

HyperMM : Robust Multimodal Learning with Varying-sized Inputs

Hava Chaptoukaev¹, Vincenzo Marciano^{1,2}, Francesco Galati¹, and Maria A. Zuluaga¹

¹ EURECOM, Data Science Department, France

² School of Biomedical Engineering & Imaging Sciences, King's College London, UK
{chaptouk,marciano,galati,zuluaga}@eurecom.fr

Abstract. Combining multiple modalities carrying complementary information through multimodal learning (MML) has shown considerable benefits for diagnosing multiple pathologies. However, the robustness of multimodal models to missing modalities is often overlooked. Most works assume modality completeness in the input data, while in clinical practice, it is common to have incomplete modalities. Existing solutions that address this issue rely on modality imputation strategies before using supervised learning models. These strategies, however, are complex, computationally costly and can strongly impact subsequent prediction models. Hence, they should be used with parsimony in sensitive applications such as healthcare. We propose HyperMM, an end-to-end framework designed for learning with varying-sized inputs. Specifically, we focus on the task of supervised MML with missing imaging modalities without using imputation before training. We introduce a novel strategy for training a *universal* feature extractor using a conditional hypernetwork, and propose a permutation-invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase *task-agnostic* framework. We experimentally demonstrate the advantages of our method in two tasks: Alzheimer's disease detection and breast cancer classification. We demonstrate that our strategy is robust to high rates of missing data and that its flexibility allows it to handle varying-sized datasets beyond the scenario of missing modalities. We make all our code and experiments available at github.com/robustml-eurecom/hyperMM.

Keywords: Multimodal learning · Missing modalities · Multi-resolution data

1 Introduction

Multimodal imaging techniques are widely used both in clinical practice and medical research. Simultaneous acquisition and analysis of multiple imaging modalities, such as Emission Tomography (PET), Computed Tomography (CT), or Magnetic Resonance Imaging (MRI), has shown to be beneficial in the diagnosis of Alzheimer's disease [24], or detection of cancers [25], among others. Accordingly, deep learning methods designed to learn from multimodal medical

images [4, 16], and more generally multimodal medical data [22], have seen rapid growth. This development has been favored by the emergence of multimodal learning (MML), a field of machine learning combining modalities from various sources that depict a single subject from multiple views, thus providing both shared and complementary information. MML has shown considerable advantages in multiple domains [1, 29]. However, most current models [28, 6, 31, 26, 35] assume completeness of the training and testing data, which is rare for real-world datasets. In particular, in routine clinical practice obtaining several modalities for the same subject is not a standard. Incomplete datasets can occur for multitudes of reasons including databases fusion, unavailability of acquisition material, or simply patient refusal to partake in specific examinations. As a result, having varying numbers of modalities per patient is common, which results in multimodal datasets where one or more modalities can be missing. This makes MML challenging as it prevents the straightforward use of the existing methods. Moreso, multimodal models trained on complete datasets become unusable (without complex additional processing steps) if modalities are missing at testing time, which severely restricts their usage to complete samples only. Therefore, the robustness of multimodal models to missing modalities is of paramount importance for the use of MML in real-life applications.

1.1 Related work

MML aims to build models that process and combine information from multiple sources [1], i.e. multiple modalities. The most prominent way to combine multi-source information resides in fusion methods that can be classified in three categories: early fusion, mid-level fusion, and decision-level fusion of modalities [29]. In practice, summation and averaging are common and straightforward techniques used for fusion. However, when modalities are missing, these operations are impossible for early and mid-level fusion in classical multimodal architectures. They are usually not designed to handle varying-sized inputs and fail to account for missing data.

A vast majority of existing solutions to missing modalities in supervised learning consists of first training a generative model on a complete dataset, and using it to impute missing modalities before learning a discriminative model for prediction [2, 10, 21, 32]. This approach has considerable limitations in practice. First, an unreasonable number of samples may be needed for training a good missing-modality imputation model. For instance, generative adversarial networks (GANs) [8, 34], often used for image generation and imputation, can typically require up to 10^6 samples for efficient training [9]. This considerably limits their uses in medical applications where data is often scarce. Second, the complexity of the prediction model strongly depends on the choice of the imputation model. The imputer and predictor networks need to be adapted to each other [11, 13], which can be difficult to ensure in practice. Some studies [23, 27] address this limitation by focusing on jointly learning the modality imputation and prediction tasks, but these models rely on complex and computationally

costly training strategies. Lastly, poorly imputed data can compromise the interpretability of subsequent predictors [18], which is a crucial aspect to consider in sensitive applications, such as healthcare, where it can lead to incorrect conclusions about the impact of a feature on the outcome.

Some recent works have proposed handling missing data without using imputation [17, 33, 3]. Instead of directly imputing the missing modalities, they replace them with *dummy* inputs, such as a constant or generated data (e.g., zeros or Gaussian noise), and then learn to ignore these during training. In contrast, we propose to learn with varying-sized inputs to avoid model degradation caused by poor imputations or the presence of *dummy* data.

1.2 Contributions

In this work, we address the limitations of existing methods by proposing an end-to-end *imputation-free* strategy for multimodal supervised learning with missing imaging modalities. Building on conditional hypernetworks [5], we formulate a novel strategy for training a *universal* modality-agnostic feature extractor using a large pre-trained network. We then reformulate the problem of predicting multimodal observations with missing modalities as one of predicting *sets* of observations of varying size, thus relaxing the requirement of fixed-dimensional data inputs of most machine learning models. We implement this approach through a permutation-invariant neural network [30], allowing the mid-level fusion of varying-sized multimodal inputs, hence eliminating the need to impute [2, 21, 32] or mask missing modalities using dummy data [17, 33, 3] as done in previous works. By combining these elements into a two-step training framework, we formulate HyperMM, a novel *task* and *model agnostic* strategy for MML from incomplete datasets, without the use of imputation or dummy data in the training process. To the best of our knowledge, our work is the first proposing such an approach for multimodal learning with missing modalities.

2 Methodology

2.1 Overview of the method

We consider a dataset \mathcal{D} of $n \in \mathbb{N}$ independent input and output pairs such that $\mathcal{D} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, and for which the goal is to predict Y given X . Each $X := \{x_1, \dots, x_d\}$ corresponds to a d -modal observation, where each x_i represents one of the available modalities. Let us now introduce the indicator vector $v \in \{0, 1\}^d$ to denote the positions of missing modalities in X , such that $v_i = 1$ if x_i is missing, and 0 otherwise. The observed data of X can be expressed as $X_{obs} = (1 - v) \odot X + v \odot \mathbf{na}$, where \odot is the term-by-term product. In this setting, the learning goal becomes the prediction of Y given X_{obs} .

We intend on learning without the use of any form of imputation of missing modalities, and therefore, with entries of different dimensions. However, standard machine learning models, including MML models, are built to handle data inputs

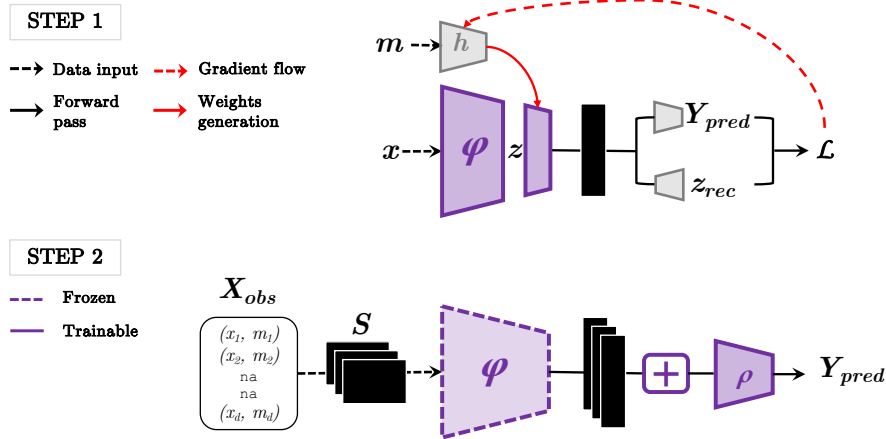


Fig. 1: Overview of our HyperMM framework. A network φ is trained to extract features from any modality in \mathcal{D} by jointly optimizing feature reconstruction and unimodal prediction (step 1). The learned φ is frozen, and used to process multimodal inputs, the latent features are then aggregated and processed through a network ρ for prediction (step 2).

of a fixed size. In contrast, we aim to learn a sum-decomposable function f of the form $f = \rho(\sum \varphi(x_i))$, operating on *sets* and thus relaxing the requirement of fixed-dimensional data. We propose a two-step framework that we call HyperMM to implement our method. Figure 1 presents an overview of our strategy. In a first step, we learn a neural network φ that can extract features from any modality present in \mathcal{D} . Then in a second step, we freeze the learned φ , use it to encode each element of X_{obs} , and feed the combination of the encoded inputs to a classifier ρ through a permutation-invariant architecture.

2.2 Universal Feature Extractor

A single network φ that can encode all observed modalities in \mathcal{D} is a requirement for learning a set function as described in Sec. 2.1. We propose to achieve this by first learning such a universal feature extractor φ using a conditional hypernetwork [5]. In this first step, we train a network on all available images x in the dataset, without any modality pairing. As illustrated in Figure 1, we introduce an *auxiliary* network h that takes as input m , the modality identifier corresponding to the image x , and generates conditional weights for the last layer of the encoder φ . By doing so, the last feature extraction step is different for each modality but still performed by the same network. Specifically, modality-specific layers are generated through a common hypernetwork, which facilitates information sharing across modality-specific layers.

In practice, our universal feature extractor φ can be implemented using transfer learning and networks pre-trained on natural images such as VGGs [19]. First, we use the pre-trained encoding layers of a VGG to extract features from our dataset. Then, we adapt the obtained general features into medical ones by training an additional layer on top of the VGG extractor, that is conditioned using the auxiliary network h . By stacking these elements together, we obtain our universal feature extractor φ that is adapted to the modalities of our dataset.

To ensure that the features learned by φ are relevant, the network is trained to both predict y from the single modality images, and reconstruct z , the features outputted by the second-to-last layer of φ . This is achieved by optimising a loss function of the form $\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE}$, where \mathcal{L}_{MSE} denotes the mean squared error between z and z_{rec} , and \mathcal{L}_{CE} the cross-entropy loss between y and y_{pred} . This optimisation loss has been chosen by cross-validation, as it yielded better performances than optimising on the classification or reconstruction only.

2.3 Permutation Invariant Architecture

Once we have learned φ , we freeze it, and use it to implement a permutation invariant network for supervised MML with missing modalities. To do so, we define S , the set representation of the $q = |S|$ observed elements of X_{obs} , such that $S := \{s_1, \dots, s_q\}$, with $q \leq d$. Each element s_j is represented as a tuple (x_i, m_i) consisting of an observed modality x_i , and the corresponding modality identifier m_i . This reformulation allows observations of varying dimensions. Thereby, it does not require nor expects all observations to have the same number of elements and it fully allows observations with missing modalities. A d -modal observation X_{obs} containing `na` values can simply be expressed as a set S of size $q \leq d$ where the `na` values are not represented anymore.

Using this definition, we leverage on the findings of [30], who proposed a learning framework that considers permutation invariant functions operating over sets. We reformulate our learning goal as one of learning a set function f of the form

$$f(X_{obs}) = \rho \left(\sum_{s_k \in S} \varphi(s_k) \right), \quad (1)$$

where the function $\varphi : \mathbb{R} \times \{r \times r\}^d \rightarrow \mathbb{R}^{d_l}$ corresponds the encoder obtained from the pre-training phase, the function $\rho : \mathbb{R}^{d_l} \rightarrow \mathbb{R}$ is implemented as neural network, r is the size of each image and $d_l \in \mathbb{N}^+$ is the dimensionality of the latent space of φ .

As illustrated in Figure 1, a given observation X_{obs} with missing modalities is encoded as a set S . Each element $s_k \in S$ is then transformed into a representation $\varphi(s_k) := \varphi(x_i|m_i)$ through the frozen network φ conditioned by the modality identifier m . The representations $\varphi(s_k)$ are aggregated using a permutation invariant operation such as the sum, the mean or the maximum. The aggregation is processed through the network ρ , which allows to predict the target Y corresponding to the input X_{obs} . The proposed architecture interprets

each observation S of a dataset as a set of unordered modalities, where all information available in X_{obs} is conserved and no new information, such as imputed images, is added. By transforming individual elements s_k of S at a time and then aggregating the transformations, our network encodes sets of arbitrary sizes into a fixed representation $\sum \varphi(s_k)$. This aspect is particularly relevant and further justifies handling our dataset with missing modalities as unordered sets.

Our permutation invariant model is learned by optimising the loss function

$$\mathcal{L}(\theta) := \mathbb{E}_{(S,Y) \in \mathcal{D}} \left[\ell \left(Y, \rho_{\theta} \left(\sum_{s_k \in S} \varphi(s_k) \right) \right) \right], \quad (2)$$

where ρ is parametrised by θ , and ℓ is the cross-entropy loss. As φ is optimised in the pre-training step, its weights are not updated in this step.

3 Experiments

3.1 Alzheimer’s Disease Detection

In a first application, we illustrate the performances of HyperMM and its robustness to missing modalities on the task of binary classification of Alzheimer’s disease (AD) using multimodal images from the ADNI dataset [15]. We select a subset of 300 patients for which both T1-weighted MRIs and FDG-PET images are available, resulting in 165 cognitively normal (CN) and 135 AD observations. Before learning, all the samples are skull stripped using HD-BET [7], resampled through bicubic interpolation to set an uniform voxel size, standardised, and normalised using min-max scaling.

Baselines. We first evaluate the advantages of our strategy for MML with complete data. We compare the performances of HyperMM against: **Uni-CNN**, a unimodal CNN as implemented by [12]; **Multi-CNN**, a multimodal CNN as proposed by [26]; and **Multi-VAE**, a multimodal VAE [28] that we adapt for classification. Then, we compare our method against state-of-the-art techniques for MML with missing modalities in two scenarios: complete MRIs +50% of PETs available for training and testing, and complete PETs +50% of MRIs available. Specifically we compare to: **pix2pix**, a strategy where an image-to-image translation model [8] is trained on the subset of the training data containing only modality-complete samples, is then used to impute the missing modality of the incomplete data, and once imputed the data is classified using a Multi-CNN; and **cycleGAN**, the same strategy, only using a cycleGAN [34] for imputation.

Implementation details. We randomly split the data into train, validation and test sets with a 6:1:3 ratio on the patient-level, and repeat all experiments 3 times. For simplicity and fairness, we use the same feature extraction strategy (Figure 2) in all baselines, following [12]. Specifically, 3D MRI and PET images are processed as batches of 2D slices that are each fed to a pre-trained frozen VGG11 [19] feature extractor. We feed all 2D slices of a 3D volume to the VGG,

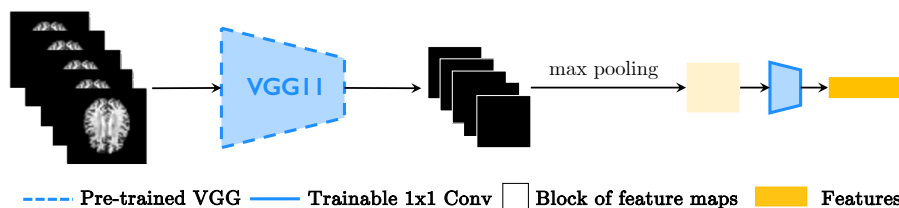


Fig. 2: Feature extraction strategy used in the ADNI baselines (see [12]). All 2D slices of one 3D volume are fed to a VGG11. A 1D max pooling on the slice dimension is applied to the resulting feature blocks to obtain a single block per 3D image. The latter is passed through a 1×1 convolution layer to obtain AD-specific features that can then be fed to a classifier.

and apply a 1D max pooling on the slice dimension to the resulting feature blocks to obtain a single block per 3D image. The resulting block is passed through a 1×1 convolution layer after the pre-trained VGG encoder, allowing us to adapt the pre-trained features into AD-specific ones. This corresponds to the training of the φ network in the step 1 of our framework, where we simply make the last 1×1 convolution layer conditional. In step 2, the ρ network is implemented by 3 linear layers separated by ReLU activations. All models are implemented with PyTorch, and trained on an Nvidia TITAN Xp GPU for a maximum of 100 epochs using an early stopping strategy, where training stops after 10 iterations without a decrease in the validation loss. We use a batch size of 1 and an Adam optimiser with an initial learning rate of $1e-4$.

Results. Performances of all models are reported in Table 1. Several observations can be drawn from these results. First, MML shows considerable improvements over unimodal baselines. In particular, HyperMM achieves the best performances for binary classification of AD using complete multimodal data and considerably improves the F1-score, recall metric, and precision/recall balance. Second, MML with missing modalities still achieves better results than unimodal models. Notably, HyperMM trained on MRIs available even for only 50% of the patients performs better than an unimodal model trained on PETs only. Inversely, having access to PETs for 50% of the patients improves the F1-score and recall of learning from MRIs only. Third, HyperMM outperforms state-of-the-art strategies on MML with missing modalities. While GAN-based strategies can handle missing PETs in the input data, they are considerably less efficient in terms of precision/recall balance when the missing modality is MRI. In this scenario, the missing high-resolution MRIs need to be translated from the available low-resolution PETs before learning. This limitation is further illustrated in Figure 3. While PET imputation yields realistic images, the imputed MRIs are of poor quality: they suffer from important structural deformations and a great loss of information (as highlighted by the SSIM and PSNR scores between the imputations and the original images). In contrast, as HyperMM does not rely

Table 1: Performances (mean±std) on the ADNI dataset. **Bold** means best.

	Acc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)	Time (↓)
Complete unimodal						
Uni-CNN PET	0.61±.05	0.58 ±.05	0.58±.06	0.65±.06	0.31±.05	< 20 min
Uni-CNN MRI	0.71±.02	0.69±.02	0.58±.02	0.85±.03	0.43±.05	< 20 min
Complete multimodal						
Multi-VAE classifier	0.66±.03	0.65±.03	0.54±.04	0.74±.04	0.41 ±.03	< 30 min
Multi-CNN	0.70±.02	0.70±.01	0.67±.01	0.67±.02	0.68±.02	< 30 min
HyperMM w/o 2-steps (ours)	0.62±.03	0.61±.02	0.53±.02	0.61±.03	0.46±.03	< 20 min
HyperMM w/ 2-steps (ours)	0.74±.02	0.73±.02	0.70±.01	0.70±.02	0.70±.02	< 1 h
100% MRI + 50% PET						
pix2pix	0.65±.02	0.64±.02	0.62±.02	0.62±.03	0.61±.02	> 14+1 h
cycleGAN	0.62±.09	0.60±.07	0.57±.07	0.61±.08	0.54±.08	> 30+1 h
HyperMM (ours)	0.67±.02	0.66±.02	0.61±.03	0.61±.03	0.61±.03	< 1 h
100% PET + 50% MRI						
pix2pix	0.62±.04	0.62±.03	0.53±.03	0.61±.05	0.48±.05	> 14+1 h
cycleGAN	0.62±.09	0.59±.1	0.47±.07	0.60±.07	0.39±.07	> 30+1 h
HyperMM (ours)	0.64±.02	0.63±.02	0.61±.02	0.61±.03	0.61±.03	< 1 h

on any imputation, it performs well in both scenarios, and trains in significantly less time than competitors. Lastly, these results highlight the importance of the pre-training and conditioning step of the HyperMM framework.

Discussion. The results illustrate how HyperMM tackles the main limitations of existing methods. First, as our model does not require training an imputation model prior to prediction, it does not call for the large amounts of data typically required for training GANs efficiently. The results observed in Table 1 highlight the poor performances of cycleGAN for translating PETs into MRIs, which could be due to insufficient training data. Second, our model is agnostic to the missing modality, whereas the prediction and imputation quality in other approaches strongly depends on it, as highlighted by our experiments. Indeed, because HyperMM bypasses the imputation step altogether, our approaches eliminates the need to ensure that the imputer and predictor are adapted to each other. This, in turn, leads to drastically reduced computing time and learning complexity. Lastly, as our method does not employ any imputed or dummy data, it avoids model degradation caused by poor imputations or noisy data.

3.2 Breast Cancer Classification

In a second application, we demonstrate the flexibility of HyperMM and its benefits for learning with varying-sized datasets, beyond the scenario of missing modalities. We investigate the usage of HyperMM for the slightly different task of analysing multi-resolution histopathological images. Because potential tumors are typically acquired at multiple magnification levels, the numbers of samples per patients in histopathology datasets are often highly varying. We per-

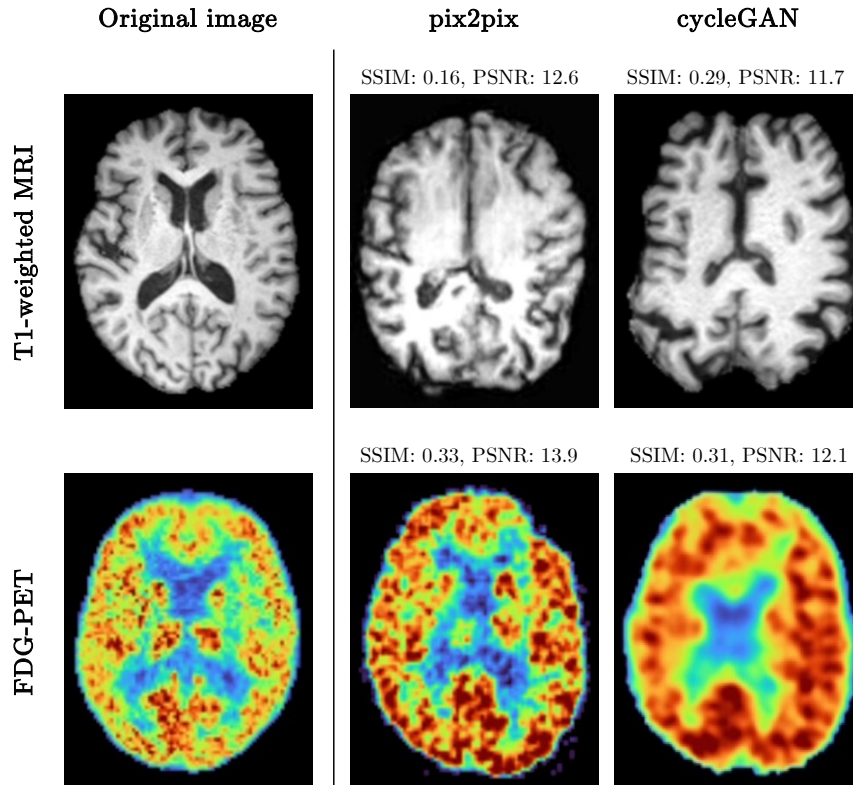


Fig. 3: Examples of real and imputed slices of MRI and PET images for one patient. While the PET reconstructions (bottom right) translated from the corresponding MRI (top left) are reasonably similar to the original PET image (bottom left), the MRI reconstructions (top right) translated from the low-resolution PET (bottom left) are much less consistent with reality (top left).

form binary classification of breast cancer using histopathological images from the BreakHis dataset [20]. BreakHis contains multiple images per sample (i.e. patient) of benign or malignant tumors observed through different microscopic magnifications: $40\times$, $110\times$, $200\times$, and $400\times$. We select a balanced subset of the data composed of samples of 24 benign and 29 malignant tumors, resulting in 5,575 images in total. We use the images as they are for learning, and do not perform any pre-processing or data augmentation.

In clinical practice, pathologists combine the complimentary information present in images captured under different magnifications in order to make a patient-level decision. Nonetheless, most current learning approaches consist of magnification-specific models, due to the difficulty of processing images of different natures with a single model. Moreover, because the number of available

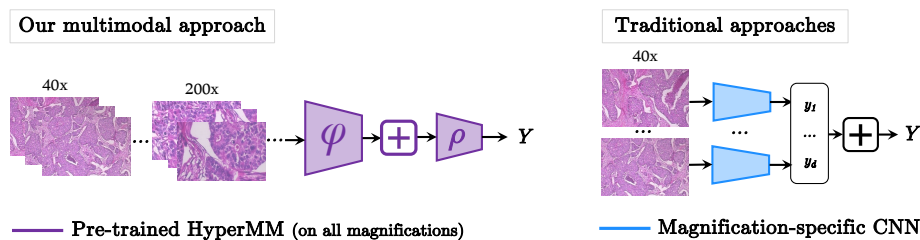


Fig. 4: Comparison of decision strategies for patient-level tumor classification. Our method (left) enables the combination of a subject’s available images during training, regardless of the magnification level to obtain a patient-level decision. In opposition, traditional approaches (right) make prediction on the image-level, and combine the final predictions to obtain a patient-level decision.

images can vary a lot from one patient to another, traditional algorithms cannot be applied at the patient-level. Existing methods rather predict from individual images, and later combine the predictions in order to form a global decision. Instead, we propose to tackle this problem using HyperMM, conditioning the universal feature extractor on the different magnification levels. We classify tumors at patient-level by combining all available images during training directly.

Baselines. We evaluate the benefits of HyperMM for learning from histopathology data, and compare its performances with: **CNN**, where a magnification-specific CNN is trained to classify tumor types from individual images, and patient-level prediction is obtained by averaging the classification scores of individual images [20]; and **Incremental-CNN**, in which a magnification-agnostic CNN is trained by incrementally updating its weights on successive batches of $40\times$, $100\times$, $200\times$ then $400\times$ magnifications, as proposed in [14]. The patient-level decision is obtained similarly to the previous baseline. The differences between our approach and traditional ones are further illustrated in Figure 4.

Implementation details. We randomly split the data into train and test sets with a 8:2 ratio at the patient-level, and repeated all experiments 5 times. We use a pre-trained VGG11 [19] feature extractor for all baselines, and adapted the features to our application by adding a 1×1 convolution block on top of the frozen VGG encoder. All models are trained for a maximum of 50 epochs using an early stopping strategy such that training stops after 10 iterations without a decrease in the validation loss. We train the model with an Adam optimiser with an initial learning rate of $1e-4$. We use a batch size of 16 for image-level baselines (i.e. CNN and Incremental-CNN) and 1 for HyperMM.

Results. All performances averaged over 5 repetitions are reported in Table 2. They underline the clear benefits of HyperMM for cancer classification from histopathological images. In particular, our method outperforms magnification-

Table 2: Performances (mean \pm std) on the BreakHis dataset. **Bold** means best.

	Acc. (\uparrow)	AUC (\uparrow)	F1 (\uparrow)	Prec. (\uparrow)	Rec. (\uparrow)
Magnification-specific					
CNN 40 \times	0.83 \pm 0.07	0.81 \pm 0.07	0.83 \pm 0.06	0.85 \pm 0.08	0.83 \pm 0.08
CNN 100 \times	0.85 \pm 0.08	0.85 \pm 0.08	0.87 \pm 0.06	0.85 \pm 0.07	0.90 \pm 0.07
CNN 200 \times	0.84 \pm 0.07	0.84 \pm 0.09	0.84 \pm 0.05	0.80 \pm 0.11	0.90 \pm 0.09
CNN 400 \times	0.83 \pm 0.09	0.83 \pm 0.09	0.85 \pm 0.10	0.88 \pm 0.11	0.83 \pm 0.15
Magnification-agnostic					
Incremental-CNN	0.89 \pm 0.11	0.88 \pm 0.12	0.90 \pm 0.10	0.88 \pm 0.12	0.93\pm0.09
HyperMM (ours)	0.92\pm0.06	0.91\pm0.07	0.90\pm0.08	0.94\pm0.09	0.88 \pm 0.10

specific models, and is closely followed by Incremental-CNN, which highlights the benefits of combining the information carried by different magnifications. Moreover, while Incremental-CNN maximises the recall score of the task, HyperMM maximises precision, and overall improves upon Incremental-CNN. This shows that learning to predict an early latent combination of features (i.e. combining multiple images of a same patient during model training directly) yields better performances than combining predictions made on individual images.

Discussion. While the analysis of multi-resolution images is not a multimodal application by definition, our method is designed to enable mid-level fusion of latent features of varying-sized inputs, and is therefore adapted for this use case. Because of the varying number of images per patient in histopathology datasets, traditional approaches are not equipped to combine multiple resolutions directly during training to make patient-level decisions, and instead rely on the late fusion of image-level decisions. In contrast, HyperMM offers this possibility. It opens a new and different way to classify cancer patients. Moreover, our experiments suggest that mid-level fusion even considerably improves the performances of existing late fusion models.

4 Conclusion

We have demonstrated the advantages of HyperMM for robust MML with missing modalities: our method eliminates the need to use complex and computationally costly imputation strategies, thus significantly decreasing model training time; and unlike competitors, its performances are not dependant on which modality is missing in the data. In particular, unlike imputation-based methods, our approach is end-to-end: HyperMM eliminates the time-consuming steps of manually imputing the missing modalities using a previously trained imputation model, before finally training a prediction model. On the contrary, our two-step model is trained without interruption or human intervention. Moreover, by only utilising the observed images of the incomplete dataset, HyperMM avoids prediction bias caused by poor imputation or the presence or generated dummy

data. In addition, we have shown that the flexibility of HyperMM alleviates the constraints usually met in applications with varying-sized datasets and opens up a whole new range of possible learning strategies. Our framework is *task-agnostic*, and can be easily used beyond the two applications we have presented. For instance, it could be extended to multivariate time series analysis, where incomplete data is common (e.g. damaged channels in EEG recordings). Moreover, while we used pre-trained feature extractors in all our experiments for simplicity, HyperMM is *model-agnostic* and adaptable to any neural network-based feature extractor or predictor.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d’Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19- P3IA-0002).

References

- [1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
- [2] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018.
- [3] B. Chen, J. Oliva, and M. Niethammer. A unified model for longitudinal multi-modal multi-view prediction with missingness. *arXiv preprint arXiv:2403.12211*, 2024.
- [4] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim. Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3507–3516, 2021.
- [5] D. Ha, A. M. Dai, and Q. V. Le. Hypernetworks. *CoRR*, 2016.
- [6] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, and A. D. N. I. (ADNI). Diagnosis of alzheimer’s disease via multi-modality 3d convolutional neural network. *Frontiers in neuroscience*, 13:509, 2019.
- [7] F. Isensee, M. Schell, I. Pfueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.
- [8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [9] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

- [10] J.-C. Kim and K. Chung. Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE Access*, 8:104933–104943, 2020.
- [11] M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What’s a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.
- [12] G. Liang, X. Xing, L. Liu, Y. Zhang, Q. Ying, A.-L. Lin, and N. Jacobs. Alzheimer’s disease classification using 2d convolutional neural networks. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3008–3012, 2021.
- [13] Z. Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [14] M. S. Mayouf and F. Dupin de Saint-Cyr. Curriculum incremental deep learning on breakhis dataset. In *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, pages 35–41, 2022.
- [15] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimer’s disease: the alzheimer’s disease neuroimaging initiative (adni). *Alzheimer’s & Dementia*, 1(1):55–66, 2005.
- [16] M. Odusami, R. Maskeliūnas, R. Damaševičius, and S. Misra. Machine learning with multimodal neuroimaging data to classify stages of alzheimer’s disease: a systematic review and meta-analysis. *Cognitive Neurodynamics*, pages 1–20, 2023.
- [17] S. Parthasarathy and S. Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020.
- [18] T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torné, E. Sala, P. Lió, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.
- [19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [20] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015.
- [21] W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang. Semi-supervised multimodal image translation for missing modality imputation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4320–4324, 2021.
- [22] Z. Sun, M. Lin, Q. Zhu, Q. Xie, F. Wang, Z. Lu, and Y. Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*, page 104482, 2023.
- [23] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, volume 3534, page 3540, 2019.

- [24] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavado, G. B. Frisoni, W. Hoffmann, et al. Multimodal imaging in alzheimer’s disease: validity and usefulness for early detection. *The Lancet Neurology*, 14(10):1037–1053, 2015.
- [25] C. M. Tempany, J. Jayender, T. Kapur, R. Bueno, A. Golby, N. Agar, and F. A. Jolesz. Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer*, 121(6):817–827, 2015.
- [26] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multi-modal deep learning models for early detection of alzheimer’s disease stage. *Scientific reports*, 11(1):3254, 2021.
- [27] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.
- [28] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.
- [29] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [30] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.
- [31] T. Zhang and M. Shi. Multi-modal neuroimaging feature fusion for diagnosis of alzheimer’s disease. *Journal of Neuroscience Methods*, 341:108795, 2020.
- [32] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou. Unified multi-modal image synthesis for missing modality imputation. *arXiv preprint arXiv:2304.05340*, 2023.
- [33] Q. Zhou, H. Zou, H. Jiang, and Y. Wang. Incomplete multimodal learning for visual acuity prediction after cataract surgery using masked self-attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 735–744. Springer, 2023.
- [34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [35] Q. Zuo, B. Lei, Y. Shen, Y. Liu, Z. Feng, and S. Wang. Multimodal representations learning and adversarial hypergraph fusion for early alzheimer’s disease prediction. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*, pages 479–490. Springer, 2021.