



Enhancing Model Generalization through Aggregated Small Scale Image Dataset Training and Cross Dataset Evaluation

A study of Alzheimer disease classification from structural MRI

Author

Yasir Usman

Email: uyasir85@gmail.com

Faculty of Computer Science and Information Technology
The Superior University Lahore Pakistan

Supervisor

H.M. Shahzad

Email: shahzad.dar@superior.edu.pk

Third Author

Abdullah Ahmad Arslan

Email: abdullah.ahmad.arslan125@gmail.com

Faculty of Computer Science and Information Technology
The Superior University Lahore Pakistan

Meer Usman Amjad

Faculty of Computer Science and Information Technology, The Superior University Lahore

Abstract

Small medical imaging datasets produce good performing models when tested internally; however, such models underperform on other medical imaging datasets from a different site. In our work, we explore the possibility of improving representation and transfer learning by combining two independent datasets containing Alzheimer's MRI images. The source datasets contained 86,437 and 44,000 two-dimensional MRI images belonging to four disease classes. After applying label harmonization and deduplication through the use of MD5 fingerprints, 127,437 unique images were found. Patient-wise split was conducted wherever possible followed by stratified split, normalization, and augmentation during the training process. Moreover, the DenseNet-121 architecture pretrained with ImageNet was fine-tuned on the combined datasets with weighted cross-entropy through two stages. The combined data of I-JEPA gave an accuracy of 88.10%, 89.25%, and 91.80% for Dataset 1, Dataset 2, and the combined test data of the two datasets, respectively. The obtained accuracies were better than the accuracies of the best single datasets by 5.65%, 4.30%, and 13.40%, respectively. Training the DenseNet-121 architecture with the combined data yielded an accuracy of 89.40% and 90.20% for Datasets 1 and 2, respectively. This shows that training on multiple datasets has helped in obtaining higher accuracies in this study. Since the experiments have been done only one time without having a third dataset from other sites, this shows cross-source performance improvement in this study.

Keywords Alzheimer disease; dataset aggregation; domain generalization; I-JEPA; self-supervised learning; DenseNet-121; magnetic resonance imaging

1 Introduction

But deep learning algorithms have the ability to learn discriminative features on structural magnetic resonance imaging, and the success of such approaches depends on the characteristics of the training set. In publicly available datasets associated with Alzheimer's disease, there is data scattered in different sources and imbalanced with regard to the class, preprocessing steps, scanner type, and subject. This causes source-specific feature extraction that results in overly optimistic internal performance and poor external performance [1-4].

Combining datasets is the way to address small sample size. It increases the number of examples in the dataset and makes the algorithm see a wide range of features. Combining datasets introduces many problems including domain



shift, label disagreement, image duplication, and even overlap of the same patient's slices in both training and test datasets.

In particular, the self-supervised learning approach may come in handy due to the fact that it is a representation learning algorithm which does not rely on all labels during the pre-training step. I-JEPA predicts the latent representations of the masked regions in the image based on the visible context without any image reconstruction [5]. The linear evaluation of the encoder proves that the learned representations are able to distinguish between the classes downstream. The transfer learning with the use of DenseNet-121 acts as a supervisory baseline.

The questions for this research are connected to the potential for improvement in the model performance by combining two small sets of Alzheimer MRI data sets compared to training the same sets individually and the sustainability of this improvement when tested on the separate sources. The main hypothesis is associated with the decrease in source dependency and increase in accuracy and macro F1.

1.1 Research Questions and Contributions

This research was conducted under the guidance of the following hypotheses. First, whether pre-training on a composite dataset of images will improve performance in classification for all the sources. Second, the amount of drop in performance due to the transfer of the encoder between the sources and the amount of reduction in the drop through joint pre-training. Finally, whether the performance of the supervised transfer-learning model can be sustained on the combined dataset from both sources.

In terms of experimental setup, the study has made it possible to test the research questions posed above. The experimental setup includes the following procedures: label harmonization, exact file deduplication, patient-wise split, self-supervised representation learning, frozen encoder performance evaluation, supervised transfer learning, and source-specific performance evaluation. The main value of the experimental setup lies in the ability to compare the three conditions of equivalent pre-training in a consistent way. This is due to the fact that the effect of the training data configuration will become more noticeable.

In this case, the generalization will be seen as a change in the performance of the model across the data sources and not as an intrinsic property of the model. It is important to note that in the field of medicine imaging, good performance of the model inside an internal random split may suggest something similar to the scanners, processing or patients involved.

2 Related Work

Classification work on Alzheimer's disease is now moving from manual generation of volumetric and textural features to neural network models like convolutions and transformers. Yet reproducibility analysis showed that evaluation mistakes such as data leakage, shortage of large and independent cohorts can produce grossly inflated performance [2,7]. Deep learning for Alzheimer neuroimaging received similar criticism in literature [8,9].

Transfer learning in medicine reduces the number of labeled cases by using features trained on big image collections. The DenseNet model was applied to classify Alzheimer patients as well as to other tasks because of skip connections ensuring feature reutilization and stable gradients [6,10]. Pretrained features do not compensate for dataset shift, therefore holding source-specific test set is vital.

The self-supervised learning methods like contrastive learning, masked autoencoders, and joint-embedding prediction have been found effective in making label utilization more efficient in computer vision and medical imaging applications [5,11–14]. In the review of medical imaging literature, it has been said that improvements in performance are dependent on the nature of task, scale of pre-training, and the way evaluation process has been designed [13,14].

2.1 Dataset Aggregation and Domain Shift

The combining of data sets is not only about increasing sample size. Every sample will be equipped with certain recognizable features like the resolution of images, scaling of intensity, field of view, compression, cropping, and



preprocessing. The demographic makeup and diagnostic technique used will contribute further variability in the relation of the images and labels. A network with enough capacity may leverage the existing differences for each class of image labels without biological connection.

Aggregation can complicate the employment of source-specific shortcut tricks, since the conflicting acquisition methods will be aggregated into one optimization problem. In case of similar diagnostic classes in various samples, there will be no reason for the model to associate the class with the visual cues of one dataset alone. Nevertheless, these things cannot be expected to happen, as label noise due to the poor choice of labeling, false augmentation of sample sizes by repeated files, and unbalanced sources favoring the bigger one might occur.

2.2 Class Imbalance in Alzheimer Staging

The problem of class imbalance is extremely relevant to the task of dividing Alzheimer into four stages, because the number of scans of people who are healthy might be higher than the number of scans of people with an advanced stage of Alzheimer. This means that an algorithm trained on the accuracy metric will work great and give a good result by considering the majority class.

Dataset 2 helped to balance the very unbalanced Dataset 1 through the addition of 11,000 images to each class. Even though the resulting dataset was unbalanced, there were way more examples of Mild Demented and Moderate Demented. The weighted cross-entropy was used, as numerical balance did not necessarily mean the same learning balance.

3 Materials and Methods

3.1 Datasets and Labels

The experiments were conducted using the OASIS Alzheimer MRI dataset available on Kaggle as Dataset 1 and the balanced Alzheimer MRI multi-class dataset as Dataset 2. In these datasets, there were 2D images classified into one of the four classes, which are Non Demented, Very Mild Demented, Mild Demented, or Moderate Demented. The first dataset was found to be highly imbalanced whereas the second dataset contained 11,000 images in each class. There were 130,437 images from the mentioned source in total, out of which duplicates were removed resulting in 127,437 images.

Table 1 Reported image distribution after cross-source consolidation

Disease stage	Dataset 1	Dataset 2	Unique combined
Non Demented	67,200	11,000	75,840
Very Mild Demented	13,725	11,000	24,445
Mild Demented	5,002	11,000	15,742
Moderate Demented	510	11,000	11,410
Total	86,437	44,000	127,437

The totals for the class numbers in Table 1 are the numbers referred to in the source thesis document after consolidation. The totals are lower than the total of the two source documents since duplicates have been removed. The largest percentage change is the change in the Moderate Demented class, which has increased from 510 images to 11,410 images.

3.2 Preprocessing and Data Integrity

Class naming took place prior to the comparison of the databases. Duplicate files were identified on the basis of their MD5 hashes and removed. Splitting for class proportion balancing at the level of patients was carried out in



order to guarantee that no slice of an image from the same patient falls into different partitions; then stratification was carried out. All the images were rescaled to the size of 224×224 pixels and normalized. Training images were augmented through rotation, flipping, and intensity adjustment.

Such a sequence of actions was intentional. Harmonization of labeling preceded deduplication because it was required to analyze the duplicate images under one terminology of classes. Deduplication preceded splitting because otherwise duplicates could end up in different partitions. Grouping by patients preceded stratification because stratification at the image level would not have been sufficient for avoiding putting of the duplicate images from the same patient into training and test sets. Parameters of normalization and augmentation were calculated after partitioning.

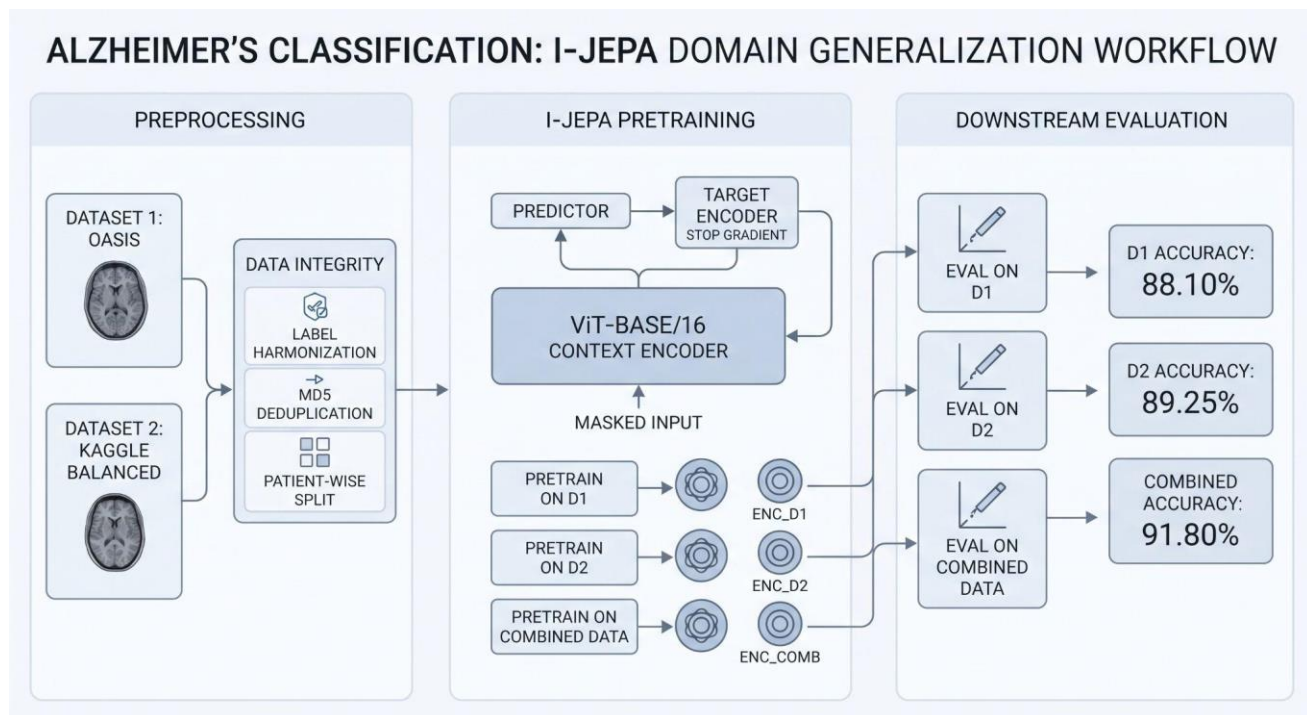


Figure 3.1: Alzheimer Classification Pipeline

MD5 hashing looks for duplicates on a byte level. MD5 hashing will not detect duplicates if an image has been manipulated by scaling, contrast, compression, or even cropped. Future auditing should include both perceptual and exact hashing as well as metadata checking. This is especially important where one publicly released data set might have been created from another data set and repackaged in another folder structure.

3.3 Self-Supervised Pretraining and Linear Evaluation

ViT-Base/16 was used as the backbone architecture for all I-JEPA designs. The input image with a size of 224 × 224 was split into 196 patches with no overlap. The four targets blocks covered 15-20% of the image while the context area skipped a single block which was 35-45% of the patches. Four attention heads and 384-dimensional hidden layers were used for the predictor. The target encoder was updated by averaging the context encoder using exponential moving average with momentum from 0.996 to 0.999. Five epochs of training were carried out using AdamW optimizer with learning rate of 1.5×10^{-4} , weight decay of 0.05 and batch size of 64.



The context encoder was frozen afterward and linear classification task was done for ten epochs using Adam optimizer with learning rate of 1×10^{-3} . The three encoders were pre-trained separately using Dataset 1, Dataset 2 and the combined datasets. All encoders were then evaluated on Dataset 1, Dataset 2 and the combined dataset. The I-JEPA objective seeks to reduce the average squared difference between the predictor output and stop-gradient target encoder representation for a batch size of B images in T, which is the set of masked targets.

$$L_{ijepa} = (1 / (|B| \cdot |T|)) \cdot \sum_i \sum_t \| p_{\phi}(f_{\theta}(x_{i, context}), t) - \text{sg}(f_{\theta}(x_i)) \|_2^2$$

Equation 1 I-JEPA latent prediction loss

The parameters of the target encoder, θ , do not undergo update during backpropagation. Instead, the parameters of the target encoder are tracked through an exponential moving average using the momentum factor m.

$$\bar{\theta}_k \leftarrow m_k \cdot \bar{\theta}_{k-1} + (1 - m_k) \cdot \theta_k, \quad m_k \in [0.996, 0.999]$$

Equation 2 Exponential moving average target update

By freezing the encoders, the quality of the representations becomes independent of its adaptation to the tasks. In case when all the weights of the encoders are fine-tuned during the test phase, even the poor representation can produce a good score through its task adaptation. Linear classifier is deliberately restricted, because it asks whether the representation is ready to arrange the data points such that there is a four-way separation boundary.

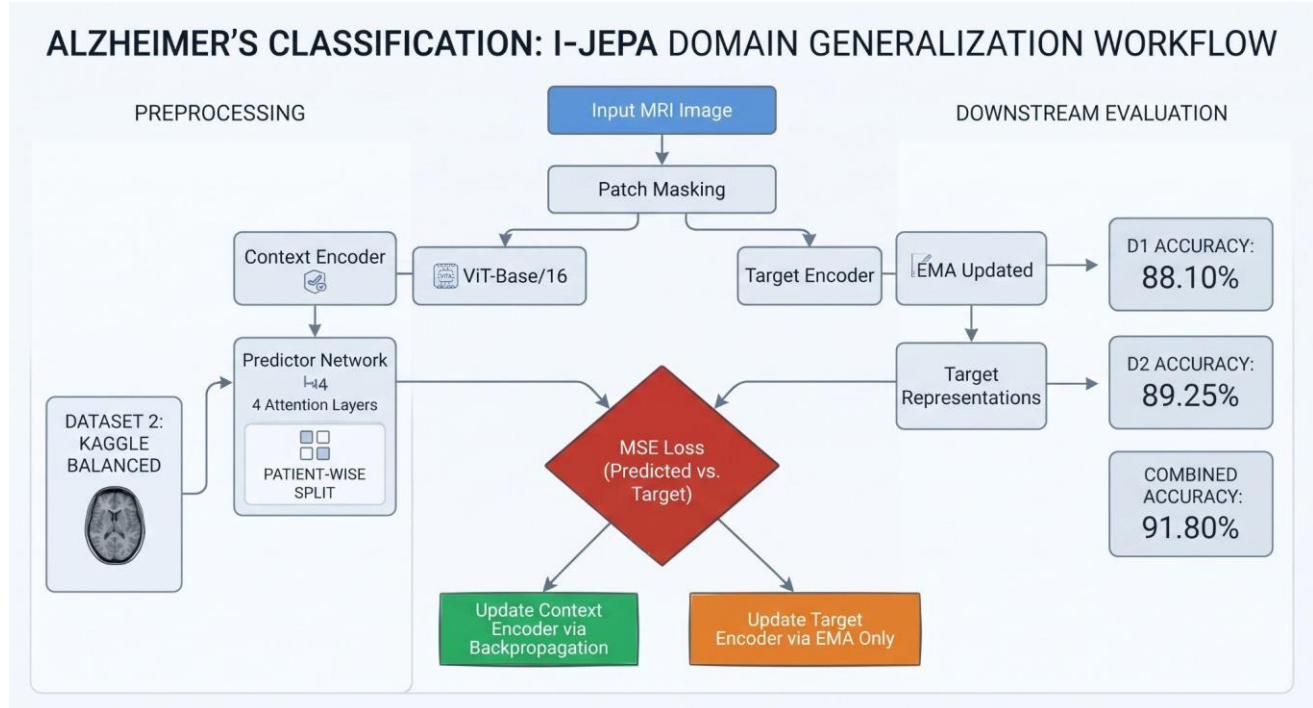


Figure 3.2: I-JEPA Self-Supervised Learning Architecture

3.4 DenseNet Transfer Learning

Pre-trained DenseNet-121 using ImageNet weights was employed to train the concatenated dataset. During epoch 1-5, the convolution layers were frozen and Adam was used as the optimizer for training the network with a learning



rate of 1×10^{-3} . From epoch 6 until epoch 45, all layers were unfrozen and trained with a learning rate of 1×10^{-5} . The factors of ReduceLROnPlateau were 0.5 and patience of 5, whereas for early stopping, the factor was 10. DenseNet-121 was trained using weighted cross-entropy loss. In the above equation, z represents the logits of the class, n_c is the number of training samples of class c , C is the number of classes, and N is the total number of training samples.

$$L_{WCE} = - (1 / B) \cdot \sum_i w_{y_i} \cdot \log [\exp(z_{i, y_i}) / \sum_c \exp(z_{i, c})] \quad \text{and} \quad w_c = N / (C \cdot n_c)$$

Equation 3 Weighted cross entropy and inverse frequency class weight

This two-step process provides a compromise between stability and adapting to a domain. Classification step is responsible for giving enough time to the newly added output layer to create a separate decision surface without interfering with the pretrained convolutional filters right away. Fine-tuning the entire network in the second step makes sure that the filters adapt to gray-scale anatomy of the brain.

3.5 Experimental Comparisons

Table 2 Experimental configurations and their analytical purpose

Configuration	Training data	Model stage	Evaluation targets	Purpose
A	Dataset 1	I-JEPA probe and linear	Dataset 1, Dataset 2, Combined	Measure in-source learning and transfer from Dataset 1
B	Dataset 2	I-JEPA probe and linear	Dataset 1, Dataset 2, Combined	Measure in-source learning and transfer from Dataset 2
C	Combined data	I-JEPA probe and linear	Dataset 1, Dataset 2, Combined	Test whether aggregation improves representation transfer
D	Combined data	DenseNet-121 fine-tuning	Dataset 1 and Dataset 2	Provide a supervised transfer-learning benchmark

The hyperparameters in the I-JEPA experiments were fixed in such a way that the intended difference between the two sets of experiments was the pre-training data. The DenseNet experiment represents another type of learning and acts as a point of contrast rather than a competitor to the architectures under study. Given that the source text speaks only about the combined-data DenseNet experiment results, the causal analysis is most relevant for the three I-JEPA pretraining experiments.

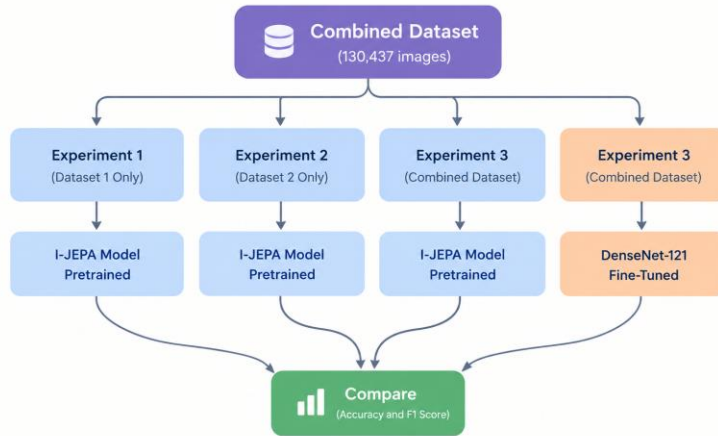


Figure 3.3: Overall Experimental Workflow

3.6 Evaluation

The evaluation of the performance of the models was done using performance measures like accuracy, macro-precision, macro-recall, macro F1-score, and One-vs-Rest ROC-AUC where applicable. Testing partitions were kept for each domain separately to test for transfer learning between both domains. Absolute improvement in accuracy was calculated by subtracting the highest accuracy of any model trained on one domain from the combined-pretrained accuracy.

Accuracy is defined as the proportion of correctly classified samples; but accuracy is extremely dependent on the majority class, which is Non Demented. Precision (macro) is the average of all positive predictive value in all classes, while recall (macro) is the average of all sensitivity in four stages. F1 score (macro) computes the harmonic mean of precision and recall for all classes before averaging them, so that bad performance in any of the minority stage penalizes the classification algorithm. ROC-AUC evaluates ranking performance across different decision thresholds. Confidence interval is not presented in this experiment.

TP, FP and FN stand for true positives, false positives and false negatives, respectively, for each class c .

$$\text{Accuracy} = \sum_c \text{TP}_c / N, \quad \text{P}_c = \text{TP}_c / (\text{TP}_c + \text{FP}_c), \quad \text{R}_c = \text{TP}_c / (\text{TP}_c + \text{FN}_c)$$

Equation 4 Accuracy precision and recall

Macro F1 gives each diagnostic class equal weight, making it more informative than accuracy when the class distribution is imbalanced.

$$\text{F1}_c = 2 \cdot \text{P}_c \cdot \text{R}_c / (\text{P}_c + \text{R}_c), \quad \text{Macro F1} = (1 / C) \cdot \sum_c \text{F1}_c$$

Equation 5 Per class F1 and macro F1

4 Results

4.1 I-JEPA Pretraining

The pre-training loss is generally lower in all cases. Loss after training ended was 0.0290 for Dataset 1, 0.0260 for Dataset 2, and 0.0190 for the combined datasets. It is clear that the model has been trained using images which are distributed differently; therefore, the above losses should be viewed in light of the following test results.

Table 3 Average I JEPA pretraining loss by epoch



Graph 4.1: I-JEPA Pretraining Loss Curves

Training data	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5	Time
Dataset 1	0.0520	0.0410	0.0350	0.0310	0.0290	21.31 min
Dataset 2	0.0480	0.0380	0.0320	0.0280	0.0260	9.57 min
Combined	0.0450	0.0320	0.0250	0.0210	0.0190	29.55 min

4.2 Linear Evaluation

The single source encoders worked well with the data set they were pre-trained on, while performing poorly on the other data set. The performance of the Dataset 1 encoder fell from 82.45% on Dataset 1 to 68.30% on Dataset 2. On the other hand, the performance of the Dataset 2 encoder was 84.95% on Dataset 2 and 71.15% on Dataset 1.

Table 4 Linear probe performance across pretraining and test sources

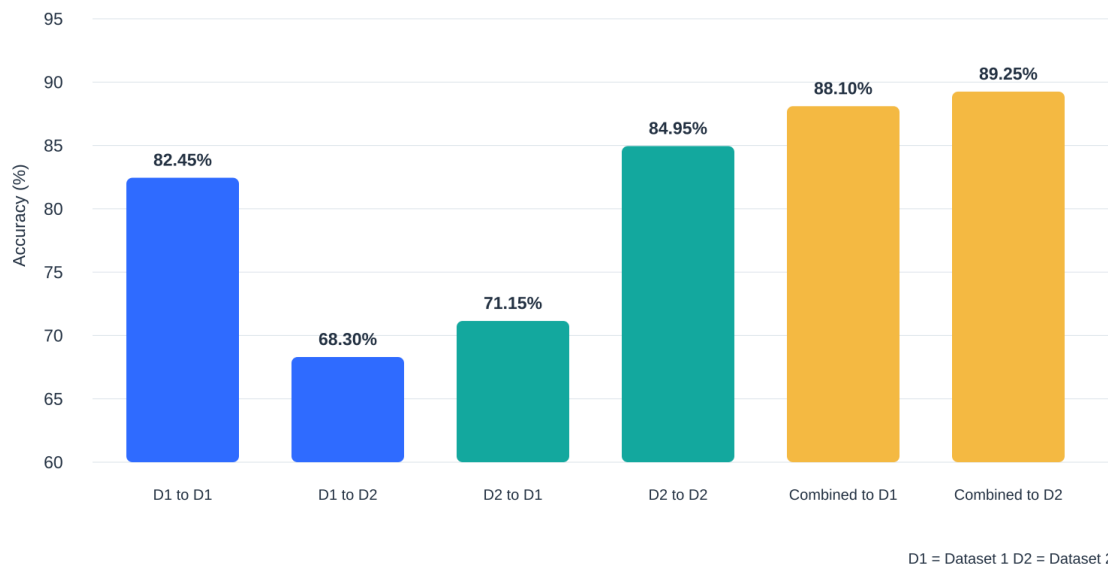
Pretraining data	Test data	Accuracy	Precision	Recall	Macro F1	ROC AUC
Dataset 1	Dataset 1	82.45%	0.8210	0.8185	0.8197	0.861
Dataset 1	Dataset 2	68.30%	0.6795	0.6750	0.6772	0.731
Dataset 2	Dataset 1	71.15%	0.7080	0.7045	0.7062	0.754
Dataset 2	Dataset 2	84.95%	0.8450	0.8490	0.8465	0.879
Combined	Dataset 1	88.10%	0.8790	0.8765	0.8777	0.913
Combined	Dataset 2	89.25%	0.8895	0.8870	0.8882	0.924



Pretraining data	Test data	Accuracy	Precision	Recall	Macro F1	ROC AUC
Combined	Combined	91.80%	0.9150	0.9125	0.9137	0.941

In comparison with the most accurate one-source encoder, the improvements in terms of percentage points in accuracy were attained by 5.65 for Dataset 1 and 4.30 for Dataset 2. For the merged testing dataset, the improvement in accuracy in terms of percentage points is from 78.40% to 91.80%, which is an improvement of 13.40%.

Linear Evaluation Accuracy by Pretraining Source



Graph 4.2: Linear Evaluation Accuracy by Pretraining Configuration



I-JEPA Confusion Matrix on Dataset 1

		Non Demented	Very Mild	Mild	Moderate
Actual class	Non Demented	12,195 95.0%	480 3.7%	150 1.2%	15 0.1%
	Very Mild	220 8.0%	2,427 88.4%	91 3.3%	7 0.3%
	Mild	38 3.8%	97 9.7%	860 85.9%	6 0.6%
	Moderate	2 2.0%	5 4.9%	12 11.8%	83 81.4%
		Predicted class			

Cells show count and row percentage

Figure 4.4: Confusion Matrix for I-JEPA Evaluated on Dataset 1

4.3 Cross-Source Performance Degradation

Table 5 Accuracy change when single-source encoders cross datasets



Cross-Dataset Accuracy Decline



Graph 4.3: Cross-Dataset Performance Degradation

Pretraining source	In-source accuracy	Other-source accuracy	Absolute decline
Dataset 1	82.45%	68.30%	14.15 pp
Dataset 2	84.95%	71.15%	13.80 pp

The transfer penalty was the same for both single-source encoders, regardless of different sample sizes and distributions of classes. It is safe to assume that the transfer penalty could not be caused by the relatively small size of Dataset 2 only. What is more likely is that each of the two encoders uses particular properties of its respective source that do not get fully transferred. As for the combined encoder, there was no in-source comparison because both of the sources served as pretraining data; however, the performance decline was significantly smaller at 1.15 points.

4.4 DenseNet Results

Model DenseNet-121 that was trained using the merged dataset achieved accuracy and macro F1 scores of 89.40% and 0.889 respectively for Dataset 1 while on Dataset

Table 6 DenseNet 121 performance after combined dataset training



DenseNet-121 Confusion Matrix on Dataset 1

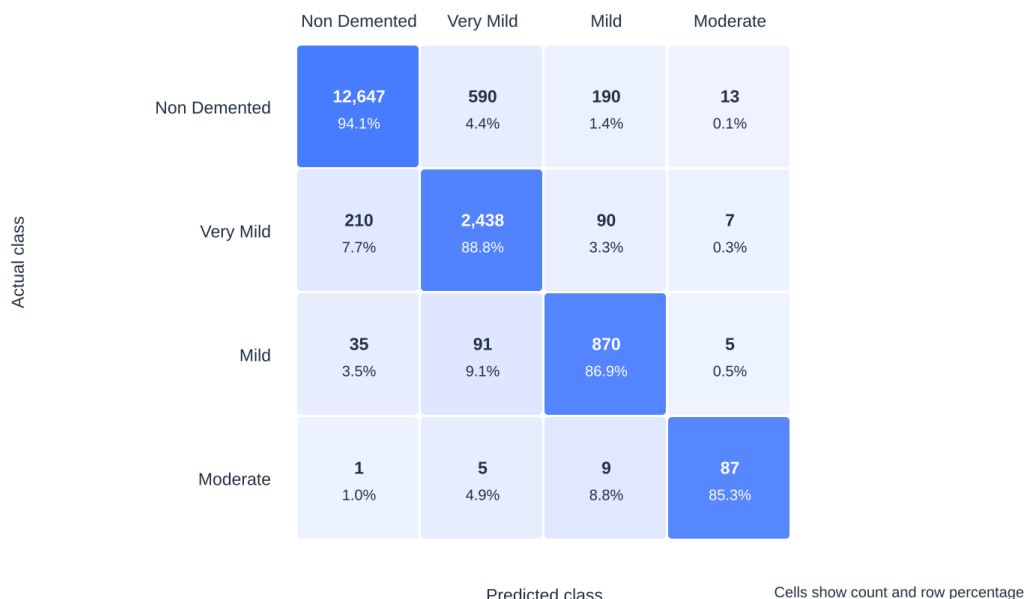


Figure 4.5: Confusion Matrix for DenseNet-121 Evaluated on Dataset 1

Test data	Accuracy	Precision	Recall	Macro F1	ROC AUC
Dataset 1	89.40%	0.891	0.888	0.889	0.902
Dataset 2	90.20%	0.898	0.895	0.896	0.911

5 Discussion

These results confirm the hypotheses suggested for the two considered datasets. Training with only one dataset produced an in-domain advantage, resulting in a drop in performance by 13.80-14.15 percentage points in the second dataset. The joint training balanced the equation: the unified model scored 88.10% for Dataset 1 and 89.25% for Dataset 2. The reduced gap of 1.15 percentage points suggests that there is less correlation with any particular distribution of a dataset.

There are three potential explanations for this performance gain. First, via the aggregation process, it becomes evident that there are more varied images available for representation learning. Second, the second dataset was considerably more balanced and had a lot of images in the minority class, which is Moderate Demented. Third, through label harmonization, deduplication, and splitting at the patient level, we have prevented the primary sources of leakage.

The benchmark that was carried out using supervision recorded comparable results in both the source-specific test partitions. This is an encouraging fact, but it cannot be referred to as near perfect or anything close to an indication of the validity of the results obtained clinically. The results recorded in terms of accuracy in the study were about 89-90%. The two test partitions were from the source-specific datasets used for aggregation in the training datasets. From the results, it has been proven the reason why just relying on accuracy is not sufficient. In dataset 1, there were 510 images of Moderate Demented people before aggregation. As per the original thesis, the use of class



weights reduced the F1 score of the class from 0.858 to about 0.614 despite the insensitivity of the aggregated models' performance.

5.1 Interpretation of the Self-Supervised Results

However, the combined-data encoder showed the smallest loss and best linear probe results during pretraining. The downstream result is much more crucial than the difference in loss because self-supervised loss is tailored specifically for the representation created by the encoder in each particular distribution in training. This does not mean that two encoders, being individually trained, can be compared using a common semantic metric based on a smaller loss.

The largest increase in the performance was seen in the case of the combined test set. The reason for such results is the fact that the combined test set represents the most varied representation of the patterns pre-trained. What matters more are the generalizable findings, that is, those that are obtained from the tests in each individual source, which means an increase of 5.65 percentage points in the performance of Dataset 1 and 4.30 points for Dataset 2.

5.2 Implications for Medical Imaging Research

The findings suggest a reasonable method of data creation for those laboratories which cannot find means to create a new large set of patients. Proper merging of available small datasets can be beneficial for modeling under the condition that these datasets share the same clinical application. In addition to merely joining images into one folder, a team needs a proper class ontology, removal of duplicates, patients' partitioning method, sources identification and an evaluation partitioning scheme, maintaining evaluation possibility per each source.

A source-oriented evaluation should be used together with the overall metrics. A single overall metric can cover a case of a model which performs poorly everywhere but at the prevalent source. Separate evaluation of sources can reveal the domain specificity and point out at places which need more data or harmonization. The same applies to class-based metrics since one good overall score may appear together with a poor detection of rare disease stages.

However, for its use in clinical settings, a much more compelling chain of evidence would be required. The assessment would involve an untainted institution, calibration, stratification, error analysis, and comparison against clinically relevant benchmarks. In the case of prospective validation, the difference between public images curated by experts and the entire pipeline of MRI processing should be considered.

6 Limitations

Five main limitations can be mentioned about this research. Firstly, each configuration was launched only once, there is no variation in multiple runs or seed values, and confidence intervals were not provided. Secondly, the I-JEPA pretraining was conducted for just five epochs. Thirdly, all datasets were made of two-dimensional slices rather than whole volumes of three-dimensional images. Fourthly, no clinical covariates such as age, cognitive score, genotype, or acquisition parameters have been provided. The most crucial limitation was the lack of a third institutional dataset that would not be affected by the experiments at all.

Very careful calculations are needed in order to determine the totals of the provided datasets. The total number of images provided by sources is 130,437, but the consolidated total equals 127,437, as 3,000 images were excluded due to duplication. In the future, the retain file list, patient identifiers, grouping, version of datasets, and splits should be provided.

7 Future Research

In future study, all scenarios should be replicated using different random seeds in order to calculate means, standard deviations and bootstrapped confidence intervals. It can provide possibilities to apply McNemar test for comparison of differences in accuracy using paired predictions on identical testing data points. Analysis plan distinguishing between confirmatory and exploratory results will prevent from misinterpreting of single run result.



Third dataset provided by an institution should be kept apart from other datasets until pipeline configuration and hyperparameters tuning are completed. Leave-one-source out cross validation based on several data sources will evaluate generalization ability of the model in the case of training on several sources and testing on new source. Information regarding scanner manufacturer, magnetic field strength, acquisition protocol, age, gender, diagnostic criteria should be collected wherever possible.

Future research on this topic will need to progress from slice-level two-dimensional modeling to three-dimensional patient-level modeling or combination of many slices into one patient-level data. Application of patient-level modeling is more natural in a clinical context and does not put too much emphasis on one patient with many slices. Evaluation of our model against ResNet, EfficientNet, masked autoencoder, and domain generalization models will show whether aggregation is useful across other model families as well.

Interpretability and calibration of the model will need to be evaluated. Saliency maps or attention weights could be used to test the plausibility of the neuroanatomy utilized in the prediction. Yet, the interpretability of visual explanation needs to be validated. Calibration plots, expected calibration error, and decision curve analysis would be helpful for assessing the appropriateness of predicted probability in communication of risks.

8 Conclusion

In terms of the performance of the I-JEPA model using the aggregated two MRI Alzheimer datasets, it outperformed the best performing individual source encoder for Datasets 1, 2, and the aggregated test set, whereas DenseNet-121 attained approximately 90% accuracy for the individual source datasets. Therefore, this calls for further exploration of the possibility of applying dataset aggregation as a possible solution for small medical imaging datasets, which will include multiple iterations, confidence intervals, split manifest, three-dimensional modeling of patients and testing by a third party not involved in training data.

References

- [1] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies OASIS. *Journal of Cognitive Neuroscience*, 19(9), 1498–1507. <https://doi.org/10.1162/jocn.2007.19.9.1498>
- [2] Wen, J., Thibeu-Sutre, E., Diaz-Melo, M., et al. (2020). Convolutional neural networks for classification of Alzheimer disease. *Medical Image Analysis*, 63, 101694. <https://doi.org/10.1016/j.media.2020.101694>
- [3] Yu, A. C., Mohajer, B., and Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis. *Radiology Artificial Intelligence*, 4(3), e210064. <https://doi.org/10.1148/ryai.210064>
- [4] El Jurdi, R., Petitjean, C., Honeine, P., Abdallah, F., and Cheplygina, V. (2024). A systematic review of generalization research in medical image classification. *Computers in Biology and Medicine*, 183, 109256. <https://doi.org/10.1016/j.combiomed.2024.109256>
- [5] Assran, M., Duval, Q., Misra, I., et al. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629. <https://doi.org/10.1109/CVPR52729.2023.01498>
- [6] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>
- [7] Willemink, M. J., Koszek, W. A., Hardell, C., et al. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4–15. <https://doi.org/10.1148/radiol.2020192224>
- [8] Alsubaie, M. G., Luo, S., and Shaikat, K. (2024). Alzheimer disease detection using deep learning on neuroimaging. *Machine Learning and Knowledge Extraction*, 6(1), 464–505. <https://doi.org/10.3390/make6010024>
- [9] Menagadevi, M., Devaraj, S., Madian, N., and Thiyagarajan, D. (2024). Machine and deep learning approaches for Alzheimer disease detection using magnetic resonance images. *Measurement*, 226, 114100. <https://doi.org/10.1016/j.measurement.2023.114100>
- [10] Ahmed, M. R., Zhang, Y., Feng, Z., Lo, B., Inan, O. T., and Liao, H. (2023). An Alzheimer disease classification model using transfer learning DenseNet. *Decision Analytics Journal*, 9, 100348. <https://doi.org/10.1016/j.dajour.2023.100348>
- [11] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597–1607.



- [12] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition, 16000–16009. <https://doi.org/10.1109/CVPR52688.2022.01553>
- [13] Huang, S. C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., and Chaudhari, A. S. (2023). Self-supervised learning for medical image classification. npj Digital Medicine, 6, 74. <https://doi.org/10.1038/s41746-023-00811-0>
- [14] Zhang, C., Zheng, H., and Gu, Y. (2023). Dive into the details of self-supervised learning for medical image analysis. Medical Image Analysis, 89, 102879. <https://doi.org/10.1016/j.media.2023.102879>