LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-854517

Enhancing Scientific Image Classification through Multimodal Learning: Insights from Chest X-Ray and Atomic Force Microscopy Datasets

D. C. Meshnick, N. Shahini, D. Ganguly, Y. Wu,
R. H. French, V. Chaudhary

September 18, 2023

IEEE BigData 2023 Conference
Sorrento, Italy
December 15, 2023 through December 18, 2023

Disclaimer

Enhancing Scientific Image Classification through Multimodal Learning: Insights from Chest X-Ray and Atomic Force Microscopy Datasets

David C. Meshnick *
Lawrence Livermore National Laboratory
Livermore, USA
meshnick1@llnl.gov

Nahal Shahini
Case Western Reserve University
Cleveland, USA
nxs814@case.edu

Debargha Ganguly
Case Western Reserve University
Cleveland, USA
dxg512@case.edu

Yinghui Wu
Case Western Reserve University
Cleveland, USA
yxw1650@case.edu

Roger H. French
Case Western Reserve University
Cleveland, USA
rxf131@case.edu

Vipin Chaudhary
Case Western Reserve University
Cleveland, USA
vxc204@case.edu

***Abstract*—In this study, we conduct an exhaustive novel evaluation of various machine learning and multimodal learning techniques on complex datasets, exploring their potential to enhance image classification in applied sciences. We utilize the CheXpert chest x-ray and Fluoropolymer Atomic Force Microscopy (AFM) datasets, replicating and augmenting these with additional images and one-hot encoded binary metadata values. A comprehensive set of pretrained and non-pretrained Convolutional Neural Network (CNN) architectures, including ResNet50, ResNet101, DenseNet121, InceptionV3, and Xception, were tested on different configurations of image and metadata. We observe that the integration of multimodal data, even simple one-hot encoded metadata, provides substantial improvements in model classification performance compared to traditional unimodal or state-of-the-art MADDi models. The results show the promising capability of multimodal learning in providing richer data representation and improved performance in image classification tasks. In particular, the Xception models demonstrated superior results in the CheXpert experiments, while almost all models enhanced the prediction of crystal structures in the AFM datasets. Our findings offer a new performance benchmark and highlight the transformative potential of multimodal learning in applied scientific research.**

***Index Terms*—Machine Learning, Multimodal Learning, Image Classification, Atomic Force Microscopy (AFM), CheXpert Dataset, Material Science**

I. Introduction

IMAGE classification, a crucial task in domains ranging from healthcare to autonomous driving, has traditionally relied on Convolutional Neural Networks (CNNs) [1] and Vision Transformers [2], particularly excelling in benchmarks like ImageNet [3]. However, the increasing complexity and diversity of big data from various sources—be it medical imaging in biomedicine or text and sensor data in social media and IoT—demand more nuanced approaches for effective learning. This paper employs multimodal learning systems to integrate disparate types of data for robust learning [4]. Initially focused on the CheXpert dataset [5], our methods are extended to explore novel applications in material science, revealing untapped potential for data-driven insights in this critical domain.

Multimodal learning in applied sciences presents challenges such as the heterogeneous nature of data sources and the intricate interactions between different modalities. These complexities risk overfitting [6] and computational inefficiency, often referred to as the 'curse of dimensionality.' Additional issues like data modality misalignment and missing modalities can degrade performance. To address these challenges, our study provides *a simple experimental setup*, proposing *diverse learning configurations* and designing multimodal learning models. Notably, our work in material science demonstrates the versatility of multimodal learning, opening new avenues for research and applications.

Our comparative performance evaluation against state-of-the-art unimodal and MADDi models shows the efficacy and robustness of our approach, mitigating many complexities associated with multimodal learning. The cross-modal synergy not only enhances predictive accuracy but also proves instrumental in unlocking new applications in material science, underscoring the transformative potential of our research.

II. Related Works

Baltrusaitis et al. [6] classifies challenges in multimodal learning into five broad areas: Representation, Translation, Alignment, Fusion, and Co-Learning. The foremost challenge is the *diverse representation* of data modalities. Two main strategies address this: joint and coordinate representation

This research was supported in part by NSF awards 2112606 (Vipin Chaudhary) and 2117439 (Vipin Chaudhary and Roger H. French). Roger H. French, Yinghui Wu and Vipin Chaudhary were partly supported by NNSA DE-NA00004104. This work was performed, in part, under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, LLNL-CONF-854517.

* This work was done while at Case Western Reserve University.

[7]. The former combines different data types into a shared space, useful when all modalities are available during training and testing [8]. The latter maintains separate spaces for each modality but links them through constraints, useful when a modality might be missing during testing [6]. *Translating* between modalities is another challenge, divided into example-based and generative methods. Example-based approaches use explicit mappings from data dictionaries to convert data between modalities [9]. These dictionaries can also combine data to represent it in a new modality [10]. Such translation tasks add complexity, requiring a deep understanding of the modalities involved. Generative translation complicates modality mapping by using learning models instead of explicit rules. Common approaches include grammar-based methods in NLP [11] and encoder/decoder models for latent representations [6]. Another challenge is *data alignment*, divided into explicit and implicit types. Explicit alignment focuses on techniques like dynamic time warping for time series [12] or CNNs for text-image similarity [13]. Implicit alignment, meanwhile, learns alignment indirectly during model training.

Fusion integrates information from various modalities for better classification and comes in two main forms: model-agnostic and model-based. Model-agnostic methods combine features directly and average unimodal results, while model-based methods adapt the architecture for specific data types. Other model-based techniques include Multiple Kernel Learning, which modifies support vector machines for each modality [14], and graphical models that use conditional probability like conditional random fields [15]. These approaches highlight the complexity and potential of multimodal learning in applied sciences. Multimodal co-learning uses a data-rich modality to enhance a less-resourced one, addressing issues like missing or noisy data. Co-learning methods can be tailored for parallel, non-parallel, or hybrid data. For parallel data like synced video and audio, weak classifiers are trained on a few labeled samples, then applied to unlabeled ones, although this risks overfitting. An alternative is transfer learning, where insights from one modality inform the training of another. [6] [16] [17] Co-learning also works with non-parallel data by using shared categories, enhancing modalities without needing to align them initially. For hybrid data with partially matched instances, a 'bridging' modality can connect the gaps, such as using images to link different languages [18] [19].

Two recent studies highlight the growing benefits of multimodal approaches. Ellen et al. enhanced plankton image classification by adding contextual metadata to a VGG-16 model, slightly boosting accuracy and underscoring the value of early metadata inclusion in CNNs. [20], [21] Golovanevsky et al. used multimodal techniques for early Alzheimer's diagnosis, combining MRI scans, genetic markers, and clinical observations in their MADDi framework. This resulted in higher accuracy compared to unimodal approaches, suggesting multimodal techniques' potential in healthcare. [22] Both studies serve as benchmarks for our own multimodal research.

III. Experimental setup

In this study, we aim to characterise the advantages of multimodal learning in enhancing image classification performance, specifically focusing on fluoropolymer crystal growth analysis and chest X-ray (CXR) pathology classification. We designed and executed experiments using CXR data to estimate the potential performance gains conferred by employing a dual-modality dataset. Building upon these discoveries, we conducted an additional series of experiments within the field of material science, focusing on fluoropolymer Atomic Force Microscopy (AFM) images.

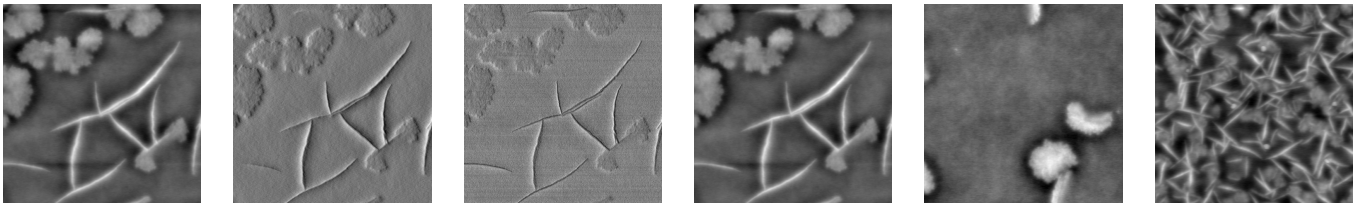
A. *CheXpert Task*

The CheXpert dataset, introduced by Stanford in 2019, is a comprehensive compilation of 224,316 chest radiographs (CXR) from 65,240 patients, each image associated with five potential pathologies: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion [5]. These labels were generated through an automated system that examined corresponding medical records. Included alongside these radiographs are metadata entries detailing the patient's age and sex, and the CXR image orientation, which can be either frontal, with additional anterior or posterior specification, or lateral. To enhance model training and predictive capabilities, we constructed an augmented version of this dataset, implementing image augmentations for increased realism and diversity, and converting metadata into one-hot encoded binary values for comprehensive integration. This approach utilizes the available metadata in conjunction with radiograph images, despite the insufficiency of metadata alone for predicting pathology presence.

B. *The Novel Fluoropolymer Crystallization Task*

Fluoropolymers, primarily composed of carbon-fluorine bonds, exhibit a range of desirable qualities such as chemical inertness when exposed to acids, bases, and solvents, aging and thermal resistance, as well as low dielectric constant, flammability, moisture absorption, and refractive index; they also show resistance to hydrolysis and oxidation [23]. These polymers, exemplified by materials like polytetrafluoroethylene (Teflon), are characterized by their insolubility, which stems from the semi-crystalline surface formed by the structure of repeated polymer subunits. A higher degree of crystallization can make the material more insoluble but also renders it brittle and challenging to work with. Furthermore, the crystalline regions within a fluoropolymer can grow over time due to repeated exposure to environmental factors, making it crucial to continually test the material's development.

This testing can be accomplished using AFM, which tracks the contact point of a sub-nanometer tip along the surface of a fluoropolymer. Regularly tapping the surface of a fluoropolymer, or some other thin material, with the tip produces a height map of its topology, along with several corresponding error and correction signals [24]. An example of the varying signals generated with AFM can be seen in Figure 1, where



(a) Height (b) Amplitude (c) Phase (d) Z-Sensor (e) Spherulites (f) Lamellar

Fig. 1: AFM images and crystal structures of fluoropolymer samples. (a-d) AFM four-channel images of fluoropolymer. (e-f) Fluoropolymer samples with spherulite and lamellar crystallinity.

Channel	Name	Description
A	Height	Direct topology of material surface
B	Amplitude	Measure of error tracking the material surface
C	Phase	Measure of error induced by material surface effect on tapping oscillation
D	Z-Sensor	Measure of motion for the Height sensor

TABLE I: Summary of AFM channels.

a single scan produces four separate images. The differences between these channels is summarized in Table I.

The sample depicted in Figure 1 originated from a partnership between the CWRU SDLE Laboratory and Lawrence Livermore National Laboratory (LLNL) [25] [26]. A part of that joint venture involved the study crystalline growth on fluoropolymer surfaces. The development of two specific types of crystal structures, spherulite and lamellar, serve as an important aspect of this research. Spherulite crystals grow spherically in all directions, producing a rounded two-dimensional image as seen in Figure 1. In contrast, lamellar crystals grow biaxially along a single axis, similar to a needle's shape. To track this growth, different material samples were recorded by an AFM device where crystalline expansions were accelerated.

The formed crystalline substructures, however, were not solely determined by the fluoropolymer composition of the material. As AFM devices scan with a resolution that have orders of magnitude less than a nanometer, the fluoropolymer being imaged must be thin enough so that the crystals do not grow too high along the z-axis and interfere with the scanning tip. This is accomplished through spin coating, where some material is mixed with a solvent and the resulting solution is spread over a rotating surface. The centripetal force and surface tension of the solution creates an even film, from which the solvent evaporates and leaves a thin layer of the original material [27]. However, before the solution is spread onto the rotating surface, it can be filtered through small pores to remove larger particles, which have been shown to change the surface behavior of polymer films [28]. If the altered behavior of a material caused by this preprocessing step is not considered, incorrect conclusions can be drawn by comparing one compound that was filtered with another that was not.

As this filtering step can impact crystal growth on fluoropolymer surfaces, understanding what type of crystal structures form in the fluoropolymer film is important. For a single image, it may be a trivial task to manually determine what crystalline structures are more prevalent. Yet a single fluoropolymer experiment can produce thousands or tens of thousands of AFM images. Furthermore, each AFM image may contain several composite images along with other metadata. This increasing scale of data necessitates the need for automation. By applying image classification techniques to the AFM images, the dominant crystal structure type can be cross-referenced with known information about the depicted materials to describe how fluoropolymers crystallize when filtered differently.

C. *Data Preprocessing*

The CheXpert dataset, comprising of chest X-ray (CXR) images along with corresponding metadata, underwent several preprocessing steps. Class labels were produced from patient medical records via an automated labeler, resulting in an imbalanced distribution across fourteen possible classes as shown in Table II [5]. Uncertainty in these labels was handled by assigning uncertain instances as negative, based on a preliminary performance test. To manage the class imbalance, data augmentation was employed, only on the image modality, due to the inherent difficulty in altering binary-valued metadata fields without losing their true information. Duplications were made for underrepresented classes ("Enlarged Cardiomediastinum", "Pneumonia", "Lung Lesion", "Pleural Other", and "Fracture") and a random half of these duplicated images underwent transformations including random rotation, cropping, brightness and contrast adjustment, and reduced image compression. Every CXR image was also resized to dimensions of 224×224 pixels. The CXR dataset has three binary metadata fields (sex, frontal/lateral, and anterior/posterior), and to avoid imparting a potentially misleading numerical representation, one-hot encoding was used, transforming each binary field into two mutually exclusive variables [29]. The combination of these preprocessing steps resulted in four model variants: the original unmodified CheXpert dataset (*CXR experiment 1*), a version with data augmentation (*CXR experiment 2*), one with one-hot encoding of binary metadata (*CXR experiment 3*), and

finally a model variant with both data augmentation and one-hot encoding (*CXR experiment 4*).

Pathology	Positive	Uncertain	Negative
No Finding	16627	0	171014
Enlarged Cardiom	9020	10148	168473
Cardiomegaly	23002	6597	158042
Lung Lesion	6856	1071	179714
Lung Opacity	92669	4341	90631
Edema	48905	11571	127165
Consolidation	12730	23976	150935
Pneumonia	4576	15658	167407
Atelectasis	29333	29377	128931
Pneumothorax	17313	2663	167665
Pleural Effusion	75696	9419	102526
Pleural Other	2441	1771	183429
Fracture	7270	484	179887
Support Devices	105831	898	80912

TABLE II: Class distribution of CheXpert data. [5]

In the context of multimodal learning with **fluoropolymer Atomic Force Microscopy (AFM) data**, preprocessing is primarily concerned with managing a greater scale of modality, as each AFM dataset includes four concurrent images (Channels A-D) and over a thousand metadata points of diverse nature. Initially, up to four 512x512 pixel TIFF images and corresponding metadata for each observation were extracted from the original aging fluoropolymer dataset. Prior work employed the YOLOv4 object detection technique to the Channel A images, yielding $1,285,204$ distinct observations related to spherulitic or lamellar crystal instances [30]. These observations were subsequently categorized based on the corresponding TIFF image, leading to an aggregated dataset wherein each Channel A image was paired with total counts of spherulitic and lamellar crystals, resulting in five mutually exclusive classes: Equal, Majority Spherulite, Majority Lamellar, Vast Majority Spherulite, and Vast Majority Lamellar, detailed in Table III. All observations, inclusive of the four images, class labels, and metadata, were collated into a unified dataset, eliminating entries with missing image channels or partial metadata values. Categorical metadata fields, identifiable by having three or fewer unique possible values, were one-hot encoded [29]. The finalized dataset encompassed 10,242 observations from 18 distinct material samples, with each observation linked to four images and 284 metadata fields. To ensure compatibility with TensorFlow, the TIFF images were converted into lossless PNG images of equivalent size.

Class	Quantity
Equal	382
Majority Spherulite	1,694
Majority Lamellar	1,806
Vast Majority Spherulite	4682
Vast Majority Lamellar	1678

TABLE III: Fluoropolymer AFM class distribution.

D. Model Training

Our study employed five CNN architectures—ResNet50, ResNet101, DenseNet121, Xception, and InceptionV3—with initial weights either randomly assigned or pretrained on ImageNet. We also explored a model solely dependent on sample metadata for classification. Ten unique implementations of the MADDi framework served as benchmarks.

The dataset was partitioned in an 80:20 split for training (D_{train}) and testing (D_{test}). All models were trained using a binary cross-entropy loss function:

$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y_i}) + (1 - y_i) \cdot log(1 - \hat{y_i})$

Performance metrics—loss, accuracy, precision ($P = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), ROC AUC, and F1-score ($F1 = \frac{2 \cdot P \cdot R}{P+R}$)—were monitored across three runs of 40 epochs each. Each model's final classification layer used a sigmoid activation function for independent class probabilities.

E. Unimodal Approach Models

To critically assess the effectiveness of multimodal learning, we incorporated two unimodal methodologies for chest pathology classification, which included either image-only data or metadata-only data.

- **Image-Only Architecture** A generic architecture for a unimodal image approach is illustrated in Figure 3.1(a). An image input I is processed through a specific backbone model to extract a feature map F. This feature map is subsequently passed through an average pooling layer, transforming it into a pooled feature P, defined as:

$$P = \frac{1}{N}\sum_{i=1}^{N} F_i \tag{1}$$

  Next, a dropout layer is introduced to prevent overfitting through selective neuron deactivation. The resulting feature D is then passed into a fully connected dense layer with a Rectified Linear Unit (ReLU) activation function and L2 kernel regularization to further mitigate overfitting. This process is mathematically represented as:

$$D = ReLU(W_d \cdot P + b_d) \tag{2}$$

  where W_d and b_d are the weights and biases of the dense layer, respectively. Finally, the output O is computed by passing D through a fully connected output layer, predicting the classification labels:

$$O = \sigma(W_o \cdot D + b_o) \tag{3}$$

  where W_o and b_o are the weights and biases of the output layer, and σ is the sigmoid activation function.
- **Metadata-Only Architecture**
  The metadata-only model is outlined in Figure 3.1(b). Metadata M is input into a shallow neural network starting with a single dense layer to produce D':

$$D' = ReLU(W_m \cdot M + b_m) \tag{4}$$

where W_m and b_m are the weights and biases of the dense layer. The output of the dense layer is passed through a ReLU activation function and then normalized using batch normalization. The normalized information is then passed to the classifier segment, constituted by two fully connected layers. The output of the network, representing the classification labels, is computed as:

$$O' = \sigma(W'_o \cdot D' + b'_o) \tag{5}$$

where W'_o and b'_o are the weights and biases of the output layer.

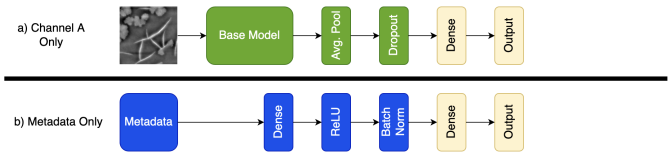


Fig. 2: Representation of unimodal model architectures evaluated on AFM data: a) single channel image-only, and b) metadata only.

F. Multimodal Approach Models

To scrutinize the potential advantages of a multimodal learning approach for chest pathology classification, we deployed the architecture depicted in Figure 3(c). This setup merges two unimodal models prior to the final feature selection stage, thereby incorporating both modalities into the decision-making process. The overarching aim is to balance the proven methodologies as employed in [20], permitting the metadata modality to interact within its domain before integrating with the image subnet.

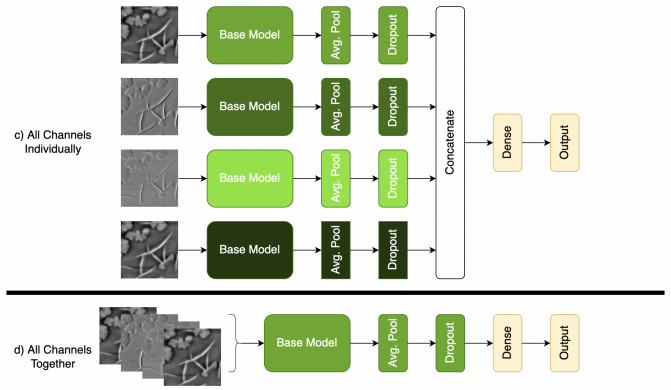


Fig. 3: Representation of multimodel architectures evaluated on AFM data: c) all channels treated as the same modality, and d) each channel has a unique image subnet.

- **Partial Metadata Context** Two distinct categories within the metadata modality of the CXR dataset were recognized: patient-related data and image-orientation data. Patient-related metadata encompasses age and sex, while image-orientation data indicates whether an image was taken from a frontal/lateral or anterior/posterior view. Partitioning the metadata into these groups and retaining only one during training allows for investigating the influence of these metadata types. As a result, two additional multimodal models were trained with the same architecture, each model exclusively containing one type of sample metadata. Consequently, an ensemble of 41 average models was trained for each CXR experiment.

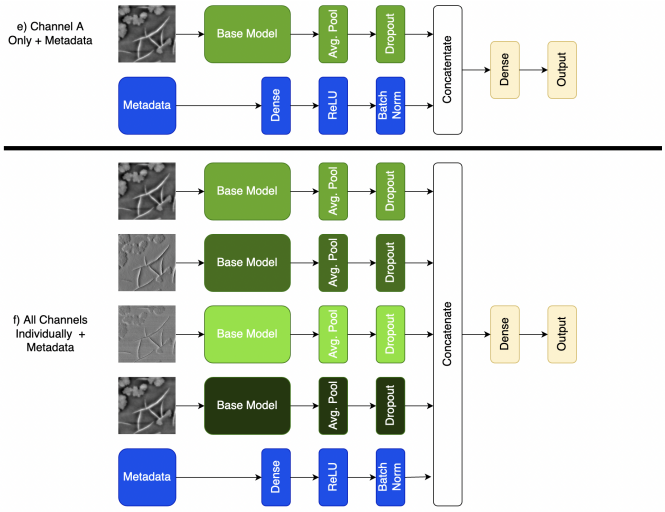


Fig. 4: Representation of multimodel architectures evaluated on AFM data: e) single channel image with metadata, and f) all channels treated as the same modality with metadata.

- **MADDi** An adapted bimodal version of the MADDi framework was instantiated for comparing our proposed multimodal approach with state-of-the-art techniques. The revised model architecture is presented in Figure 3(d), where the "Dense Group" in the metadata modality corresponds to the three layers defined in the unimodal metadata-only model (i.e., Dense, ReLU, and Batch Normalization). Given only two modalities instead of three, the cross-modal units are initially concatenated together. They are then merged with another instance of each unimodal unit prior to entering the classification layer. Given feature vector $\mathbf{f}$ from the image subnet and metadata vector $\mathbf{m}$ from the metadata subnet, the final features $\mathbf{c}$ for classification are obtained by concatenating $\mathbf{f}$ and $\mathbf{m}$:

$$\mathbf{c} = [\mathbf{f}, \mathbf{m}] \tag{6}$$

These concatenated features are then passed to the final classification layer to predict the labels. The output of the network is computed as:

$$O'' = \sigma(W''_o \cdot \mathbf{c} + b''_o) \tag{7}$$

where W''_o and b''_o are the weights and biases of the output layer, and σ is the sigmoid activation function.

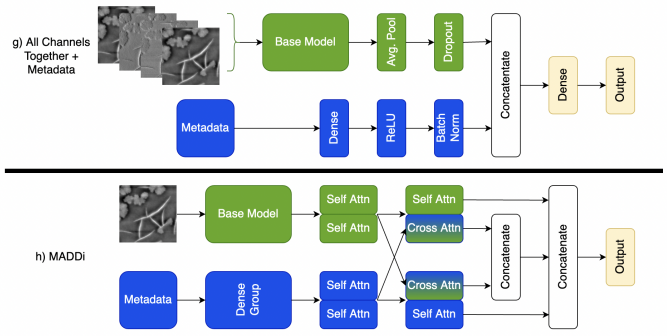


Fig. 5: Representation of multimodel architectures evaluated on AFM data: g) metadata and each channel has a unique image subnet, and h) adaptation of bimodal MADDi.

G. *Distributed Learning Framework Exploiting Parallelism*

The continuous increase in complexity of computational tasks required to train multiple models on large datasets requires the use of distributed computing systems. Such frameworks leverage parallel computing mechanisms to circumvent the limitations imposed by singular hardware configurations. We thus propose a mirrored strategy for distributed computing, facilitating efficient parallelized model training.

- **Data Partition, Gradient Computation, and Model Replication with TensorFlow Distributed MirroredStrategy ($\mathcal{D}$, ∇, $\mathcal{M}_s$):** The dataset $\mathcal{D}$ is uniformly partitioned among the N GPUs. Let $\mathcal{D}_i$ denote the subset of data allocated to the i-th GPU. Each GPU independently computes its gradient $\nabla\mathcal{L}_i(\mathbf{W}_i)$ and loss function $\mathcal{L}_i(\mathbf{W}_i)$ utilizing $\mathcal{D}_i$. The computation on each GPU can be formalized as $\nabla\mathcal{L}_i(\mathbf{W}_i) = \frac{1}{|\mathcal{D}_i|}\sum \mathbf{x} \in \mathcal{D}_i \nabla\mathcal{L}(\mathbf{x}; \mathbf{W}_i)$ where $\mathbf{x}$ is an instance in $\mathcal{D}_i$. This process results in better utilization of the computational resources of each GPU and an increase in throughput. This strategy for data partitioning and gradient computation is managed and enhanced by TensorFlow's Distributed MirroredStrategy $\mathcal{M}_s$. $\mathcal{M}_s$ is capable of robust model replication and dataset distribution across N GPUs, thereby ensuring redundancy and system resilience. It is especially well-suited for scenarios with multiple GPUs in a single machine but also extends to multi-machine configurations, catering effectively to a diverse hardware environment
- **Gradient Aggregation, Weight Updates, and Data Communication with TensorFlow Hierarchical Copy All Reduce ($\Delta\mathbf{W}$, $\mathcal{H}$):** After the completion of independent computations, the gradients $\nabla\mathcal{L}_i(\mathbf{W}_i)$ from each GPU are aggregated and synchronized across all units using TensorFlow's Hierarchical Copy All Reduce ($\mathcal{H}$). The $\mathcal{H}$ methodology hierarchically aggregates tensors, initially within a machine and then across machines, thus effectively reducing the time required for the all-reduce operation. It further ensures more efficient use of the network bandwidth, minimizing network congestion and enhancing the overall performance of the distributed system. The collective gradient obtained post-aggregation is then utilized to perform consistent network weight updates across all GPUs. This can be mathematically represented as $\Delta\mathbf{W} = -\eta\sum i = 1^N \nabla\mathcal{L}_i(\mathbf{W}_i)$, where N is the number of GPUs and η is the learning rate. The use of $\mathcal{H}$ in this process guarantees uniformity and synchronization of the learning process across the distributed network.

To illustrate, the CXR dataset was efficiently split across multiple instances without a substantial impact on model performance. With the data distributed across $N = 8$ GPUs, each instance still had over $|\mathcal{D}_i| > 27,000$ images for training. This optimized distributed learning framework was effectively instantiated on the High-Performance Computing (HPC) cluster of Case Western Reserve University (CWRU), employing eight NVIDIA A100 GPUs, which dramatically enhanced the training speed for CXR models. The results suggest a scalable and efficient methodology for handling large-scale datasets and complex models, emphasizing the feasibility of distributed learning in high-performance computing ecosystems.

IV. Results

We present experimental results of both the CXR and fluoropolymer AFM tests—the overall and relative performances of multimodal models under different data conditions. The fluoropolymer crystal classification results compare the various multimodal approaches that were implemented.

A. *Chest X-Ray*

We conducted four experiments to explore the efficacy of different machine-learning approaches and data manipulations in predictive modeling. Experiment 1 served as a baseline, using a multimodal learning approach and producing modest f1-scores across a range of robust models like Xception, InceptionV3, and DenseNet121. Experiment 2 introduced data augmentation techniques targeting underrepresented classes and saw improvements in the f1-score, although surprisingly, the multimodal versions of the models did not outperform the unimodal ones. In experiment 3, one-hot encoding was used for binary metadata, resulting in varied model performances but generally failing to surpass the f1-scores of the unimodal Xception model from previous experiments.

Experiment 4, however, integrated the data augmentation from experiment 2 and one-hot encoding from experiment 3. This integrated approach yielded exciting results: while the unimodal Xception model still led in f1-score, the multimodal version of DenseNet121 outperformed its unimodal counterpart, making a case for carefully integrating various data modalities and preprocessing techniques. These results are summarized in Table IV.

Benchmark Results: In assessing multimodal learning methods, we implemented a bimodal iteration of the MADDi framework and evaluated it alongside a Metadata Only approach. Various backbone architectures were tested with MADDi, and the results are detailed in Table V The most effective MADDi model utilized the InceptionV3 architecture

		imagenet weights						random weights					
Model	Approach	Loss	Accuracy	Precision	Recall	ROC AUC	F1	Loss	Accuracy	Precision	Recall	ROC AUC	F1
DenseNet121	Multimodal	0.6735	0.8451	0.5619	0.4429	0.671	0.2998	0.8657	0.8125	0.4241	0.2408	0.6068	0.1675
DenseNet121	Image	0.3891	0.8581	0.6211	0.445	0.7218	0.2891	0.4254	0.8514	0.5853	0.4615	0.7023	0.2755
InceptionV3	Multimodal	0.7503	0.8504	0.5787	0.4735	0.6512	0.2926	0.7287	0.8322	0.514	0.4584	0.651	0.2808
InceptionV3	Image	0.4483	0.8538	0.6036	0.4355	0.7108	0.2939	0.4301	0.853	0.5908	0.4752	0.7082	0.311
ResNet101	Multimodal	0.7749	0.8405	0.547	0.4197	0.6318	0.2584	0.8826	0.8331	0.5103	0.3594	0.6347	0.239
ResNet101	Image	0.4916	0.8507	0.5885	0.4378	0.685	0.2768	0.4597	0.8431	0.5484	0.4871	0.6837	0.2766
ResNet50	Multimodal	0.8496	0.8336	0.518	0.3714	0.6136	0.2315	0.8894	0.8149	0.432	0.3442	0.5925	0.1713
ResNet50	Image	0.4608	0.8517	0.5871	0.4589	0.6855	0.2685	0.4103	0.8538	0.5939	0.4751	0.7012	0.2729
Xception	Multimodal	0.8457	0.8498	0.5891	0.4113	0.6449	0.2826	0.936	0.8486	0.5758	0.4691	0.6326	0.2684
Xception	Image	0.5454	0.8506	0.5806	0.4681	0.6967	0.3226	0.4918	0.8552	0.6047	0.4564	0.6953	0.2854

TABLE IV: CXR experiment 4 multimodal and image-only classification results, with pre-trained imagenet weights on the left and randomized initial weights on the right.

with pretrained imagenet weights for the highest f1-score, while the DenseNet121 with pretrained weights achieved the top ROC AUC value. Interestingly, pretrained models generally incurred higher losses than their non-pretrained counterparts. For the Metadata Only models, the predictive capacity was notably lower, as demonstrated in Table VI The one-hot encoded metadata in experiment 3 yielded the highest f1-score, albeit not exceeding 0.1. Moreover, data augmentation to balance class labels resulted in reduced accuracy and f1-scores across the board.

B. Fluoropolymer AFM

The experimental results for the various image-only and multimodal models are presented, first in terms of general model results and then in terms of class-wise metrics.

Image Only Model Results: In evaluating various multimodal approaches, a baseline experiment employing just a single image from each AFM sample was analyzed. The non-pretrained InceptionV3 architecture outperformed others with an impressive ROC AUC of 0.9687±0.0118 and f1-score of 0.6413±0.0113. The non-pretrained ResNet50 also yielded better f1-scores than its pretrained variant, while ResNet101 showed mixed results. Interestingly, DenseNet121 was the sole architecture where the pretrained version surpassed the non-pretrained model across all metrics, boasting an accuracy of 0.9133±0.0099, ROC AUC of 0.9289±0.0118, and f1-score of 0.6356±0.0198. This suggests that pretraining may not always be advantageous, as the results varied based on the architecture employed. Class-wise metrics revealed that pretrained models exhibited higher recall rates for the "Equal" class and greater consistency across the "Vast Majority" classes, whereas non-pretrained models were more variable in these metrics.

C. Ablation study for fusion techniques

In this ablation study, we aim to elucidate the impact of different data fusion strategies on the performance of our models. We categorize the strategies based on well known fusion paradigms in the existing literature. We also utilize multiple architectures to gain insights into the interplay between fusion strategies and the pertaining of models.

- **Late Fusion via Feature Concatenation (Baseline)** We found that the pretrained InceptionV3 model exhibited superior performance, achieving an ROC AUC of 0.9370 ± 0.0019 and an f1-score of 0.6817 ± 0.0202. However, this superiority of pretrained models was not universal. For instance, the non-pretrained ResNet101 model outperformed the pretrained DenseNet121 in f1-score.
  Upon closer examination, we observed that pretrained models consistently had a higher recall, particularly for the "Equal" and "Vast Majority" classes. In terms of precision, both pretrained and non-pretrained models were more evenly matched. Notably, non-pretrained models came close to matching the f1-score of pretrained models in the "Majority Spherulite" class, with scores of 0.6618 ± 0.0772 and 0.6891 ± 0.0935 respectively, well within the range of standard deviation.
  The Late Fusion strategy offered a straightforward but effective method for fusing different AFM channels. While pre-trained models generally performed better, the results indicate that non-pretrained models can be competitive, especially in class-specific metrics. The variable performance across different architectures suggests that the choice of model and fusion strategy should be carefully considered depending on the specific requirements of the task.
- **Early fusion via processing all channels together**: This approach adopts an Early Fusion strategy by treating each channel of a given AFM image as an individual sample. The class labels are replicated across each channel for each sample. The pretrained InceptionV3 model continued to excel in classification, posting an accuracy of 0.9561 ± 0.0072, ROC AUC of 0.9739 ± 0.0044, and f1-score of 0.7949 ± 0.0285. It was striking that pretrained models universally surpassed their non-pretrained counterparts in class label prediction. In fact, no non-

Model	Weight	Loss	Accuracy	Precision	Recall	AUC	F1
DenseNet121	random	0.933±0.1388	0.8461±0.0038	0.5653±0.0097	0.4449±0.0366	0.6404±0.0119	0.2698±0.0056
DenseNet121	imagenet	9.3066±14.6226	0.8484±0.0068	0.5726±0.0262	0.4646±0.0551	0.6678±0.0207	0.2923±0.0277
InceptionV3	random	0.8144±0.0383	0.844±0.0084	0.554±0.0322	0.4905±0.0406	0.6526±0.0114	0.2973±0.0126
InceptionV3	imagenet	0.9898±0.0574	0.8477±0.0016	0.5694±0.0058	0.4627±0.0108	0.6449±0.0047	0.298±0.0173
ResNet101	random	0.6501±0.1477	0.8435±0.0059	0.5641±0.0277	0.3941±0.0442	0.6435±0.0246	0.2415±0.0131
ResNet101	imagenet	1.0577±0.0246	0.851±0.0008	0.5777±0.0026	0.4927±0.0078	0.6321±0.0043	0.2818±0.0097
ResNet50	random	0.9802±0.0622	0.8398±0.0005	0.5463±0.0023	0.3953±0.025	0.6145±0.0068	0.2326±0.0129
ResNet50	imagenet	1.0016±0.0267	0.8465±0.0062	0.5686±0.0209	0.437±0.0408	0.6348±0.0046	0.2696±0.0185
Xception	random	1.0555±0.1271	0.851±0.0043	0.5825±0.0184	0.4694±0.0264	0.6322±0.0054	0.2728±0.0078
Xception	imagenet	0.7698±0.3477	0.8484±0.0067	0.5818±0.0183	0.4206±0.1169	0.6371±0.0239	0.2392±0.1288

TABLE V: CXR bimodal MADDi model results.

CXR experiment	Loss	Accuracy	Precision	Recall	AUC	F1
experiment 1	0.3444±0.0002	0.8507±0.0001	0.591±0.0016	0.3052±0.0044	0.6083±0.0011	0.0915±0.0008
experiment 2	0.3702±0.0001	0.8427±0.0001	0.5845±0.0024	0.2895±0.0064	0.609±0.0006	0.0912±0.0006
experiment 3	0.3445±0.0001	0.8506±0.0001	0.5895±0.0023	0.3091±0.0061	0.6072±0.002	0.092±0.001
experiment 4	0.3703±0.0001	0.8428±0.0001	0.5897±0.0004	0.2774±0.002	0.6093±0.0011	0.0897±0.0004

TABLE VI: Unimodal metadata models for the four CXR experiments.

pretrained model managed to outperform any pretrained model in classifying crystalline structures. Distinct hierarchies of performance were noted between pretrained and non-pretrained architectures. The pretrained InceptionV3 and DenseNet121 models showed superior performance compared to the ResNet models, consistent with observations in the CXR experiments. Conversely, non-pretrained ResNet models fared better than their DenseNet121 counterpart. Specifically, the ResNet50 model had an f1-score of 0.5896 ± 0.0958, slightly edging out the non-pretrained InceptionV3 model's 0.5873 ± 0.0851, though these scores lie within their respective ranges of standard deviation.

Non-pretrained models exhibited irregular recall across classes, whereas pretrained models demonstrated more consistent recall. Coupled with generally higher precision—particularly in the "Equal" class—pretrained models maintained an edge in overall f1-scores.

The Early Fusion method capitalizes on the combined information from all channels at the input stage, allowing the model to capture potentially complex inter-channel relationships. Pretrained models consistently outperformed non-pretrained ones, yet differences in performance did emerge based on the architectural choices, emphasizing the importance of architecture in model performance.

- **Multimodal Fusion: AFM Single Channel + Metadata** This Multimodal Fusion approach combines the AFM height channel (Channel A) with sample metadata before feature extraction (see Figure 6). Pretrained InceptionV3 led in classification with ROC AUC 0.9447 ± 0.0056 and f1-score 0.6889 ± 0.0224. Like in unimodal setups, pretrained models generally outperformed non-pretrained ones in f1-scores.

  However, the non-pretrained InceptionV3 also showed competence, posting the second-highest ROC AUC score of 0.9332 ± 0.0376, indicating that pretrained models were not universally superior. Unlike the first unimodal approach where loss values were inconsistent, the multimodal models exhibited more stable loss metrics. Pretrained versions typically had about half the loss of non-pretrained models, except for InceptionV3, where the difference was marginal and within the range of standard deviation.
- **Hybrid Fusion: AFM All Channels Separate + Metadata** Hybrid Fusion combines multimodal data (metadata) with multi-view learning (separate AFM channels) as shown in Figure 6. Pretrained InceptionV3 outperformed others with ROC AUC 0.9496 ± 0.0112 and f1-score 0.7042 ± 0.0228. While pretrained models generally had higher f1-scores, non-pretrained models displayed variances, like higher ROC AUC values compared to certain pretrained models (except ResNet50). The pretrained ResNet101 struggled with low recall in "Vast Majority" and "Majority" classes, impacting its f1-scores. Conversely, high recall in the non-pretrained ResNet50 model allowed it to perform better than more robust architectures.
- **Multimodal Fusion: AFM All Channels Together + Metadata** Results of this multimodal method, fusing all image channels and metadata, are shown in Figure 6. Pretrained InceptionV3 again led the pack with accuracy 0.9549 ± 0.0005, ROC AUC 0.9718 ± 0.0035, and f1-score 0.7873 ± 0.0029. Pretrained models universally outclassed their non-pretrained counterparts across all metrics. Non-pretrained DenseNet121 exhibited anomalously low precision in "Vast Majority" classes but was on par with others in "Majority Spherulite" and "Majority Lamellar." Conversely, pretrained models were consistently strong

across classes, yielding uniformly high classification metrics.

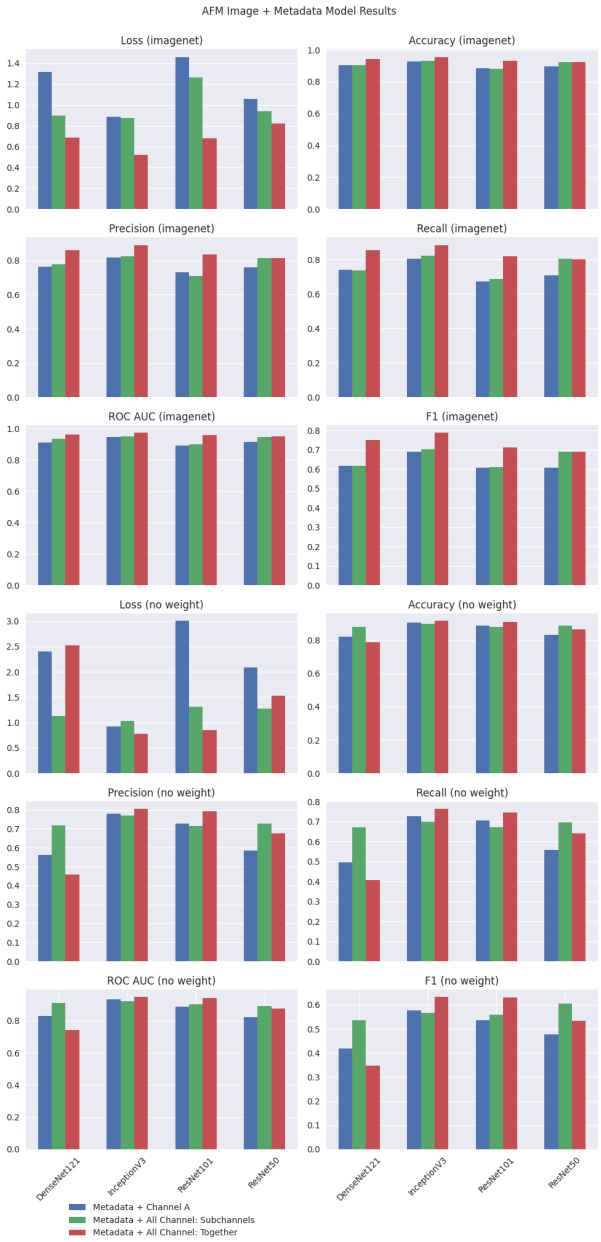


Fig. 6: Classification metrics for multimodal models trained on AFM data with images and metadata, where the top half use pretrained imagenet weights and the bottom half have randomized initial weights.

Benchmark Results: To provide reference points to the proposed image-only and multimodal learning methods, a bimodal implementation of the MADDi framework along with a metadata-only model was tested, and their results are shown in Table VII.

MADDi: In our MADDi tests, the pretrained InceptionV3 led in ROC AUC, while non-pretrained DenseNet had the best f1-score and lowest loss. Non-pretrained ResNet50 outperformed its pretrained version in f1-score. Interestingly, all MADDi models had double the recall compared to precision, and non-pretrained models generally had higher recall except for DenseNet121.

Metadata only: Lastly, a model trained for fluoropolymer AFM images achieved an accuracy of 0.8332 and a credible ROC AUC value of 0.8752. However, the model had limitations, including a low f1-score below 0.5 and poor recall at 0.2874. It failed to predict positive labels for the ”Equal,” ”Majority Spherulite,” and ”Majority Lamellar” classes.

Our results therefore underscore that the fusion of diverse data modalities can substantially enhance a model’s predictive prowess in this novel application domain.

V. Discussion

Our ablation study reveals nuanced interactions between fusion strategies and model architectures. Pretrained InceptionV3 consistently dominated across all fusion paradigms, with particularly strong f1-scores and ROC AUC values in the Multimodal Fusion setups involving all AFM channels and metadata. On the other hand, the Late Fusion and Early Fusion strategies highlighted the competitive capabilities of non-pretrained models in class-specific metrics, such as in the ”Majority Spherulite” class for Late Fusion and the overall f1-score for Early Fusion. The Hybrid Fusion strategy presented an interesting case where non-pretrained models demonstrated strength in ROC AUC, challenging the generality of pretrained models’ superiority. These observations underscore the importance of selecting an appropriate fusion strategy and model architecture based on task-specific requirements. Non-pretrained models show promise in certain scenarios and should not be universally discounted. Each fusion strategy also seems to affect the class-wise behavior of models differently, suggesting that a careful choice of fusion paradigm is crucial when class-level performance matters.

It is important to acknowledge that the scope of this study does not necessarily encompass the cutting-edge developments or the most recent techniques proposed in the multimodal learning literature. The rapidly evolving field has introduced a multitude of innovative techniques, each having its unique potential and sophistication, which are not directly incorporated in our work. However, this study serves a distinct purpose. Instead of chasing the state-of-the-art, our focus lies on building a principled and applied foundation, demonstrating that multimodal techniques can be harnessed effectively to improve models in applied science research. By utilizing established methodologies, we aim to present a practical, concrete application of multimodal learning in a real-world context. The goal is not only to illustrate the benefits and feasibility of adopting such an approach, but also to inspire further integration of multimodal learning techniques into a broader range of applied scientific disciplines, thereby stimulating advancements in these fields through interdisciplinary collaboration.

VI. Conclusion

In our comprehensive study, we meticulously evaluated the impact of multimodal learning techniques in image classification tasks, specifically using the CheXpert chest x-ray and fluoropolymer AFM datasets. Utilizing various pretrained and non-pretrained CNN architectures, we explored the benefits of fusing image and metadata for advanced data interpretation. Our key findings reveal that categorizing metadata fields led to greater performance improvements than mere image augmentation. *Multimodal models consistently outperformed their unimodal counterparts in predictive metrics.* Remarkably, our multimodal approaches, especially those using the Xception model, matched or even exceeded the performance of state-of-the-art MADDi models. Our results robustly support the significant potential of multimodal learning in enhancing analytical methods in applied sciences.

Model	Pretrained Weight	Approach	Loss	Accuracy	Precision	Recall	ROC AUC	F1
DenseNet121	random	MADDi	0.5497±0.1911	0.7025±0.0777	0.4063±0.062	0.9667±0.0279	0.9099±0.0315	0.6681±0.08
DenseNet121	imagenet	MADDi	0.7881±0.2457	0.7023±0.069	0.4029±0.0547	0.947±0.0554	0.8904±0.0318	0.6226±0.0499
InceptionV3	random	MADDi	0.9932±0.4894	0.5136±0.1562	0.2963±0.0675	0.9288±0.1209	0.8491±0.0106	0.5586±0.1806
InceptionV3	imagenet	MADDi	0.8622±0.458	0.7797±0.0745	0.4888±0.0919	0.9419±0.0466	0.9219±0.0136	0.6417±0.039
ResNet50	random	MADDi	0.7044±0.555	0.6355±0.2151	0.3859±0.1267	0.9881±0.0141	0.9125±0.0507	0.6912±0.0192
ResNet50	imagenet	MADDi	1.0362±0.6961	0.6588±0.1006	0.3748±0.0809	0.9478±0.0318	0.8521±0.0633	0.5899±0.0537
ResNet101	random	MADDi	0.7513±0.3284	0.5324±0.2422	0.3369±0.1555	0.9782±0.0299	0.8572±0.0819	0.6209±0.072
ResNet101	imagenet	MADDi	0.611±0.0583	0.7175±0.0454	0.4146±0.0377	0.9645±0.0201	0.9109±0.0173	0.6708±0.0154
NA	None	Metadata	1.0633±0.0411	0.8332±0.0135	0.6976±0.0451	0.2874±0.0676	0.8752±0.0117	0.4932±0.0421

TABLE VII: Fluoropolymer AFM results for benchmark classification techniques (bimodal MADDi and metadata-only).

REFERENCES

[1] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[4] D. C. Meshnick, "Multimodal Image Classification In Fluoropolymer AFM And Chest X-Ray Images," MS Thesis in Computer Science, Case Western Reserve University, Cleveland OH, USA, Jaunary 2023. [Online]. Available: http://rave.ohiolink.edu/etdc/view?acc$_num$ = $case$1674834757745168

[5] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019. [Online]. Available: https://arxiv.org/abs/1901.07031

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[7] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[8] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2130–2134, 2015.

[9] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, p. 2346–2352.

[10] R. Lebret, P. O. Pinheiro, and R. Collobert, "Phrase-based image captioning," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 2085–2094.

[11] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang, "Video in sentences out," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'12. Arlington, Virginia, USA: AUAI Press, 2012, p. 102–112.

[12] M. Müller, *Dynamic Time Warping*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84. [Online]. Available: https://doi.org/10.1007/978-3-540-74048-3_4

[13] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2015, pp. 19–27.

[14] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, no. 64, pp. 2211–2268, 2011. [Online]. Available: http://jmlr.org/papers/v12/gonen11a.html

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289.

[16] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, ser. COLT' 98. New York, NY, USA: Association for Computing Machinery, 1998, p. 92–100. [Online]. Available: https://doi.org/10.1145/279943.279962

[17] S. Moon, S. Kim, and H. Wang, "Multimodal transfer deep learning with applications in audio-visual recognition," 2014. [Online]. Available: https://arxiv.org/abs/1412.3121

[18] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, no. C, p. 203–239, may 2022. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.12.003

[19] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, "Bridge correlational neural networks for multilingual multimodal representation learning," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 171–181. [Online]. Available: https://aclanthology.org/N16-1021

[20] J. S. Ellen, C. A. Graff, and M. D. Ohman, "Improving plankton image classification using context metadata," *Limnology and Oceanography: Methods*, vol. 17, no. 8, pp. 439–461, 2019. [Online]. Available: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10324

[21] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[22] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for alzheimer's disease diagnosis," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 12, pp. 2014–2022, Nov 2022.

[23] F. Boschet and B. Ameduri, "(co)polymers of chlorotrifluoroethylene: synthesis, properties, and applications," *Chem. Rev.*, vol. 114, no. 2, pp. 927–980, jan 2014.

[24] P. J. de Pablo, *Introduction to Atomic Force Microscopy*. Totowa, NJ: Humana Press, 2011. [Online]. Available: https://doi.org/10.1007/978-1-61779-282-3_11

[25] C. A. Orme, G. Bordia, and J. Lewicki, "Phase changes in fluoropolymer binders," Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), Tech. Rep. LLNL-TR–765105, 1491962, Jan. 2019.

[26] C. Orme, E. Cho, and J. Lewicki, "Fluoropolymer Aging Assessments," Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), Tech. Rep. LLNL-TR-827066, 1822604, 1041997, Sep. 2021.

[27] F. Mammeri, "Chapter 3 - nanostructured flexible pvdf and fluoropolymer-based hybrid films," in *Nanostructured Thin Films*, ser. Frontiers of Nanoscience, M. Benelmekki and A. Erbe, Eds. Elsevier, 2019, vol. 14, pp. 67–101. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081025727000039

[28] A. Debot, P. Tripathi, and S. Napolitano, "Solution filtering affects the glassy dynamics of spincoated thin films of poly(4-chlorostyrene)," *Eur. Phys. J. E Soft Matter*, vol. 42, no. 8, p. 102, Aug. 2019.

[29] J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks,” *J. Big Data*, vol. 7, no. 1, Dec. 2020.
[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” 2020. [Online]. Available: https://arxiv.org/abs/2004.10934