modalities** |
| Usman and Rajpoot [103] | manual | early | concatenation (features) | RF* | RF* | image |
| Huddar et al. [43] | manual/WE | early | concatenation (features) | ensemble | score merge | audio/text/video |
| Li et al. [55] | SVM* | early | concatenation (features) | SVM* | SVM | signal |
| Liang et al. [61] | manual | early | concatenation (features) | GBT* | GBT* | image |
| Elola et al. [20] | manual | early | concatenation (features) | RF* | RF* | signal |
| Ieracitano et al. [45] | manual | early | concatenation (features) | MLP* | MLP* | signal |
| Kautzky et al. [52] | manual | early | concatenation (features) | RF* | RF* | image/tabular |
| Li et al. [60] | manual | early | concatenation (features) | LR* | LR* | image |
| Panda et al. [73] | manual | early | concatenation (features) | RF* | RF* | tabular/text |
| Syed et al. [94] | SVD/WE | early | concatenation (features) | RF* | RF* | image/text |
| Sorinas et al. [91] | manual | early | concatenation (features) | *k*NN* | *k*NN* | signal |
| Zhou et al. [125] | manual | early | concatenation (features) | RF* | RF* | image |
| Gupta et al. [33] | manual | early | concatenation (features) | MKL | MKL | image/tabular |
| Qureshi et al. [75] | manual | early | concatenation (features) | ELM* | ELM* | image |
| Lee et al. [54] | SVM* | late | concatenation (scores) | SVM* | SVM | image/tabular |
| Guggenmos et al. [31] | manual | late | concatenation (scores) | SVM | score merge | image |
| Lin et al. [63] | manual | late | concatenation (scores) | ELM | ELM | image/tabular |
| Uddin and Canavan [102] | XGBoost | cross-modality | concatenation (features) | XGBoost | XGBoost | tabular |

The * marker is used to identify cases where two of the stages are shared by the same model.

## 4.1 Traditional Machine Learning

*4.1.1 Early Fusion.* Since many of the traditional machine learning classifiers are susceptible to the Curse of Dimensionality [10], explicit feature extraction is often performed for high dimensional data. This approach allows for the use of 2-D and 3-D imaging data that shows up often in the recently proposed multimodal models. For example, Usman and Rajpoot [103] used the Random Forest classifier with extracted imaging features from four MRI modalities to predict brain tumor status, and Zhou et al. [125] used MRI-based features for genotyping brain tumors. Kautzky et al. [52] developed an **attention-deficit and hyperactivity disorder (ADHD)** diagnostic tool by using 49 **regions of interest (ROI)** features with **positron emission tomography (PET)** images and 30 **single nucleotide polymorphisms (SNP)**-based features. Syed et al. [94] built a predictive model for the automatic standardization of **Digital Imaging and Communications in Medicine (DICOM)** structure sets used in the Radiation Oncology field. The 3-D representation of each delineated structure was reduced to 50 features using **singular value decomposition (SVD)** and combined with word embedding features from the associated text annotations. Both sets of feature vectors were concatenated before using the Random Forest classifier. Random Forest was also used with data from automated external defibrillators that provided **electrocardiogram (ECG)**, **thoracic impedance (TI)**, and the capnogram information [20]. This proposed model was time series-based and manual feature engineering was performed before data fusion. Panda et al. [73] used **electroencephalogram (EEG)** and customer review text to predict emotional response. Data from both modalities were encoded into a shared feature space before being passed to the Random Forest classifier.

Although Random Forest was the most commonly used traditional classifier, other algorithms were shown to be effective. For example, Li et al. [60] extracted 396 CT and MRI radiomic features from patients with locally advanced rectal cancer using the Linear Regression classifier for predicting therapeutic response after neoadjuvant chemotherapy. With the **Gradient Boost Tree (GBT)** classifier, Liang et al. [61] predicted individuals with schizophrenia using **structural magnetic resonance imaging (sMRI)** and diffusion tensor imaging-based features. Li et al. [55] developed a **Support Vector Machine (SVM)**-based multimodal model for human activity and fall detection using an **inertial measurement unit (IMU)** and radar. However, this work used a hierarchical classification approach where sub-groups of features were grouped for predicting sub-activities. To differentiate between patients with Mild Cognitive Impairment, Alzheimer's disease, or

Table 5. An Overview of Publications Using Deep Learning Methods for Multimodal Classification

| Multimodal Classification with Deep Learning | | | | | | |
|---|---|---|---|---|---|---|
| **Reference** | **Feature Extraction** | **Fusion Architecture** | **Data Fusion Technique** | **Primary Learner** | **Final Classifier** | **Modalities** |
| Jafari et al. [47] | CNN* | early | concatenation (features) | CNN* | FCNN | signal |
| Li et al. [57] | manual | early | concatenation (features) | DBN | SVM | image |
| Garillos-Manliguez and Chiang [28] | CNN* | early | concatenation (features) | CNN* | FCNN | image |
| Chancellor et al. [13] | WE/CNN* | late | concatenation (network) | WE/CNN* | FCNN | image/text |
| Gallo et al. [24] | WE/CNN | late | concatenation (scores) | SVM/RF | score merge | image/text |
| Kang and Kang [51] | DNN/CNN | late | concatenation (network) | FCNN | FCNN | image/tabular |
| Tan et al. [98] | CNN | late | concatenation (network) | RNN | FCNN | video |
| Vielzeuf et al. [105] | CNN/LSTM* | late | concatenation (score) | CNN/LSTM* | score merge | image/signal |
| Liu et al. [66] | DNN/CNN* | late | concatenation (network) | DNN/CNN* | FCNN | image/tabular |
| Liu and Li [65] | CNN* | late | concatenation (network) | CNN* | *k*-NN | image/tabular |
| Oramas et al. [71] | CNN* | late | concatenation (network) | CNN* | FCNN | audio/image |
| Suzuki et al. [93] | CNN | late | concatenation (network) | CNN | FCNN | image |
| Vijay and Indumathi [106] | manual | late | concatenation (scores) | Multi-SVNN | score merge | image |
| Yap et al. [117] | CNN | late | concatenation (network) | FCNN* | FCNN* | image/tabular |
| Aceto et al. [1] | CNN/GRU* | late | concatenation (network) | CNN/GRU* | FCNN | tabular |
| Choi et al. [17] | CNN | late | merge (scores) | FCNN* | FCNN* | tabular |
| Illendula and Sheth [46] | CNN/BiLSTM* | late | concatenation (network) | CNN/BiLSTM* | FCNN | image/text |
| Jaiswal et al. [48] | CNN/GRU* | late | concatenation (network) | CNN/GRU* | FCNN | audio/text |
| Ma and Jia [69] | CNN* | late | concatenation (network) | CNN* | LR | image |
| Shen et al. [84] | LSTM/CNN* | late | concatenation (network) | LSTM/CNN* | FCNN | image/text |
| Tian et al. [101] | AE/CNN/WE | late | merge (network) | LSTM/WE | FCNN/CNN | audio/image/text |
| Xu et al. [111] | CNN/LSTM/DT* | late | concatenation (scores) | CNN/LSTM/DT* | score merge | tabular/text |
| Agbley et al. [3] | FCNN/CNN* | late | concatenation (network) | FCNN/CNN* | FCNN | tabular/signal |
| Alay and Al-Baity [6] | CNN* | late | concatenation (scores) | CNN* | score merge | image |
| Erickson et al. [21] | DNN/CNN | late | concatenation (network) | DNN * | DNN * | image |
| Gadiraju et al. [23] | CNN/LSTM* | late | concatenation (scores) | CNN/LSTM* | SVM | image/signal |
| Zhai et al. [121] | CNN/LSTM* | late | concatenation (scores) | CNN/LSTM* | score merge | signal |
| Ahmad et al. [4] | CNN | late | merge (network) | CNN | SVM | signal |
| Aceto et al. [2] | CNN/GRU | late | concatenation (network) | CNN/GRU | FCNN | tabular |
| Felipe et al. [22] | manual/CNN | late | concatenation (scores) | CNN/SVM | score merge | image |
| Liang et al. [62] | CNN | late | concatenation (network) | CNN | FCNN | text |
| Liu et al. [67] | CNN/DNN* | late | concatenation (scores) | CNN/DNN* | score merge | image/tabular |
| Song et al. [90] | CNN* | late | concatenation (network) | CNN* | FCNN | image |
| Syed et al. [95] | CNN/WE | late | concatenation (scores) | LR | score merge | audio/image/text |
| Venugopalan et al. [104] | CNN*/manual | late | concatenation (network) | FCNN/CNN* | FCNN | tabular/image |
| Vijay and Indumathi [107] | manual | late | concatenation (scores) | Multi-SVNN | DBN | image |
| Xu et al. [113] | CNN | late | concatenation (scores) | CNN/GBT | GBT | image/tabular |
| Zehtab-Salmasi et al. [120] | CNN* | late | concatenation (network) | CNN* | FCNN | image/tabular |
| Said et al. [82] | AE* | cross-modality | merge (network ) | AE* | FCNN | signal |
| Liu et al. [64] | DNN/CNN | cross-modality | concatenation (network) | DNN | FCNN | image/tabular |
| Xu et al. [112] | BiLSTM/CNN | cross-modality | concatenation (network) | GRU | FCNN | image/text |
| Yu et al. [119] | CNN/LSTM* | cross-modality | concatenation (network) | CNN/LSTM* | FCNN | image/text |
| Hong et al. [39] | CNN | cross-modality | merge (network) | CNN | FCNN | image |
| Huddar et al. [44] | BiLSTM/GGA | cross-modality | concatenation (network) | BiLSTM | FCNN | audio/text/video |
| Yang et al. [116] | CNN/WE | cross-modality | concatenation (network) | AN | DNN/CNN | image/text |
| Gao et al. [26] | CNN* | cross-modality | concatenation (network) | CNN* | FCNN | image |

The * marker is used to identify cases where two of the stages are shared by the same model.

without a neurological condition, Ieracitano et al. [45] used both **Continuous Wavelet Transform (CWT)** and **bispectrum (BiS)** data from EEG recordings. The experimental results showed that MLP outperformed AE, LR, and SVM classifiers with these concatenated features. Instead of using a single classifier, Huddar et al. [43] built an ensemble of five individual classifiers. This work used transcription, audio, and video data for which a significant amount of preprocessing and feature selection was performed. Because the primary learning was done with the fused multimodal data, we have chosen to treat this as an early fusion architecture.

Using the **Extreme Learning Machine (ELM)** classifier, Qureshi et al. [75] built a model to assist in the diagnosis of schizophrenia. Features extracted from structural and functional MRI scans were grouped together to create new data modalities that were trained individually. Based on their respective predictive power, each modality was given a weight used at the final multimodal classification step. Gupta et al. [33] created a predictive model for distinguishing healthy patients from those with Alzheimer's disease or mild cognitive impairment where transformed image and **apolipoprotein (APOE)** genotype-based features were shaped into kernel space before classification with the **multiple kernel learning (MKL)** algorithm [5]. Sorinas et al. [91] also used the **$k$-Nearest Neighbor ($k$NN)** classifier with manually extracted features from EEG, ECG, and skin temperature data to predict emotional response to videos.

*4.1.2 Late Fusion.* In the work by Guggenmos et al. [31], five neuroimaging modalities were used for classifying psychiatric disorders. Each modality was trained independently using both SVM and the **weighted robust distance (WeiRD)** classifiers. This model also used an optimization approach where the best classifier and hyperparameters were chosen for each modality. The final classification was performed using a weighted average of the results from each modality-specific classifier. Lee et al. [54] also used SVMs for feature extraction and classification for predicting clinical pain states using brain imaging and heart rate data. To differentiate between healthy patients and those with either Alzheimer's disease or mild cognitive impairment, Lin et al. [63] developed a predictive model using MRI, **fluorodeoxyglucose positron emission tomography (FDG-PET)**, **cerebrospinal fluid (CSF)**, and APOE $\epsilon4$ gene data. Unlike the previous two methods, this model used the ELM classifier [41], which is a variant of the traditional SVM algorithm.

*4.1.3 Cross-modality Fusion.* Using a multi-layer model with XGBoost, Uddin et al. [102] built a predictor for the presence of chronic back pain as part of the EmoPain 2020 Challenge [8]. At each layer, some aspect of back pain was classified and the resulting class probabilities were then merged with the existing feature vectors. This fusion method shares some similarity to both early and late fusion architectures but has the unique property of progressively updating the feature vectors.

### 4.2 Deep Learning

*4.2.1 Early Fusion.* Using both **application-specific integrated circuit (ASIC)** and **field-programmable gate array (FPGA)**-based hardware platforms, Jafari et al. [47] used time series data from wearable sensors to detect human activity with a CNN. The sensor data was converted into a single channel image with one row per sensor and each column as a time series point. Using visible light and hyperspectral imaging, Garillos-Manliguez and Chiang [28] created a multimodal network for classifying papaya fruit maturity. The resulting RGB and hyperspectral image data was stacked as separate columns before being trained using a CNN. Li et al. [57] used multimodal satellite imagery to predict land cover types. Preprocessing was performed on each image modality, and the resulting pixels were stacked into a single input vector for a DBN. Classification was then performed using SVM with the learned deep features.

*4.2.2 Late Fusion.* Biometric identification systems can improve security with multiple verification methods. In the work by Vijay and Indumathi [106], ear and palm vein images were used, and feature extraction was performed with the **Multi-Support Vector Neural Network (Multi-SVNN)** classifier for each modality. The final identification check was performed with the sum of the modality-specific scores after optimizing the model weights. This work was later extended [107] using finger knuckle, ear, and iris image data. Although Mutli-SVNN was used again for learning each modality, a DBN was used for classification. Alay and Al-Baity [6] also built a biometric classifier with iris, face, and finger vein images. All three modalities were trained with CNNs and their outputs passed through a softmax layer. These values were normalized before being combined using either an arithmetic mean or product rule, and the highest value was chosen for the predicted class.

In the work of Aceto et al. [1], the authors developed a multimodal framework called MIMETIC for classifying network traffic. From each modality, data was extracted from network traces and independently trained with a CNN or **gated recurrent unit (GRU)** model. Final layers were concatenated and classified with FCNN layers. This work was later extended to support multi-task classification with a new framework called DISTILLER [2]. This approach trained on shared FCNN layers after the initial modal concatenation but the features were split again for training task-specific layers.

The field of medicine uses data from many potential sources, including imaging, textual notes, and discrete information, making it a natural domain for multimodal learning. In the work of Said et al. [82], an AE network was built for classifying EEG and **electromyography (EMG)** data. Each modality had its own AE that were merged, and classification was performed by fine-tuning a softmax function as the network bottleneck. Tan et al. [98] predicted cognitive events from EEG and optical flow temporal data. Both modalities are trained with CNNs and reshaped into a 2-D feature vector before being classified with an RNN. Venugopalan et al. [104] used MRI, SNP, and clinical data for predicting cognitive disorders. The modalities were trained with CNNs and AEs with their output layers concatenated and classified with a two-layer FCNN.

Using cardiovascular and actigraphy sensing, Zhai et al. [121] built an ensemble model for predicting sleep cycles. Three different time windows were chosen for both sensor types, and the combined data was trained on CNN and LSTM models, respectively, resulting in an ensemble of six total classifiers. All posterior probabilities were added to a classification matrix, and the final classification was made with the highest average or argmax value. Using three derived image modalities from ECG data, Ahmad et al. [4] developed a model for heartbeat classification. Each modality was trained on a CNN, the results were summed, and an SVM was chosen for the final classification. Clinical and cough audio data were used by Agbley et al. [3] to predict COVID-19 infections. The audio data was converted to a 2-D scalogram and trained with a CNN, while the clinical data was encoded using an FCNN. The final layers were merged, and classification was performed with a dense layer and softmax.

Liu et al. [67] used genomic data and pathology images to predict breast cancer subtypes. The genomic data was trained with an FCNN and the image data with a CNN after **principal component analysis (PCA)** performed feature reduction. Song et al. [90] used two image modalities from **contrast-enhanced spectral mammography (CESM)** scans for breast cancer detection. Each modality was trained with a CNN and the final layers were concatenated for classification with two FCNN layers. For skin lesion classification, Yap et al. [117] combined CNN-generated features with tabular clinical data and performed classification with a three-layer FCNN. Ma and Jia [69] used MRI and pathology images to predict the cancer stage of brain tumors. Both modalities were classified with CNNs and their final layers were concatenated and classified using an LR classifier.

Using ground-based cloud cover images and weather information, Liu et al. [66] created a **joint fusion convolutional neural network (JFCNN)**. Image data was trained with a CNN, and the weather data was trained with a decoder style FCNN for feature learning. The final layers were concatenated and classified with a joint FCNN layer. JFCNN was later used with GAN-generated artificial examples to increase the training dataset size [65]. Suzuki et al. [93] used airborne imagery and geospatial features to classify forest cover. Each modality was trained with a CNN, and the results were classified with CNN and FCNN layers. Xu et al. [113] used image and tabular visit data for urban functional area classification. Learned features from the image and area visit information were trained with a GBT, and their class probabilities were classified with a softmax layer.

Chancellor et al. [13] used text and image data to detect pro-eating disorder policy violations on Tumblr. The text was trained using tag embedding with a fully connected layer and image data with a CNN. The resulting layers were concatenated and classified with a two-layer FCNN. In the work by Illendula and Sheth [46], a model was built to predict emotion from social media posts using image and text. The image data was classified with ResNet [37], the textual data with BiLSTM, and the final classification was performed with a softmax layer. Syed et al. [95], used audio, text, and image data to predict levels of public trust in politicians. Using CNN and word embedding models for each modality, the final classification was performed using either majority vote or the summed confidence scores.

Gallo et al. [24] predicted real-world objects using image and text tags. A CNN was used to learn image features, and the bag-of-words method was used for text. Experiments showed that SVM with images and Random Forest for text led to best performance. Visual and near-infrared spectroscopy images were used by Erickson et al. [21] for helping robots interact with objects. The image data was trained with an FCNN and the final layers were concatenated before classification with another two-layer FCNN.

Vielzeuf et al. [105] used video frames and audio data to classify the expressed emotion in video clips. Using CNN and RNN-based networks, the final values for each modality were scored with a weighted mean. In the work by Oramas et al. [71], audio and album cover art were used to predict music genres. Each modality was trained with a CNN and two FCNN layers, with the final results classified with a cosine loss function. Tian et al. [101] used audio, video frames, and text descriptions to identify different types of natural disasters. Feature extraction was performed with AENet [97] for audio, Inception v3 [96] for video data, and text embedding with GloVe [74]. The audio and video features were trained using an LSTM with the SVM-based **Sequential Minimal Optimization (SMO)** algorithm, the text features were trained using a 1-D CNN, and a softmax layer for classification. Because some video concepts were not present in many of the video frames, the textual model was used for less common concepts instead of the joint audio-video.

An FCNN multimodal network was built by Kang et al. [51] to predict crime occurrences with spatial, temporal, and environmental data. Each modality was trained with an FCNN, and the final layers were concatenated and classification was performed using two dense layers with softmax. Using text and the visual presentation of Wikipedia documents, Shen et al. [84] built a model to predict document quality. Text data was classified using BiSTLM and image data with Inception v3 for the image data. Final results were combined using a dense layer and softmax. Zehtab-Salmasi et al. [120] predicted smartphone prices with images of the devices and their discrete properties. Each modality was trained with CNNs, and the output layers were flattened, concatenated, and classified with three more FCNN layers. The EmbraceNet [17] framework was designed to accept data from any kind of input modality. This method embeds data from modality-specific networks in a common length vector using *docking* layers. A single *embraced* vector is built from the *docking* output using multinomial distribution so each feature is only populated by a single modality.

Another approach to multimodal learning is to create different feature sets from the same training data. For example, Liang et al. [62] represented text as different modalities by splitting the data into phrases, words, n-grams, or other granularities using the **Spatial View Attention Convolutional Neural Network (SVA-CNN)** framework. This architecture was designed for preserving the relationships between these textual representations. Using context attention, parallel connection, and serial connection CNN sub-networks, the output convolutional layers were concatenated and classified with an FCNN. Similarly, Felipe et al. [22] also created several data modalities from Enteric Nervous System images of rats. Using a combination of handcrafted and model-generated image features, several chronic degenerative diseases were classified.

*4.2.3 Cross-modality Fusion.* Using both image and discrete weather station data, Liu et al. [64] built a system to predict cloud types. Low-level features were learned with a CNN for images and a fully connected network for the discrete data. These features were combined for further learning with another fully connected network, and the resulting new features were combined with prior learned features before final classification. This network design included a multimodal skip connection, similar to what is found within neural network architectures like ResNet [37].

Yu et al. [119] built a classification network for social media-based sentiment analysis using image and text. First, an LSTM with average pooling was used to extract target entity information from the text. These results were combined with a CNN trained on the image data and with the textual context information trained on two other LSTMs. Several different combinations of fused features were concatenated for softmax classification. This approach allowed for different modalities to learn with information from the other modalities.

Different image modalities were used by Hong et al. [39] to predict land cover and they tested multiple fusion network architectures. In addition to versions of early and late fusion, more advanced architectures were also investigated using both encoder-decoder and modality sharing schemes. The best results came from the latter methods, which also used the merge fusion style that compacted the multimodal data instead of concatenation.

A text-and-image-based sentiment analysis network was built by Yang et al. [116]. Text embedding and extracted image features were combined and trained on multi-modal CNN-LSTM-based attention networks. Results from these networks were concatenated and classified with softmax to predict the emotion expressed in a social media post.

For predicting brain diseases such as Alzheimer's, Gao et al. [26] used imaging data with a pathwise transfer network where partially extracted features were shared between modalities. At each layer of the network, weights from each modality were concatenated and convoluted before being concatenated again with the original modality outputs. This process allowed for the continuous sharing of information between each modality-specific network. At the classification step, the final results from each modality were again concatenated and fine-tuned with CNN, dense, and softmax layers. In addition, this work used a GAN to create artificial examples to address cases where one of the image modalities was missing.

In one work by Huddar et al. [44], a multimodal network was created for sentiment analysis using audio, video, and text data. Feature selection was performed on each modality independently with a **greedy search-based genetic algorithm (GGA)** followed by context extraction with a BiLSTM model. The unimodal results were then concatenated to form three new bimodal feature vectors representing the audio-video, text-video, and text-audio combinations. The same GGA/BiLSTM process was performed on those three modalities, and the resulting vectors were again concatenated into a single feature vector. Finally, the combined vector was processed again with a GGA/BiLSTM followed by a softmax classifier.

The **Multi-Interactive Memory Network (MIMN)** [112] was developed to predict sentiment labels from associated image and text information. Feature extraction was performed with word and phrase embedding for text data and a CNN for image data. Features from each modality were further processed by their own LSTM before being used by two parallel memory networks, one for text and another for image data. Each network had multiple blocks that included a GRU and an attention mechanism. The first block accepts its matching image or text LSTM feature vector as well as the average pool from the aspect vector and returns the output of the GRU. In the following blocks, the input is the matching LSTM vector and the GRU output vector from the other parallel network, providing context from the other modality. The final classification is performed by concatenating the output of both networks followed by a softmax layer.

### 4.3 Observations

From our literature search, we have identified some trends including the use of feature extraction, fusion architectures, and model sharing. It was also clear that the current focus of multimodal classification is with deep learning, although new research based on traditional machine learning is still being produced. Table 6 provides some comparisons between machine learning and

Table 6. A Comparison between Recent ML and DL Multimodal Classification Architectures Based on Some of Their Primary Features

**Comparison of ML and DL Multimodal Architectures**

**Occurrence of Modalities (%)**

| | Audio | Image | Signal | Tabular | Text | Video |
|---|---|---|---|---|---|---|
| ML | 5.6 | 55.6 | 22.2 | 33.3 | 16.7 | 5.6 |
| DL | 10.9 | 71.7 | 15.2 | 30.4 | 28.3 | 2.2 |

**Fusion Architecture Frequency (%)**

| | Early | Late | Cross-modality |
|---|---|---|---|
| ML | 77.8 | 16.6 | 5.6 |
| DL | 6.5 | 76.1 | 17.4 |

**Data Fusion Technique Frequency (%)**

| | Feature Concatenation | Score Concatenation | Feature Merge | Score Merge |
|---|---|---|---|---|
| ML | 83.3 | 16.7 | 0.0 | 0.0 |
| DL | 60.9 | 26.1 | 10.7 | 2.3 |

**Model Sharing Frequency (%)**

| | Extraction/Primary Learner | Primary Learner/Final Classifier | No Model Sharing |
|---|---|---|---|
| ML | 11.1 | 52.2 | 27.8 |
| DL | 61.1 | 6.5 | 41.3 |

deep learning models with the frequency that each modality, fusion method, and model sharing approach was used.

Compared to deep learning, the traditional machine learning models primarily use the manual feature extraction. The early fusion type was used in 14 of the 18 traditional machine learning models, likely because these classifiers expect a single feature vector or matrix as input and do not provide a way to further augment the data during training. For the same reason, those models all used simple multimodal feature concatenation. Most works also used the same algorithm for learning and classification unless it was a multi-task problem or used an ensemble, as *Primary Learner - Final Classifier* stages were shared in 11 of the 18 models compared to only twice for the *Feature Extraction - Primary Learner* stages. Tree base algorithms, such as Random Forest, GBT, and XGBoost, were the most popular classifiers. All of the major data types were used in the machine learning models but images were most common.

Among the deep learning–related publications, 3 used early fusion, 8 used cross-modality fusion, and 35 used late fusion, a reversal from the traditional machine learning works. Feature extraction was usually data-type dependent, and CNN was by far the most popular method followed by RNNs, such as LSTM and GRU. The architectural design of the deep learning early fusion models were similar to those using machine learning, with the biggest difference being the individual algorithms used.

Models using late fusion perform the bulk of modality training independently, which allows for the use of pre-trained models in deep learning solutions. This also supports the use of fundamentally different models for each specific modality, such as CNN for images and LSTM for text. The concatenation was still the most popular data fusion technique, but node merging was also used. In most late fusion models, a shallow neural network followed by a softmax layer was used as the final classifier.

While cross-modality fusion is still not the most common method, its popularity may be increasing, as it was more prominent in the deep learning models. These architectures tend to be more complicated than early or late fusion, but it has been shown that the performance may be superior.

At this time, multimodal learning lacks large pre-trained networks that are available for unimodal learning such as ResNet, Inception v3, or VGG [87]. If the best performing cross-modality architectures could be pre-trained on very large datasets, then the resulting models would provide a significant benefit for future work.

Several sub-network patterns have also emerged from these prior works. The most common architecture was a shared CNN for *Feature Extraction - Primary Learner* stages, followed by an FCNN and softmax classifier. Unlike traditional machine learning, the deep learning architectures tended to use the *Feature Extraction - Primary Learner* stage sharing, which occurred 24 times compared to only 3 times for the *Primary Learner - Final Classifier* combination. In cases where different network types were used for each modality, CNN and RNN networks were often paired.

Architectures using the *Primary Learner* stage model as full classifiers were more likely to use score merge as the *Final Classifier* method instead of an FCNN. This approach may have been often used because it is convenient to simply apply a single softmax layer to the pre-trained model outputs. However, future work is needed to determine if that method is superior to using earlier layers as input for the fusion stage.

## 5 APPLYING THE TAXONOMY

### 5.1 Challenges with Model Descriptions

While depictions of early and late fusion styles have been relatively consistent across multiple papers [9, 18, 38, 42, 77, 86, 123], there are still cases where other terms have been used. In the work by Guo et al. [32], one network architecture described as *multi-view-one-network* is essentially early fusion, and *one-view-one-network* could be considered late fusion. Li et al. [58] provided a view-wise feature extraction architecture that could be mapped to our late fusion taxonomy with the use of different learner and classifier models. Gao et al. [25] described early fusion as *shadow multiple modality* and late fusion as *deep multiple modality*. For the cross-modality style architectures, Ramachandram and Taylor [77] and Syed et al. [94] used the term *intermediate*, while Gao et al. [25] used *deep shared modality*. In the survey by Gao et al. [25], the authors also used *deep cross-modality* to describe multitask architectures, but these terms were used in the context of generative models such as **restricted Boltzmann machines (RBM)**. Even more specific models were described by Oloyede and Hancke [70] including *Fusion at the Decision Level* (late fusion, score merge), *Fusion at the Matching Score Level* (late fusion, using different learners/classifiers), *Biometric Traits at the Sensor Level* (early fusion), and *Fusion at the Feature Level* (late fusion, feature concatenation). In a similar manner, Yaman et al. [114] used the terms *data fusion* (early fusion), *feature fusion* (late fusion, feature merge), and *score fusion* (late fusion, score merge).

The term *intermediate* has been used for cross-modality fusion but can also refer to fusion occurring somewhere between feature extraction and classification [17]. Terms such as *middle* [39], *joint* [42], and *hybrid* [123] have also been used to describe this kind of fusion. Since our proposed taxonomy is based on the five processing stages, the middle fusion concept can be captured by early or late fusion style architectures. This allows for the cross-modality fusion to describe only the inter-modality data sharing concept. These examples only cover a small portion of all existing works, and so it is likely that many other depictions of multimodal classification architectures exist. Using our proposed taxonomy, we are able to label a wide range of models with consistent terms.

### 5.2 Describing Multimodal Classification Models

One of the challenges we faced when reviewing the previous works is that in addition to inconsistent terms, many publications did not provide enough information to confidently recreate their process. To address this issue, we have provided a checklist in Table 7 that could help ensure that the critical aspects of a multimodal classification model are fully presented with well-defined terms.

Table 7. A Checklist of Topics that Should Be Discussed when Describing Multimodal Classification Architectures

| **Checklist for Describing Multimodal Classification Architectures** | |
|---|---|
| **Model Property** | **Description** |
| Input Data | Describe the data type and properties of each modality and its potential benefit to model performance |
| Preprocessing | On a case-by-case basis, detail any preprocessing performed on each modality |
| | Discuss any dataset-level modifications such as class balancing, normalization, or imputation |
| Feature Extraction | List the method of feature extraction (e.g., manual feature engineering, model learned features) and the purpose of this choice |
| | Description of the model used for the primary learner step, including any relevant setting such as hyperparameters, dropout rate, or regularization |
| Data Fusion | Describe when data fusion is performed within the architecture (early, late, cross-modality) |
| | Describe how data fusion is performed (concatenation, merge) |
| Cross-Fusion | If cross-modality fusion is performed, then provide a detailed description of this process with a matching diagram |
| Primary Learner | Description of the model used for the primary model step, including any relevant setting such as hyperparameters, dropout rate, or regularization |
| Final Classifier | Description of the model used for the classifier step, including any relevant setting such as hyperparameters, dropout rate, or regularization |
| | Explain the output format (binary or multi-class labels, class probabilities) |
| Shared Stages | Describe any models shared between multiple architecture stages |

In addition to describing model architectures with text, visual depictions can also be useful. In Figures 8 and 9, we show how the taxonomy defined in Section 3 could be applied to previously published models using a network architecture diagram. Figure 8 shows an ML architecture developed by Syed et al. [94] that predicted radiotherapy structure set names based on 3-D volumes and physician provided labels. The *Preprocessing* and *Feature Extraction* steps show different techniques for each modality. *Data Fusion* was then performed using concatenation, and the Random Forest algorithm was shared between the *Primary Learner* and *Final Classifier* steps. Figure 9 shows a DL method presented by Song et al. [90] for classifying breast cancer. All four input data modalities were different representations of a patient's medical imaging, and they all received the same *Preprocessing* operations. The *Feature Extraction* and *Primary Learner* steps were performed independently for each modality using Res2Net50 models. The output nodes were concatenated for the *Data Fusion* step before applying the *Final Classifier* with a shallow full connected neural network. When presenting models like these in a manuscript, additional details applicable to the specific case should be provided, as suggested in Table 7.

## 6 DISCUSSION OF OPEN PROBLEMS

While there has been much progress in the recent years with multimodal data classification, there are still several important areas that have not been adequately addressed. In this section, we discuss open problems related to the ever growing size of datasets, difficult classification tasks, and the lack of general tools for multimodel classification.

### 6.1 Big Data

The emergence of Big Data has led to new opportunities and challenges. While there have been many effective solutions for classification on large datasets, little has been done specifically for

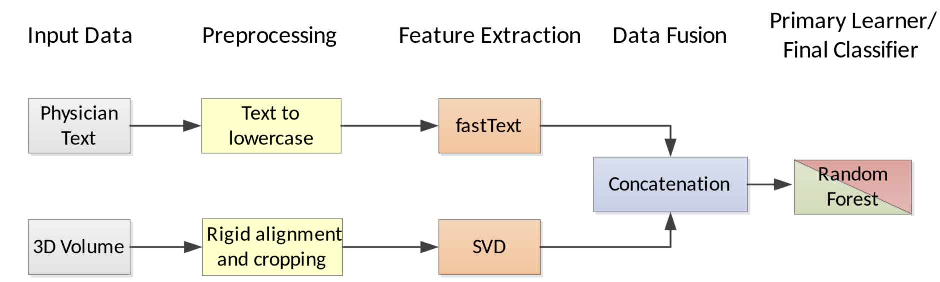


Fig. 8. The model architecture by Syed et al. [94] as described by our multimodal taxonomy.

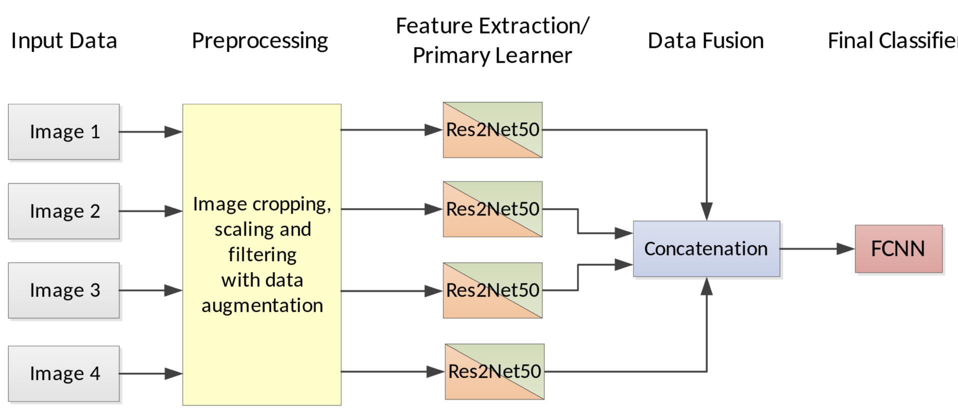


Fig. 9. The model architecture by Song et al. [90] as described by our multimodal taxonomy.

multimodal data, as unimodal has been the focus up to this point. One of the major limitations with multimodal learning research is the lack of large, publicly available datasets. The vast majority of experiments carried out in the reviewed papers used small datasets with only tens to thousands of examples. The private datasets, often containing medical imaging, also were small, as collecting healthcare data can be time-consuming, expensive, and often comes with ethical and legal restrictions on its use.

Deep learning networks are currently the most popular way of performing multimodal classification, and this technology works best with large training datasets. Although Illendula and Sheth [46] used a dataset with approximately 500,000 examples, the majority of other works used significantly fewer. The challenge of limited data is further compounded when presented with multiclass datasets, as the total number of examples per class will be even further reduced. The creation of large benchmark datasets that cover different modality combinations will be a significant benefit for future multimodal research.

While the lack of data may negatively affect these models, its impact on multimodal classification has not been fully investigated. Transfer learning with pre-trained models may help with limited training data, and this approach has been used in a number of the reviewed papers. Data augmentation is a common method for inserting new training examples such as image shifts or rotations. GANs have also been used to add synthetic examples to increase the size of multimodal datasets [56, 65].

Large datasets are often not fully curated and can have missing or erroneous values. Since these issues can exist independently for each modality, it is likely that a higher percentage of training examples will require data cleaning. In additional to traditional data imputation, it was also shown that GANs can be used to add data missing from one of the modalities [26]. However, even if GAN-generated or cleaned examples appear to be realistic, it is important to ensure that inter-modality relationships are also representative of the original dataset. Future work is needed to provide methods that discover these relationships and how the training data can be safely modified.

### 6.2 Imbalanced Data

Many machine learning algorithms were designed with the assumption that class distributions are balanced in the training dataset. However, this is often not the case and class imbalance may introduce a bias that could negatively affect classifier performance [36]. In many real-world datasets, there may be a majority class that is significantly larger than the other minority class. The same problem is magnified with multiclass datasets, as there may be many majority-minority class relationships.

Class imbalanced data can be addressed by modifying the data itself, using algorithms inherently designed for addressing this issue, or a combination of both [53]. Data-level methods have been most popular with unimodal datasets and was the only method found within the multimodal models included in this literature review. Random oversampling is one of the most common data-level methods, where random examples are simply replicated until the desired level of class balance is achieved. Random undersampling has been used but may run the risk of removing useful examples or producing a final training dataset that is too small overall. Another popular method is the **synthetic minority over-sampling technique (SMOTE)** [16], which generates new examples from a combination of nearby examples of the same class. This approach has inspired the development of many other oversampling algorithms and has been shown to outperform random sampling methods in many cases.

While there has been limited work done on multimodal imbalanced learning, it was addressed in a few of the reviewed papers. For example, two works [51, 103] used random undersampling while another [20] used random oversampling. The more advanced SMOTE method was also used by Li et al. [60] and Uddin and Canavan [102]. Since SMOTE is based on the $k$-NN algorithm, it is important that the individual features are normalized or weighed by importance. Without performing this step, the generation of synthetic examples may be over-influenced by noisy or less impactful features and thus resulting in lower-quality examples.

GANs have become a popular method for data augmentation [85] and have already been used with multimodal datasets. In one work [72], a GAN was used for adding missing text modality data that was paired with image data. Moreover, Li et al. [56] generated completely new synthetic examples instead of just imputing missing modality data. Future work is needed to better understand how methods such as SMOTE and GANs could be best utilized with multimodal datasets.

### 6.3 Instance-level Difficulty

Poor classifier performance cannot always be blamed on class imbalance, especially if classes are well separated. Some examples are harder to learn from, because they exist in a region contaminated with examples from other classes or are close to decision boundaries [81]. The difficulty of learning from such examples increases with more classes or a higher level of class imbalance, as both can lead to more regional contamination. These challenges have been well researched in traditional unimodal learning, and a number of solutions have been proposed to address this issue, especially when presented with class imbalance.

Popular oversampling methods such as random oversampling and SMOTE have been shown to do well at addressing class imbalance but do not consider which examples are more important for classification. Algorithms such as Save Level SMOTE [12] and Borderline SMOTE [34] favor specific types of examples during oversampling to add more emphasis on certain parts of the feature space, such as safe class regions and decision boundaries. However, no such method has been designed that includes these learning concepts with multimodal datasets.

These unimodal solutions may not directly translate to multimodal problems, as individual modalities could exhibit different levels of learning difficulty. Embedding each modality to a shared latent space may help solve this problem, but more experimentation is required to identify the best solutions.

### 6.4 Parallel and Distributed Computing

The relatively small datasets used in current multimodal research has not required distributed computing solutions. However, classification models on larger datasets will need more resources than are available with a single CPU or GPU. Distributed systems such as Apache Hadoop, Apache Spark, and GPU clusters are potential solutions but currently do not have direct multimodal learning support. Although this omission may have been caused by the lack of historical use cases, there will be a need for multimodal frameworks compatible with large datasets in the coming years.

Distributed computing provides a new challenge for multimodal learning, because model performance may be affected by how data is shared between computational nodes. If the distribution of class instances or sub-concepts are not carefully considered, then partial results generated at each node may not properly reflect the global properties of the training data [88]. While this has been shown to be a potential issue with unimodal learning, it has not yet been determined how this would affect multimodal models.

Two possible patterns for multimodal distributed solutions include each modality being processed together (early fusion) or processed independently and then reduced at the end (late fusion). In the early fusion case, data at a particular node may be well proportioned for one modality but not for the other. This can introduce learning challenges similar to class imbalance or instance-level difficulty as found in traditional unimodal models. The late fusion architectures could have additional issues if the data distributions between the modality-specific models are different. Further work is needed to better understand the potential challenges with distributed multimodal learning and the best architectures for addressing them.

### 6.5 Evaluation Metrics

Classification models must be evaluated to determine their performance, and the specific metrics used are dependent on the desired results. True positive rate, precision, recall, and $F_1$ scores are often used, especially for binary class datasets. The macro and micro average of $F_1$ is also used for multiclass datasets, but these metrics may provide over-optimistic results when presented with imbalanced data, as it may hide poor performance on minority classes. As reported by Branco et al. [11], there are a number of other metrics that are better suited for dealing with imbalanced data, such as **Average Accuracy (AvAcc)** of classes, the Geometric **Macro Average of recall in each class (MAvG)**, and **Class Balance Accuracy (CBA)**.

These existing metrics were designed for unimodal problems, and we are unaware of any metrics designed specifically for evaluating multimodal classifiers. While these metrics are addressing performance related to the predicted classes, there may be value in knowing how well each model does with each individual modality. Going further, the performance of each class may be affected differently by each modality. New metrics designed for multimodal classification could help identify areas in which a modality-specific model is under-performing as well as provide a better understanding of the relationships between classes and each modality.

### 6.6 Universal Models and Benchmarks

As shown in previous research, traditional machine learning algorithms and deep learning architectures can be successfully used for multimodal classification. However, the general usability of these models is limited as most are tailored for their domain-specific input modalities and may not work directly on different data combinations. Although the EmbraceNet [17] partially addresses this issue, it was only designed for late fusion style networks and may lose some shared context between modalities, depending on how the fused feature vector is constructed. Ideally, future multimodal frameworks will be configurable to support arbitrary input types using well-defined architectural rules. This will allow for a straightforward manual or automated construction of a complicated model or network. With the success of transfer learning and an ever-increasing number of pre-trained models, providing plug-and-play support for these networks will simplify the process of constructing powerful multimodal models.

Unimodal learning has benefited from large, publicly available datasets such as those from the **Modified National Institute of Standards and Technology (MNIST)** and ImageNet. These datasets are needed for training large models that could be used for transfer learning and providing benchmarks for evaluating different approaches. The creation of such datasets will be an important step for advancing multimodal learning.

## 7 CONCLUSIONS

Although unimodal learning has dominated the machine learning field, there is a growing interest in multimodal problems. New methods for combining data from multiple sources, the large collection of social media and customer reviews, and the aggregation of healthcare-related information are all providing more valuable use cases for multimodal learning. We have also seen cases where unimodal datasets can be treated as multimodal problems and thus utilize these novel learning methods. The general consensus from the reviewed papers is that multimodal-based architectures have the potential of outperforming the traditional unimodal models. However, the lack of consistent terminology for the primary aspects of multimodal-based learning and classification architectures has made it difficult to compare or evaluate different approaches. Many of the most difficult problems facing classification, such as big data, class imbalance, and instance-level difficulty, have not yet been fully addressed in this context.

As discussed in Section 1, we were motivated to address four outstanding issues related to multimodal classification problems. First, a taxonomy specific to multimodal classification was presented in Section 3 to make it easier to describe such models. In Section 4, the taxonomy was applied to 64 previously published multimodal models to highlight recent trends in these architectures. Section 5 gave examples of how to apply this taxonomy to new models and provided a checklist that could be used to guide model descriptions in manuscripts. Finally, future challenges were discussed in Section 6.

There are a number of other important challenges not addressed in this article that will require further research. Regression models have successfully been used for mutlimodal datasets [54, 108, 109] but have gotten much less focus than classification. A future survey on this topic could help identify if the general challenges and our proposed taxonomy for classification translates well to regression models. Generative models, such as GANs and AEs, have been well represented in multimodal survey papers but were used in only a few of the reviewed classification models. In a similar manner, **Canonical Correlation Analysis (CCA)** has been commonly discussed in the context of multimodal learning but is less often used with classification models. An in-depth discussion of how or when to use such methods for multimodal classification would be a benefit to the research community.

In summary, we have proposed a new multimodal classification taxonomy for describing both the overall model architectures and the style in which data fusion is performed. Unlike previous taxonomies, we focus solely on classification problems and guided our definitions based on common patterns from prior works. We believe this kind of taxonomy will be helpful when describing multimodal models that tend to be more complicated than their unimodal counterparts.

## REFERENCES

[1] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, and Antonio Pescapè. 2019. MIMETIC: Mobile encrypted traffic classification using multimodal deep learning. *Comput. Netw.* 165 (2019), 106944.

[2] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, and Antonio Pescapé. 2021. DISTILLER: Encrypted traffic classification via multimodal multitask deep learning. *J. Netw. Comput. Applic.* 183 (2021), 102985.

[3] Bless Lord Y. Agbley, Jianping Li, Aminul Haq, Bernard Cobbinah, Delanyo Kulevome, Priscilla A. Agbefu, and Bright Eleeza. 2020. Wavelet-based cough signal decomposition for multimodal classification. In *17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 5–9.

[4] Zeeshan Ahmad, Anika Tabassum, Ling Guan, and Naimul Mefraz Khan. 2021. ECG heartbeat classification using multimodal fusion. *IEEE Access* (2021). https://doi.org/10.1109/ACCESS.2021.3097614

[5] Fabio Aiolli and Michele Donini. 2015. EasyMKL: A scalable multiple kernel learning algorithm. *Neurocomputing* 169 (2015), 215–224.

[6] Nada Alay and Heyam H. Al-Baity. 2020. Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits. *Sensors* 20, 19 (2020), 5523.

[7] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. 2019. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geosci. Rem. Sens. Mag.* 7, 2 (2019), 159–173.

[8] Min S. H. Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew Kemp, Moshen Shafizadeh, et al. 2015. The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal emopain dataset. *IEEE Trans. Affect. Comput.* 7, 4 (2015), 435–451.

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Patt. Anal. Mach. Intell.* 41, 2 (2018), 423–443.

[10] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. 1957. *Dynamic Programming*. Princeton University Press.

[11] Paula Branco, Luís Torgo, and Rita P. Ribeiro. 2017. Relevance-based evaluation metrics for multi-class imbalanced domains. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 698–710.

[12] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2009. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'09)*. 475–482.

[13] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *CHI Conference on Human Factors in Computing Systems*. 3213–3226.

[14] Ganesh Chandrasekaran, Tu N. Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisc. Rev.: Data Mining Knowl. Discov.* (2021), e1415. http://doi.org/10.4018/IJSSMET.2019040103

[15] Guoqing Chao, Shiliang Sun, and Jinbo Bi. 2021. A survey on multiview clustering. *IEEE Trans. Artif. Intell.* 2, 2 (2021), 146–168.

[16] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16 (2002), 321–357.

[17] Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Inf. Fus.* 51 (2019), 259–270.

[18] Sven Dähne, Felix Biessmann, Wojciech Samek, Stefan Haufe, Dominique Goltz, Christopher Gundlach, Arno Villringer, Siamac Fazli, and Klaus-Robert Müller. 2015. Multivariate machine learning methods for fusing multimodal functional neuroimaging data. *Proc. IEEE* 103, 9 (2015), 1507–1530.

[19] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *J. Comput. Assist. Learn.* 34, 4 (2018), 338–349.

[20] Andoni Elola, Elisabete Aramendi, Unai Irusta, Per Olav Berve, and Lars Wik. 2020. Multimodal algorithms for the classification of circulation states during out-of-hospital cardiac arrest. *IEEE Trans. Biomed. Eng.* 68, 6 (2020), 1913–1922.

[21] Zackory Erickson, Eliot Xing, Bharat Srirangam, Sonia Chernova, and Charles C. Kemp. 2020. Multimodal material classification for robots using spectroscopy and high resolution texture imaging. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10452–10459.

[22] Gustavo Z. Felipe, Jacqueline N. Zanoni, Camila C. Sehaber-Sierakowski, Gleison D. P. Bossolani, Sara R. G. Souza, Franklin C. Flores, Luiz E. S. Oliveira, Rodolfo M. Pereira, and Yandre M. G. Costa. 2021. Automatic chronic degenerative diseases identification using enteric nervous system images. *Neural Comput. Applic.* (2021), 1–23. https://doi.org/10.1007/s00521-021-06164-7

[23] Krishna Karthik Gadiraju, Bharathkumar Ramachandra, Zexi Chen, and Ranga Raju Vatsavai. 2020. Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3234–3242.

[24] Ignazio Gallo, Alessandro Calefati, and Shah Nawaz. 2017. Multimodal classification fusion in real-world scenarios. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 36–41.

[25] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computat.* 32, 5 (2020), 829–864.

[26] Xingyu Gao, Feng Shi, Dinggang Shen, and Manhua Liu. 2021. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in alzheimers disease. *IEEE J. Biomed. Health Inform.* (2021). https://doi.org/10.1109/JBHI.2021.3097721

[27] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J. Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervas. Mob. Comput.* 51 (2018), 1–26.

[28] Cinmayii A. Garillos-Manliguez and John Y. Chiang. 2021. Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors* 21, 4 (2021), 1288.

[29] David Griffiths and Jan Boehm. 2019. A review on deep learning techniques for 3D sensed data classification. *Rem. Sens.* 11, 12 (2019), 1499.

[30] Yanfeng Gu, Jocelyn Chanussot, Xiuping Jia, and Jon Atli Benediktsson. 2017. Multiple kernel learning for hyperspectral image classification: A review. *IEEE Trans. Geosci. Rem. Sens.* 55, 11 (2017), 6547–6565.

[31] Matthias Guggenmos, Katharina Schmack, Ilya M. Veer, Tristram Lett, Maria Sekutowicz, Miriam Sebold, Maria Garbusow, Christian Sommer, Hans-Ulrich Wittchen, Ulrich S. Zimmermann, et al. 2020. A multimodal neuroimaging classifier for alcohol dependence. *Sci. Rep.* 10, 1 (2020), 1–12.

[32] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access* 7 (2019), 63373–63394.

[33] Yubraj Gupta, Ji-In Kim, Byeong Chae Kim, and Goo-Rak Kwon. 2020. Classification and graphical analysis of Alzheimer's disease and its prodromal stage using multimodal features from structural, diffusion, and functional neuroimaging data and the APOE genotype. *Front. Aging Neurosci.* 12 (2020), 238.

[34] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing*, De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang (Eds.). Springer Berlin, 878–887.

[35] Grant Haskins, Uwe Kruger, and Pingkun Yan. 2020. Deep learning in medical image registration: A survey. *Mach. Vis. Applic.* 31, 1 (2020), 1–18.

[36] Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 9 (2009), 1263–1284.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[38] Nathan Henderson, Jonathan Rowe, Luc Paquette, Ryan S. Baker, and James Lester. 2020. Improving affect detection in game-based learning with multimodal data fusion. In *International Conference on Artificial Intelligence in Education*. Springer, 228–239.

[39] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Rem. Sens.* (2020). https://doi.org/10.1109/TGRS.2020.3016820

[40] Bing Huang, Feng Yang, Mengxiao Yin, Xiaoying Mo, and Cheng Zhong. 2020. A review of multimodal medical image fusion techniques. *Computat. Math. Meth. Med.* 2020 (2020). https://doi.org/10.1155/2020/8279342

[41] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. 2011. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst., Man, Cyber., Part B (Cyber.)* 42, 2 (2011), 513–529.

[42] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* 3, 1 (2020), 1–9.

[43] Mahesh G. Huddar, Sanjeev S. Sannakki, and Vijay S. Rajpurohit. 2018. An ensemble approach to utterance level multimodal sentiment analysis. In *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. IEEE, 145–150.

[44] Mahesh G. Huddar, Sanjeev S. Sannakki, and Vijay S. Rajpurohit. 2020. Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. *Computat. Intell.* 36, 2 (2020), 861–881.

[45] Cosimo Ieracitano, Nadia Mammone, Amir Hussain, and Francesco C. Morabito. 2020. A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. *Neural Netw.* 123 (2020), 176–190.

[46] Anurag Illendula and Amit Sheth. 2019. Multimodal emotion classification. In *World Wide Web Conference*. 439–449.

[47] Ali Jafari, Ashwinkumar Ganesan, Chetan Sai Kumar Thalisetty, Varun Sivasubramanian, Tim Oates, and Tinoosh Mohsenin. 2018. SensorNet: A scalable and low-power deep convolutional neural network for multimodal data classification. *IEEE Trans. Circ. Syst. I: Reg. Pap.* 66, 1 (2018), 274–287.

[48] Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. 2019. Controlling for confounders in multimodal emotion classification via adversarial learning. In *International Conference on Multimodal Interaction*. 174–184.

[49] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. 2021. A review of multimodal image matching: Methods and applications. *Inf. Fus.* (2021). https://doi.org/10.1016/j.inffus.2021.02.012

[50] Adán José-García, Julia Handl, Wilfrido Gómez-Flores, and Mario Garza-Fabre. 2021. An evolutionary many-objective approach to multiview clustering using feature and relational data. *Appl. Soft Comput.* 108 (2021), 107425.

[51] Hyeon-Woo Kang and Hang-Bong Kang. 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PloS One* 12, 4 (2017).

[52] A. Kautzky, T. Vanicek, C. Philippe, G. S. Kranz, W. Wadsak, M. Mitterhauser, A. Hartmann, A. Hahn, M. Hacker, D. Rujescu, et al. 2020. Machine learning classification of ADHD and HC by multimodal serotonergic data. *Translat. Psychiat.* 10, 1 (2020), 1–9.

[53] Bartosz Krawczyk. 2016. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* 5, 4 (2016), 221–232.

[54] Jeungchan Lee, Ishtiaq Mawla, Jieun Kim, Marco L. Loggia, Ana Ortiz, Changjin Jung, Suk-Tak Chan, Jessica Gerber, Vincent J. Schmithorst, Robert R. Edwards, et al. 2019. Machine learning-based prediction of clinical pain using multimodal neuroimaging and autonomic metrics. *Pain* 160, 3 (2019), 550.

[55] Haobo Li, Aman Shrestha, Francesco Fioranelli, Julien Le Kernec, and Hadi Heidari. 2018. Hierarchical classification on multimodal sensing for human activity recogintion and fall detection. In *IEEE SENSORS Conference*. IEEE, 1–4.

[56] Qing Li, Guanyuan Yu, Jun Wang, and Yuehao Liu. 2020. A deep multimodal generative and fusion framework for class-imbalanced multimodal data. *Multim. Tools Applic.* 79, 33 (2020), 25023–25050.

[57] Xianju Li, Zhuang Tang, Weitao Chen, and Lizhe Wang. 2019. Multimodal and multi-model deep fusion for fine classification of regional complex landscape areas using ZiYuan-3 imagery. *Rem. Sens.* 11, 22 (2019), 2716.

[58] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. 2018. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinf.* 19, 2 (2018), 325–340.

[59] Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE Trans. Knowl. Data Eng.* 31, 10 (2018), 1863–1883.

[60] Zheng-Yan Li, Xiao-Dong Wang, Mou Li, Xi-Jiao Liu, Zheng Ye, Bin Song, Fang Yuan, Yuan Yuan, Chun-Chao Xia, Xin Zhang, et al. 2020. Multi-modal radiomics model to predict treatment response to neoadjuvant chemotherapy for locally advanced rectal cancer. *World J. Gastroent.* 26, 19 (2020), 2388.

[61] Sugai Liang, Yinfei Li, Zhong Zhang, Xiangzhen Kong, Qiang Wang, Wei Deng, Xiaojing Li, Liansheng Zhao, Mingli Li, Yajing Meng, et al. 2019. Classification of first-episode schizophrenia using multimodal brain features: A combined structural and diffusion imaging study. *Schizoph. Bull.* 45, 3 (2019), 591–599.

[62] Yunji Liang, Huihui Li, Bin Guo, Zhiwen Yu, Xiaolong Zheng, Sagar Samtani, and Daniel D. Zeng. 2021. Fusion of heterogeneous attention mechanisms in multi-view convolutional neural network for text classification. *Inf. Sci.* 548 (2021), 295–312.

[63] Weiming Lin, Qinquan Gao, Jiangnan Yuan, Zhiying Chen, Chenwei Feng, Weisheng Chen, Min Du, and Tong Tong. 2020. Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. *Front. Aging Neurosci.* 12 (2020), 77.

[64] Shuang Liu, Linlin Duan, Zhong Zhang, and Xiaozhong Cao. 2019. Hierarchical multimodal fusion for ground-based cloud classification in weather station networks. *IEEE Access* 7 (2019), 85688–85695.

[65] Shuang Liu and Mei Li. 2018. Multimodal GAN for energy efficiency and cloud classification in Internet of Things. *IEEE Internet Things J.* 6, 4 (2018), 6034–6041.

[66] Shuang Liu, Mei Li, Zhong Zhang, Baihua Xiao, and Xiaozhong Cao. 2018. Multimodal ground-based cloud classification using joint fusion convolutional neural network. *Rem. Sens.* 10, 6 (2018), 822.

[67] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng. 2021. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM* (2021). https://doi.org/10.1016/j.irbm.2020.12.002

[68] Xiaonan Liu, Kewei Chen, Teresa Wu, David Weidman, Fleming Lure, and Jing Li. 2018. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Translat. Res.* 194 (2018), 56–67.

[69] Xiao Ma and Fucang Jia. 2019. Brain tumor classification with multimodal MR and pathology images. In *International MICCAI Brainlesion Workshop*. Springer, 343–352.

[70] Muhtahir O. Oloyede and Gerhard P. Hancke. 2016. Unimodal and multimodal biometric sensing systems: A review. *IEEE Access* 4 (2016), 7532–7555.

[71] Sergio Oramas, Francesco Barbieri, Oriol Nieto Caballero, and Xavier Serra. 2018. Multimodal deep learning for music genre classification. *Trans. Int. Societ. Music Inf. Retr.* 1, 1 (2018) 4–21.

[72] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. 2021. Multimodal prototypical networks for few-shot learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. 2644–2653.

[73] Debadrita Panda, Debashis Das Chakladar, and Tanmoy Dasgupta. 2020. Multimodal system for emotion recognition using EEG and customer review. In *Global AI Congress*. Springer, 399–410.

[74] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[75] Muhammad Naveed Iqbal Qureshi, Jooyoung Oh, Dongrae Cho, Hang Joon Jo, and Boreom Lee. 2017. Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine. *Front. Neuroinf.* 11 (2017), 59.

[76] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proc. ACM Inter., Mob., Wear. Ubiq. Technol.* 1, 4 (2018), 1–27.

[77] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Sig. Process. Mag.* 34, 6 (2017), 96–108.

[78] Vishal Raman and Madhu Kumari. 2018. Multimodal deep learning in semantic image segmentation: A review. In *International Conference on Cloud Computing and Internet of Things*. 7–11.

[79] Mohammad Naim Rastgoo, Bahareh Nakisa, Andry Rakotonirainy, Vinod Chandran, and Dian Tjondronegoro. 2018. A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Comput. Surv.* 51, 5 (2018), 1–35.

[80] Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155 (2017), 530–548.

[81] José A. Sáez, Bartosz Krawczyk, and Michał Woźniak. 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Patt. Recog.* 57 (2016), 164–178.

[82] Ahmed Ben Said, Amr Mohamed, Tarek Elfouly, Khaled Harras, and Z. Jane Wang. 2017. Multimodal deep learning approach for joint EEG-EMG data compression and classification. In *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 1–6.

[83] Nonita Sharma, Geeta Sikka, et al. 2020. Multimodal sentiment analysis of social media data: A review. In *International Conference on Recent Innovations in Computing*. Springer, 545–561.

[84] Aili Shen, Bahar Salehi, Timothy Baldwin, and Jianzhong Qi. 2019. A joint model for multimodal document quality assessment. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 107–110.

[85] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data* 6, 1 (2019), 1–48.

[86] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. 2019. Multimodal music information processing and retrieval: Survey and future challenges. In *International Workshop on Multilayer Music Representation and Processing (MMRP)*. IEEE, 10–18.

[87] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[88] William C. Sleeman IV and Bartosz Krawczyk. 2021. Multi-class imbalanced big data classification on Spark. *Knowl.-based Syst.* 212 (2021), 106598.

[89] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image Vis. Comput.* 65 (2017), 3–14.

[90] Jingqi Song, Yuanjie Zheng, Muhammad Zakir Ullah, Junxia Wang, Yanyun Jiang, Chenxi Xu, Zhenxing Zou, and Guocheng Ding. 2021. Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. *Int. J. Comput. Assist. Radiol. Surg.* 16, 6 (2021), 979–988.

[91] Jennifer Sorinas, Jose Manuel Ferrández, and Eduardo Fernandez. 2020. Brain and body emotional responses: Multimodal approximation for valence classification. *Sensors* 20, 1 (2020), 313.
[92] Shiliang Sun. 2013. A survey of multi-view machine learning. *Neural Comput. Applic.* 23, 7 (2013), 2031–2038.
[93] Kumiko Suzuki, Utei Rin, Yoshiko Maeda, and Hiroshi Takeda. 2018. Forest cover classification using geospatial multimodal data. *Int. Archiv. Photogram., Rem. Sens. Spatial Inf. Sci.* 42, 2 (2018).
[94] Khajamoinuddin Syed, William C. Sleeman, Michael Hagan, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. Multi-view data integration methods for radiotherapy structure name standardization. *Cancers* 13, 8 (2021), 1796.
[95] Muhammad Shehram Shah Syed, Elena Pirogova, and Margaret Lech. 2021. Prediction of public trust in politicians using a multimodal fusion approach. *Electronics* 10, 11 (2021), 1259.
[96] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition.* 2818–2826.
[97] Naoya Takahashi, Michael Gygli, and Luc Van Gool. 2017. AENet: Learning deep audio features for video analysis. *IEEE Trans. Multim.* 20, 3 (2017), 513–524.
[98] Chuanqi Tan, Fuchun Sun, Wenchang Zhang, Jianhua Chen, and Chunfang Liu. 2017. Multimodal classification with deep convolutional-recurrent neural networks for electroencephalography. In *International Conference on Neural Information Processing*. Springer, 767–776.
[99] Chang Tang, Jiajia Chen, Xinwang Liu, Miaomiao Li, Pichao Wang, Minhui Wang, and Peng Lu. 2018. Consensus learning guided multi-view unsupervised feature selection. *Knowl.-based Syst.* 160 (2018), 49–60.
[100] Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, Jing Zhang, Jian Xiong, and Lizhe Wang. 2021. Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection. *IEEE Trans. Knowl. Data Eng.* (2021). https://doi.org/10.1109/TKDE.2020.3048678
[101] Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. 2019. Multimodal deep representation learning for video classification. *World Wide Web* 22, 3 (2019), 1325–1341.
[102] Md Taufeeq Uddin and Shaun Canavan. 2020. Multimodal multilevel fusion for sequential protective behavior detection and pain estimation. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG'20).* IEEE Computer Society, 467–471.
[103] Khalid Usman and Kashif Rajpoot. 2017. Brain tumor classification from multi-modality MRI using wavelets and machine learning. *Patt. Anal. Applic.* 20, 3 (2017), 871–881.
[104] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* 11, 1 (2021), 1–13.
[105] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *19th ACM International Conference on Multimodal Interaction.* 569–576.
[106] M. Vijay and G. Indumathi. 2018. Multimodal biometric system using ear and palm vein recognition based on GwPeSOA: Multi-SVNN for security applications. In *International Conference on Computational Vision and Bio Inspired Computing*. Springer, 215–231.
[107] M. Vijay and G. Indumathi. 2021. Deep belief network-based hybrid model for multimodal biometric system for futuristic security applications. *J. Inf. Secur. Applic.* 58 (2021), 102707.
[108] Ke Wang, Mohit Bansal, and Jan-Michael Frahm. 2018. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In *IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 1842–1851.
[109] Shangfei Wang, Longfei Hao, and Qiang Ji. 2019. Knowledge-augmented multimodal deep regression Bayesian networks for emotion video tagging. *IEEE Trans. Multim.* 22, 4 (2019), 1084–1097.
[110] Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634* (2013).
[111] Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, et al. 2019. Multimodal machine learning for automated ICD coding. In *Machine Learning for Healthcare Conference.* PMLR, 197–215.
[112] Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence.* 371–378.
[113] Xiujuan Xu, Yulin Bai, Yu Liu, Xiaowei Zhao, and Yuzhi Sun. 2021. MM-UrbanFAC: Urban functional area classification model based on multimodal machine learning. *IEEE Trans. Intell. Transport. Syst.* (2021). https://doi.org/10.1109/TITS.2021.3083486
[114] Dogucan Yaman, Fevziye Irem Eyiokur, and Hazim Kemal Ekenel. Multimodal age and gender classification using ear and profile face images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*
[115] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. 2021. Deep multi-view learning methods: A review. *Neurocomputing* (2021). https://doi.org/10.1016/j.neucom.2021.03.090
[116] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multim.* (2020). https://doi.org/10.1109/TMM.2020.3035277

[117] Jordan Yap, William Yolland, and Philipp Tschandl. 2018. Multimodal skin lesion classification using deep learning. *Experim. Dermatol.* 27, 11 (2018), 1261–1267.

[118] Naoto Yokoya, Pedram Ghamisi, Junshi Xia, Sergey Sukhanov, Roel Heremans, Ivan Tankoyeu, Benjamin Bechtel, Bertrand Le Saux, Gabriele Moser, and Devis Tuia. 2018. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest. *IEEE J. Select. Topics Appl. Earth Observ. Rem. Sens.* 11, 5 (2018), 1363–1377.

[119] Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 28 (2019), 429–439.

[120] Aidin Zehtab-Salmasi, Ali-Reza Feizi-Derakhshi, Narjes Nikzad-Khasmakhi, Meysam Asgari-Chenaghlu, and Saeideh Nabipour. 2021. Multimodal price prediction. *Ann. Data Sci.* (2021), 1–17. https://doi.org/10.1007/s40745-021-00326-z

[121] Bing Zhai, Ignacio Perez-Pozuelo, Emma A. D. Clifton, Joao Palotti, and Yu Guan. 2020. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proc. ACM Inter., Mob., Wear. Ubiq. Technol.* 4, 2 (2020), 1–33.

[122] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Select. Topics Sig. Process.* 14, 3 (2020), 478–493.

[123] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. 2020. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.* (2020), 104042. https://doi.org/10.1016/j.imavis.2020.104042

[124] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. 2017. Multi-view learning overview: Recent progress and new challenges. *Inf. Fus.* 38 (2017), 43–54.

[125] Hao Zhou, Ken Chang, Harrison X. Bai, Bo Xiao, Chang Su, Wenya Linda Bi, Paul J. Zhang, Joeky T. Senders, Martin Vallières, Vasileios K. Kavouridis, et al. 2019. Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low-and high-grade gliomas. *J. Neuro-oncol.* 142, 2 (2019), 299–307.

Received 14 September 2021; revised 4 June 2022; accepted 7 June 2022