



# Multimodal data fusion for systems improvement: A review

Nathan Gaw, Safoora Yousefi & Mostafa Reisi Gahrooei

To cite this article: Nathan Gaw, Safoora Yousefi & Mostafa Reisi Gahrooei (2022) Multimodal data fusion for systems improvement: A review, IISE Transactions, 54:11, 1098-1116, DOI: 10.1080/24725854.2021.1987593

To link to this article: https://doi.org/10.1080/24725854.2021.1987593

4	1	(	1

Iransactio

Design & Manufacturing

Data Science, Quality & Reliability

INDUSTRA A SYSTEM

Published online: 03 Dec 2021.



Submit your article to this journal





View related articles



View Crossmark data 🗹

Citing articles: 2 View citing articles 🗹

Check for updates

# Multimodal data fusion for systems improvement: A review

Nathan Gaw<sup>a</sup>, Safoora Yousefi<sup>b</sup>, and Mostafa Reisi Gahrooei<sup>c</sup>

<sup>a</sup>Department of Operational Sciences, Graduate School of Engineering Management, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA; <sup>b</sup>Microsoft Corporation, Seattle, WA, USA; <sup>c</sup>Department of Industrial and Systems Engineering, University of Florida, Gainsville, FL, USA

#### ABSTRACT

In recent years, information available from multiple data modalities has become increasingly common for industrial engineering and operations research applications. There have been a number of research works combining these data in unsupervised, supervised, and semi-supervised fashions that have addressed various issues of combining heterogeneous data, as well as several existing open challenges that remain to be addressed. In this review paper, we provide an overview of some methods for the fusion of multimodal data. We provide detailed real-world examples in manufacturing and medicine, introduce early, late, and intermediate fusion, as well as discuss several approaches under decomposition-based and neural network fusion paradigms. We summarize the capabilities and limitations of these methods and conclude the review article by discussing the existing challenges and potential research opportunities.

## ARTICLE HISTORY

Received 27 December 2020 Accepted 2 July 2021

**KEYWORDS** Multimodal data; heterogeneous data; data fusion

# **1. Introduction**

Evaluating, analyzing, and modeling complex systems require a multi-perspective data acquisition framework. Such a framework gathers data through different instruments, sensors, and experiments to be translated into information and knowledge about the system. Like Rumi's elephant in the dark room, each instrument or sensor collects partial information about a few characteristics of the system. However, when the partial information is combined, one may be able to discern a more complete picture. We refer to information collected by each instrument as a data mode and the full dataset obtained from the multi-perspective acquisition framework as a multimodal dataset. A multivariate dataset may also contain features collected from multiple instruments; however, there is no distinction of particular perspectives represented by one or more of the features. In contrast, a multimodal dataset is organized in such a way that data can be grouped into multiple perspectives, for which each perspective consists of at least one feature. Improvements in sensing technology have created opportunities for the acquisition of multimodal data, which can be used for analysis that goes beyond separate evaluation of each mode of the data. In this article, data fusion is defined as the process of integrating different modes of multimodal data to achieve a more comprehensive or accurate understanding of the system that could not be achieved otherwise by analysis of each mode.

The relationship between multiple data sets was first analyzed in the breakthrough work by Hotelling (1936). Since then, the number of analysis methods for multimodal datasets has increased rapidly. Techniques such as multiset canonical correlation analysis, parallel factor analysis (PARAFAC), and tensor decomposition were introduced in the 1960s and 1970s (Tucker, 1964; Ilarshman, 1970; Kettenring, 1971). Nevertheless, only a limited number of domains, including chemometrics, benefited from these developments. With the recent insurgence of multimodal datasets, an increasing number of domains, including manufacturing, healthcare, and renewable energy, demonstrate interest in exploiting the potential benefits of these datasets.

A systematic analysis and fusion of multimodality datasets results in increased understanding that can facilitate decision making and result in systems improvement. We define systems improvement as the practice of using datadriven actions that increase the efficiency of the system's processes. This increase of efficiency can be due to more effective descriptive and predictive models, more reliable abnormality detection methods, or more accurate and interpretable features extracted from data that facilitate decision making. Examples of such systems improvement include the following applications: in systems prognostics, where the remaining useful lifetime of a system component is predicted for better maintenance scheduling; in healthcare, where medical imaging and other patient data are used for more accurate identification of a disease; in renewable energy, where consumption data is used to improve energy management; and in agriculture, where imaging and weather data can be used for crop yield predictions and crop health management. Even though the benefits of analyzing multimodal data sets is evident, the knowledge of how to exploit the similarities and differences of modalities is still limited.

Problems such as heterogeneity of data (i.e., modes are in different dimension or structure), differences in scale, resolution, accuracy, conflicting modes, and redundant modes create significant challenges that hinder the advancements of multimodal data analysis.

We group the multimodal data fusion algorithms into two classes: (i) algorithms that do not use neural networks and mainly focus on decomposition techniques; (ii) methods that use neural networks to perform data fusion. Tensor data analysis, factor analysis, and generalized principal complement analysis belong to the first group of algorithms, and are suitable when the sample size is small compared with the number of variables. Deep neural networks with architectures that contain several layers of fusion belong to second class of algorithms. In this article, we refer to the first class of algorithms as decomposition-based fusion algorithms and to the second class of algorithms as neural network-based algorithms. Within both of these categories, three ways of fusion are available: (i) early fusion (low-level fusion), (ii) late fusion (high-level fusion), and (iii) intermediate fusion. Some application areas tend to take one way over others. For example, whereas in healthcare applications late fusion is more common (Zhang and Ma, 2012; Suk et al., 2017; Khasha et al., 2019; Liu, Chen, Wu, Weidman, Lure, Li and Alzheimer's Disease Neuroimaging Initiative, 2020), in systems monitoring and prognostics early fusion is more prevalent (Liu et al., 2013; Liu and Huang, 2014; Liu et al., 2015; Fang et al., 2017; Chehade et al., 2018; Song and Liu, 2018).

Early fusion (or low-level fusion) is the process of fusing modalities by only using information from the predictors (i.e., independent variables). Early fusion can either occur as a preprocessing task before incorporation into the main model, or as a purely unsupervised task to generate features that best describe the underlying patterns across modalities. In feature preprocessing, the main goal is to combine raw features from different modalities to generate new features that combine complementary information of the raw features from different modalities. These new features are then inputted to a supervised model for a training task. Early fusion as a purely unsupervised task has the goal of combining features across modalities to discern underlying patterns present across different modalities or generate visualization that aptly describes information from the different modalities (i.e., combining different types of medical imaging to generate another image that displays complementary information) (He et al., 2010; Moin et al., 2016; Rajalingam and Priya, 2017). Principal component regression is an example of early fusion, in which Principal Component Analysis (PCA) is performed to extract input features that are then employed to predict an output value.

Late fusion (or high-level fusion) is fusion at the decision level. When modalities have been processed and modeled separately, the individual predictions from each modality can be combined in a number of ways depending on the importance of each modality for the prediction task, the appropriateness of the modality combination (whether the fusion should be modeled as an element-wise summation (Brentan *et al.*, 2017), weighted average (Kahou *et al.*, 2016), bi-linear product (Chen and Irwin, 2017), etc.), the noise level present in each modality, and/or other considerations deemed appropriate by the practitioner. Some popular examples of late fusion include ensemble learning (Yokoya *et al.*, 2017; Sagi and Rokach, 2018; Samareh *et al.*, 2018) and deep late fusion (Simonyan and Zisserman, 2014; Kahou *et al.*, 2016; Wu *et al.*, 2016; Ramachandram and Taylor, 2017).

Intermediate fusion is the process of incorporating features from different modalities in the model training process, using both predictors (i.e., independent variables) and response (i.e., dependent variables). These methods incorporate fusion directly in the model training process and make decisions on fusion in a way to optimize the objective (e.g., accuracy, detection rate, etc.). Partial Least Square (PLS) is an example of intermediate fusion, in which the fused features are extracted in a supervised fashion to best explain the output (Zhao *et al.*, 2012). Another example of intermediate fusion can be found in tensor regression, which can extract features from tensors that contain multiple sources to estimate the output (Gahrooei *et al.*, 2020). Deep learning architectures can also be designed to perform intermediate fusion (Karpathy *et al.*, 2014).

To better understand the intuition behind early, late, and intermediate fusion, let us consider an example in prognostics where the remaining useful lifetime (or Time To Failure (TTF)) of an asset is predicted based on the available sensors data (Song and Liu, 2018; Song et al., 2019; Li et al., 2021). Consider a rotary machine whose condition is monitored by three different sensors that produce vibration signals (A), noise signals (B), and infrared thermal images (C). Early fusion (see Figure 1(a)) focuses on generating latent variables by combining these signals and then using the latent variables for model training and prediction. For example, early fusion first performs PCA on the merged dataset and then uses the PC scores to create a model that predicts TTF. In this scenario, f and g are trained separately. Late fusion (see Figure 1(b)) focuses on developing models fA, fB, and fC separately from sensor types A, B, and C to estimate separate TTF values, yA, yB, and yC, then fuses the separate predictions into one overall prediction, for example by taking the average or median of the TTFs. Intermediate fusion (see Figure 1(c)) performs fusion during the model training process and simultaneously finds latent information shared between the different sensor modalities, while also determining the best settings to accurately predict the TTF.

Let us also consider another data fusion example in healthcare. In a research hospital, there is a cohort of patients with a brain disease along with healthy controls for which three different types of neuroimaging were collected: (i) structural magnetic resonance images (sMRI), (ii) functional magnetic resonance images (fMRI), and (iii) magnetoencephalography (MEG). sMRI conveys brain structure and provides the highest spatial resolution, but has no temporal resolution. fMRI indicates blood oxygen level and provides an acceptable spatial resolution along with a lower temporal resolution. MEG records magnetic fields generated by brain





Figure 1. Illustration of different levels of fusion.

electrical activity, and provides a higher temporal resolution at the cost of a lower spatial resolution. The task is to build a statistical model capable of optimally combining the information available in images A, B, and C to accurately quantify patient disease severity and highlight particular brain locations or functions that may be causing impairment. Early fusion (see Figure 1(a)) can use an algorithm, such as Independent Component Analysis (ICA), to identify spatially independent signals that convey underlying brain networks/ structures (f), from which features can be extracted and used to predict disease severity (g). Late fusion (see Figure 1(b)) can develop separate machine learning models trained on each image type, fA, fB, and fC, to make predictions of disease severity, yA, yB, and yC, that can then be fused into an overall prediction score by g (ex., through averaging the scores). Intermediate fusion (see Figure 1(c)) performs neuroimaging fusion during the model training process (e.g., via fused group lasso) and can simultaneously fuse information from images A, B, and C while building an accurate model that can predict the patient disease severity.

# 1.1. Contribution and article organization

This article reviews data fusion algorithms as categorized earlier into the decomposition-based algorithms and neural network-based algorithms. Within each category, this article covers a broad range of algorithms in multimodality fusion that are mainly developed in recent years with the focus on the Industrial Engineering (IE) applications, such as healthcare and prognostics. Other review papers are available in the area of multimodality data analysis (Atrey et al., 2010; Khaleghi et al., 2013; Lahat et al., 2015) with different points of focus. Atrey et al. (2010) concentrates on multimodal fusion techniques for multimedia analysis. They group the available techniques into rule-based, classification-based, and estimation-based algorithms and provide several traditional techniques such as linear modeling, Support Vector Machines (SVMs), entropy maximization, and Kalman and particle filtering. This review article is distinct from the one by Atrey et al. (2010), both in terms of domain application and the algorithms covered. Khaleghi et al. (2013) provides a review of multi-sensor fusion with information theoretic perspective. The main focus of that review is on low-level fusion and discusses available frameworks in addressing multi-sensor fusion challenges such as imperfections, correlation, inconsistency, and disparateness. The frameworks introduced in that paper are related to probability theory, fuzzy set theory, possibility theory, rough set theory, and Dempster-Shafer evidence theory. Although, that paper is a great source for understanding formal definitions of data fusion challenges, it does not provide practical understanding of data fusion and available algorithms, particularly to the IE audience.

One of the more recent review papers on multimodal data analysis is authored by Lahat *et al.* (2015). The main focus of that paper is on early fusion using decomposition-based techniques such as ICA, canonical analysis, and tensor analysis. Our manuscript extends Lahat *et al.* (2015) in several ways. First, we will introduce methods beyond matrix/tensor decomposition and canonical analysis, as well as discuss techniques that perform intermediate fusion. Second, we introduce recent developments that were published after 2015. For example, tensor and factor analysis gained a significant attention in the IE community in the past few years, and are covered in this article. Finally, we tailor the methods to IE applications through examples as our main audience is this community.

Section 2 discusses recent developments in decompositionbased algorithms with a specific focus on IE applications; Section 3 presents neural network-based fusion; Section 4 covers a brief discussion on data and domain knowledge integration; Section 5 discusses current challenges and future research directions; and Section 6 concludes the article.

# 2. Decomposition-based fusion algorithms

In this article, we define decomposition-based fusion as fusion techniques that decompose data matrices or tensors to extract patterns and features. These methods do not employ neural networks. Factor analysis, tensor analysis, and dimensionality reduction methods, are examples of decomposition-based fusion that are covered in the following discussion. The focus of this section is not particularly on supervised or unsupervised methods, and we assume that a knowledgeable reader can distinguish them from the provided

Table 1. Summary of the decomposition-based methods and their corresponding capabilities (C) and limitations (L).

Framework	Descriptions	Capabilities (C) and Limitations (L)
Tensor Analysis (Bro,1996; Zhao et al., 2012; Zhou et al., 2013; Acar et al., 2014; Lock,2018; Mou et al., 2019; Fang et al., 2019; Yan et al., 2019; Yue et al., 2020; Gahrooei et al., 2020)	Extract common and uncommon structures among modalities via tensor decomposition and tensor calculus	<ul> <li>(C1) Suitable for heterogeneous datasets with different dimensions</li> <li>(C2) Suitable for extracting interpretable patterns between and within the modes</li> <li>(L1) Lack of identifiability and uniqueness</li> <li>(L2) Tensor rank selection</li> </ul>
Factor Analysis (Bro et al., 1997; Li et al., 2003; Wang et al., 2012; Virtanen et al., 2012; Klami et al., 2013; Acar et al., 2015; Argelaguet et al., 2018; Li and Li, 2019)	Describes covariance among observed modes and variables in terms of a potentially smaller set of latent variables (factors)	<ul> <li>(C2) Tensor Tank Selection.</li> <li>(C1) Interpretable model for understanding the underlying between-mode relationships</li> <li>(C2) Generates a lower-dimension representation of the original data</li> <li>(L1) Limited inference approaches are available</li> <li>(L2) Most estimation algorithms are based on Expectation-Maximization (EM) algorithms that are computationally expensive</li> <li>(L3) Unknown number of latent variables</li> <li>(L4) Models need to be better understood for particular applications for accurate factor recovery (Acar et al. 2015)</li> </ul>
Generalized PCA and beyond (Maaten and Hinton, 2008; Lampertand and Krömer, 2010; Candès et al., 2011; Zhang et al., 2011; White et al., 2012; White and Schuurmans, 2012; Li et al., 2017; Xiao et al., 2018)	Finds a low-dimensional representation across modalities	<ul> <li>(C1) Suitable for visualization</li> <li>(C2) Dimension reduction of both paired and unpaired modalities (Lampertand and Krömer, 2010)</li> <li>(C3) Linear and nonlinear dimensionality reduction (Maatenand and Hinton, 2008)</li> <li>(L1) Lack of scalability to incorporate into parallel and distributed computing structures (Candès <i>et al.</i>, 2011)</li> <li>(L2) Incorporation of decision-level fusion is lacking (Xiao <i>et al.</i>, 2018)</li> <li>(L3) t-SNE not guaranteed to converge to the global optimum of its cost function (Maatenand Hinton, 2008)</li> <li>(L4) t SNE connect differentiate modalities</li> </ul>
<i>Regularization</i> (Xiang <i>et al.</i> , 2014; Paynabar <i>et al.</i> , 2015; Zhang <i>et al.</i> , 2018a; Si <i>et al.</i> , 2020)	Selects informative modes and informative features within a mode	<ul> <li>(C4) First cannot unreference inductives.</li> <li>(C1) Suitable for automatic extraction of relevant modes and features</li> <li>(C2) Outputs interpretable, parsimonious models</li> <li>(L1) Increases parameters estimation complexity</li> <li>(L2) Highly-correlated modes and features may cause unstable estimations</li> </ul>

context. This section does not also explicitly mention whether a technique is suitable for early, late, or intermediate fusion as it should be clear from the context. As an example, tensor regression techniques are supervised and intermediate fusion methods, whereas coupled tensor decomposition are unsupervised methods that can be employed in both early or late fusion. Among the described methods, tensor analysis mainly focuses on the fusion of heterogeneous datasets that are different in their form (e.g., images and profiles). The other approaches mainly take each observation as a vector. Several regularization techniques are also discussed in this section. These techniques are complementary to decompostion-based fusion and are often used to improve the data fusion by pinpointing and eliminating noninformative modes or features. Therefore, regularization methods are subtractive. That is, while most of the methods combine different modes to obtain a feature, regularization techniques omit the uninformative modes/features. Table 1 reports a brief summary of decomposition-based fusion literature.

#### 2.1. Factor and canonical analysis

Factor Analysis (FA) is a method that describes covariance among observed variables in terms of a potentially smaller set of latent variables (called factors). FA takes the following basic form:

# $\mathbf{X} = \mathbf{Z}\mathbf{W}^T + \mathbf{E}$

where **X** is an  $n \times p$  matrix, where *n* represents the number of instances and *p* represents the number of observed variables; **Z**, the latent factor matrix, is an  $n \times p'$  matrix, where  $p' \leq p$ ; **W**, the factor loading matrix, is an  $p \times p'$  matrix and performs the transformation between latent and observed variables; and **E** is an error matrix. FA can be solved by traditional Maximum Likelihood Estimation (MLE) (Gaskin and Happell, 2014).

One of the first known instances of using FA in the context of multimodality fusion is CANDECOMP/PARAFAC (CP) decomposition (Bro *et al.*, 1997; Li, Choi, Perros, Sun, and Vudue, 2017). CP decomposition is often used for three datasets, but has the capability to be expanded to more. For simplicity, this article will describe the three-way formulation, which is summarized as follows:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + \epsilon_{ikj}; \ i = 1, ..., I; j = 1, ..., J, k = 1, ..., K$$

With associated sum of squares loss:

$$\min_{ijk}\sum_{ijk}\left\|x_{ijk}-\sum_{r=1}^{R}a_{ir}b_{jr}c_{kr}\right\|^2,$$



Figure 2. Group factor analysis (adapted from Virtanen et al. (2012)).

where  $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_R)$ ,  $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_R)$ , and  $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_R)$  denote the  $I \times R$ ,  $J \times R$ , and  $K \times R$  matrices containing the R different factor loadings in the three datasets. The model may also be written as

$$\sum_{r=1}^R \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r,$$

where  $\mathbf{a}_r$ ,  $\mathbf{b}_r$ , and  $\mathbf{c}_r$  are the rth columns of A, B, and C, respectively. Under assumptions of Gaussian noise, CP decomposition can be solved via MLE. CP decomposition has previously been applied to raw, high-dimensional electronic health records to identify useful phenotypes or medical concepts that can be utilized for patient diagnosis, prognosis and treatment (Ho *et al.*, 2014; Wang *et al.*, 2015; Li, Cerise, Yang and Han, 2017). CP has also been used for change detection and monitoring systems with image or multichannel data (Yan *et al.*, 2014; Li *et al.*, 2015)

Bayesian Group FA (Virtanen *et al.*, 2012) is a method that is capable of finding factors of different types–namely, those specific to all sources, a combination of some data sources, a single data source, or "noise" factors. The main task of Bayesian Group FA is to find a set of factors that explains dependencies between all possible subsets of the data sources. Bayesian Group FA takes the following form:

$$[\mathbf{X}_1 | \mathbf{X}_2 | ... | \mathbf{X}_M] \approx \mathbf{Z} \mathbf{W}^T$$

where  $\mathbf{X}_m$ , m = 1, ..., M, are  $n \times p_m$  matrices represent the different set of possible data sources; pm is the number of features in modality m; Z is a  $n \times p'$  matrix that represents the latent components; W, the factor loading matrix, is an  $p \times p'$  matrix that is group-wise sparse (i.e., a sparsity constraint is applied with respect to the variables in each modality), so each factor is active only in some subset of data sources, all of them, or only one. If the factor is active in only one modality, it is associated with noise or independent variation of that particular modality. W is made sparse by a group-wise Automatic Relevance Determination (ARD) prior. Figure 2 demonstrates an illustration of this method. Bayesian Group FA has been applied in the area of drug discovery by identifying systems-level drug-response phenotypes from genome-wide transcriptomic profiles (Khan et al., 2014; Yadav et al., 2015; Kibble et al., 2016).

Bayesian Interbattery FA (BIBFA) is a method motivated by Canonical Correlation Analysis (CCA) (Klami *et al.*, 2013; Acar *et al.*, 2015). However, instead of merely extracting correlated components between datasets like CCA, BIBFA takes into consideration both shared and unshared components (originally based on the non-probabilistic form,

InterBattery FA (IBFA) (Tucker, 1958). Given two data sources,  $X_1$  and  $X_2$ , BIBFA relies on the following probabilistic form of IBFA:

$$\mathbf{Z}_0 \sim N(\mathbf{0}, \mathbf{I}); \, \mathbf{Z}_m \sim N(\mathbf{0}, \mathbf{I}); \mathbf{X}_m \sim N(\mathbf{A}_m \mathbf{Z}_0 + \mathbf{B}_m \mathbf{Z}_m, \mathbf{\Sigma}_m),$$

where  $N(\mu, \Sigma)$  is the normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$ ;  $\Sigma_m$  is a diagonal matrix that describes the covariance of modality m; **1** and **0** are the identity and zero matrices;  $Z_0$  denotes the factors shared between the data sources;  $Z_m$  denotes the unshared factors in each data source  $X_m$ , m = 1, 2.

To derive an effective method to solve the model, it can be reformulated as follows:

$$\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}); \mathbf{X} \sim N(\mathbf{W}\mathbf{Z}, \mathbf{\Sigma})$$

where

$$\mathbf{W} = egin{pmatrix} \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{0} \ \mathbf{A}_2 & \mathbf{0} & \mathbf{B}_2 \ \end{pmatrix}, \mathbf{Z} = egin{pmatrix} \mathbf{Z}_0 \ \mathbf{Z}_1 \ \mathbf{Z}_2 \ \end{pmatrix}, \mathbf{\Sigma} = egin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \ \mathbf{0} & \mathbf{\Sigma}_2 \ \end{pmatrix}.$$

An appropriate structure in **W** is accomplished by imposing group-wise sparsity via an ARD prior. For inference, variational approximation is used (based on priors that assume maximally orthogonal latent factors). See Klami *et al.* (2013) for more information on solving BIBFA. BIBFA has been applied to cancer gene prioritization via DNA copy number data and integrative analysis of mRNA expression (Lahti *et al.*, 2013).

Cross-modal Factor Analysis (CFA) combines two data sources,  $\mathbf{X}_1$  and  $\mathbf{X}_2$  by finding two linear transformations,  $\mathbf{W}_1$  and  $\mathbf{W}_2$  for each data source (Li *et al.*, 2003).  $\mathbf{X}_m$  is a  $n \times p_m$  matrix, whereas  $\mathbf{W}_m$  is a  $p_m \times p'_m$  matrix, where  $p'_m \leq p_m$  and m = 1, 2. CFA is formulated as follows:

$$\min_{\mathbf{W}_1,\mathbf{W}_2} \|\mathbf{X}_1\mathbf{W}_1^T - \mathbf{X}_2\mathbf{W}_2^T\|_F^2$$
  
s.t.  $\mathbf{W}_m^T\mathbf{W}_m = \mathbf{I}, m = 1, 2,$ 

where  $\|\cdot\|_{F}^{2}$  is the Frobenius norm and **I** is an identity matrix of appropriate dimensions. Li *et al.* (2003) showed that the formulation can be reduced to

$$\max_{\mathbf{W}_1, \mathbf{W}_2} \operatorname{Tr} \mathbf{X}_1 \mathbf{W}_1 \mathbf{W}_2^T \mathbf{X}_2^T$$
  
s.t.  $\mathbf{W}_m^T \mathbf{W}_m = \mathbf{I}, m = 1, 2,$ 

which can be solved via Singular Value Decomposition (SVD) to obtain the final transformed latent factors,  $\mathbf{Z}_m$  as,

$$\hat{\mathbf{Z}}_m = \mathbf{X}_m \mathbf{W}_m^T, m = 1, 2$$

Unlike CCA, CFA does not need to calculate the inverse of the covariance matrices, provides orthogonal transformations, and does not require independence of vectors,  $\mathbf{X}_m$ . However, CFA cannot provide correct information associations if the modalities are not linearly related. An extension of CFA, called Kernel CFA (KCFA) handles this issue by mapping the  $\mathbf{X}_m$  vectors in the original space to a high-dimensional space via the kernel trick (Wang *et al.*, 2012). One application for CFA is in integration of multimedia sources (i.e., audio and video); CFA can compensate for missing or noisy media sources and effectively integrate multiple streams of information together (Li *et al.*, 2003).

Multi-Omics Factor Analysis (MOFA) (Argelaguet *et al.*, 2018) is a recently developed method made specifically for integrating multiple omics data modalities, but capable to be applied to other applications as well. MOFA builds upon Bayesian Group FA (Virtanen *et al.*, 2012), by (i) enabling fast inference via variational approximation, (ii) inducing sparse solutions to help interpretation, (iii) handling missing values in an efficient manner, and (iv) allowing for flexibility in combining different likelihood models for each data modality. MOFA takes on the following form:

$$\mathbf{X}_m = \mathbf{Z}\mathbf{W}_m^T + \mathbf{E}_m, m = 1, ..., M$$

where  $\mathbf{X}_m$  denotes the original feature matrix for modality m,  $\mathbf{Z}$  is the factor matrix (common for all data modalities),  $\mathbf{W}_m$  is the weight matrix corresponding to modality m, and  $\mathbf{E}_m$  is the error term for the particular modality m. The MOFA model is formulated in a probabilistic Bayesian framework, in which a prior distribution is applied on all unobserved variables. MOFA utilizes two levels of regularization: (i) view- and factor-wise sparsity via an ARD prior, which helps identify which factor is active in which view, and (ii) feature-wise sparsity via a spike-and-slab prior that usually results in a smaller number of features with active weights.

Statistical inference for FA of multimodal data has mainly remained unexplored; however, one recent work examines (i) how to infer the significance of one data source given other sources in the model, (ii) how to infer the significance of a combination of variables across different modalities, or from a single modality, and (iii) how to quantify the contribution of one data source given the other data sources, using a goodness-of-fit measure (Li and Li, 2019).

Multimodal datasets have also been employed for more effective clustering of the data. One key challenge in designing clustering algorithms is to identify which modes and which features within each mode are informative for distinguishing clusters. Regularization techniques combined with factor analysis have been introduced for this purpose. For example, Si *et al.* (2020) introduced a hierarchical clustering approach that uses an  $L_{12}$  penalty to effectively identify informative modes and features. Specifically, let  $\mathbf{x}_{m,i}$  denote the vector of features of *m*th mode in the *i*th sample. Then, the following factor analysis is considered,

$$\mathbf{x}_{m,i} = \mathbf{H}_m \mathbf{f}_{m,i} + \mathbf{B}_m \mathbf{z}_i + \mathbf{e}_{m,i},$$

where  $\mathbf{f}_{m,i}$  is a latent factor,  $\mathbf{z}_i$  is a vector of known covariates,  $\mathbf{H}_m$  and  $\mathbf{B}_m$  are the loading matrices, and  $\mathbf{e}_{m,i}$  represents the model errors that follow a multivariate normal

distribution with mean zero and covariance matrix  $\Sigma_m$ . The latent factors are then linked to clusters  $\mathbf{s}_i$ :

$$\mathbf{f}_{m,i} = \mathbf{A}_m \mathbf{s}_i + \mathbf{v}_{m,i},$$

where  $\mathbf{A}_m$  is a loading matrix and  $\mathbf{v}_{m,i}$  is an error term that follows a multivariate normal distribution with mean zero and covariance matrix  $\Psi_m$ . The goal is to estimate the model parameters,  $\mathbf{\Theta} = \{\mathbf{H}_m, \mathbf{B}_m, \mathbf{A}_m, \boldsymbol{\Sigma}_m, \Psi_m\}$  while imposing sparsity on loading matrices  $\mathbf{H}_m$  and  $\mathbf{A}_m$ . Imposing sparsity facilities the selection of informative modes and features. The following objective function has been minimized through Expectation-Maximization (EM) technique to achieve this goal:

$$-l(\mathbf{\Theta}) + \sum_{j} \sum_{m} (||\mathbf{h}_{m}^{j}||_{2}) + \sum_{m} (||\mathbf{A}_{m}||_{2}),$$

where  $\mathbf{h}_m^j$  denotes the *j*th column of  $\mathbf{H}_m$  and  $||.||_2$  represents the  $L_{21}$  norm. The model was applied to a multimodal MRI dataset of brain cortical area, thickness and volume to identify subgroups of migraine patients.

#### 2.2. Tensor and functional data analysis

One of the main challenges in integrating multimodal datasets is identifying a unified and flexible representation of each mode of a dataset without losing information. Defining such representation allows the use of mathematical tools that apply to all modes of data. A straightforward approach is to model all the data as vectors. However, this approach breaks down and loses the structural information of data instances such as images. Recent advances in multi-linear algebra (Kolda, 2006) create an unprecedented opportunity for the use of tensors (i.e., higher-order arrays) and tensor calculus for multimodal data fusion. In this section, we describe different existing approaches used for integrating data using tensor analysis.

#### 2.2.1. Tensor regression approaches

Tensor regression is an early or intermediate data fusion approach that uses tensor decomposition techniques, including CP and Tucker decomposition (Kolda, 2006), to extract features out of an input tensor that contains data from multiple sources to estimate the output. Most approaches focus on single tensor input (Bro, 1996; Zhao *et al.*, 2012; Zhou *et al.*, 2013; Lock, 2018; Fang *et al.*, 2019; Yan *et al.*, 2019; Yue *et al.*, 2020) but multiple tensor-input regression frameworks are also available (Gahrooei *et al.*, 2020). Scalar-ontensor, tensor-on-scalar, and tensor-on-tensor models are different forms of tensor regression modeling that have been introduced in the literature. Let us denote a tensor by calligraphy font. For example,  $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$  is a tensor with *d* modes, whose *k*th mode is of size *Ik*. A linear scalar-on-tensor model (Yan *et al.*, 2019) is usually formulated as

$$\mathcal{Y} = \mathcal{B} \times_{d+1} \mathbf{X} + \mathcal{E},$$

where  $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ ;  $\mathbf{X} \in \mathbb{R}^{I_1 \times p}$ ;  $\mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_d \times p}$ ;  $\mathcal{E}$  is the tensor of errors; and  $\times_i$  denotes the *i*th-mode multiplication of a tensor and a matrix. The linear tensor-on-scalar (Zhou

*et al.*, 2013; Fang *et al.*, 2019) models are those with scalar output and tensor inputs as follows:

$$y = < \mathcal{B}, \mathcal{X} > +e,$$

where y is a scalar;  $\mathcal{X} \in \mathbb{R}^{Q_1 \times \cdots \times Q_l}$ ;  $\mathcal{B} \in \mathbb{R}^{Q_1 \times \cdots \times Q_l}$ ; e is the error term; and  $\langle ..., \rangle$  denotes the tensor inner product, which is equivalent to the inner product of the vectorized version of the tensors. Finally the tensor-on-tensor regression model (Lock, 2018; Gahrooei *et al.*, 2020) is defined as

$$\mathcal{Y} = < \mathcal{B}, \mathcal{X} >_l + \mathcal{E},$$

where  $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ ;  $\mathcal{X} \in \mathbb{R}^{Q_1 \times \cdots \times Q_l}$ ;  $\mathcal{B} \in \mathbb{R}^{Q_1 \times \cdots \times Q_l \times I_1 \times \cdots \times I_d}$ ; and  $\mathcal{E}$  is the tensor of errors. The operation  $\langle ., . \rangle_l$  is called the tensor contraction operation over *l* modes. In all the cases, the tensor of parameters  $\mathcal{B}$  is assumed to be low-rank and has a decomposition form that results in a lower number of parameters to be estimated. Depending on the learning algorithm, these models can be viewed as either early or intermediate fusion. In particular, if the decomposition of the parameter tensor is learned during the training, the model is of intermediate fusion type. Tensor regression models are suitable for the fusion of heterogeneous sources of data when multi-channel and multi-dimensional data sets are available (Gahrooei et al., 2020). Tensor regression models have been applied in modeling and optimization of lathe-turning process (Yan et al., 2019), in semiconductor manufacturing to predict overlay error of the lithographic process based on wafer shape (Gahrooei et al., 2020) as well as prediction of the aged state of Ni-based superalloys (Gorgannejad et al., 2019). It has also been applied in prognostics applications for prediction of useful remaining lifetime based on thermal images (Fang et al., 2019)

#### 2.2.2. Coupled decomposition

Modes of the multimodality data may contain common and uncommon features to be discovered. The coupled decomposition technique decomposes the tensors assuming they share a common subspace spanned by a set of bases. Simultaneous matrix and tensor factorization is an approach for discovering the common bases. Let  $\mathbf{X} \in \mathbb{R}^{P \times Q}$  and  $\mathcal{Y} \in \mathbb{R}^{P \times I_1 \times I_2}$ . By minimizing the following objective function, Acar *et al.* (2013) identify the common space, spanned by basis matrix **U**, between the tensor and matrix modes:

$$L = ||\mathcal{Y} - [\lambda; \mathbf{U}, \mathbf{V}_1, \mathbf{V}_2]||^2 + ||\mathbf{X} - \mathbf{U}\mathbf{W}^{\mathrm{T}}||^2,$$

where  $[\![\lambda; \mathbf{U}, \mathbf{V}_1, \mathbf{V}_2]\!]$  denotes CP decomposition of the tensor with basis matrices  $\mathbf{U}, \mathbf{V}_1$ , and  $\mathbf{V}_2$ .

Zhao *et al.* (2012) proposed Higher-order PLS (HOPLS) as an approach for identifying the common features between two tensors that represent different modalities. Let  $\mathcal{X} \in \mathbb{R}^{P \times Q_2 \cdots \times Q_l}$  and  $\mathcal{Y} \in \mathbb{R}^{P \times I_2 \cdots \times I_d}$ . Then, the coupled Tucker decomposition of these two tensors are written as:

$$\mathcal{X} = \mathcal{C}_1 \times_1 \mathbf{U} \times_2 \mathbf{U}_2 \cdots \times_l \mathbf{U}_l$$

and

$$\mathcal{Y} = \mathcal{C}_2 \times_1 \mathbf{U} \times_2 \mathbf{V}_2 \cdots \times_d \mathbf{V}_d.$$

Here,  $C_1$  and  $C_2$  are decomposition core tensors and  $U_2, ..., U_l, V_2, ..., V_d$  are bases that do not coincide between the

two tensors. The matrix **U** denotes the common basis that spans the common subspace between the two modes. Zhao *et al.* (2012) proposed an approach based on the Singular Value Decomposition of the covariance matrices to estimate the basis matrices. The approach was applied to decoding electrocorticography (ECoG) signals in relation to 3D hand trajectories of monkeys performing movement tasks. Potentially, this approach can be applied to before and after medical intervention data, including the MRI and Electroencephalography (EEG) signals of patients before and after an intervention.

#### 2.2.3. Structural revealing decomposition

The main limitation of coupled decomposition is that it only finds the common bases for the shared mode. However, the shared mode may contain common and uncommon bases. The structural revealing technique is designed to resolve this issue. Nevertheless, it is only designed for situations where the data modes are a 3-D tensor and a matrix. More specifically, let  $\mathbf{X} \in \mathbb{R}^{P \times Q}$  and  $\mathcal{Y} \in \mathbb{R}^{P \times I_2 \times I_3}$ . Then the structural revealing technique will identify the common and uncommon features related to the shared dimension of the tensor and matrix by minimizing the following objective function:

$$L = ||\mathcal{Y} - [\lambda; \mathbf{U}, \mathbf{V}_1, \mathbf{V}_2]||^2 + ||\mathbf{X} - \mathbf{U}\mathbf{\Sigma}\mathbf{W}||^2 + ||\lambda||_1 + ||\sigma||_1,$$

where  $[\![\lambda; \mathbf{U}, \mathbf{V}_1, \mathbf{V}_2]\!]$  denotes CP decomposition of the tensor with basis matrices  $\mathbf{U}, \mathbf{V}_1$ , and  $\mathbf{V}_2$ ;  $\mathbf{U}\Sigma\mathbf{W}$  denotes the SVD decomposition of the matrix with singular values  $\sigma = diag(\Sigma)$ . In this formulation  $\mathbf{U}$  contains common and uncommon bases, which are identified by the lasso penalization over  $\lambda$  and  $\sigma$ . The utility of this method was tested on a dataset of mixtures with known chemical composition (inferred from nuclear magnetic resonance and mass spectrometry) and it was found that the method can successfully determine the chemicals in the mixtures, as well as their relative concentrations. For more details, please refer to Acar *et al.* (2014).

#### 2.2.4. Self-expressive models

Subspace clustering has been used in many applications, including systems monitoring and diagnosis based on profile data (Zhang *et al.*, 2020). One main approach for subspace clustering is the self-expressive model, which clusters a set of signals into subspaces (Elhamifar, 2016). Let  $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]$  to be a matrix containing *n* signals of length *D*. Then, the self expressive model assumes that each signal can be explained by other signals within the matrix:

$$Y = YC + E,$$

where diag(C) = 0 and E is a matrix of errors. To find clusters of signals, i.e., a set of similar signals, one may impose sparsity or low-rankness on *C*. Therefore computing the set of model parameters *C* requires solving the following optimization algorithm:

$$\frac{1}{2}||\mathbf{Y} - \mathbf{Y}\mathbf{C}||_{2}^{2} + \lambda||\mathbf{C}||_{1} + ||\mathbf{C}||_{*},$$

where  $||.||_1$  and  $||.||_*$  refer to  $L_1$  and nuclear norms, respectively. The nuclear norm of a matrix is the sum of its singular values and is used for imposing low-rankness.

The extension of self-expressive models has been introduced for Multimodal datasets (Abavisani and Patel, 2018). Let  $\{\mathbf{Y}_1, ..., \mathbf{Y}_m\}$  be an *m*-mode dataset where each mode contains a set of signals similar to the case of uni-mode self-expressive model. Then the goal of Multimodal selfexpressive models is to simultaneously cluster the signals in distinct modalities according to their subspaces. For this purpose, one can minimize the following objective function:

$$\frac{1}{2}\sum_{i=1}^{m}||\mathbf{Y}_{i} - \mathbf{Y}_{i}\mathbf{C}||_{2}^{2} + \lambda||\mathbf{C}||_{1} + ||\mathbf{C}||_{*}$$

The model has been applied to cluster face images using various facial components (i.e., eyes, nose and mouth) (Elhamifar, 2016).

#### 2.2.5. Regularization for functional and multimodal data

Regularization techniques have been used for identification of the informative modes of data, as well as informative features within each mode. The most common regularization techniques used for this purpose are based on non-negative Garrote and  $L_{21}$  norm. Paynabar *et al.* (2015) integrated profiles data to predict a scalar output. The informative profiles and features within the selected profiles were identified using a hierarchical non-negative Garrote technique. Let  $\mathbf{y} \in \mathbb{R}^n$  denote the vector of outputs,  $\mathbf{C}_k$  be the design matrix of the kth(k = 1, 2, ..., K) group (profile), and  $\beta_k^{ols}$  be the ordinary least square parameter. Then by minimizing

$$L = ||\mathbf{y} - \sum \mathbf{C}_k \beta_k^{ols} d_k||_2^2 + \lambda \sum_k^K d_k,$$

the shrinkage term  $d_k \ge 0$  is estimated. If  $d_k = 0$ , then the group is considered as uninformative. By only considering selected groups (profiles) and imposing a lasso penalty on the feature shrinkage terms, the informative features can be identified. This model was applied to develop a predictive model that can estimate vehicle design comfort from three-dimensional motion signals of test drivers.

Zhang *et al.*, (2018b) employed sparsity penalty to monitor a weakly correlated set of profiles (multi-channel data). Specifically, they expanded the multivariate functional eigendecomposition technique to the situation where not all eigen-functions are informative for spanning a given profile. Let  $\mathbf{Y}_i \in \mathbb{R}^{n \times p}$  denote the *i*th (i = 1, ..., N) sample that contains p profiles (channels), each observed at n points;  $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K]$  be a matrix whose columns are of eigen-functions; and  $\Psi_i = [\psi_{i1}, ..., \psi_{iK}]$  be the spanning coefficients. Then by minimizing

$$\sum_{i=1}^{N} ||\mathbf{Y}_{i} - \mathbf{V} \Psi_{i}||_{2}^{2} + \sum_{i} \sum_{j=1}^{K} ||\psi_{ij}||,$$

subject to

$$\mathbf{V}^{T}\mathbf{V}=\mathbf{I},$$

the sparse multi-channel decomposition is achieved. This technique is used for monitoring semiconductor manufacturing, where a large number of weekly correlated sensors are available. Gahrooei, Payanbar, Pacella and Shi (2019) also combined a large number of profile inputs to estimate a profile output using a functional group lasso penalty, and applied the model to retrieve joint motion trajectory based on other joint trajectories from sensors located on the the human body (e.g., hip, neck, elbows, knees, etc.).

A generalized sparse model for multimodality data (assuming every instance has all modalities) was formulated by Xiang *et al.* (2014) using a multi-task framework. The two stages of this model are (i) to learn different models for each data-modality, and (ii) combine the learned models appropriately. The formulation is summarized below:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\gamma}} \frac{1}{2} \| \mathbf{y} - \sum_{m=1}^{M} \boldsymbol{\gamma}_m \cdot \mathbf{X}_m \boldsymbol{\alpha}_m \|_2^2 + \sum_{m=1}^{M} \frac{\lambda_m}{p} \| \boldsymbol{\alpha}_m \|_p^p + \sum_{m=1}^{M} \frac{\eta_m}{q} | \boldsymbol{\gamma}_m |^q,$$

where **y** is the response variable;  $\mathbf{X}_m$  are the features for modality *m*;  $\boldsymbol{\alpha}_m$  are the weights of the linear model learned for the *m*th modality;  $\gamma$  are the weights that combine the learned models together; *p* and *q* are adjustable integers that can induce the desired sparsity on the feature (*p*) and instance level (*q*); and finally,  $\lambda m$  and  $\eta m$  are tuning parameters. This model can be reduce to common regularization methods and solved by standard multi-task learning algorithms.

A more realistic (and challenging) scenario occurs when not every modality is available for every instance. Xiang *et al.* (2014) developed a bi-level multi-source learning algorithm for heterogeneous block-wise missing data to handle missing medical modalities in an Alzheimer's Disease dataset. That paper developed methods of handling situations in multimodality datasets when not all instances have all data sources. First, the instances are divided into different groups according to which data modalities are available. Then, using a similar strategy as the generalized sparse model for multimodality data (the model presented previously) the two stages of this model were (i) to learn different models for each data-group, and (ii) combine the learned models appropriately. The formulation is as follows:

$$\begin{split} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{|\boldsymbol{p}\boldsymbol{f}|} \sum_{m \in \boldsymbol{p}\boldsymbol{f}} \frac{1}{n} \mathcal{L}\left(\sum_{m=1}^{M} \alpha_m X_m \boldsymbol{\beta}, \mathbf{y}_m\right) + \lambda \mathbf{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \\ \text{s.t. } \mathbf{R}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_m) \leq 1, \forall m \in \boldsymbol{p}\boldsymbol{f} \end{split}$$

where  $\beta$  corresponds to the coefficients of features across all modalities;  $\alpha$  is the weighted combination across the different modalities; pf is an *n*-dimensional vector that encodes binary indicators for which modalities are present for each of the *n* instances;  $X_m$  are the features for modality *m* (m = 1, ..., M);  $\mathbf{y}_m$  is the response;  $\mathcal{L}$  is any convex loss function (e.g., least squares, logistic loss, etc.); and  $\mathbf{R}_{\alpha}$ ,  $\mathbf{R}_{\beta}$ are regularizations on  $\alpha$  and  $\beta$ , respectively. The solution to this model can be approximated via alternating optimization between  $\alpha$  and  $\beta$ . The benefit of this method is that out-ofsample test instances with different modality combinations can still be predicted, as the model is designed to have a generalized  $\beta$  across all modality combinations.

# 2.3. Generalized PCA and beyond

The goal of dimension reduction methods such as PCA is to find a low-dimensional representation across multiple heterogeneous data modalities that can be utilized for a variety of tasks (e.g., data compression, clustering, and model training). Dimension reduction falls under one of two categories (Lampert and Krömer, 2010): (i) inductive and (ii) noninductive. Inductive methods inherently include a function that can be used for future data. Non-inductive methods do not include such a function, and must be re-applied any time new data is obtained. There is not much work performed in non-inductive multimodality dimension reduction (Fernandez-Beltran *et al.*, 2018a, 2018b), so the focus of the discussion will be on inductive methods.

One of the first known examples of multimodality reduction is Maximum Covariance Analysis (MCA), a generalization of PCA (Tucker, 1958). Given two centered data sources,  $X_1$  and  $X_2$ , with dimensions  $n \times p_m$  (m = 1, 2), MCA performs multimodal dimension reduction by solving for  $W_1$  and  $W_2$  in the objective function below:

$$\max_{\mathbf{W}_1,\mathbf{W}_2} \mathrm{Tr} \mathbf{W}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{W}_2.$$

where  $W_1$  and  $W_2$  are orthogonal matrices with dimensions  $p_m \times p'_m$ , where  $p'_m < p_m$ , m = 1, 2. One of the main disadvantages of MCA is that it requires instances to be completely paired, i.e., both data sources have identicial instances. This scenario is often unrealistic in larger datasets.

Weakly Paired MCA (WMCA) allows for data sources to have differing instances, which provides greater flexibility with using all the available data (Lampert and Krömer, 2010). Given two centered data sources,  $X_1$  and  $X_2$ , with dimensions  $n_m \times p_m$  (m = 1, 2), MCA performs multimodal dimension reduction by solving for  $W_1$  and  $W_2$  in the objective function below:

$$\max_{\mathbf{W}_1,\mathbf{W}_2} \mathrm{Tr} \mathbf{W}_1^T \mathbf{X}_1^T \mathbf{\Pi} \mathbf{X}_2 \mathbf{W}_2.$$

where  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are orthogonal matrices with dimensions  $p_m \times p'_m$ , where  $p'_m < p_m$ , m = 1, 2;  $\boldsymbol{\Pi}$  is an  $n_1 \times n_2$  binary matrix that encodes the different grouping structure to take into account the weakly paired data. There is no closed form solution for this objective function, but it can be estimated via an alternating maximization of  $\mathbf{W}_1$ ,  $\mathbf{W}_2$ , and  $\boldsymbol{\Pi}$ . WMCA has been used to discriminate between different textures of materials (e.g., styrofoam, bricks, wallpaper, etc.) based on images and audio signals recorded over the surfaces (Lampert and Krömer, 2010). Additional description for the mechanics of  $\boldsymbol{\Pi}$ , an extension to the kernelized space, and expansion to handling more than two modalities can be found in Lampert and Krömer (2010).

Comparing traditional CCA (from Section 2.1) to these dimension reduction methods, CCA is limited in that it is restricted to a square loss under a particular normalization. In the single data modality case, there have been several papers that have been able to relax these constraints and allow for more generalizable cases that can incorporate several convex losses while maintaining a reasonable computational complexity (Candès *et al.*, 2011; Zhang *et al.*, 2011). One proposal for the Multimodality setting is convex multi-view subspace learning (White *et al.*, 2012). Assuming that two data sources,  $\mathbf{X}_1$  ( $n \times p_1$ ) and  $\mathbf{X}_2$  ( $n \times p_2$ ) are conditionally independent given their shared latent representation  $\mathbf{H}$  ( $n \times p'$ ), where *n* is the number of instances and  $p' < p_1 + p_2$  is the reduced dimension, an optimal data reconstruction can be found via utilizing an implicit convex regularizer that recovers  $\mathbf{H}$  jointly. One formulation of this method is summarized below:

$$\begin{split} \min_{\mathbf{W}_{1},\mathbf{W}_{2},\mathbf{H}} \mathcal{L}(\mathbf{H}\mathbf{W};\mathbf{X}) + \alpha \|\mathbf{H}^{T}\|_{2,1} \\ \text{s.t.} \left[ \mathbf{W}_{1(:,i)} \ \mathbf{W}_{2(:,i)} \right] \in \mathcal{C}, \ \forall i \\ \text{where } \mathcal{C} := \{ \|\mathbf{w}_{1}\|_{2} \leq \beta_{1}, \|\mathbf{w}_{2}\|_{2} \leq \beta_{2} \}, \mathbf{W} = \left[ \mathbf{W}_{1} \ \mathbf{W}_{2} \right], \mathbf{X} = \left[ \mathbf{X}_{1} \ \mathbf{X}_{2} \right] \end{split}$$

where  $\mathcal{L}$  is the convex loss function between HW and Z (some examples can be found in White and Schuurmans (2012));  $W_m$  represent the loading matrices such that  $HW_m \approx X_m$ ,  $m = 1, 2; \alpha, \beta_1$  and  $\beta_2$  are tuning parameters of the objective function. This method was applied to classify a face image dataset consisting of various poses and lighting conditions (Georghiades et al., 2001).t-distributed Stochastic Neighborhood Embedding (t-SNE) is a relatively recent dimension reduction and visualization technique that can be used to combine data features from single or multiple modalities and reduce to a two- or three-dimensional dataset (Maaten and Hinton, 2008; Li, Cerise, Yang and Han, 2017; Xiao et al., 2018). Using conditional probability theory, t-SNE assumes that coordinates in the low dimension follow a t-distribution, which has the effect of increasing the distance between formed clusters allowing for greater distinction between different instances. One instance of t-SNE for Multimodalities performs misalignment fault diagnosis of wind turbines by fusing time and frequency features from the vibration, temperature, and stator current signals, and generates two information-dense features (Xiao et al., 2018). The two features were then used as input for a least square SVM that was optimized by the artificial bee colony algorithm. This application of t-SNE to multiple modalities does not take into account differences between modalities (i.e., covariance), but rather inputs all features from the modalities as a single unit. More work will need to be performed in this area to better incorporate individual modality differences.

## 3. Neural network-based fusion

Neural networks, in particular deep learning models, have demonstrated great promise within several applications including medical, manufacturing, internet of things, remote sensing, and urban big data. There have also been several recent implementations proposed for multimodality fusion in each of these application areas (Calhoun and Sui, 2016; Schmitt and Zhu, 2016; Li *et al.*, 2018; Wang *et al.*, 2018; Liu, Li, Xie, Du, Teng and Yang, 2020; Qi *et al.*, 2020). A handful of review papers have also been written on this topic (Ramachandram and Taylor, 2017; Gao *et al.*, 2020). Multimodal neural network approaches have the goal of

Table 2. Summary of the neural network-based fusion methods and their corresponding capabilities (C) and limitations (L).

Framework	Descriptions	Capabilities (C) and Limitations (L)
Early Fusion (Ng, 2011; Srivastava and Salakhutdinov, 2012; Liu <i>et al.</i> , 2017; Zhang <i>et al.</i> , 2018a)	Extracts and fuses information from each modality before model training (commonly through the use of autoencoders)	<ul> <li>(C1) Generates a lower-dimensional representation of the original data</li> <li>(C2) Flexibility to use generated features for any statistical learning algorithm</li> <li>(L1) Incorporation of decision-level fusion is lacking</li> <li>(L2) Interpretability of generated features is not straightforward</li> <li>(L3) Most models assume conditional independence between modalities; but in practice, modalities tend to be highly correlated, e.g., multimodal medical images, video/audio, etc. (Ramachandran and Tavlor 2017)</li> </ul>
Late Fusion (Simonyan and Zisserman, 2014; Kahou et al., 2016; Wu et al., 2016)	Fuses predictions between multiple neural networks trained on different modalities using averaging, maximum value, Bayes decision rule, metaclassifiers, etc.	<ul> <li>(C1) Easy to implement since fusion is performed at a high-level</li> <li>(C2) Errors from multiple neural networks tend to be uncorrelated, making the late fusion feature independent</li> <li>(L1) Less flexibility in regards to when multimodal representations are learned and where multimodality fusion occurs</li> </ul>
		(L2) No conclusive evidence that late fusion is better than early fusion (Ramachandran and Taylor, 2017).
Intermediate Fusion (Karpathy et al., 2014; Neverova et al., 2015; Gao et al., 2018)	Fuses different modalities at various levels of the neural network during a model training task	<ul> <li>(C1) Offers additional flexibility beyond early or late fusion as to where multimodality fusion occurs in the network</li> <li>(C2) Allows for fusion of modalities at different levels and can generate many multimodal representations that can be used at the decision level</li> <li>(L1) Requires careful design for when and where to apply modality fusion in the network</li> </ul>

training an end-to-end architecture that achieves both high accuracy and informative modality fusion.

There are several advantages that multimodal neural networks have over conventional multimodal learning (Ramachandram and Taylor, 2017). Neural networks can learn both inter- and intra-modality representations with minimal preprocessing of input data, whereas decomposition-based methods often require manual design and are more sensitive to preprocessed data. Neural networks also provide implicit dimensionality reduction within the architecture, whereas with other techniques conventional feature selection methods (i.e., filter, wrapper, or embedded) must be incorporated. Additionally neural networks allow for the inter-modality fusion architecture to be learned during training, whereas other approaches usually must resort to hand-crafted fusion methods. Some challenges involved with using neural network methods include handling the high number of hyperparameters that must be tuned, which can require high computation and need powerful computer processing units (CPUs) or graphics processing units (GPUs) to train the model in a reasonable amount of time (this problem is especially found in deep learning models). On the other hand, other learning models typically do not suffer from as many hyperparameters and having a CPU or GPU cluster is typically not necessary.

Multimodal neural network-based fusion can be divided into early, late, and intermediate fusion. Table 2 summarizes the capabilities and limitations of the different deep fusion methods.

#### 3.1. Early fusion

Oftentimes it can be difficult to fuse multiple modalities, due to disparities between the modalities. For instance, two sensors being fused for a prediction model may be in different forms (e.g., have different sampling rates, one is analog and another is digital, etc.). To ameliorate some of these disparities, early fusion can be used to extract information from each modality and fuse before model training.

One of the common forms of early fusion comes in the form of autoencoders. An autoencoder is an unsupervised neural network that sets the target values to be equal to the input values (Ng, 2011). One of the hidden layers in the neural network (which has less elements than the input/ output) serves as an information bottleneck, from which a compressed representation of the input features can be derived. This concept can easily be implemented to find underlying shared representations in a multimodality setting. Zhang et al., (2018a) obtained eigenvectors from video and audio sources, which were then transformed via autoencoding into reconstructed eigenvectors that have a shared representation. Srivastava and Salakhutdinov (2012) developed two Boltzmann machines to combine text and image feature vectors in to a new feature vector that was used as input for a SVM classifier. Liu et al. (2017) integrated eigenvectors from different views of a face to improve face recognition by unifying the eigenvectors of the different views into a more descriptive and integrative eigenspace.

## 3.2. Late fusion

Late fusion involves the integration of predictions from multiple neural networks trained separately on different modalities. This method is attractive to many practitioners, as the combination of predictions from different modalities is more straightforward, especially when there are very different dimensionalities or sampling rates between the modalities (Ramachandram and Taylor, 2017). Various fusion rules are available in neural networks including averaging, maximum value, Bayes decision rule, and metaclassifiers (Ramachandram and Taylor, 2017). Simonyan and Zisserman (2014) fuse Convolutional Neural Networks (CNNs) trained on image and optic flow data for the purpose of action recognition. Kahou et al. (2016) combine audio and video data using CNNs, recursive neural networks, SVM and autoencoders. Wu et al. (2016) fuse output from a deep belief network and CNN trained on skeletal and image features to provide a posterior estimate of gesture recognition.

#### 3.3. Intermediate fusion

Due to their hierarchical nature, neural networks allow for the fusion of features at all intermediate levels and offer flexibility beyond early or late fusion. For example in Vielzeuf *et al.* (2018), uni-modal features, as well as a central joint representation at every layer, are trained toward a multi-task objective. In Joze *et al.* (2020), multi-modal Squeeze-and-Excitation (SE) modules are used to perform fusion at any intermediate level. Guided by the end learning objective, the SE modules adaptively adjust the contribution of the features of each modality, explicitly encouraging modalities to collaborate.

Deciding which features of each modality to fuse in neural networks can be a combinatoric search problem. Pérez-Rúa *et al.* (2019) propose a sequential, model-based architecture search approach to find the optimal fusion architecture, instead of empirically deciding what layer to fuse intermediate features. A recurrent surrogate model takes candidate model descriptions as input and predicts their performance on the end task, guiding the sampling process from the architecture search space. Ramachandram *et al.* (2018) employ Bayesian optimization using a graph-induced kernel for the same purpose.

The goal of intermediate fusion is to combine early and late fusion into a single framework. The typical workflow of an intermediate fusion network involves (i) transforming features from modalities into latent representations, (ii) fusing the representations from each modality into a single hidden layer, and (iii) learning a joint representation across the modalities to make a single prediction. Additionally, there is great flexibility with this framework since one can develop a neural network architecture that fuses various representations of the multimodal data at varying depths. Neverova *et al.* (2015) implemented a progressive fusion approach with visual, audio and motion capture data by fusing highly correlated modalities, then moving to fusion of less correlated modalities later in the architecture. Gao *et al.* (2018)

used a combination of shallow and deep CNN architectures to perform image reconstruction of different views of breast cancer images and generate two different types of feature representation sets that were combined for classification using a gradient boosting tree. Karpathy *et al.* (2014) introduced a model that fuses video stream representations in a gradual manner, using multiple fusion layers, and were able to show the superiority of intermediate fusion to early and late fusion approaches.

# 4. Discussion on domain knowledge and data fusion

One increasingly popular multimodal fusion research area we would like to highlight is data and domain knowledge integration. These hybrid models combine both objective information collected from a real application as well as theoretical knowledge of the underlying process via a mathematical/physical model. This process known as hybridization can occur in a variety of manners including arithmetic combination, mathematical model parameter estimation, Bayesian estimation, and feature input. These areas are summarized in Table 3 and briefly described below:

- Arithmetic combination: Arithmetic combination refers to the fusion of mathematical and machine learning model outputs in an arithmetic fashion. There is no change to the inner workings of either model, but instead both are treated as black boxes and their output is combined in a posthoc fashion. Brentan *et al.* (2017) added the outputs of Support Vector Regression (SVR) and a Fourier time series model for foreasting urban water demand. Chen and Irwin (2017) multiplied the outputs of a machine learning and physical model to improve solar forecasting.
- Mathematical model parameter estimation: Another class of hybridization methods uses machine learning to provide an estimate of some parameters for a mathematical/ physical model to better inform the prediction capabilities. To make the model more patient-specific, Clifton et al. (2017) used statistical linear estimation to fine-tune parameters for a mechanistic model of mobile health intervention in chronic pain. Meng et al. (2019) used machine learning to estimate parameters of a mechanistic model of cane sugar crystallization. Dong et al. (2016) utilized a statistical model to estimate parameters for a mechanistic model to improve forecasting of residential electricity. Mak et al. (2018) utilized Gaussian processes to estimate parameters in a mechanistic model to quantify turbulent flows in swirl injectors with varying geometries.
- *Bayesian framework*: A Bayesian framework can also be used to incorporate prior information from the mathematical model to the machine learning model and *vice versa*. Mascheroni *et al.* (2020) used a mechanistic model of tumor growth to inform the prior of a Bayesian model that incorporates additional empirical information to better inform the prediction. Albers *et al.* (2018) leveraged a Bayesian methodology to integrate physiologic knowledge

to phenotype. Li and Shi (2007) used Bayesian networks along with manufacturing domain knowledge to discover the causal relationships between the process quality and process variables.

Feature input: Another way to implement a hybrid model is to include output from the mathematical model as input to the training of a machine learning model. Liu, Clemente, Poirier, Ding, Chinazzi, Davis, Vespignani and Santillana (2020) utilized a mechanistic model to inform a machine learning model's prediction of future Covid-19 cases in China. Gaw et al. (2019) used the output of a mechanistic model of brain tumor growth and integrated it with a machine learning model as an input feature. They also encoded differences in the mechanistic model in the form of a graph to regularize tumor cell density predictions. Liu and Guo (2018) utilized output from a mechanistic model as a feature in a tree-based gradient boosting method (along with process conditions, such as cutting speed, feed per tooth, etc.) to predict the specific cutting energy of steel.

#### 5. Challenges and future research directions

There are several challenges associated with multimodality fusion. Atrey *et al.* (2010), Khaleghi *et al.* (2013), Lahat *et al.* (2014) and Ramachandram and Taylor (2017) have performed comprehensive reviews of these challenges and we provide these as additional references for the reader. Below, we highlight the main multimodality data fusion challenges that demand further research to be fully addressed:

• Missing modalities: In many instances, data from all modalities are not available across all instances or phases of model training (i.e., training, validation, and test sets). For example, in prognostics missing sensor data is prevalent (Fang et al., 2015). There have been a handful of methods developed in recent years to address these issues (Xiang et al., 2014; Liu et al., 2016; Galán et al., 2017; He et al., 2017; Adhikari et al., 2019; Liu, Chen, Wu, Weidman, Lure, Li and Alzheimer's Disease Neuroimaging Initative, 2020), but methodological progress is limited. Galán et al. (2017) and He et al. (2017) handle incomplete modality datasets via imputation algorithms. There are also deep learning analogues to imputation, for example, in the medical field, to transform images from one modality to another (ex., transforming MRI to Positron Emission Tomography (PET) images (Li et al., 2014). However, the current imputation approaches are limited when there are too many missing values to impute. Separate modeling is another solution, in which different models are trained for different cohorts (determined by the available data modalities). Nevertheless, this approach is also limited because the sample size in each cohort may be small and prevent construction of a generalizable model. One may improve this problem by including data modalities from some cohorts to append the instances in other cohorts (i.e., if one cohort has data modalities 1 and 2, and another cohort only has data modality 2, the instances in the first cohort for modality 2 can be appended to the instances in the second cohort). However, these methodologies still do not use all available information for model training. Another method addresses this issue by developing a transfer learning model that has flexibility to train on instances with differing missing modalities and predict out-of-sample instances with a different combination of modalities (Liu, Chen, Wu, Weidman, Lure, Li and Alzheimer's Disease Neuroimaging Initative, 2020). The model accomplishes this task via EM using the assumption that input features follow a normal distribution. More work needs to be performed to consider features and response variables that follow different distributions.

- Noncommensurability: Additionally, modalities may be difficult to combine when they are at different resolutions or aggregation levels (i.e., not commensurate with each other). However, there can be a large advantage to combining multiple modalities to make use of their strengths. For example, in medical imaging, functional imaging techniques (e.g., fMRI, EEG) are capable of collecting information about a patient's brain function over time, which allows an additional dimension of temporal resolution (Lahat et al., 2015). Unfortunately, having temporal resolution comes at the cost of a reduced spatial resolution. However, information from these images can be complemented by higher resolution medical images that do not have a time component, such as structural MRI and diffusion tensor imaging (Lahat et al., 2015). Gaw et al. (2018) created a machine learning framework that can combine features from structural and functional imaging at different aggregation levels. Additionally, in the area of meteorology, radar and satellite images provide large spatial coverage, but at the cost of not being able to measure precipitation at the ground level (Seyyedi, 2010). However, this can be overcome by utilizing information from rain gauges and microwave links to enhance resolution of actual amount of ground precipitation (Seyyedi, 2010; Liberman et al., 2014). Many of the challenges found in noncommensurability problems lie in the specific applications themselves. More work will need to be performed by practitioners to best understand ways to overcome specific issues that arise in practice.
- Noise: There may also be issues with different sources of noise across data modes. Each mode often has a different type of measurement tool or device, which can subsequently produce different magnitudes and kinds of error (Van Mechelen and Smilde, 2010; Lahat *et al.*, 2014). There has been some work performed to address discrepancies between noise in these different modes and how to properly weigh them (Khaleghi *et al.*, 2013; Şimşekli *et al.*, 2013). One work developed a wavelet transformbased fusion method that can combine multiple medical images (i.e., computed tomography, MRI, and PET) that is resilient to Gaussian or speckle noise (Prakash *et al.*, 2019). Additional work also needs to be performed in

Table 3.	Summarv	of the	data and	domain	knowledge	integration	as well as	s their	corresponding	capabilities (	C) ar	nd limitations	(L).
													(-/-

Framework	Descriptions	Capabilities (C) and Limitations (L)			
Arithmetic Combination (Brentan et al., 2017; Chen and Irwin, 2017)	Fuses mathematical and machine learning model outputs in an arithmetic fashion (e.g., addition,	<ul> <li>(C1) Easy to implement since fusion is performed at a high-level</li> <li>(L1) Fusion is performed superficially and does not fully consider the intricacies of either model</li> </ul>			
	subtraction, multiplication, division, etc.)				
		(L2) Less flexibility in regards to what aspects of each model to fuse and where multimodality fusion occurs			
Mathematical Model Parameter Estimation (Dong et al., 2016: Clifton et al., 2017: Mak et al., 2018:	Uses machine learning to provide an estimate of some parameters for a mathematical/	(C1) Can make mathematical model estimation more efficient (Mak <i>et al.</i> , 2018)			
Meng <i>et al.</i> , 2019)	physical model	(C2) Can introduce direct influence from empirical data directly into the estimation of the mathematical model			
		(L1) Can introduce unnecessary noise in mathematical model estimation (if empirical data is not relevant to the prediction/ classification task) and make mathematical model estimation more imprecise			
		(L2) Methods are generally limited to specific mathematical models and cannot be easily translated to other problems			
Bayesian Framework (Li and Shi, 2007; Albers et al.,	Incorporates prior information from the	(C1) Allows more flexibility as to where fusion			
2018; Mascheroni <i>et al.</i> , 2020)	mathematical model to the machine learning	occurs between the models			
	model and vice versa	(C2) Enables fusion of modalities in multiple ways (ex., hierarchical, graphical, etc.)			
		(L1) Most estimation algorithms are based on Markov Chain Monte Carlo methods that are computationally expensive			
		(L2) Requires careful design for when and where			
		to apply modality fusion in the network			
<i>Feature Input</i> (Liu and Guo, 2018; Gaw <i>et al.</i> , 2019; Liu, Clemente, Poirier, Ding, Chinazzi, Davis,	Includes output from the mathematical model as input to training of a machine learning model	(C1) Can significantly improve machine learning model accuracy (Gaw <i>et al.</i> , 2019)			
Vespignani and Santillana, 2020)		(C2) Flexibility to use features for any statistical learning algorithm of interest			
		(L1) Model fusion is indirect/imprecise because it			
		does not fully utilize the inner-mechanics of the mathematical model			
		(L2) Communication is only one-way from the mathematical model to the machine learning algorithm (no influence of machine learning on mathematical model estimation)			

examining the correlation between the noise of different modalities to improve predictive performance (Chlaily *et al.*, 2016).

- *Discordance*: It is possible that there is conflicting information in the modalities that are being fused, due to different views conveyed by the data in each modality. This may cause discrepancy in modality fusion, resulting in less confident fusion and/or prediction. Methods need to be developed that can be robust to such a phenomenon, and focus on features in each modality that are complementary to each other. In early or intermediate fusion, this may result from inconsistency in multimodal sensors in regards to their mutual information (Tmazirte *et al.*, 2013) or random events that may be a result of the nature of data collection or type of the sensor (Kumar *et al.*, 2007). In late fusion, this situation can be improved by a voting rule (Van Mechelen and Smilde, 2010).
- *Correlation considerations*: Correlation between modalities can be seen between different individual features and also between modalities as a whole. Constructing a model to take advantage of these connections still remains a challenging task. Highly correlated modalities may lead to models with high collinearity, requiring additional consideration with how to handle high correlation

(Gaw *et al.*, 2018). Additionally, some sensors in multimodal problems may also be subject to the same external noise, causing bias in their measurements that may lead to over or under confident predictions (Khaleghi *et al.*, 2013). In contrast, independent modalities (correlation = 0) may also present challenges with modality fusion. In these cases, one cannot rely on correlation between modalities to help with fusion, and must rely on other ways (e.g., confidence of individual modalities (Castellano *et al.*, 2008), etc.).

• Intermodality correlation: Often the incorporation of features from multiple modalities only indirectly considers intermodality correlation to inform model-building. There is a greater need to build models that consider intermodality correlation explicitly and incorporate it directly into the model building process. The authors have only found a limited number of works that harness intermodality correlation. In early fusion, Guo *et al.* (2018) proposes a canonical correlation analysis algorithm that performs joint intermodal and intramodal fusion for semi-paired scenarios. The advantage of this method is that it considers scenarios for which not all modalities have a strong pairing with each other ("semi-paired" scenarios), while also performing fusion that considers intermodality correlations that preserve intramodal correlations. Additionally, there are a couple methods that develop neural network architectures that more explicitly consider intermodality correlations (Peng *et al.*, 2017; Said *et al.*, 2017). Said *et al.* (2017) demonstrated a deep learning approach that utilizes intermodality correlation to improve classification of EEG and EMG signals. Peng *et al.* (2017) utilized both intramodality and intermodality correlation in a hierarchical classification network that takes as input texts and images to inform image and text retrieval tasks.

- Varying confidence levels: Each modality will often have a different level of confidence (Siegel and Wu, 2004), e.g., due to noise level, nature of the data collection task, correlation with response variable, etc. As an example, if given an audio and video data of a person crying, one may have higher confidence in predicting this event using audio instead of video (Atrey et al., 2010). In another example, Rankawat and Dubey (2017) fuse noisy ECG and atrial blood pressure signals by defining a beat Signal Quality Index (SQI) that indicates the level of noise in each modality, and incorporating the SQI into a majority voting scheme. A model that can take these aforementioned factors into account will be able to emphasize the "higher confidence" modalities, while still utilizing useful information from the less confident modalities that can improve the fusion quality. There is a need to more optimally quantify confidence in specific modalities for a particular task (e.g., in regards to measure of information content, relevance to prediction task, etc.) for the purpose of improving flexibility and adaptivity in modality fusion. Having this information would enhance capability to choose the degree of fusion and which modalities to fuse for each instance in the dataset (e.g., whether to place a higher emphasis on some modalities over others, not select particular modalities entirely).
- Negative transfer reduction: There is limited work in reducing negative transfer between different modalities. Developing models that can successfully integrate modality covariances (i.e., positive transfer), while also preventing fusion of conflicting information or correlations that are irrelevant model training (i.e., negative transfer), will provide valuable insights into how to best integrate interactions between modalities while also producing more accurate models. In one example, Yoon and Li (2018) develop a Positive Transfer Learning (PTL) model on telemonitoring data of Parkinson's Disease patients that is robust to negative transfer between patients' individual sub-models. The PTL model is built on the premise that not all information from patient data is useful for building accurate models of other patients, and it is necessary to identify the conditions for which negative transfer can happen and negatively affect the model. However this work does not handle multiple modalities, highlighting a need for more methods such as this one to be incorporated into multimodal fusion.
- Computation: Because of the complexity of some multimodality fusion algorithms, they often cannot be solved

analytically and approximation algorithms are needed. There have been studies that have successfully made approximation techniques that will eventually converge (Virtanen *et al.*, 2012; Xiang *et al.*, 2014; Zhang *et al.*, 2018b). However, there is still need to make more efficient approximation methods and improve efficiency to bring convergence rate to a more acceptable level.

- Theoretical criteria and verification: Even though extra modes of data provide additional information in most circumstances, integration of modes is not always beneficial and may cause deterioration in the performance of a model. For example, when integrating multi-accuracy data in applications such as geometric inspection and metrology (Gahrooei, Payanbar, Pacela and Colosimo, 2019) or building simulation (Safarzadegan Gilan et al., 2016), a data mode with high non-stationary bias or variance may harm the development of a surrogate model. Establishing a set of criteria for identifying suitable modes is a challenging task and requires theoretical analysis. Other theoretical questions includes: what is the best quantitative measure of success? What is the measure that quantifies the gain of integrating several modalities? Can we obtain a theoretical lower-bound of gain? What is the best error that is achievable by fusion of data?
- Multimodal data collection: Although several techniques are available for integration of multimodal data, the literature is very limited in how to design the collection of multimodal data to minimize the data collection cost while obtaining the adequate level of information. In telemedicine applications, for example, a visit from a doctor (accurate mode of health data) is an expensive means of data collection in comparison with wearable devices (low accuracy mode). How often a patient should visit a doctor given the low-accuracy data is the question to be addressed by the multimodal data collection policies. How to select the right modes, collect data, and integrate the modes is a major challenge. Gahrooei, Payanbar, Pacella and Colosimo (2019) proposed an adaptive sampling of high-accuracy data when low-accuracy data is available, and applied it to vehicle engine calibration application. However, this approach is limited to a static case and assumes the high-accuracy mode is already known.
- *Dynamic data fusion*: Limited work is available for dynamic data fusion of multimodal data. Such techniques can be applied in smart cities and telemedicine applications, where several multi-accuracy sensors are collecting information over time. Appropriate synchronization (i.e., when and how much data should be processed from each modality) should also be taken into consideration (Atrey *et al.*, 2010).
- *High-dimensional data*: Due to the nature of multimodal data, there is often a high number of features from which to choose for fusion. Sparse methods (Xiang *et al.*, 2014; Argelaguet *et al.*, 2018) can be used to select individual features or latent components can also be derived to infer the underlying patterns expressed in the data

(Kibble *et al.*, 2016; Li Choi, Perres, Sun and Vudue, 2017). Additional consideration should be made for determining when to reduce the dimension of the data (whether it should be as a preprocessing step within individual modalities before fusion, as a part of the fusion process across all modalities, or a combination of the two).

- Deep learning: The literature in IE-related applications has been limited in regards to incorporating deep learning models (Said *et al.*, 2017; Peng *et al.*, 2017; Gao *et al.*, 2018; Ramachandram *et al.*, 2018; Pérez-Rúa *et al.*, 2019; Joze *et al.*, 2020). The applications of deep learning are limited in this area due to lack of instances available for model training. However, there is possibility to use pre-trained neural networks that have reasonable degree of discriminative ability which can be transferred to new datasets (DenseNet121, 2018; InceptionV3, 2019). Additionally, generative shallow learning models that have the capability to produce new artificial instances can augment the training set and overcome the issues of small sample size.
- *Privacy*: Due to regulatory constraints from a particular application (e.g., medical, military, industry etc.), one may be limited in ability to fuse different modalities because of concerns of compromising confidentiality. There are only a limited number of works in this area (Kefayati *et al.*, 2007a, 2007b; Gao, 2020; Kumar and Diwakar, 2021). One such work utilizes the nonsubsampled shearlet transform combined with noise reduction to transform, share, and fuse imaging data through a secure environment (Kumar and Diwakar, 2021). Another work focuses on the privacy and preservation of information fused between multiagent systems in regards to synchronization, information fusion, decentralized control and load balancing (Gao, 2020).
- Objective measures of modality fusion: More work should be performed in finding objective metrics that evaluate the degree to which information from modalities are fused (Zhu *et al.*, 2018), along with measures of positive or negative transfer (for example, Yoon and Li (2018)). This will better assist researchers in this area to compare methods they are developing by providing objective benchmarks of "fusion performance".

# 7. Conclusion

As systems have become more advanced, there has been an increasing abundance of available data from different modality types. To advance the performance of statistical learning algorithms, it is crucial to understand how to best incorporate the relationships of these different modalities, while also avoiding a negative transfer of knowledge (i.e., transfer of discordant information between modalities). The key understanding that is necessary to implement and advance such algorithms is discerning the connections between the different data modes and how to best exploit them. Because multimodal datasets can be incredibly diverse, there is no one-size-fits-all model, which requires the practitioner to understand the particular application while developing a multimodal approach. As this work has demonstrated, there is potential for huge performance improvement from training a model on a single data mode if relationships between different available data modes are appropriately considered. It will be important to focus efforts in both decomposition-based models and neural networks, as some applications will require decompositionbased models when fewer training instances are at hand, while others with be able to employ the advantages of neural networks (especially deep learning) when many training instances are available. Progress in this area will span a broad number of applications, including systems monitoring/prognostics, healthcare, renewable energy, and many others since multimodal measurement technologies are emerging everywhere.

#### Data availability statement

There is no data set associated with this article.

#### Notes on contributors

*Nathan Gaw* is an Assistant Professor in the Department of Operational Sciences at Air Force Institute of Technology. He received his BS and MS in biomedical engineering and a PhD in industrial engineering from Arizona State University (ASU), Tempe, AZ, USA, in 2013, 2014, and 2019, respectively. Nathan's research focuses on multimodality fusion in healthcare and military applications fusing imaging, genetics, and telemonitoring data. He is a member of IISE, INFORMS, and IEEE.

*Safoora Yousefi* is an applied scientist at Microsoft. Safoora received their BSc in computer science from University of Tehran, and their PhD in computer science from Emory University. Prior to Microsoft, they interned at Google Brain and Roche. Their research interests include machine learning, multi-task, adversarial, and self-supervised learning and their applications to real world scenarios such as natural language processing and cancer genomics.

*Mostafa Reisi Gahrooei* received the master's degree in computational science and engineering and the PhD degree (2019) in industrial and systems engineering from the Georgia Institute of Technology, Atlanta, GA, USA, and the MSc degrees in transportation engineering and applied mathematics from the Southern Illinois University Edwardsville, Edwardsville, IL, USA. He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, the University of Florida, Gainesville, FL, USA. His research focuses on modeling, monitoring, and control of complex systems with multimodal, functional, and high-dimensional data. Dr. Reisi Gahrooei is a member of the Institute for Operations Research and the Management Sciences (INFORMS) and the Institute of Industrial and Systems Engineers (IISE).

#### References

- Abavisani, M. and Patel, V.M. (2018) Multimodal sparse and low-rank subspace clustering. *Information Fusion*, **39**, 168–177.
- Acar, E., Bro, R. and Smilde, A.K. (2015) Data fusion in metabolomics using coupled matrix and tensor factorizations. *Proceedings of the IEEE*, **103**(9), 1602–1620.

- Acar, E. Papalexakis, E.E., Gürdeniz, G., Rasmussen, M.A., Lawaetz, A.J., Nilsson, M. and Bro, R. (2014) Structure-revealing data fusion. BMC Bioinformatics, 15(1), 1–17.
- Acar, E., Rasmussen, M.A. Savorani, F., Naes, T. and Bro, R. (2013) Understanding data fusion within the framework of coupled matrix and tensor factorizations. *Chemometrics and Intelligent Laboratory Systems*, **129**, 53–63.
- Adhikari, S. Lecci, F., Becker, J.T., Junker, B.W., Kuller, L.H., Lopez, O.L. and Tibshirani, R.J. (2019) High-dimensional longitudinal classification with the multinomial fused lasso. *Statistics in Medicine*, 38(12), 2184–2205.
- Albers, D.J., Levine, M.E., Stuart, A., Mamykina, L., Gluckman, B. and Hripcsak, G. (2018) Mechanistic machine learning: How data assimilation leverages physiologic knowledge using Bayesian inference to forecast the future, infer the present, and phenotype. *Journal* of the American Medical Informatics Association, 25(10), 1392–1401.
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W. and Stegle, O. (2018) Multi-Omics Factor Analysis-A framework for unsupervised integration of multiomics data sets. *Molecular Systems Biology*, 14(6), e8124.
- Atrey, P.K., Hossain, M.A., El Saddik, A. and Kankanhalli, M.S. (2010) Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6), 345–379.
- Brentan, B.M., Luvizotto Jr, E., Herrera, M., Izquierdo, J. and Pérez-García, R. (2017) Hybrid regression model for near real-time urban water demand forecasting. *Journal of Computational and Applied Mathematics*, **309**, 532–541.
- Bro, R. (1996) Multiway calibration. multilinear PLS. Journal of Chemometrics, 10(1), 47–61.
- Bro, R. et al. (1997) Parafac. tutorial and applications. Chemometrics and Intelligent Laboratory Systems, 38(2), 149-172.
- Calhoun, V.D. and Sui, J. (2016) Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(3), 230-244.
- Candès, E.J., Li, S., Ma, Y. and Wright, J. (2011) Robust principal component analysis? *Journal of the ACM (JACM)*, **58**(3), 1–37, 2011.
- Castellano, G., Kessous, L. and Caridakis, G. (2008) Emotion recognition through multiple modalities: Face, body gesture, speech, in *Affect and Emotion in Human-Computer Interaction*, Springer, Milwaukee, WI, pp. 92–103.
- Chehade, A., Song, C., Liu, K., Saxena, A. and Zhang, X. (2018) A data-level fusion approach for degradation modeling and prognostic analysis under multiple failure modes. *Journal of Quality Technology*, 50(2), 150–165.
- Chen, D. and Irwin, D. (2017) Black-box solar performance modeling: Comparing physical, machine learning, and hybrid approaches. *ACM SIGMETRICS Performance Evaluation Review*, **45**(2), 79–84.
- Chlaily, S., Amblard, P.-O., Michel, O. and Jutten, C. (2016) Impact of noise correlation on multimodality, in 2016 24th European Signal Processing Conference (EUSIPCO), IEEE Press, Piscataway, NJ, pp. 195–199.
- Clifton, S.M., Kang, C., Li, J.J., Long, Q., Shah, N. and Abrams, D.M. (2017) Hybrid statistical and mechanistic mathematical model guides mobile health intervention for chronic pain. *Journal of Computational Biology*, 24(7), 675–688.
- DenseNet121. (2018) Densenet121. https://github.com/keras-team/ keras-applications/blob/master/keras\_applications/densenet.py (accessed 27 December 2020).
- Dong, D., Li, Z., Rahman, S.M.M. and Vega, R. (2016) A hybrid model approach for forecasting future residential electricity consumption. *Energy and Buildings*, 117, 341–351.
- Elhamifar, E. (2016) High-rank matrix completion and clustering under self-expressive models, in *Advances in Neural Information Processing Systems*, Barcelona, Spain, pp. 73–81.
- Fang, X., Gebraeel, N.Z. and Paynabar, N. (2017) Scalable prognostic models for large-scale condition monitoring applications. *IISE Transactions*, 49(7), 698–710.

- Fang, X., Paynabar, K. and Gebraeel, N. (2019) Image-based prognostics using penalized tensor regression. *Technometrics*, 61(3), 369–384.
- Fang, X., Zhou, R. and Gebraeel, N. (2015) An adaptive functional regression-based prognostic model for applications with missing data. *Reliability Engineering & System Safety*, 133, 266–274.
- Fernandez-Beltran, R., Haut, J.M., Paoletti, M.E., Plaza, J., Plaza, A. and Pla, F. (2018a) Multimodal probabilistic latent semantic analysis for sentinel-1 and sentinel-2 image fusion. *IEEE Geoscience and Remote Sensing Letters*, 15(9), 1347–1351.
- Fernandez-Beltran, R., Haut, J.M., Paoletti, M.E., Plaza, J., Plaza, A. and Pla, F. (2018b) Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4982–4993.
- Gahrooei, M.R., Paynabar, K., Pacella, M. and Colosimo, M.B. (2019) An adaptive fused sampling approach of high-accuracy data in the presence of low-accuracy data. *IISE Transactions*, 51(11), 1251–1264.
- Gahrooei, M.R., Paynabar, K., Pacella, M. and Shi, J. (2019) Process modeling and prediction with large number of high-dimensional variables using functional regression. *IEEE Transactions on Automation Science and Engineering*, **17**(2), 684–696.
- Gahrooei, M.R., Yan, H., Paynabar, K. and Shi, J. (2020) Multiple tensor-on-tensor regression: An approach for modeling processes with heterogeneous sources of data. *Technometrics*, 63(2), 147–159.
- Galán, C.O., Lasheras, F.S., de Cos Juez, F.J. and Sánchez, A.B. (2017) Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. *Journal of Computational and Applied Mathematics*, **311**, 704–717.
- Gao, F., Wu, T., Li, J., Zheng, B., Ruan, L., Shang, D. and Patel, B. (2018) SD:CNN: A shallow-deep CNN for improved breast cancer diagnosis. *Computerized Medical Imaging and Graphics*, **70**, 53–62.
- Gao, H. (2020) Coordination and Privacy Preservation in Multi-Agent Systems (Doctoral dissertation, Clemson University).
- Gao, J., Li, P., Chen, Z. and Zhang, J. (2020) A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5), 829–864.
- Gaskin, C.J. and Happell, B. (2014) On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, **51**(3), 511–521.
- Gaw, N., Hawkins-Daarud, A., Hu, L.S., Yoon, H., Wang, L., Xu, Y., Jackson, P.R., Singleton, K.W., Baxter, L.C., Eschbacher, J. et al. (2019) Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric mri. Scientific Reports, 9(1), 1–9.
- Gaw, N., Schwedt, T.J., Chong, C.D., Wu, T. and Li, J. (2018) A clinical decision support system using multi-modality imaging data for disease diagnosis. *IISE Transactions on Healthcare Systems Engineering*, 8(1), 36–46.
- Georghiades, A.S., Belhumeur, P.N. and Kriegman, D.J. (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis* and Machine Intelligence, **23**(6), 643–660.
- Gorgannejad, S., Gahrooei, M.R., Paynabar, K. and Neu, R.W. (2019) Quantitative prediction of the aged state of ni-base superalloys using pca and tensor regression. *Acta Materialia*, 165, 259–269.
- Guo, H., Wang, S., Tie, Y., Qi, K. and Guan, L. (2018) Joint intermodal and intramodal correlation preservation for semi-paired learning. *Pattern Recognition*, 81, 36–49.
- He, C., Liu, Q., Li, H. and Wang, H. (2010) Multimodal medical image fusion based on ihs and pca. *Procedia Engineering*, 7, 280–285.
- He, D., Wang, Z., Yang, L. and Dai, W. (2017) Study on missing data imputation and modeling for the leaching process. *Chemical Engineering Research and Design*, **124**, 1–19.
- Ho, J.C., Ghosh, J. and Sun, J. (2014) Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 115–124.
- Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**(3/4), 321-377.

- Ilarshman, RA. (1970) Foundations of the parafac procedure: Models and methods for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84.
- InceptionV3. Inceptionv3. https://github.com/keras-team/keras-applications/blob/master/keras\_applications/inception\_v3.py (accessed 27 December 2020).
- Joze, H.R.V., Shaban, A., Iuzzolino, M.L. and Koishida, K. (2020) MMTM: Multimodal transfer module for CNN fusion, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE/CVF, Seattle, WA, pp. 13289–13299.
- Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., *et al.* (2016) Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, **10**(2), 99–111.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) Large-scale video classification with convolutional neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE/CVF, Columbus, OH, pp. 1725–1732.
- Kefayati, M., Talebi, M.S., Khalaj, B.H. and Rabiee, H.R. (2007a) Secure consensus averaging in sensor networks using random offsets, in 2007 IEEE International Conference on Telecommunications and Malaysia International Conference on Communications, IEEE, Penang, Malaysia, pp. 556–560.
- Kefayati, M., Talebi, M.S., Rabiee, H.R. and Khalaj, B.H. (2007b) On secure consensus information fusion over sensor networks, in 2007 IEEE/ACS International Conference on Computer Systems and Applications, IEEE/ACS, Amman, Jordan, pp. 108–115.
- Kettenring, J.R. (1971) Canonical analysis of several sets of variables. *Biometrika*, 58(3), 433–451.
- Khaleghi, B., Khamis, A., Karray, F.O. and Razavi, S.N. (2013) Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, **14**(1), 28–44.
- Khan, S.A., Virtanen, S., Kallioniemi, O.P., Wennerberg, K., Poso, A. and Kaski, S. (2014) Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. *Bioinformatics*, **30**(17), i497–i504.
- Khasha, R., Sepehri, M.M. and Mahdaviani, S.A. (2019) An ensemble learning method for asthma control level detection with leveraging medical knowledge-based classifier and supervised learning. *Journal* of Medical Systems, 43(6), 1–15.
- Kibble, M., Khan, S.A., Saarinen, N., Iorio, F., Saez-Rodriguez, J., Mäkelä, S. and Aittokallio, T. (2016) Transcriptional response networks for elucidating mechanisms of action of multitargeted agents. *Drug Discovery Today*, 21(7), 1063–1075.
- Klami, A., Virtanen, S. and Kaski. S. (2013) Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr), 965–1003.
- Kolda, T.G. (2006) Multilinear operators for higher-order decompositions. Technical report, Sandia National Laboratories.
- Kumar, M., Garg, D.P. and Zachery, R.A. (2007) A method for judicious fusion of inconsistent multiple sensor data. *IEEE Sensors Journal*, 7(5), 723–733.
- Kumar, P. and Diwakar, M. (2021) A novel approach for multimodality medical image fusion over secure environment. *Transactions on Emerging Telecommunications Technologies*, **32**(2), e3985.
- Lahat, D., Adalý, T. and Jutten, C. (2014) Challenges in multimodal data fusion, in 2014 22nd European Signal Processing Conference (EUSIPCO), IEEE Aerospace and Electronic Systems Society (AESS), Lisbon, Portugal, pp. 101–105.
- Lahat, D., Adali, T. and Jutten, C. (2015) Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, **103**(9), 1449–1477.
- Lahti, L., Schäfer, M., Klein, H.-U., Bicciato, S. and Dugas, M. (2013) Cancer gene prioritization by integrative analysis of MRNA expression and DNA copy number data: A comparative review. *Briefings in Bioinformatics*, 14(1), 27–35.
- Lampert, C.H. and Krömer, O. (2010) Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer

learning, in *European Conference on Computer Vision*, Springer, Heraklion, Crete, Greece, pp. 566–579.

- Li, D., Dimitrova, N., Li, M. and Sethi, I.K. (2003) Multimedia content processing through cross-modal association, in *Proceedings of the Eleventh ACM International Conference on Multimedia*, ACM, Berkeley, CA, pp. 604–611.
- Li, J., Choi, J., Perros, L., Sun, J. and Vuduc, R. (2017a) Model-driven sparse CP decomposition for higher-order tensors, in 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE Press, Piscataway, NJ, pp. 1048–1057.
- Li, J. and Shi, J. (2007) Knowledge discovery from observational data for process control using causal Bayesian networks. *IIE Transactions*, 39(6), 681–690.
- Li, N., Gebraeel, N., Lei, Y., Fang, X., Cai, X. and Yan, T. (2021) Remaining useful life prediction based on a multi-sensor data fusion model. *Reliability Engineering & System Safety*, 208, 107249.
- Li, Q. and Li, L. (2019) Integrative factor regression and its inference for multimodal data analysis. arXiv preprint arXiv:1911.04056.
- Li, R., Zhang, W., Suk, H.-Il, Wang, L., Li, J., Shen, D. and Ji, S. (2014) Deep learning based imaging data completion for improved brain disease diagnosis, In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI, Boston, MA, pp. 305–312.
- Li, S., Wang, W., Qi, H., Ayhan, B., Kwan, C. and Vance, S. (2015) Low-rank tensor decomposition based anomaly detection for hyperspectral imagery, in 2015 IEEE International Conference on Image Processing (ICIP), IEEE Press, Piscataway, NJ, pp. 4525–4529.
- Li, W., Cerise, J.E., Yang, Y. and Han, H. (2017) Application of t-sne to human genetic data. *Journal of Bioinformatics and Computational Biology*, 15(04),1750017.
- Li, Y., Wu, F.-X. and Ngom, A. (2018) A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2), 325–340.
- Liberman, Y., Samuels, R., Alpert, P. and Messer, H. (2014) New algorithm for integration between wireless microwave sensor network and radar for improved rainfall measurement and mapping. *Atmospheric Measurement Techniques*, 7(10), 3549–3563.
- Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J.T., Vespignani, A. and Santillana, M. (2020a) A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models. arXiv preprint arXiv:2004.04019.
- Liu, K., Chehade, A., and Song, C. (2015) Optimize the signal quality of the composite health index via data fusion for degradation modeling and prognostic analysis. *IEEE Transactions on Automation Science and Engineering*, 14(3), 1504–1514.
- Liu, J., Li, T., Xie, P., Du, S., Teng, F. and Yang, X. (2020) Urban big data fusion based on deep learning: An overview. *Information Fusion*, 53, 123–133.
- Liu, K., Gebraeel, N.Z. and Shi, J. (2013) A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. *IEEE Transactions on Automation Science and Engineering*, **10**(3), 652–664.
- Liu, K. and Huang, S. (2014) Integration of data fusion methodology and degradation modeling process to improve prognostics. *IEEE Transactions on Automation Science and Engineering*, **13**(1), 344–354.
- Liu, M., Zhang, J., Yap, P.-T. and Shen, D. (2016) Diagnosis of Alzheimer's disease using view-aligned hypergraph learning with incomplete multi-modality data, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI, Athens, Greece, pp. 308–316.
- Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F. YM, Li, J., Alzheimer's Disease Neuroimaging Initiative (ADNI), *et al.* (2020c) A novel transfer learning model for predictive analytics using incomplete multimodality data. medRxiv.
- Liu, Z. and Guo, Y. (2018) A hybrid approach to integrate machine learning and process mechanics for the prediction of specific cutting energy. *CIRP Annals*, 67(1), 57–60.

- Liu, Z., Zhang, W., Quek, T.Q.S. and Lin, S. (2017) Deep fusion of heterogeneous sensor data, in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, Piscataway, NJ, pp. 5965–5969.
- Lock, E.F. (2018) Tensor-on-tensor regression. Journal of Computational and Graphical Statistics, 27(3), 638–647.
- Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V.R., Yang, V. and Wu, C.F.J. (2018) An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal* of the American Statistical Association, 113(524), 1443–1456.
- Mascheroni, P., Alfonso, J.C.L., Meyer-Hermann, M. and Hatzikirou, H. (2020) Bayesian combination of mechanistic modeling and machine learning (BAM3): Improving clinical tumor growth predictions. bioRxiv.
- Meng, Y., Yu, S., Zhang, J., Qin, J., Dong, Z., Lu, G. and Pang, H. (2019) Hybrid modeling based on mechanistic and data-driven approaches for cane sugar crystallization. *Journal of Food Engineering*, 257, 44–55.
- Moin, A., Bhateja, V. and Srivastava, A. (2016) Weighted-PCA based multimodal medical image fusion in contourlet domain, in *Proceedings of the International Congress on Information and Communication Technology*, Springer, IEEE, Udaipur, India pp. 597–605.
- Neverova, N., Wolf, C., Taylor, G. and Nebout, F. (2015) Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(8), 1692–1706.
- Ng, A. (2011) Sparse autoencoder. CS294A Lecture Notes, 72(2011), 1–19.
- Paynabar, K., Jin, J. and Reed, M.P. (2015) Informative sensor and feature selection via hierarchical nonnegative garrote. *Technometrics*, 57(4), 514–523.
- Peng, Y., Qi, J., Huang, X. and Yuan, Y. (2017) CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2), 405–420.
- Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M. and Jurie, F. (2019) MAFS: Multimodal fusion architecture search, in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, Piscataway, NJ, pp. 6966–6975.
- Prakash, O., Park, C.M., Khare, A., Jeon, M. and Gwak, J. (2019) Multiscale fusion of multimodal medical images using lifting scheme based biorthogonal wavelet transform. *Optik*, **182**, 995–1014.
- Qi, J., Yang, P., Newcombe, L., Peng, X., Yang, Y. and Zhao, Z. (2020) An overview of data fusion techniques for internet of things enabled physical activity recognition and measure. *Information Fusion*, 55, 269–280.
- Rajalingam, B. and Priya, R. (2017) Multimodality medical image fusion based on hybrid fusion techniques. *International Journal of Engineering and Manufacturing Science*, 7(1), 22–29.
- Ramachandram, D., Lisicki, M., Shields, T.J., Amer, M.R. and Taylor, G.W. (2018) Bayesian optimization on graph-structured search spaces: Optimizing deep multimodal fusion architectures. *Neurocomputing*, 298, 80–89.
- Ramachandram, D. and Taylor, G.W. (2017) Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6), 96–108.
- Rankawat, S.A. and Dubey, R. (2017) Robust heart rate estimation from multimodal physiological signals using beat signal quality index based majority voting fusion method. *Biomedical Signal Processing and Control*, 33, 201–212.
- Gilan, S.S., Goyal, N. and Dilkina, B. (2016) Active learning in multiobjective evolutionary algorithms for sustainable building design, in *Proceedings of the Genetic and Evolutionary Computation Conference* 2016, ACM, Denver, CO, pp. 589–596.
- Sagi, O. and Rokach, L. (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.
- Said, A.B., Mohamed, A., Elfouly, T., Harras, K. and Wang, Z.J. (2017) Multimodal deep learning approach for joint EEG-EMG data compression and classification, in 2017 IEEE Wireless Communications

and Networking Conference (WCNC), IEEE Press, Piscataway, NJ, pp.1–6.

- Samareh, A., Jin, Y., Wang, Z., Chang, X. and Huang, S. (2018) Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces. *IISE Transactions on Healthcare Systems Engineering*, 8(3), 196–208.
- Schmitt, M. and Zhu, X.X. (2016) Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4), 6–23.
- Seyyedi, H. (2010) Comparing satellite derived rainfall with ground based radar for Northwestern Europe. University of Twente Faculty of Geo-Information and Earth Observation (ITC).
- Si, B., Schwedt, T.J., Chong, C.D., Wu, T. and Li, J. (2020) A novel hierarchically-structured factor mixture model for cluster discovery from multi-modality data. *IISE Transactions*, 53(7), 1–13.
- Siegel, M. and Wu, H. (2004) Confidence fusion, in *IEEE International Workshop on Robot Sensing*, IEEE Press, Piscataway, NJ, pp. 96–99.
- Simonyan, K. and Zisserman, A. (2014) Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27, 568–576.
- Şimşekli, U., Ermiş, B., Cemgil, A.T. and Acar, E. (2013) Optimal weight learning for coupled tensor factorization with mixed divergences, in 21st European Signal Processing Conference (EUSIPCO 2013), IEEE Press, Piscataway, NJ, pp. 1–5.
- Song, C. and Liu, K. (2018) Statistical degradation modeling and prognostics of multiple sensor signals via data fusion: A composite health index approach. *IISE Transactions*, **50**(10), 853–867.
- Song, C., Liu, K. and Zhang, X. (2019) A generic framework for multisensor degradation modeling based on supervised classification and failure surface. *IISE Transactions*, 51(11), 1288–1302.
- Srivastava, N. and Salakhutdinov, R. (2012) Idquo, multimodal learning with deep Boltzmann machines, in *Proceedings of Neural Information and Processing Systems*. Neural Information Processing Systems, Lake Tahoe, Nevada, pp. 2222–2230.
- Suk, H.-Il, Lee, S.-W., Shen, D., Alzheimer's Disease Neuroimaging Initiative, et al. (2017) Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical Image Analysis*, 37, 101–113.
- Tmazirte, N.A., El Najjar, M.E., Smaili, C. and Pomorski, D. (2013) Dynamical reconfiguration strategy of a multi sensor data fusion algorithm based on information theory, in 2013 IEEE Intelligent Vehicles Symposium (IV), IIE Press, Piscataway, NJ, pp. 896–901.
- Tucker, L.R. (1958) An inter-battery method of factor analysis. *Psychometrika*, **23**(2),111–136.
- Tucker, L.R. (1964) The extension of factor analysis to three-dimensional matrices. *Contributions to Mathematical Psychology*, 110119.
- Van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.
- Van Mechelen, I. and Smilde, A.K. (2010) A generic linked-mode decomposition model for data fusion. *Chemometrics and Intelligent Laboratory Systems*, **104**(1), 83–94.
- Vielzeuf, V., Lechervy, A., Pateux, S. and Jurie, F. (2018) Centralnet: A multilayer approach for multimodal fusion, in *Proceedings of the European Conference on Computer Vision (ECCV)*, ECCV, Munich, Germany, pp. 1–15.
- Virtanen, V., Klami, A., Khan, S. and Kaski, S. (2012) Bayesian group factor analysis, in *Artificial Intelligence and Statistics*, PMLR, La Palma, Canary Islands, pp. 1269–1277.
- Wang, J., Ma, Y., Zhang, L., Gao, R.X. and Wu, D. (2018) Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48, 144–156.
- Wang, Y., Chen, R., Ghosh, J., Denny, J.C., Kho, A., Chen, Y., Malin, B.A. and Sun, J. (2015) Rubik: Knowledge guided tensor factorization and completion for health data analytics, in *Proceedings of the* 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1265–1274.
- Wang, Y., Guan, L. and Venetsanopoulos, A.N. (2012) Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3), 597–607.

- White, M. and Schuurmans, D. (2012) Generalized optimal reverse prediction, in Artificial Intelligence and Statistics, PMLR, La Palma, Canary Islands, pp. 1305–1313.
- White, M., Zhang, X., Schuurmans, D. and Yu, Y.-l. (2012) Convex multi-view subspace learning, in Advances in Neural Information Processing Systems, Neural Information Processing Systems, Lake Tahoe, Nevada, pp. 1673–1681.
- Wu, D., Pigou, L., Kindermans, P.-J., Le, N.D.-H., Shao, L., Dambre, J. and Odobez, J.-M. (2016) Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(8), 1583–1597.
- Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., Alzheimer's Disease Neuroimaging Initiative, *et al.* (2014) Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, **102**, 192–206.
- Xiao, Y., Wang, Y. and Ding, Z. (2018) The application of heterogeneous information fusion in misalignment fault diagnosis of wind turbines. *Energies*, 11(7), 1–15.
- Yadav, B., Gopalacharyulu, P., Pemovska, T., Khan, S.A., Szwajda, A., Tang, J., Wennerberg, K. and Aittokallio, T. (2015) From drug response profiling to target addiction scoring in cancer cell models. *Disease Models & Mechanisms*, 8(10),1255–1264.
- Yan, H., Paynabar, K. and Pacella, M. (2019) Structured point cloud data analysis via regularized tensor regression for process modeling and optimization. *Technometrics*, 61(3), 385–395.
- Yan, H., Paynabar, K. and Shi, J. (2014) Image-based process monitoring using low-rank tensor decomposition. *IEEE Transactions on Automation Science and Engineering*, **12**(1), 216–227.
- Yokoya, N., Ghamisi, P. and Xia, J. (2017) Multimodal, multitemporal, and multisource global data fusion for local climate zones classification based on ensemble learning, in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE Press, Piscataway, NJ, pp. 1197–1200.

- Yoon, H. and Li, J. (2018) A novel positive transfer learning approach for telemonitoring of Parkinson's disease. *IEEE Transactions on Automation Science and Engineering*, **16**(1), 180–191.
- Yue, S., Park, J.G., Liang, Z. and Shi, J. (2020) Tensor mixed effects model with application to nanomanufacturing inspection. *Technometrics*, 62(1), 116–129.
- Zhang, C. and Ma, Y. (2012) *Ensemble Machine Learning: Methods and Applications*. Berlin/Heidelberg, Germany: Springer.
- Zhang, C., Yan, H., Lee, S. and Shi, J. (2018) Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis. *IISE Transactions*, 50(10), 878–891.
- Zhang, C., Yan, H., Lee, S. and Shi, J. (2020) Dynamic multivariate functional data modeling via sparse subspace learning. *Technometrics*, 1–33.
- Zhang, L., Xie, Y., Xidao, L. and Zhang, X. (2018a) Multi-source heterogeneous data fusion, in 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE Press, Piscataway, NJ, pp. 47–51.
- Zhang, C., Yan, H., Lee, S. and Shi, J. (2018b) Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis. *IISE Transactions*, 50(10), 878–891.
- Zhang, X., Yu, Y., White, M., Huang, R. and Schuurmans, D. (2011) Convex sparse coding, subspace learning, and semi-supervised extensions, in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI, San Francisco, CA.
- Zhao, Q., Caiafa, C.F., Mandic, D.P., Chao, Z.C., Nagasaka, Y., Fujii, N., Zhang, L. and Cichocki, A. (2012) Higher order partial least squares (hopls): A generalized multilinear regression method. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 35(7), 1660–1673.
- Zhou, H., Li, L. and Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, **108**(502), 540–552.
- Zhu, Z., Yin, H., Chai, Y., Li, Y. and Qi, G. (2018) A novel multimodality image fusion method based on image decomposition and sparse representation. *Information Sciences*, 432, 516–529.