Multimodal Deep Learning for Low-Resource Settings: A Vector Embedding Alignment Approach for Healthcare Applications

David Restrepo[1,2*†], Chenwei Wu[3†], Sebastián Andrés Cajas[4, 5], Luis Filipe Nakayama[1, 6], Leo Anthony Celi[1, 7, 8], Diego M López[2]

[1]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America.
[2]Departamento de Telemática, Universidad del Cauca, Popayán, Cauca, Colombia.
[3]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, United States of America.
[4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Boston, Massachusetts, United States of America.
[5]School of Computer Science, University College Dublin, Belfield, Dublin, Ireland.
[6]Department of Ophthalmology, São Paulo Federal University, São Paulo, São Paulo, Brazil.
[7]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, United States of America.
[8]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America.

*Corresponding author(s). E-mail(s): davidres@mit.edu;
Contributing authors: chenweiw@umich.edu; sebasmos@mit.edu; luisnaka@mit.edu; lceli@mit.edu; dmlopez@unicauca.edu.co;
†These authors contributed equally to this work.

Abstract

Objective: Large-scale multi-modal deep learning models and datasets have revolutionized various domains such as healthcare, underscoring the critical role of computational power. However, in resource-constrained regions like Low and

Middle-Income Countries (LMICs), GPU and data access is limited, leaving many dependent solely on CPUs. To address this, we advocate leveraging vector embeddings for flexible and efficient computational methodologies, aiming to democratize multimodal deep learning across diverse contexts.

Background and Significance: Our paper investigates the computational efficiency and effectiveness of leveraging vector embeddings, extracted from single-modal foundation models and multi-modal Vision-Language Models (VLM), for multimodal deep learning in low-resource environments, particularly in healthcare applications. Additionally, we propose an easy but effective inference-time method to enhance performance by further aligning image-text embeddings.

Materials and Methods: By comparing these approaches with traditional multimodal deep learning methods, we assess their impact on computational efficiency and model performance using accuracy, F1-score, inference time, training time, and memory usage across 3 medical modalities such as BRSET (ophthalmology), HAM10000 (dermatology), and SatelliteBench (public health).

Results: Our findings indicate that embeddings reduce computational demands without compromising the model's performance, and show that our embedding alignment method improves the performance of the models in medical tasks.

Discussion: This research contributes to sustainable AI practices by optimizing computational resources in resource-constrained environments. It highlights the potential of embedding-based approaches for efficient multimodal learning.

Conclusion: Vector embeddings democratize multimodal deep learning in LMICs, especially in healthcare. Our study showcases their effectiveness, enhancing AI adaptability in varied use cases.

Keywords: foundation Models, Efficient Deep Learning, Embeddings, Multimodal Data

1 Introduction

In the era of data-driven decision-making in healthcare, deep learning has emerged as a pivotal methodology for extracting meaningful information from the vast amounts of data from different modalities such as clinical notes, vital signs, lab values, and medical images, among others. The increase of multimodal data, which integrates disparate data formats such as text, image, or audio, requires developing sophisticated computational techniques to process and integrate these heterogeneous data types [1–3]. This integration, known as multimodal data fusion, leverages mainly deep learning techniques [4, 5] and is critical for building systems that can interpret complex data in a manner akin to human cognition, thereby enhancing decision-making processes in clinical applications [6–13]: with fundus photos [14], Chest X-rays [15], or even public health applications using remote sensing techniques [16–23].

However, the computational exigencies of such advanced methods pose a formidable barrier, especially in low-resource settings [24] environments characterized mainly by limited computational power [25, 26] and medical data scarcity [27–32]. Addressing these constraints requires innovative approaches that optimize computational efficiency without compromising the efficacy of the learning algorithms.

Vector embeddings represent a promising concept in the domain of data efficiency, particularly concerning high-dimensional data like images and text. The embeddings are high-dimensional vectors that encapsulate the essential features of data entities (e.g., words, or images) in a continuous vector space [33]. By transforming raw data into a more abstract and computationally manageable form, embeddings facilitate significant reductions in the complexity and dimensionality of data, which is paramount in resource-constrained environments[34].

The concept of foundation models, representing deep learning architectures primarily based on the transformer framework, has garnered significant attention in recent years [35]. These models exhibit remarkable capabilities across diverse domains, such as natural language processing (NLP), exemplified by BERT [36], GPT [37], and LLAMA 2 [38]; as well as Computer Vision with Vision Transformer (ViT) [39], or DINO v2 [40]; and even multimodal tasks such as Vision Language Models (VLM) with models like CLIP [41], or BLIP 2 [42]. Leveraging pre-training on extensive datasets, foundation models offer a versatile starting point for various tasks by providing pre-learned representations that capture a wide array of data characteristics. When applied to multimodal data fusion, embeddings extracted from such models can dramatically lessen the computational load, making it feasible to deploy sophisticated deep-learning models in low-resource settings.

On the other hand, although the use of foundation models such as computer vision models, Large Language Models (LLMs), or VLMs provide us with a robust way of extracting embeddings, we must also take into account that these embeddings may be biased depending on the distribution of data learned during training, mostly from overrepresented groups, perpetuating health biases [43]. Liang et al. [44] demonstrated that the intrinsic structure of deep learning models generates embedding representations biased to a very small latent space (cone effect). The cone effect generates representations where the image and text embeddings are distant and confined to a very small region of the latent space. Liang et al. also demonstrated how modifying the gap between the embeddings of different modalities improves the fairness and performance of the models [44].

In this paper, we'll comprehensively examine the computational efficiency gleaned from leveraging vector embeddings extracted from foundation models in multimodal data fusion tasks. We will compare the results with the conventional transfer learning approach using the raw data. Additionally, we'll demonstrate how the cone effect can be amplified in medical data and propose an embedding alignment method to close the modality gap in medical data.

This comparison, grounded in a series of methodical experiments across diverse benchmark datasets, aims to elucidate the trade-offs regarding model performance, processing time, memory utilization, and convergence rates. Through this analysis, our research contributes to the broader dialogue on sustainable AI practices, advocating for efficient computational resource utilization in an era marked by the escalating ecological footprint of AI technologies [45].

2 Methods

This study investigates the computational efficiency of using embeddings extracted from foundation models and VLMs versus processing raw data in multimodal deep learning, particularly for healthcare applications. We conducted a series of experiments across three image-text medical datasets to compare the performance of three distinct approaches: 1. Unimodal embedding extraction using a foundation computer vision model, and an LLM for image and text individual embedding extraction; a VLM for image and text embedding extraction; and a transfer learning approach to fine-tune pre-trained transformer-based models using raw data directly. This comparison focuses on key metrics such as accuracy, F1 score, inference time, training time, and memory usage, providing insights into the effective use of computational resources in multimodal data fusion. Additionally, we'll provide a tool to close the embedding gap between modalities generated in medical data.

2.1 Datasets

The evaluations encompassed four multimodal datasets across healthcare applications: diabetic retinopathy, skin lesion classification, and dengue prediction using satellite imagery. The datasets include:

- BRSET (Brazilian Multilabel Ophthalmological Dataset) [46, 47]: A multi-labeled ophthalmological Brazilian dataset. In this case we use BRSET focusing only on binary diabetic retinopathy disease classification. The dataset comprises 16,266 retinal photos from 8,524 patients with metadata corresponding to patient demographic and disease information.
- HAM10000 [48]: The HAM10000 dataset, an acronym for "Human Against Machine with 10,000 training images," encompasses a comprehensive collection of 10,015 dermatoscopic images for the automated diagnosis of pigmented skin lesions. The images have been sourced from diverse populations and were captured using various dermatoscopic imaging techniques. The categories include Actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage, vasc).
- Satellite Images for Public Health (SatelliteBench) [49, 50]: Adapted from 12-band satellite images to RGB images, the dataset comprises 12,636 satellite images from 81 Colombian municipalities. The task in this dataset involves binary dengue outbreak classification: '1' is assigned to instances with Dengue cases surpassing the median (indicating higher risk), while '0' is assigned to those below. The dataset contains more than 156 images per municipality taken between 2015 and 2018.

The study used an 80-20 split for training and testing to ensure integrity. Text prompts were generated for datasets without text using prompt templates. with the information of each patient and image. The final number of train and test samples per dataset were defined as:

- BRSET [46, 47]: 13012 training samples, and 3254 testing samples.
- HAM10000 [48]: 8012 training samples, and 2003 testing samples.
- SatelliteBench [49, 50]: 936 training samples, and 312 testing samples.

2.2 Models

2.2.1 Single-Modal foundation Models as Embeddings Extractor

As shown in Figure 1B, this approach leverages pre-trained foundation models to extract embeddings, hypothesizing that these models, having been trained on extensive datasets, can generate rich, informative representations without further fine-tuning. Image embeddings were obtained using Meta's foundation computer vision model DINO V2 [40], and text embeddings were extracted from Meta's Large Language Model (LLM) LLAMA 2 [38]. To alleviate computational resources, for LLAMA 2, we used the smaller version of LLAMA 2-7B with 7 Billion parameters. We also used a version quantized to 8 bits to allow faster inference and lower usage of computational resources. It is important to mention that this same methodology can be applied to bigger models like LLAMA 2-70B, or GPT 4, to extract better-quality information. The embeddings derived from these models were then archived into individual CSV files, ready for subsequent model training and evaluation processes.

2.2.2 Vision Language foundation Models(VLMs) as Embeddings Extractors

This method based in Figure 1B, uses VLMs as embedding extractors, assuming that the model learned a joint representation of both modalities, text, and images. In this case, we selected a CLIP model [41], widely used in the community due to its simplicity and good performance. CLIP was selected also due to its ability to extract uni-modal embeddings instead of a joint embedding representation, allowing us more flexibility during the following experiments. The embeddings extracted from the CLIP model were stored in a CSV file for subsequent model training and evaluation.

2.2.3 Vector Embeddings Multimodal Fusion Learning

For the modeling tasks, two fusion techniques, an early (Figure 1C-1) and late (Figure 1C-2) fusion methods were employed [51]:

- Early Fusion: The embeddings from both modalities were concatenated at the input level of our classifier. The classifier consists of a feature extraction block composed of a dense layer with ReLu activation, dropout, and batch normalization. Finally, the output of the previous block was passed to a dense layer with the number of neurons as to the number of output classes for the classification. This approach can be seen in Figure 1C-1.
- Late-Joint Fusion: In the Late-Joint Fusion approach, the embeddings of each modality were passed separately through two feature extraction blocks composed of a dense layer with ReLu activation, a dropout, and a batch normalization of each one. These feature representations were merged later and passed through a final classification head. The approach can be seen in Figure 1C-2.

2.3 Raw Data Multimodal Fusion Learning

In this method, represented in Figure 1A, raw data were directly fed into pre-trained models based on transformers. In this case, a transfer learning approach was used by fine-tuning two transformer-based backbones pre-trained on image and text. Text data were tokenized using a BERT tokenizer and processed through a BERT model architecture [39], while images were inputted into a ViT base architecture [40] pre-trained on ImageNet. The outputs from these models were then integrated using a classification head with the same fusion techniques as the embedding approach in Figure 1C-1 and Figure 1C-2, ensuring a consistent comparison between the two methods.

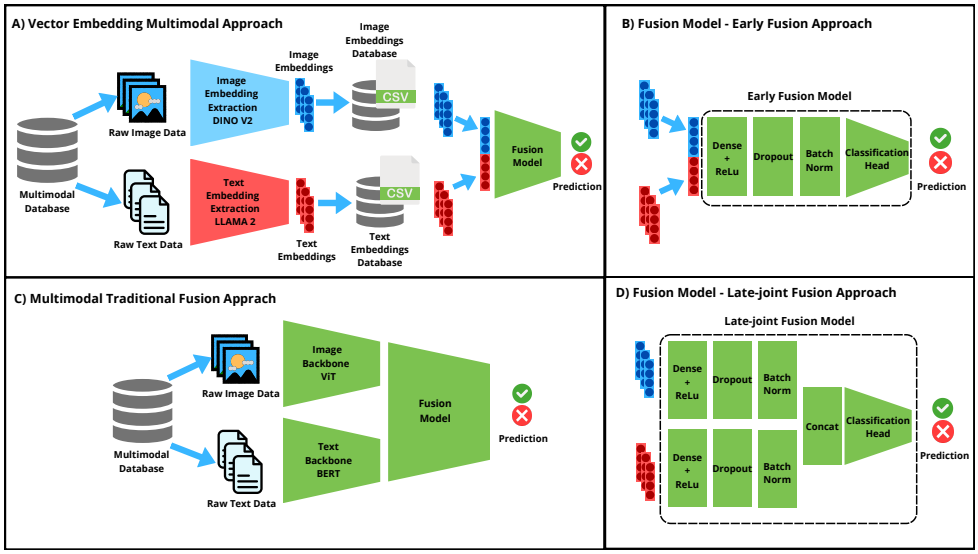

Fig. 1 Schematic Representation of Multimodal Fusion Approaches. (1A) depicts the traditional multimodal fusion approach using raw data. The approach processes text and images through BERT and ViT models respectively. (1B) shows our embedding multimodal modeling approach, illustrating the extraction of image and text embeddings from foundation models and their subsequent utilization in multimodal learning. (1C) shows the two distinct approaches for data fusion. C-1 represents an early fusion approach, where embeddings are concatenated at the input and passed through a feature extraction block, followed by a classification layer. C-2 presents the late-joint fusion approach, highlighting the separate feature extraction from each modality and their integration at a later stage.

2.4 Experimental Setup

2.4.1 Hardware

Experiments ran on Oracle's Standard.E4.Flex platform, with 2 CPU cores, 64GB each, no GPU, mimicking low-resource settings. PyTorch executed experiments independently, ensuring consistent model architectures and preventing memory leaks.

2.4.2 Training and evaluation details

The AdamW optimizer was employed with default settings from the PyTorch library. We used BCEWithLogitsLoss for binary classification on BRSET, and Satellite Bench and CrossEntropyLoss for multi-class classification on the Ham 10000 dataset. Each loss function was adjusted with class weights inversely proportional to the samples per class to avoid overfitting. A batch size was set to 64 for train and test data loaders, and all the models were trained during 30 epochs.

For the classifiers, the initial dense layers were set to have 64 neurons in each initial block for late-joint fusion; and 128 neurons in the initial block for early fusion followed by a ReLu activation function. A dropout was defined as 0.0 for the three medical datasets.

Accuracy and F1 score were selected as the performance metrics. The use of F1 as a complement was selected to present the performance of each binary or multi-class classification model while avoiding biased results due to class imbalance. These metrics were reported based on the best test set performance at the end of each epoch during the models' training. In this case, the epoch when the model reached the best performance in the test set, was reported to have a notion of the time to convergence of the model. The models' training and inference times were recorded to compare the model's efficiency in terms of computational resources alongside the memory usage.

To calculate the amount of memory used during training and testing, we iterate over each modality (like 'text', 'images', 'labels') in a batch, determining the memory consumption by multiplying the total number of elements in each tensor by the size of each element and summing these values to get the total memory usage for the batch. For the model, it calculates memory usage by summing the number of elements across all model parameters (weights and biases), and then multiplying by the size of each element, accounting for the data type used (32-bit floats).

2.4.3 Reducing the Modality Embedding Gap

The phenomenon known as the "cone effect" in deep neural networks, notably described by Liang et al. [44], highlights how embeddings tend to be concentrated within a narrow region of the high-dimensional space, primarily due to the network's architecture and activation functions. This effect is particularly pronounced in the context of medical data, where the variance in medical images and texts is inherently lower compared to general datasets used for pre-training models. In this section, we provide a formal mathematical demonstration of how the cone effect, induced by random initialization, is amplified in medical datasets. These experiments were carried out for the VLM model CLIP due to its efficiency in performance and efficiency metrics, as well as its ability to extract image and text embeddings independently. To understand how the cone effect in medical data, we need to understand 3 components:

2.4.4 Increase of Cone Effects in Deeper Models & Contrastive Models

Deep learning models are composed of a set of layers, each defined by a non-linear transformation of the input vector $\mathbf{X} = (x_1, x_2, \ldots, x_{d_{\text{in}}})$; $\mathbf{X} \in \mathbb{R}^{d_{\text{in}}}$. This transformation involves a linear transformation specified by a weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, and a bias vector $\mathbf{b} \in \mathbb{R}^{d_{\text{out}}}$. The resulting linear transformation is given by 1:

$$\mathbf{Z} = \mathbf{W}\mathbf{X} + \mathbf{b} \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{d_{\text{out}}}$. Each linear transformation is followed by a non-linear activation function. One commonly used activation function is the ReLU (Rectified Linear Unit), defined as equation 2:

$$\phi(x) = \max(x, 0) \tag{2}$$

The ReLU function is applied element-wise to the vector from Equation 1. The output embedding of a layer in a neural network is thus defined as 3:

$$\text{emb} = \phi(\mathbf{W} \cdot \mathbf{X} + \mathbf{b}) \tag{3}$$

The CLIP model is trained using a contrastive learning approach where they try to minimize the cosine similarity distance between the image embeddings and text embeddings, represented using the equation as:

$$\mathbf{u} = \phi(\mathbf{W}_{\text{image}}\mathbf{X}_{\text{image}} + \mathbf{b}_{\text{image}}) \tag{4}$$

and

$$\mathbf{v} = \phi(\mathbf{W}_{\text{text}}\mathbf{X}_{\text{text}} + \mathbf{b}_{\text{text}}) \tag{5}$$

The cosine similarity of the embeddings is defined as equation 6:

$$\cos(u, v) = \frac{\phi(\mathbf{W}_{\text{image}}\mathbf{X}_{\text{image}} + \mathbf{b})^\top \phi(\mathbf{W}_{\text{text}}\mathbf{X}_{\text{text}} + \mathbf{b})}{\|\phi(\mathbf{W}_{\text{image}}\mathbf{X}_{\text{image}} + \mathbf{b})\| \|\phi(\mathbf{W}_{\text{text}}\mathbf{X}_{\text{text}} + \mathbf{b})\|} \tag{6}$$

In this case, the activation function ensures that all negative components of the vectors are set to zero, thus restricting the vectors to the positive quadrant of the n-dimensional space increasing the similarity in deeper layers as stated in 7 by Liang et al. [44].

$$\frac{\phi(\mathbf{W}_{\text{image}}\mathbf{X}_{\text{image}} + \mathbf{b})^\top \phi(\mathbf{W}_{\text{text}}\mathbf{X}_{\text{text}} + \mathbf{b})}{\|\phi(\mathbf{W}_{\text{image}}\mathbf{X}_{\text{image}} + \mathbf{b})\| \|\phi(\mathbf{W}_{\text{text}}\mathbf{X}_{\text{text}} + \mathbf{b})\|} \geq \frac{\mathbf{X}_{\text{image}}^\top \mathbf{X}_{\text{text}}}{\|\mathbf{X}_{\text{image}}\| \|\mathbf{X}_{\text{text}}\|} \tag{7}$$

Additionally, the second part of the CLIP loss is a repulsive structure that further preserves the modality gap [52].

$$-\log\left(\frac{\exp(\text{sim}(x_i, z_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_i, z_j)/\tau)}\right) = -\text{sim}(x_i, z_i)/\tau + \log \sum_{j=1}^{N} \exp(\text{sim}(x_i, z_j)/\tau) \tag{8}$$

The first term in the equation pulls the positive examples closer, whereas the second term pushes the negative examples away, effectively managing the modality gap.

2.4.5 Variance Considerations

Medical data typically exhibit lower variance σ^2_{med} in their embeddings compared to embeddings derived from general natural domain datasets σ^2_{gen} as can be see in 9. This lower variance means that medical data embeddings are more tightly clustered even before any training, making them more susceptible to the cone effect. The reason of this is that, given D, which is the set of all natural domain data and $M \subseteq D$ represent the medical domain data as a subset. This relationship is expressed as $M \subseteq D$. This effect can be demonstrated empirically when we compare the variance of the embeddings of the 3 medical datasets BRSET [46, 47] (text embedding = 5.4e-4, image embedding = 7.919e-5), HAM 10000 [48] (text embedding = 5.6e-3, image embedding = 6.8e-5), and SatelliteBench [49, 50] (text embedding = 1.7e-5, image embedding = 3.2e-3) compared with the variance of 3 non-medical benchmarks datasets: COCO-QA [53] (text embedding = 3e-3, image embedding = 1.7e-3), Fakeddit [54] (text embedding = 1.9e-3, image embedding = 8e-3), and Recipes 5K [55] (text embedding = 1.9e-3, image embedding = 8e-3). This difference can also be seen graphically in Figure 2A and Figure 2B.

$$\sigma^2_{\text{gen}} \geq \sigma^2_{\text{med}} \tag{9}$$

Additionally, as stated by Liang et al. [44], the variance of the hidden state and layer depends directly on the random initialization and the variance of the data, represented as the variance due to the model weights $\text{Weights}_{\text{variance}} = \text{VAR}[\mathbb{E}[h_\theta(\text{embed})]]$ and the variance due to data $\text{Data}_{\text{variance}} = \mathbb{E}[\text{VAR}[h_\theta(\text{embed})]]$. So the total model's variance is represented as:

$$\text{VAR}[h_\theta(\text{embed})] = \mathbb{E}[\text{VAR}[h_\theta(\text{embed})]] + \text{VAR}[\mathbb{E}[h_\theta(\text{embed})]] \tag{10}$$

The variance of an intermediate layer in medical contexts can be articulated as:

$$\text{VAR}[h_\theta(\text{embed}_{\text{medical}})] = \mathbb{E}[\text{VAR}[h_\theta(\text{embed}_{\text{medical}})]] + \text{VAR}[\mathbb{E}[h_\theta(\text{embed}_{\text{medical}})]] \tag{11}$$

This contrasts with the variance in more general contexts, where:

$$\text{VAR}[h_\theta(\text{embed}_{\text{general}})] = \mathbb{E}[\text{VAR}[h_\theta(\text{embed}_{\text{general}})]] + \text{VAR}[\mathbb{E}[h_\theta(\text{embed}_{\text{general}})]] \tag{12}$$

Finally, since we know that $\sigma^2_{\text{med}} < \sigma^2_{\text{gen}}$, we can say that:

$$\text{VAR}[h_\theta(\text{embed}_{\text{medical}})] < \text{VAR}[h_\theta(\text{embed}_{\text{general}})] \tag{13}$$

2.4.6 Embedding Alignment & Shift, Semantic Robustness

In the preceding discussions, we highlighted the complexities associated with aligning medical image-text data pairs due to the intricate nature of the data and the challenges posed by contrastive learning's deep network requirements and repulsive loss formulation. These factors contribute to a misalignment of extracted embeddings, a problem that not only persists but also intensifies during cross-modal alignment,

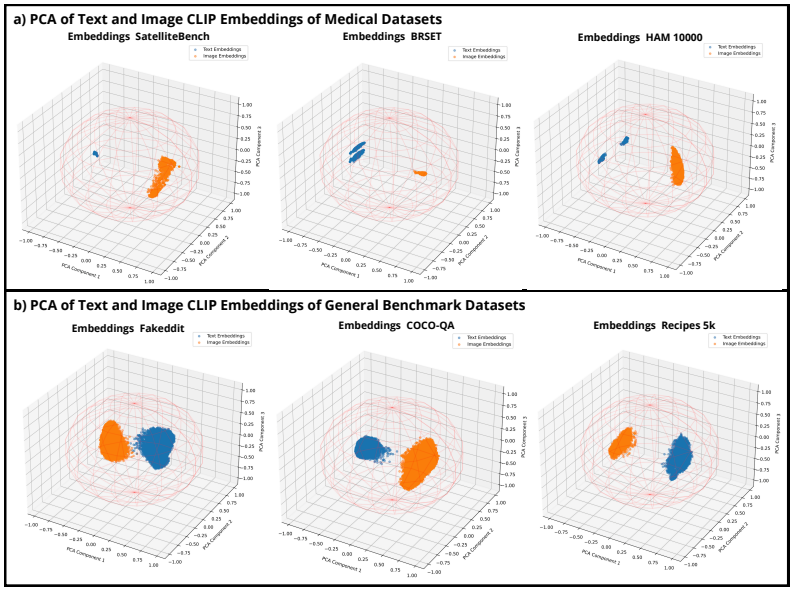

Fig. 2 Embedding modality gap between image and text embeddings for medical and non-medical datasets. (A) Represents the medical image (orange), and text (blue) embeddings generated using CLIP. (B) Represents the general image (orange), and text (blue) embeddings generated using CLIP from non-medical benchmark datasets. The embedding representations were normalized to fit inside on a unit sphere, and PCA method was used to reduce the dimensionality for visualization.

thereby undermining the robustness of the alignment process. To bridge the modality gap and bolster the semantic robustness of the embedding pairs, we propose the following approach:

1. Inject standard normal noise into the embeddings. As each embedding can be viewed as a point on the unit sphere's surface, the addition of Gaussian noise can transform the point into a small sphere. Hence, aligning two embeddings with noise forces the model to acquire the ability to align all the embeddings within the two circles and makes the semantics represented by the circle more robust than the original embedding:

$$E'_{\text{Text}} = E_{\text{Text}} + \theta_t;$$
$$E'_{\text{Image}} = E_{\text{Image}} + \theta_i.$$

   Where $\theta_t \sim \mathcal{N}(0,1)$ and $\theta_i \sim \mathcal{N}(0,1)$.
2. To further refine the cross-modal embedding alignment, we calculate the embedding gap and adjust the embeddings via a shift controlled by the hyperparameter λ, followed by renormalization to the unit hypersphere:

$$\text{Gap} = \mathbb{E}[\|E_{\text{Text}} - E_{\text{Image}}\| \mid X, Y];$$

$$E'_{\text{Text}} = E_{\text{Text}} - \frac{\lambda}{2} \times \text{Gap};$$
$$E'_{\text{Image}} = E_{\text{Image}} - \frac{\lambda}{2} \times \text{Gap}.$$

3. Moreover, we introduce an additional layer of regularization to the modality alignment process through a hyperparameter-controlled intra-modal alignment loss. This loss function is derived from the outputs of our decoupled late-fusion encoder's branch, aiming to draw paired samples closer and thereby narrow the modality gap. The regularization loss is defined as:

$$L_{\text{reg}} = \frac{1}{2N} \sum_{j} \| \| E'_{\text{Text}_j} - E'_{\text{Image}_j} \| \|_2^2 \tag{14}$$

As result of this process, Figure 3 here visually displays the original embedding representation versus the aligned embeddings.

3 Results

Our evaluation extensively demonstrates the performance and efficiency gains achieved by embedding utilization in multimodal deep learning, particularly in resource-constrained environments. Metrics measuring accuracy and computational demands underscore the advantages of embedding-based methods in low-resource scenarios.

3.1 Performance Metrics

To calculate the performance metrics of the evaluation of on the three distinct medical datasets, BRSET, HAM 10000, and SatelliteBench, different methodologies were employed to measure their effectiveness in terms of accuracy, F1-score, and convergence epoch. The approaches included embeddings using Dino v2 + Llama 2, CLIP, and direct use of raw data, each tested under early and late-joint fusion methods using our modifications. As can be seen in 1, in the BRSET dataset, the Dino v2 + Llama 2 embedding with early fusion achieved the highest accuracy of 0.987 and an F1-score of 0.944 by the fourth epoch, indicating a rapid convergence and superior model performance. This was closely followed by its late-joint counterpart, which also showed high efficacy with a slightly lower accuracy and F1-score. Similarly, the CLIP approach demonstrated robust performance with both fusion methods, though it peaked slightly later in epochs compared to Dino v2 + Llama 2. The raw data approach lagged behind the embedding methods, indicating the added value of sophisticated feature extraction techniques in handling medical datasets.

3.2 Efficiency Metrics

Table 2 provides a comprehensive analysis of the memory efficiency of the different embedding and fusion methodologies. The results delineate a contrast in memory utilization between models that utilize embeddings such as Dino v2 + Llama 2 and CLIP versus those that employ raw data directly. Notably, in the BRSET dataset,

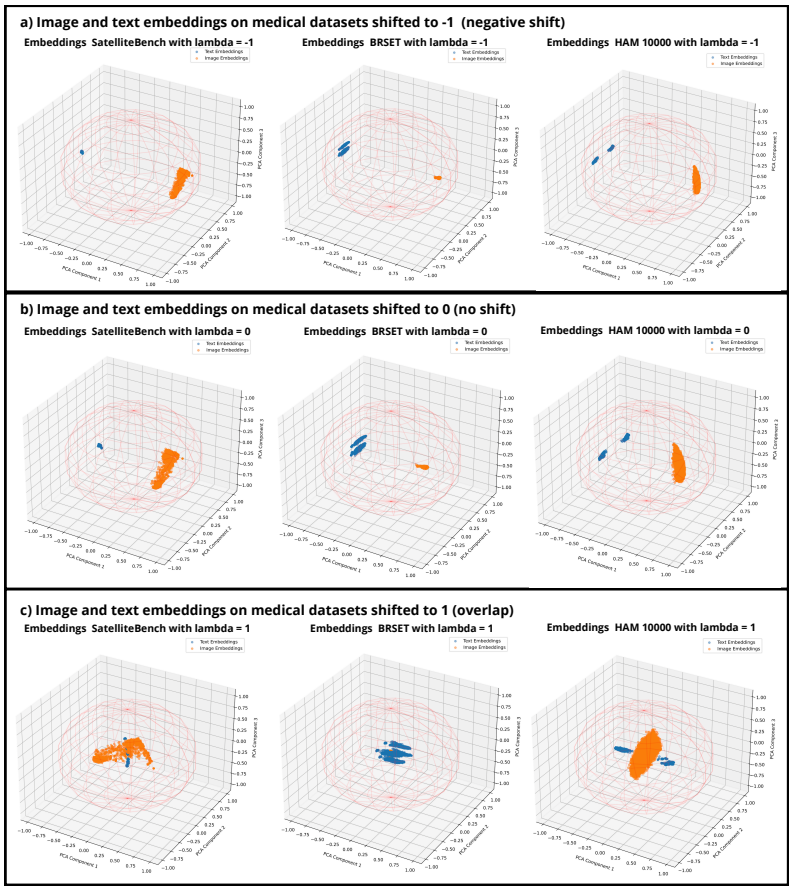

Fig. 3 Embedding alignment in the medical datasets represented as image embeddings (orange), text embeddings (blue). 3A shows the original embedding representation of each dataset with no shift. 3B Shows the embedding alignment process pooling together both embedding modalities into the same space.

the raw data approach consumed substantially more memory (approximately 747.94 MB for model size and over 7471.78 MB per epoch for training data) compared to the Dino v2 + Llama 2 and CLIP methods, which required significantly less memory (2.38 MB and 0.50 MB for model sizes respectively). Similar trends are observed in both HAM 10000 and SatelliteBench datasets where raw data approaches consistently show higher memory footprints, indicating the efficiency of embedding techniques in reducing model and data handling requirements.

Aditional to the memory and performance, the table 3 highlights the intference and training time. In the table 3 we can see how for BRSET the traditional raw data processing required significantly more time both for training (over 538 seconds) and inference (around 134 seconds) per epoch compared to the more advanced embedding techniques using Dino v2 + Llama 2 and CLIP, which drastically reduced these times (ranging from 0.95 to 1.85 seconds for training and 0.28 to 0.72 seconds for inference).

Table 1 Performance Metrics Across Medical Datasets (Max Epoch 30)

Dataset	Approach	Method	Accuracy	F1-Score	Epoch
BRSET	**Embedding Dino v2 + Llama 2**	**Early**	**0.987**	**0.944**	**4**
		Late-joint	0.984	0.929	14
	Embedding CLIP	Early	0.974	0.886	19
		Late-joint	0.975	0.885	14
	Raw Data	Early	0.952	0.760	27
		Late-joint	0.952	0.758	8
HAM 10000	Embedding Dino v2 + Llama 2	Early	0.798	0.697	28
		Late-joint	0.815	0.715	12
	Embedding CLIP	**Early**	**0.818**	**0.715**	**21**
		Late-joint	0.811	0.712	5
	Raw Data	Early	0.739	0.545	8
		Late-joint	0.743	0.617	14
SatelliteBench	**Embedding Dino v2 + Llama 2**	Early	0.752	0.751	13
		Late-joint	**0.758**	**0.758**	**18**
	Embedding CLIP	Early	0.734	0.733	30
		Late-joint	0.728	0.725	24
	Raw Data	Early	0.574	0.570	11
		Late-joint	0.571	0.565	21

Table 2 Memory Consumption Across Medical Datasets (Max Epoch 30)

Dataset	Approach	Fusion Method	Model Size	Training Dataset Size per Epoch	Test Dataset Size per Epoch
BRSET	Embedding Dino v2 + Llama 2	Early	2.38 MB	241.48 MB	15.10 MB
		Late-joint	1.19 MB	241.48 MB	15.10 MB
	Embedding CLIP	Early	0.50 MB	50.88 MB	3.18 MB
		Late-joint	0.25 MB	50.88 MB	3.18 MB
	Raw Data	Early	747.94 MB	7471.78 MB	467.13 MB
		Late-joint	747.57 MB	7471.78 MB	467.13 MB
HAM 10000	Embedding Dino v2 + Llama 2	Early	2.38 MB	148.87 MB	9.45 MB
		Late-joint	1.19 MB	148.87 MB	9.45 MB
	Embedding CLIP	Early	0.50 MB	31.51 MB	2.00 MB
		Late-joint	0.25 MB	31.51 MB	2.00 MB
	Raw Data	Early	747.95 MB	4600.85 MB	292.12 MB
		Late-joint	747.57 MB	4600.85 MB	292.12 MB
SatelliteBench	Embedding Dino v2 + Llama 2	Early	2.38 MB	17.37 MB	1.93 MB
		Late-joint	1.19 MB	17.37 MB	1.93 MB
	Embedding CLIP	Early	0.50 MB	3.66 MB	0.41 MB
		Late-joint	0.25 MB	3.66 MB	0.41 MB
	Raw Data	Early	747.94 MB	537.48 MB	59.72 MB
		Late-joint	747.57 MB	537.48 MB	59.72 MB

Similar trends are observed in the HAM 10000 and SatelliteBench datasets, where raw data approaches consistently consume more computational resources. In particular,

the SatelliteBench dataset shows the most substantial efficiency in embedding methods, especially with CLIP, achieving training times as low as 0.16 seconds and inference times around 0.09 seconds per epoch. These results underscore the effectiveness of embedding-based approaches in reducing computational load, thus enhancing the feasibility of deploying these models in real-world applications where quick processing times are crucial.

Table 3 Training and Inference Times Across Medical Datasets (Max Epoch 50)

Dataset	Approach	Fusion Method	Average Training Time Per Epoch [s]	Average Inference Time Per Epoch [s]
BRSET	Embedding Dino v2 + Llama 2	Early	1.54	0.40
		Late-joint	1.85	0.72
	Embedding CLIP	Early	0.95	0.28
		Late-joint	1.64	0.50
	Raw Data	Early	538.38	134.14
		Late-joint	543.11	132.89
HAM 10000	Embedding Dino v2 + Llama 2	Early	1.02	0.28
		Late-joint	1.20	0.42
	Embedding CLIP	Early	0.65	0.20
		Late-joint	1.12	0.42
	Raw Data	Early	260.08	64.66
		Late-joint	263.79	66.08
SatelliteBench	Embedding Dino v2 + Llama 2	Early	0.25	0.11
		Late-joint	0.28	0.13
	Embedding CLIP	Early	0.16	0.09
		Late-joint	0.22	0.12
	Raw Data	Early	28.64	9.63
		Late-joint	30.34	10.12

3.3 Embedding Alignment

The embedding alignment was tested by adding variations of the λ value to the best performing fusion models on each dataset and plotting the changes on F1-score and accuracy.

The results for BRSET using the early fusion embeddings extracted from Dino v2 and Llama 2 can be seen in figure 4, where we can see minor improvements in the F1-score from 0.944 to 0.949 using a lambda $\lambda = 0.8$, and no substantial change for accuracy.

The embedding alignment showed an increase in the model performance for SatelliteBench from an F1 score of 0.75 without data shifting, to 0.80 increasing the lambda shifting for the late fusion approach as can be seen in figure 6. Similar tendency can be seen for the F1 values in Figure 5 for the Ham 10000 dataset where the value of the clip embeddings for early fusion increased from 0.715 with no lambda shift, to 0.745.

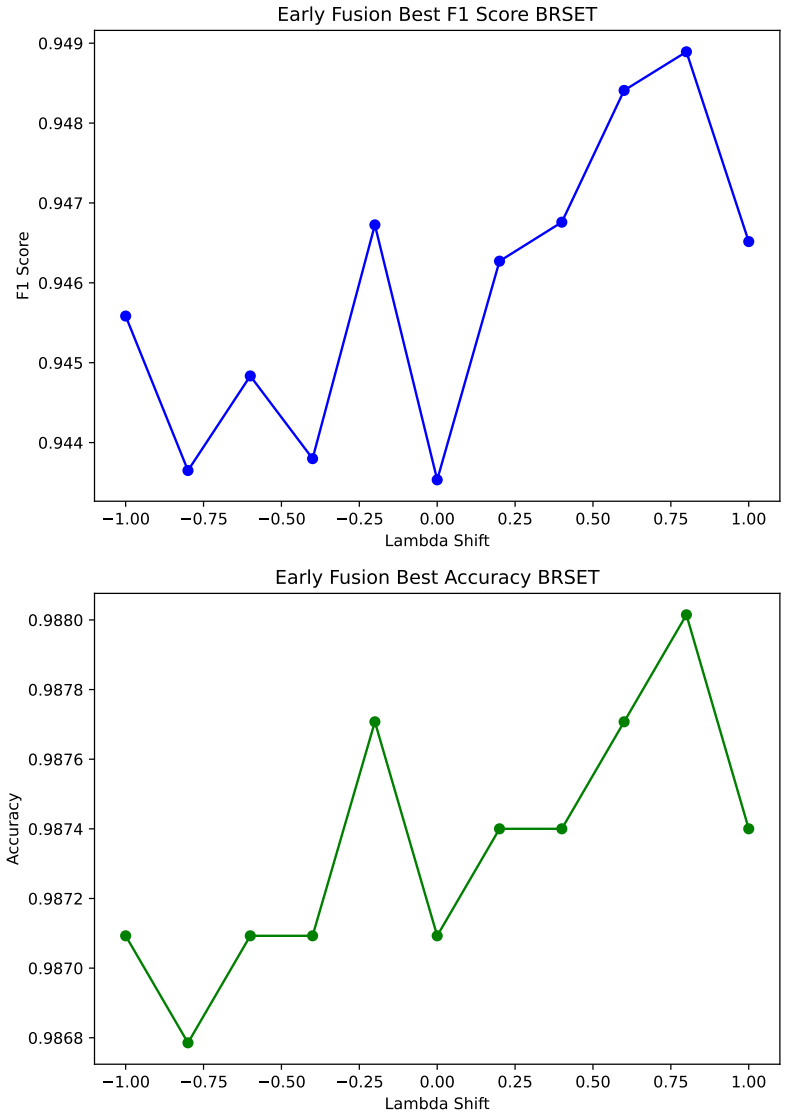

Fig. 4 Metrics calculated over shifts from negative shift -1, to positive shift 1 for BRSET Dataset.

4 Discussion

Multimodal vector embeddings present a promising avenue for computationally efficient research, particularly in low-resource settings. Our findings underscore the potential benefits of this approach, notably its simplicity and effectiveness in harnessing the power of pre-trained foundation models without the substantial computational overhead typically associated with fine-tuning or training raw data input models from scratch.

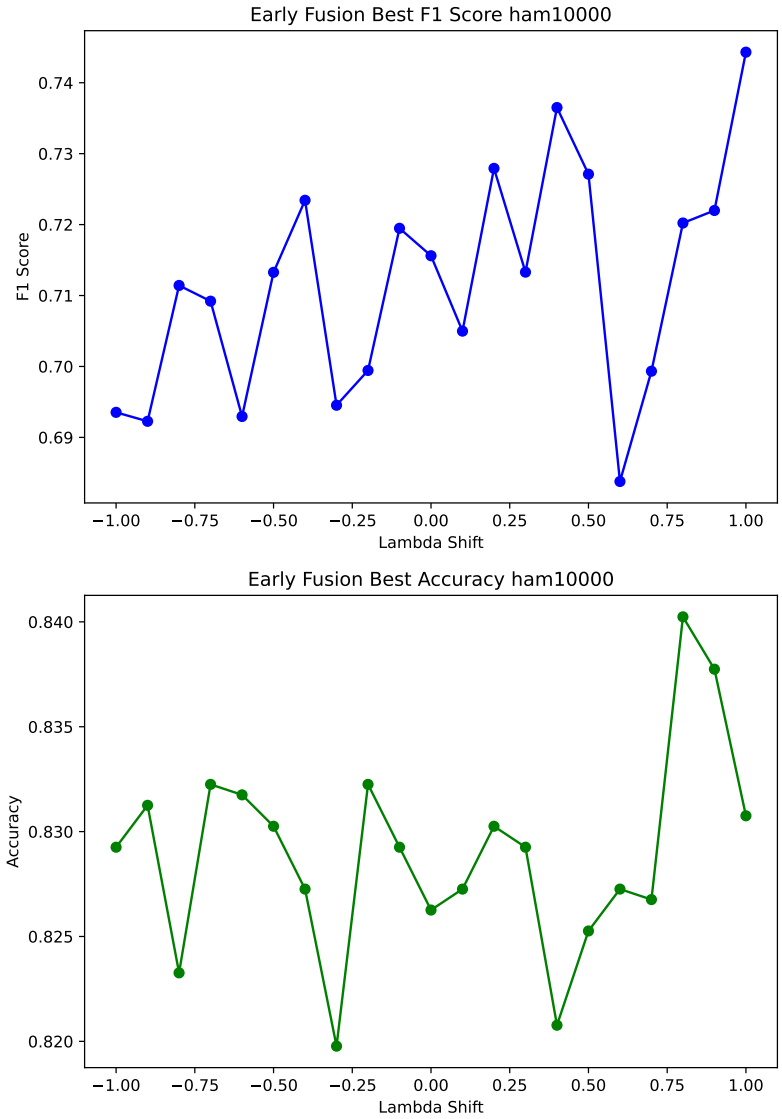

Fig. 5 Metrics calculated over shifts from negative shift -1, to positive shift 1 for HAM 10000 Dataset.

4.1 Benefits

The primary advantage of using embeddings lies in their ability to condense complex data into more manageable representations, thereby reducing the computational load and memory requirements. This is particularly beneficial in low-resource environments where computational constraints and specific expertise might limit the deployment of advanced deep-learning models. Our results indicate that embeddings can provide a rich source of pre-encoded information, enabling models to achieve competitive

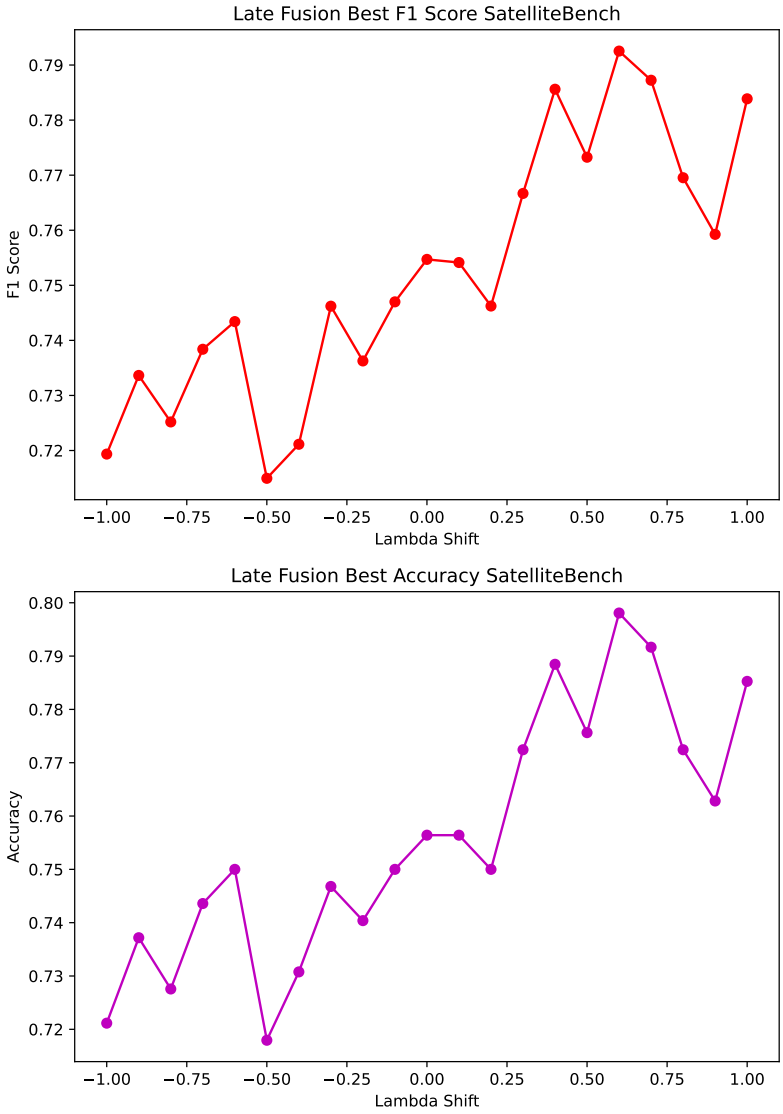

Fig. 6 Metrics calculated over shifts from negative shift -1, to positive shift 1 for SatelliteBench Dataset.

performance levels with significantly less computational demand. This is evident in the reduced training and inference times across all evaluated datasets, highlighting the approach's suitability for real-time applications and environments with limited computational capabilities.

Moreover, the simplicity of the embedding-based approach facilitates ease of implementation and adaptation to various multimodal tasks. By leveraging the generalizability of foundation models like DINO V2 [40] and LLAMA 2 [38], we can

extract high-quality embeddings that capture essential features from both images and text, enabling effective multimodal fusion without the need for extensive model customization or hyperparameter tuning.

4.2 Performance Metrics

The performance metrics—accuracy and F1-score—demonstrated that the embedding approach generally outperforms the traditional raw data approach in multimodal tasks even using pre-trained models. This superiority can be attributed to the embeddings' ability to condense complex, high-dimensional data into more manageable, semantically rich representations. These compact representations facilitate more efficient learning processes, allowing models to capture the nuances of multimodal data with fewer computational resources.

4.3 Memory Consumption

The embedding approach's reduced memory requirements underscore its computational efficiency and practical applicability in low-resource settings. This aspect is crucial for deploying advanced AI models on devices with limited memory capacity, such as mobile devices and embedded systems. Furthermore, the lower memory consumption aligns with sustainable AI practices, reducing the environmental impact associated with data storage and processing.

4.4 Training and Inference Time Insights

The significant reduction in training and inference times with the embedding approach directly impacts the practical deployment of deep learning models, especially in real-world scenarios where rapid decision-making is essential. The efficiency gains observed in the study highlight the potential for embeddings to enable advanced deep-learning models on devices with limited computational capabilities, such as mobile phones or embedded systems.

4.5 Implications

The discussion extends beyond the immediate findings to consider the broader implications for sustainable AI practices. The embedding approach's ability to deliver competitive performance with reduced computational demands aligns with the growing need for environmentally sustainable AI methodologies. By minimizing the energy and hardware requirements for training and deploying deep learning models, the embedding approach contributes to the development of more eco-friendly AI solutions.

Furthermore, the study's insights into the role of data simplicity and task complexity in model optimization processes underscore the importance of dataset selection and task design in AI research. Understanding how these factors influence model performance and resource efficiency can guide future studies in developing more effective and efficient deep learning algorithms.

4.6 Limitations

A notable challenge arises in domain-specific tasks, where the data might significantly deviate from the content typically encountered by foundation models during their training. For instance, in specialized fields such as healthcare, the images and text may encompass highly technical information that is underrepresented in the training corpora of general-purpose models like DINO V2 [40] or LLAMA 2 [38]. This can result in embeddings lacking crucial domain-specific features, leading to suboptimal performance.

While foundation model embeddings offer rich information for common data, they might miss unique characteristics in specialized datasets. Task-specific models or advanced pre-training techniques like self-supervised learning could address this, albeit with added computational costs, potentially offsetting efficiency gains in general applications.

While the use of embeddings from foundation models offers a compelling strategy for improving computational efficiency in multimodal deep learning, particularly in low-resource settings, it is not a one-size-fits-all solution. The effectiveness of this approach is contingent upon the nature of the task and the characteristics of the data involved. In domain-specific contexts where the data diverge from these norms, alternative strategies, potentially involving task-specific model training or fine-tuning, may be more appropriate. Future research should prioritize developing adaptive, domain-aware embedding strategies and exploring trade-offs between computational efficiency and task-specific performance across various application contexts.

5 Conclusion

This paper has presented a comprehensive evaluation of the use of vector embeddings extracted from foundation models for multimodal data fusion in low-resource settings, comparing it against the traditional approach of processing raw data through end-to-end models. The results highlight the potential of using aligned embeddings to significantly reduce the computational burden while retaining, and in some cases enhancing, model performance.

Our findings contribute to the ongoing discourse on sustainable AI practices by offering a viable solution for efficient computational resource utilization. By demonstrating the effectiveness of embeddings in multimodal learning, this work provides a foundation for developing more resource-efficient methodologies in AI, particularly beneficial in resource-limited environments.

However, even when this research shows promising results, further research into task-specific embeddings and advanced pre-training techniques should be carried out. Future work should explore these areas to extend the benefits of efficient multimodal learning across a broader spectrum of domains.

Acknowledgements. Random Acknowledgements

Declarations

The authors declare that they have no conflict of interest.

References

[1] Dai, Y., Yan, Z., Cheng, J., Duan, X., Wang, G.: Analysis of multimodal data fusion from an information theory perspective. Information Sciences **623**, 164–183 (2023)

[2] Pawłowski, M., Wróblewska, A., Sysko-Romańczuk, S.: Effective techniques for multimodal data fusion: A comparative analysis. Sensors **23**(5), 2381 (2023)

[3] Xu, T., Li, I., Zhan, Q., Hu, Y., Yang, H.: Research on intelligent system of multimodal deep learning in image recognition. Journal of Computing and Electronic Information Management **12**(3), 79–83 (2024)

[4] Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. Neural Computation **32**(5), 829–864 (2020)

[5] Jabeen, S., Li, X., Amin, M.S., Bourahla, O., Li, S., Jabbar, A.: A review on methods and applications in multimodal deep learning. ACM Transactions on Multimedia Computing, Communications and Applications **19**(2s), 1–41 (2023)

[6] Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M., Falk, T.H.: A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Information Fusion **76**, 355–375 (2021)

[7] Xiao, M., Li, Y., Yan, X., Gao, M., Wang, W.: Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. arXiv preprint arXiv:2404.08279 (2024)

[8] Yan, X., Wang, W., Xiao, M., Li, Y., Gao, M.: Survival prediction across diverse cancer types using neural networks. arXiv preprint arXiv:2404.08713 (2024)

[9] Shaik, T., Tao, X., Li, L., Xie, H., Velásquez, J.D.: A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom **102**, 102040 https://doi.org/10.1016/j.inffus.2023.102040

[10] Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J.: Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics **23**(2), 569 (2022)

[11] Nathan Gaw, S.Y., Gahrooei, M.R.: Multimodal data fusion for systems improvement: A review **54**(11), 1098–1116 https://doi.org/10.1080/24725854.2021.1987593 . Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24725854.2021.1987593

[12] Zhang, J., Xiao, L., Zhang, Y., Lai, J., Yang, Y.: Optimization and performance evaluation of deep learning algorithm in medical image processing. Frontiers in Computing and Intelligent Systems **7**(3), 67–71 (2024)

[13] Yuan, J., Wu, L., Gong, Y., Yu, Z., Liu, Z., He, S.: Research on intelligent

aided diagnosis system of medical image based on computer deep learning. arXiv preprint arXiv:2404.18419 (2024)

[14] Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., *et al.*: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)

[15] Shelke, A., Inamdar, M., Shah, V., Tiwari, A., Hussain, A., Chafekar, T., Mehendale, N.: Chest x-ray classification using deep learning for automated covid-19 screening. SN computer science **2**(4), 300 (2021)

[16] Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B.: More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 4340–4354 (2020)

[17] Li, M., Zhu, Z., Xu, R., Feng, Y., Xiao, L.: Research on image classification and semantic segmentation model based on convolutional neural network. Journal of Computing and Electronic Information Management **12**(3), 94–100 (2024)

[18] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J.: Deep learning in multimodal remote sensing data fusion: A comprehensive review. International Journal of Applied Earth Observation and Geoinformation **112**, 102926 (2022)

[19] Zhao, W., Liu, X., Xu, R., Xiao, L., Li, M.: E-commerce webpage recommendation scheme base on semantic mining and neural networks. Journal of Theory and Practice of Engineering Science **4**(03), 207–215 (2024)

[20] Xu, R., Yang, Y., Qiu, H., Liu, X., Zhang, J.: Research on multimodal generative adversarial networks in the framework of deep learning. Journal of Computing and Electronic Information Management **12**(3), 84–88 (2024)

[21] Restrepo, D.S., Pérez, L.E., López, D.M., Vargas-Cañas, R., Osorio-Valencia, J.S.: Multi-dimensional dataset of open data and satellite images for characterization of food security and nutrition. Frontiers in Nutrition **8**, 796082 (2022)

[22] Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G.: Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing **163**, 82–97 (2020)

[23] Cheng, L., Wang, L., Feng, R., Yan, J.: Remote sensing and social sensing data fusion for fine-resolution population mapping with a multimodel neural network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 5973–5987 (2021)

[24] Yang, J., Dung, N.T., Thach, P.N., Phong, N.T., Phu, V.D., Phu, K.D., Yen, L.M., Xuan Thy, D.B., Soltan, A.A., Thwaites, L., et al.: Generalizability assessment of ai models across hospitals: a comparative study in low-middle income and high income countries. medRxiv, 2023–11 (2023)

[25] Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D.: A survey on recent approaches for natural language processing in low-resource scenarios. arXiv preprint arXiv:2010.12309 (2020)

[26] López, D.M., Rico-Olarte, C., Blobel, B., Hullin, C.: Challenges and solutions for transforming health ecosystems in low- and middle-income countries through artificial intelligence **9** https://doi.org/10.3389/fmed.2022.958097

[27] Diab, M.: Data paucity and low resource scenarios: Challenges and opportunities. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3612–3612 (2020)

[28] Seastedt, K.P., Schwab, P., O'Brien, Z., Wakida, E., Herrera, K., Marcelo, P.G.F., Agha-Mir-Salim, L., Frigola, X.B., Ndulue, E.B., Marcelo, A., *et al.*: Global healthcare fairness: We should be sharing more, not less, data. PLOS Digital Health **1**(10), 0000102 (2022)

[29] Li, Y., Yan, X., Xiao, M., Wang, W., Zhang, F.: Investigation of creating accessibility linked data based on publicly available accessibility datasets. In: Proceedings of the 2023 13th International Conference on Communication and Network Security, pp. 77–81 (2023)

[30] Restrepo, D., Quion, J., Vásquez-Venegas, C., Villanueva, C., Anthony Celi, L., Nakayama, L.F.: A scoping review of the landscape of health-related open datasets in Latin America. Public Library of Science San Francisco, CA USA (2023)

[31] Dai, W., Tao, J., Yan, X., Feng, Z., Chen, J.: Addressing unintended bias in toxicity detection: An lstm and attention-based approach. In: 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 375–379 (2023). IEEE

[32] Restrepo, D., Quion, J.M., Do Carmo Novaes, F., Azevedo Costa, I.D., Vasquez, C., Bautista, A.N., Quiminiano, E., Lim, P.A., Mwavu, R., Celi, L.A., *et al.*: Ophthalmology optical coherence tomography databases for artificial intelligence algorithm: A review. In: Seminars in Ophthalmology, pp. 1–8 (2024). Taylor & Francis

[33] Grohe, M.: word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 1–16 (2020)

[34] Restrepo, D., Wu, C., Vásquez-Venegas, C., Nakayama, L.F., Celi, L.A., López,

D.M.: Df-dm: A foundational process model for multimodal data fusion in the artificial intelligence era. arXiv preprint arXiv:2404.12278 (2024)

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \., Polosukhin, I.: Attention is all you need **30**

[36] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding

[37] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[38] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models

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

[40] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision

[41] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[42] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR

[43] Celi, L.A., Cellini, J., Charpignon, M.-L., Dee, E.C., Dernoncourt, F., Eber, R., Mitchell, W.G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J.G., Yao, S., Data, f.M.C.: Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review **1**(3), 1–19 https://doi.org/10.1371/journal.pdig.0000022 . Publisher: Public Library of Science

[44] Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)

[45] Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., *et al.*: Sustainable ai: Environmental

implications, challenges and opportunities. Proceedings of Machine Learning and Systems **4**, 795–813 (2022)

[46] Nakayama, L.F., Goncalves, M., Zago Ribeiro, L., Santos, H., Ferraz, D., Malerbi, F., Celi, L.A., Regatieri, C.: A Brazilian multilabel ophthalmological dataset (BRSET). PhysioNet (2023)

[47] Nakayama, L.F., Restrepo, D., Matos, J., Ribeiro, L.Z., Malerbi, F.K., Celi, L.A., Regatieri, C.S.: Brset: A brazilian multilabel ophthalmological dataset of retina fundus photos. medRxiv, 2024–01 (2024)

[48] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)

[49] Kuo, K.-T., Moukheiber, D., Ordonez, S.C., Restrepo, D., Paddo, A.R., Chen, T.-Y., Moukheiber, L., Moukheiber, M., Moukheiber, S., Purkayastha, S., et al.: Denguenet: Dengue prediction using spatiotemporal satellite imagery for resource-limited countries. arXiv preprint arXiv:2401.11114 (2024)

[50] Cajas, S.A., Restrepo, D., Moukheiber, D., Kuo, K.T., Wu, C., Garcia Chicangana, D.S., Paddo, A.R., Moukheiber, M., Moukheiber, L., Moukheiber, S., Purkayastha, S., Lopez, D.M., Kuo, P.-C., Celi, L.A.: A Multi-Modal Satellite Imagery Dataset for Public Health Analysis in Colombia. PhysioNet. https://doi.org/10.13026/XR5S-XE24 . https://physionet.org/content/multimodal-satellite-data/1.0.0/ Accessed 2024-01-30

[51] Pereira, L.M., Salazar, A., Vergara, L.: A comparative analysis of early and late fusion for the multimodal two-class problem. IEEE Access (2023)

[52] Wang, Z., Zhao, Y., Huang, H., Liu, J., Yin, A., Tang, L., Li, L., Wang, Y., Zhang, Z., Zhao, Z.: Connecting multi-modal contrastive representations. Advances in Neural Information Processing Systems **36**, 22099–22114 (2023)

[53] Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Advances in neural information processing systems **28** (2015)

[54] Nakamura, K., Levy, S., Wang, W.Y.: r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854 (2019)

[55] Bolaños, M., Ferrà, A., Radeva, P.: Food ingredients recognition through multi-label learning. In: New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19, pp. 394–402 (2017). Springer