Disentangling Detail from Structure in Brain MRIs 2

Daniela Layer

Introduction

Variational Autoencoders (VAEs) are a tool for unsupervised representation learning with wide applications in biomedical research [1, 2]. Introduced by Kingma and Welling (2013) [1], VAEs consist of an encoder and decoder working together, where the encoder links the input data x to a latent representation z and the decoder converts z back to the original input space [3]. In biomedical information, VAEs are primarily used for data augmentation and representation learning in fields such as molecular design, sequence data set analysis, and medical imaging. Specific uses in medical imaging include image classification, segmentation, restoration, and reconstruction [2].

However, VAEs tend to focus on the detection of low-frequency features, a phenomenon called spectral bias [4]. This characteristic poses challenges in detecting subtle pathological changes, particularly in early-stage disease detection where high-frequency variations are crucial [5].

In the process called disentanglement, complex data is broken down into features that are independently meaningful [6], while the factors in the raw data appear to be tightly intertwined. The challenge here is to construct the representation of the data in such a way that they provide features that are suitable for a variety of possible tasks and can deal with the reality of the intertwined variation factors [7].

In our previous work, we explored whether integrating brain segmentation mask as additional input to VAEs could improve the model's ability to learn more detailed features. Using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, we implemented and compared three VAE architectures, integrating brain segmentation masks in different ways, and evaluated their performance on a dementia prediction downstream task. The detailed report for this previous work can be found in the attachment.

Alzheimer disease is the most common form of dementia. The International Classification of Diseases 11 (ICD-11) defines dementia as the presence of significant impairment in one or more cognitive domains. The impairment is related to age and the expected level of cognitive function and represents a deterioration from the previous functional level. Depending on the extent of neurocognitive and functional impairment, the severity of dementia can be categorized as mild, moderate or sever [8].

Despite promising results, we identified overfitting as an issue, which limited the generalizability of the models to unseen data. Building on this foundation, we extended the project to address this issue and explored additional downstream tasks. After setting aside a test set, we employed k-fold cross-validation for hyperparameter tuning, and expanded the project to include UK Biobank (UKBB) data to predict several additional traits. The UKBB is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants, with plans to conduct brain imaging on 100,000 participants [9]. We focused on predicting several traits known to correlate with brain structure: diastolic blood pressure, systolic blood pressure, body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL) cholesterol, testosterone, and pulse rate [10–12].

By comparing results across different VAE architectures and downstream prediction models, we aimed to evaluate both accuracy and robustness using multiple seeds to ensure reliable performance.

Research Question

The original research question asked whether integrating brain segmentation masks into VAEs could improve feature representation in medical imaging, specifically for dementia prediction. In this extended project, we aim to further explore whether we can improve results not only in dementia prediction but also in other medical imaging tasks such as predicting multiple additional traits from UKBB.

Solution and Implementation[1]

Data Preparation For the ADNI dataset, the available imaging data was filtered for T1-MPRAGE scans and skull-stripping was performed using HD-BET [13] tool. With a 1mm resolution a non-linear registration was conducted on the MNI152 templates using FLRT and FNIRT commands of the FSL tool [14]. Finally the segmentation masks of brain regions were extracted using the Synthseg software [15]. For the UKBB dataset, we utilized MNI-registered brain scans that were pre-processed by the UKBB team. These scans were already registered to the MNI152 template [16]. The segmentation masks for these scans were obtained using the same method as employed for the ADNI dataset [15].

[1] Link to GitHub repository: https://github.com/danilayDH/AMML-Disentangling-Structure-from-Detail

For both datasets, a fixed test set was set aside to ensure the VAE models never saw this data during training. Pre-processing steps for both datasets involved:

- Select the middle 2D slice on the coronal axis
- Crop the image to 176x176
- Clamp the values to the 0.98 percentile
- Normalize both the image and demographic features (age and sex) to have a mean of zero and a standard deviation of one

The segmentation masks are stored as one hot encodings.

Model Architectures The VAE architectures utilized in this study were consistent with those from the previous work, all based on the models provided by PyTorch Lightning Bolts. These included a baseline VAE without segmentation masks, a VAE that incorporated segmentation masks into the encoder by adding brain segmentation masks as additional input layers, and a VAE that featured a separate encoder for segmentation masks, allowing for the independent processing of images and segmentation masks.

Hyperparameter Tuning In the previous work, we used Bayesian optimization to fine-tune the hyperparameters. For this study, we performed grid-search optimization, applying 5-fold cross-validation to the remaining data after setting aside the test set. The key hyperparameters optimized were learning rate, batch size, and Kullback-Leibler (KL) divergence coefficient. The goal was to identify the best model configuration based on reconstruction metrics such as Mean Squared Error (MSE) and Signal-to-Noise Ratio (SNR). By using cross-validation, the model was evaluated across different splits of the data, reducing the likelihood of overfitting.
It is worth noting that our focus was not on developing the best state-of-the-art model, but rather on understanding the relative performance differences between the three VAE architectures.

Evaluation based on downstream task For the downstream tasks, we trained three versions of the predictive models:

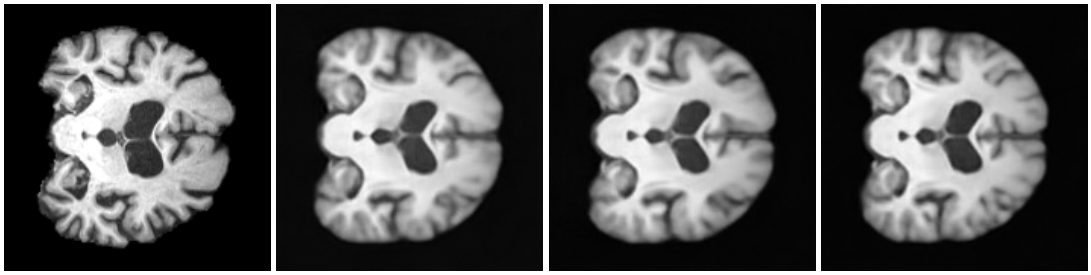
1. Using only demographic data (sex and age)
2. Using the latent representation from the VAE
3. Combining both the latent representation and demographic data

The downstream task differed for each dataset. For ADNI, logistic regression (LogisticRegressionCV) was used to predict dementia. For UKBB, ridge regression (RidgeRegression) was used to predict several continuous traits, including diastolic blood pressure, systolic blood pressure, BMI, LDL, HDL cholesterol, testosterone, and pulse rate. Both models were implemented using scikit-learn.

Unified Script for Model Training and Evaluation To streamline the training and evaluation process, we implemented a unified script that integrates the entire pipeline from training the VAE models to performing the downstream task. Each VAE architecture was trained for 30 epochs, after which the latent representation was extracted and used for downstream predictions, using logistic regression for ADNI data and ridge regression for UKBB data.
For each VAE architecture the entire pipeline was trained three times, each with a different seed, resulting in varied train-test splits for more robust evaluation. Balanced accuracy (ADNI) and R^2 scores (UKBB) were computed for each run, and the mean and standard deviation were calculated across the three seeds. This process helps to account for variations in the distribution of data sets, which is particularly important when working with relatively small data sets such as ADNI.

Results

Hyperparameter Tuning Results The grid-search optimization with 5-fold cross-validation revealed consistent optimal hyperparameters across all three VAE architectures. The best performance was achieved with a learning rate of 0.0001, KL divergence coefficient of 0.0001, and batch size of 32. However, the latent dimension differed slightly: for the standard VAE and masks in encoder architectures, it was set to 128, while for the separate encoder architecture, it was set to 64 for each encoder (summing up to 128 when concatenated). In terms of reconstruction metrics, the separate encoder architecture showed marginally better performance with a SNR of 11.329 and MSE of 0.074. The standard VAE followed closely with SNR of 11.3 and MSE of 0.075, while in encoder architecture showed slightly lower performance with SNR of 11.244 and MSE of 0.076. These results suggest that incorporating segmentation masks, particularly through a separate encoder, may lead to subtle improvements in reconstruction quality. In addition to the performance metrics, we created a number of reconstructions for images from the test set which are shown in Figure 1.



(a) Original **(b)** Standard VAE **(c)** Masks in encoder **(d)** Masks in separate enc.

Figure 1. Reconstructions of one of the test images for our three different architectures.

Performance on ADNI Dataset For the ADNI dataset, the downstream task involved predicting dementia (DX = doctors diagnoses) using balanced accuracy as the performance metric. Across all VAE architectures, using the latent representation from the VAE resulted in better performance than using only demographics (sex and age). Moreover, combining both the latent representation and demographics consistently yielded the best results across all models. The standard VAE achieved a test balanced accuracy of 0.823 ± 0.017 when using only the latent representation and 0.838 ± 0.034 when combining both demographics and latent representation. The masks in encoder architecture demonstrated similar performance, with a test balanced accuracy of 0.814 ± 0.036 for latent representation alone and 0.821 ± 0.035 when combining both. The separate encoder architecture outperformed both, achieving 0.822 ± 0.045 for latent representation and 0.842 ± 0.036 when combining both inputs. All three VAE architectures substantially improved upon using demographics alone (0.541 ± 0.016), indicating the value of incorporating imaging data. Combining VAE features with demographic information ("Both" column) yielded the best results across all architectures, with the separate-encoder achieving 0.842 ± 0.036 balanced accuracy compared to 0.541 ± 0.016 for demographics alone. The relatively small standard deviations (≤ 0.045) across all models suggest consistent performance across different data splits. The detailed results for all models and configurations can be seen in Table 1.

		Standard VAE		Masks in encoder		Masks separate encoder	
Label	Demographics	Latent Repr.*	Both	Latent Repr.*	Both	Latent Repr.*	Both
DX	0.541 ± 0.016	0.823 ± 0.017	0.838 ± 0.034	0.814 ± 0.036	0.821 ± 0.035	0.822 ± 0.045	0.842 ± 0.036

*Latent Representation

Table 1. Performance of VAE architectures on ADNI dataset for dementia prediction

Performance on UKBB Dataset For the UKBB dataset, our VAE models showed varying degrees of predictive power across different traits, with consistent performance across seeds as indicated by the small standard deviations. Similar to the ADNI dataset, using both latent representation and demographics generally yielded better performance than using either input alone. For the standard VAE, the best performance was observed when using both latent representation and demographics, with an R^2 score of 0.063 ± 0.009 for diastolic blood pressure, 0.140 ± 0.009 for systolic blood pressure, and 0.094 ± 0.006 for BMI on the test set. The masks in encoder architecture produced slightly lower R^2 scores compared to the standard VAE. For example, in the test set, BMI prediction achieved an R^2 of 0.087 ± 0.006 when using both inputs.

However, the overall trend remained consistent, with combining latent representation and demographics outperforming the individual inputs. The separate encoder architecture once again provided the best results for most traits. For BMI, it achieved a test R^2 score of 0.098 ± 0.010 when using both latent representation and demographics. Similarly, for diastolic blood pressure, the test R^2 score reached 0.063 ± 0.009 when using both inputs. For certain traits, such as testosterone, the performance remained relatively unchanged across the models and configurations. Testosterone consistently produced high R^2 scores based solely on demographics, with minimal improvement from the latent representation. For instance, all the prediction models using both inputs achieved 0.822 ± 0.006 on the test set testosterone prediction, indicating that this trait is predominantly determined by demographics rather than latent features extracted from brain scans. In summary, for the UKBB dataset, the masks in separate encoder architecture again provided the best performance, particularly for BMI, systolic blood pressure, and diastolic blood pressure. This supports the hypothesis that the separation of structural information via segmentation masks improves the VAE's ability to capture relevant features. The detailed results for all models and configurations can be seen in Table 2.

		Standard VAE		Masks in encoder		Masks separate encoder	
Label	Demographics	Latent Repr.*	Both	Latent Repr.*	Both	Latent Repr.*	Both
Diastolic BP**	0.037 ± 0.003	0.047 ± 0.006	0.063 ± 0.009	0.042 ± 0.005	0.060 ± 0.007	0.048 ± 0.007	0.063 ± 0.009
Systolic BP**	0.129 ± 0.008	0.087 ± 0.011	0.140 ± 0.009	0.085 ± 0.010	0.138 ± 0.008	0.089 ± 0.009	0.139 ± 0.009
BMI	0.016 ± 0.001	0.089 ± 0.009	0.094 ± 0.006	0.082 ± 0.006	0.087 ± 0.006	0.090 ± 0.009	0.098 ± 0.010
LDL	0.009 ± 0.002	0.007 ± 0.001	0.013 ± 0.004	0.005 ± 0.002	0.012 ± 0.003	0.006 ± 0.002	0.014 ± 0.005
HDL	0.194 ± 0.012	0.079 ± 0.005	0.205 ± 0.012	0.081 ± 0.008	0.203 ± 0.014	0.085 ± 0.007	0.206 ± 0.013
Testosterone	0.822 ± 0.006	0.344 ± 0.005	0.822 ± 0.006	0.335 ± 0.002	0.822 ± 0.006	0.352 ± 0.007	0.822 ± 0.006
Pulse rate	0.009 ± 0.000	0.016 ± 0.003	0.023 ± 0.003	0.017 ± 0.002	0.025 ± 0.002	0.019 ± 0.002	0.026 ± 0.002

*Latent Representation
**blood pressure

Table 2. Performance of VAE architectures on UKBB dataset for different traits prediction

Conclusion

This study investigated the integration of brain segmentation masks into VAEs to enhance feature representation for medical imaging tasks, specifically dementia prediction using the ADNI dataset, and the prediction of multiple traits from the UKBB dataset. By comparing three VAE architectures — standard, masks in the encoder, and separate encoders for segmentation masks — we aimed to assess their performance in learning latent representation and improving downstream predictions.

The results show that across both datasets, combining demographic data with the latent representation from VAEs consistently improved predictive performance, especially for traits such as BMI, blood pressure, and dementia. The separate encoder architecture, where brain segmentation masks are processed independently from the imaging data, yielded the best overall performance in most tasks, confirming the advantage of disentangling structural information from raw brain images.

For the ADNI dataset, balanced accuracy scores were highest when latent representation was combined with demographic data, with the separate encoder achieving a balanced accuracy of 0.842 ± 0.036. Similarly, in the UKBB dataset, the best R^2 scores for traits like BMI and systolic blood pressure were obtained using the separate encoder model, reaching 0.098 ± 0.010 and 0.139 ± 0.009, respectively.

Future work should focus on expanding the evaluation to more diverse datasets and phenotypes, with a particular emphasis on understanding the generalizability of learned latent features across different medical domains.

References

1. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*.
2. R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," *Ieee Access*, vol. 9, pp. 4939–4956, 2020.
3. D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691
4. N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 5301–5310.
5. D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M. Emberton, S. Esener, R. C. Fitzgerald, S. S. Gambhir, P. Kuhn *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, p. eaay9040, 2022.
6. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
7. G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," *arXiv preprint arXiv:1210.5474*.
8. W. H. Organization, *ICD-11: International Classification of Diseases 11th Revision : the Global Standard for Diagnostic Health Information*. World Health Organization. [Online]. Available: https://books.google.de/books?id=H8WFzgEACAAJ
9. K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson *et al.*, "Multimodal population brain imaging in the uk biobank prospective epidemiological study," *Nature neuroscience*, vol. 19, no. 11, pp. 1523–1536, 2016.
10. S. J. Heany, J. van Honk, D. J. Stein, and S. J. Brooks, "A quantitative and qualitative review of the effects of testosterone on the function and structure of the human social-emotional brain," *Metabolic brain disease*, vol. 31, pp. 157–167, 2016.
11. U. Jin, S. J. Park, and S. M. Park, "Cholesterol metabolism in the brain and its association with parkinson's disease," *Experimental neurobiology*, vol. 28, no. 5, p. 554, 2019.
12. C. Morrison, M. D. Oliver, F. Kamal, M. Dadar, A. D. N. Initiative *et al.*, "Beyond hypertension: Examining variable blood pressure's role in cognition and brain structure," *medRxiv*, 2024.
13. F. Isensee, M. Schell, I. Pflueger, 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*, vol. 40, no. 17, pp. 4952–4964, 2019.
14. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.
15. B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, 2023.
16. F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank," *Neuroimage*, vol. 166, pp. 400–424, 2018.

Disentangling Structure from Detail for Dementia Prediction

Daniela Layer* **and Jacob Schäfer***

*these authors contributed equally to this work

Introduction

Variational autoencoders (VAEs) are a tool for unsupervised representation learning and have applications in various fields, including biomedical research [1, 2]. Due to their nature, they focus primarily on the detection of low-frequency features [3], which poses a challenge for the detection of subtle pathological changes [4, 5].
In this technical report, the impact of integrating brain segmentation masks into VAEs is investigated. Using the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6], our research aims to improve the model's ability to capture detailed features relevant to the assessment of cognitive decline.
We propose three VAE architectures, compare them and evaluate their performance in terms of hyperparameter optimisation and downstream task evaluation.

Context and Related Work

VAEs, introduced in [1], are a form of unsupervised representation learning and consist of two main parts: The encoder and the decoder. These models work together, with the encoder linking the input data x to a latent representation z and the decoder converting z back to the original input space. Furthermore, the decoder serves as a guide or constraint for the encoder to capture meaningful representations of the data [7].
Current research on biomedical information using VAEs focuses primarily in two directions: Data augmentation and representation learning in the fields molecular design, sequence data set analyses and medical imaging and image analyses. VAE use on medical imaging dataset includes image classification, segmentation, restoration and reconstruction [2].
With VAEs, the focus of training is on the detection of low-frequency features, primarily reconstructing the largest number of pixels. This phenomenon is referred to as spectral bias [4]. However, this predominant focus on low-frequency features poses a challenge, especially in the context of disease detection, where high-frequency variations indicate an early stage of pathology [5].
Disentanglement is a process that involves breaking down complex data into features, which are independently meaningful [8]. In raw data, however, factors often appear to be closely intertwined. The challenge here is to construct the representations of the data in such a way that they provide features that are suitable for a variety of possible tasks and can deal with the reality of the intertwined variation factors [9].

Research Question

Our research focuses on investigating whether integrating brain segmentation masks as additional input to VAEs can enhance feature representation and enable the model to learn more detailed features in medical imaging. We are not aware of any publications that report to use segmentation mask as additional input to the VAE.
To investigate this we use the ADNI database which was established in 2004. The goal of ADNI is to investigate the use of biological markers and imaging to determine the decline of cognition in Cognitively Normal (CN), Mild Cognitive Impairment (MCI) and Alzheimer Disease (AD) [6, 10].
Alzheimer disease is the most common form of dementia. The International Classification of Diseases 11 (ICD-11) defines dementia as the presence of significant impairment in one or more cognitive domains. The impairment is related to age and the expected level of cognitive function and represents a deterioration from the previous functional level. Depending on the extent of neurocognitive and functional impairment, the severity of dementia can be categorized as mild, moderate or severe [11].
To create the used dataset, the imaging data available in ADNI was filtered for T1-MPRAGE scans and skull-stripping was performed using the HD-BET [12] tool. With a 1mm resolution a non-linear registration was conducted on the MNI152 templates using the FLIRT and FNIRT commands of the FSL tool [13]. Finally, the segmentation masks of brain regions were extracted using the Synthseg software [14].

Solution and Implementation[1]

To answer the research question we compared three different model architectures. All of them are based on the VAEs provided by Pytorch Lightning Bolts. The first architecture serves as a baseline and does not make use of additional segmentation masks. The second one adds the segmentation masks as additional layers to the encoder. Thus, the images and segmentation masks are entangled in this case. Lastly, the third architecture adds a second encoder following the same architecture which now separates the images and segmentation masks. For all those architectures we perform the following steps:

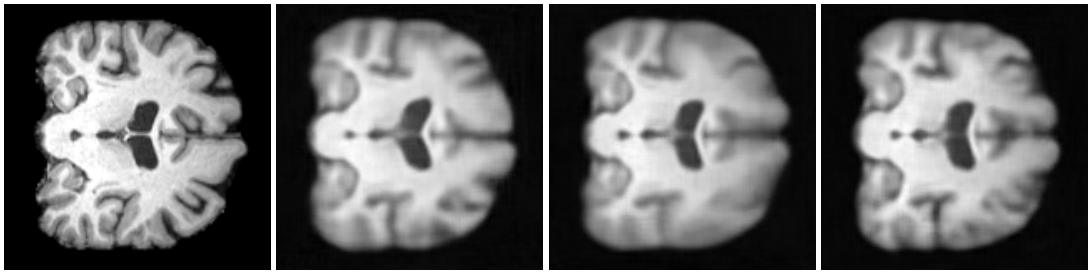
Data Preparation We use the ADNI dataset which contains 12791 3D brain scans and the corresponding segmentation masks of 1820 distinct patients. The dataset splitting is done based on the patients to avoid data leakage from training to validation and test sets. We use a split of 0.7, 0.15, 0.15 for the three datasets respectively. Since the number of scans per patient is variable, this results in dataset sizes 8847 (training), 1876 (validation), and 2014 (test).
Also, we perform several preprocessing steps per scan:

- Select the middle 2D slice on the coronal axis
- Crop the image to 160x160
- Clamp the values to the 0.98 percentile
- Normalize the image to mean zero and standard deviation one

The segmentation masks are stored as one hot encodings.

Hyperparameter Sweeps For each architecture, we optimize hyperparameters individually. We perform a bayesian sweep optimizing learning rate, batch size, latent dimensions, Kullback Leibler divergence coefficient, and whether to use an additional convolutional layer at the beginning and include a max pool layer. The sweep optimizes the validation reconstruction loss and includes 64 runs trained for 50 epochs each. With the best parameters, we train a final model for 400 epochs while saving the checkpoint with the lowest validation reconstruction loss.

Evaluation based on downstream task Finally, we evaluate for each architecture if a simple Logistic Regression Model can differentiate between cognitively normal and dementia based on the encodings. To do so, we keep the same dataset split but remove all entries which are marked as mild cognitive impairment or don't include a diagnosis. Note that this reduces the number of scans to 4642 (training), 998 (validation), and 1005 (test). As people age, their risk of developing dementia increases and on average, women tend to live longer than men [15, 16]. Therefore, we analyzed whether there could be a correlation in the data between dementia, age, and gender in predicting the onset of dementia. However, demographic factors like sex and age did not turn out to be predictive and we did not take additional measures for that (see Table 2). We use balanced accuracy as a metric for the classification task.



(a) Original **(b)** Standard VAE **(c)** Masks in encoder **(d)** Masks in separate enc.

Figure 1. Reconstructions of one of the validation images for our three different architectures.

[1] Link to GitHub repository: https://github.com/jacob271/Disentangling-Structure-from-Detail-for-Dementia-Prediction

Evaluation

As a result from the hyperparameter sweeps, we obtained one final model per architecture. See Table 1 for an overview of the final parameters as well as the performance of the models. The performance metrics are measured for the step where validation reconstruction loss was lowest.

Parameter/Metric	Standard VAE	Masks in encoder	Masks in separate encoder
learning rate	0.0001	0.0002	0.0003
batch size	64	64	32
latent dimensions	234	232	242
kl coefficient	0.048	0.049	0.042
Use first conv. layer	true	true	true
Use max pool layer	true	false	true
train recon. loss	0.062	0.072	0.056
val. recon. loss	0.0847	0.0864	0.0844
test recon. loss	0.1053	0.108	0.098

Table 1. This table shows the final configuration of our three VAEs as well as their performance at the step with minimal validation reconstruction loss.

In addition to the performance metrics, we created a number of reconstructions for images from the validation set which are shown in Figure 1.
As preparation for the downstream task, we first evaluated if sex and age are predictive for dementia. The logistic regression model achieved a balanced accuracy of 0.49 in this case, so we did not take additional measures.
With the best checkpoints from above, we obtained the following results for the downstream task of dementia diagnosis:

	Standard VAE	Masks in encoder	Masks in separate encoder	Sex and age as input
train balanced accuracy	0.798	0.784	0.821	0.502
val. balanced accuracy	0.706	0.709	0.702	0.494
test balanced accuracy	0.749	0.737	0.782	0.492

Table 2. Performance of the three different architectures for dementia prediction. The last column shows the predicitve performance based on demographic factors sex and age.

Conclusion

From the results of the evaluation we conclude that providing segmentation masks in a separate encoder improves dementia prediction. However, we also observed issues with overfitting which should be addressed in future work.
Despite evaluating the model checkpoints with lowest validation reconstruction loss, we still observe overfitting for the final VAE models. In an attempt to mitigate this, we ran an additional grid search hyperparameter sweep. Choosing 16 as number of latent dimensions and 0.1 for the Kullback-Leibler divergence coefficient did mitigate overfitting. However, the previous validation reconstruction loss was still better, so we decided to keep the final VAEs from the first sweeps.
Interestingly, the test validation reconstruction loss is even worse compared to the validation reconstruction loss. This indicates, that our final models also overfit on the validation set.
The impact of this can be seen in the results of the downstream task. While the validation balanced accuracy is similar across all architectures and up to 0.119 worse than the training balanced accuracy, the models perform significantly better on the test set. Also, the test balanced accuracy differs between the three architectures. The version with a separate encoder for the segmentation masks has a 0.033 higher test balanced accuracy than the baseline architecture. We suspect that the difference between validation results and test results for the downstream task is due to the previously mentioned overfitting on the validation set.
Overall, adding information about the structure through segmentation masks in a separate encoder does improve the predictive capabilities for dementia. However, future work should further investigate issues regarding overfitting.

References

1. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*.
2. R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," *Ieee Access*, vol. 9, pp. 4939–4956, 2020.
3. G. Bredell, K. Flouris, K. Chaitanya, E. Erdil, and E. Konukoglu, "Explicitly minimizing the blur error of variational autoencoders," *arXiv preprint arXiv:2304.05939*.
4. N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 5301–5310.
5. D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M. Emberton, S. Esener, R. C. Fitzgerald, S. S. Gambhir, P. Kuhn *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, p. eaay9040, 2022.
6. S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, 2005.
7. D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691
8. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
9. G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," *arXiv preprint arXiv:1210.5474*.
10. M. Weiner, R. Petersen, P. Aisen, M. Rafii, L. Shaw, J. Morris, and W. Jagust, "Alzheimer's disease neuroimaging initiative 3 (adni3) protocol," *Retrieved May*, vol. 24, p. 2016, 2016.
11. W. H. Organization, *ICD-11: International Classification of Diseases 11th Revision : the Global Standard for Diagnostic Health Information*. World Health Organization. [Online]. Available: https://books.google.de/books?id=H8WFzgEACAAJ
12. F. Isensee, M. Schell, I. Pflueger, 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*, vol. 40, no. 17, pp. 4952–4964, 2019.
13. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.
14. B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, 2023.
15. J. E. Seifarth, C. L. McGowan, and K. J. Milne, "Sex and life expectancy," *Gender medicine*, vol. 9, no. 6, pp. 390–401, 2012.
16. M. J. Katz, R. B. Lipton, C. B. Hall, M. E. Zimmerman, A. E. Sanders, J. Verghese, D. W. Dickson, and C. A. Derby, "Age-specific and sex-specific prevalence and incidence of mild cognitive impairment, dementia, and alzheimer dementia in blacks and whites: a report from the einstein aging study," *Alzheimer Disease & Associated Disorders*, vol. 26, no. 4, pp. 335–343, 2012.