NLDL Winter School Report
Explaining the decision process of a multi-view mammography image classifier

Anonymous Full Paper
Submission ###

1 Abstract

This report presents the development in an ongoing project that is focused on explainability in multi-view mammography classification. The objective is to develop a deep learning model that integrates four mammography views as distinct modalities to classify cancer risk and analyze the decision-making process of the model. The study employs a ResNet18-based classifier, saliency maps for interpretability, and the Angle-Based Joint and Individual Variation Explained (AJIVE) method to decompose learned representations. The adoption of AJIVE, while promising for disentangling shared and modality-specific information, presented technical challenges due to unmaintained implementations and the need for careful parameter selection. Additionally, dataset imbalance and overfitting remain key concerns in model training. Despite these challenges, the findings indicate that with further improvements it should be possible to attain better results as extreme BI-RADS categories (1 and 5) are better distinguished than intermediate ones. Initial saliency map results suggest distinct patterns of information processing between the two extremal classes as well. Future work will focus on improving model generalization and improving classification performance as well as refining AJIVE-based feature extraction and visualization techniques for decomposed representations.

2 Introduction

The main objective of this report is to present the current state of this project, which also forms one of the main research areas of my PhD project on developing methods for integrating and analyzing multimodal data in deep learning. The whole project is still in early stages, therefore this report does not contain all of the outputs optimistically stated in the proposal. However, to put the presented work into context and to showcase motivation behind it, the project is introduced in a similarly optimistic way in the following paragraph.

The main focus lies on implementing a multiview neural network for mammography image classification, treating each of the four mammography views that are taken during a standard screening, as a distinct modality. The network first processes each image separately and then integrates inner representations across views to classify images by cancer risk. To gain insight into the information flow within the network, a multi-modal data integration method AJIVE is be used. This algorithm processes individual hidden layer feature blocks, decomposing the learned representations into shared and individual components across modalities.

To further enhance interpretability and shed light onto the motivation behind networks decision, explainable AI methods like saliency maps are intended to be used to visualize the distinct and shared modes of variation directly on the original mammography images. This could help identify critical features and regions in the original images that influence predictions and provide insights into cross-view information interactions.

3 Literature Review

When fusing different blocks or modalities, it is essential to preserve semantic similarities. This is challenging because different modalities often represent similar information in very different ways. Data integration methods such as Canonical Correlation Analysis (CCA), address this issue by identifying common structure through linear transformations of the blocks so that their correlation is maximised. While CCA-based methods focus only on shared structures, Joint and Individual Variation Explained (JIVE) [1], and its extension, Angle-Based JIVE (AJIVE) [2], estimate both joint and individual structure by providing factorisations that partition the data blocks into joint and individual components. The separation of joint and individual structure could potentially enhance cross-modal interactions, allowing the dfferent modalities to better complement each other. More recently, Data Integration Via Analysis of Subspaces (DIVAS) [3] improved on JIVE and AJIVE by allowing for decompositions that capture structure shared by some, but not all, blocks. This approach was already applied on a somewhat related task of analysing breast cancer histologic images and genomic covariates in [4]. This work also used a CNN to extract features

from the images, they relied on a pretrained VGG16 model architecture [11]. In this project we base our models on a ResNet18 architecture [10] and use its ImageNet-1K pretrained weights as initialisation. For saliency map creation we follow the classical approach presented in [9].

4 Results and Discussion

All code related to this project is found on this GitHub repository. Training of models was logged using Weights&Biases and the relevant projects can be found here.

Dataset In this work we use VinDr-Mammo dataset [5]. It consists of 5000 four-view examinations with breast-level annotations. The four views are Left CC, Left MLO, Right CC and Right MLO, that are annotated using BI-RADS scores [6] laterality-wise. An example of a four-view observation from the test split is displayed in figure 1. We are working with an already split version of dataset with train:validation:test ratio of 16:4:5. The dataset is heavily imbalanced - the occurence of higher BI-RADS scores is rare, see figure 2, this causes a lot of issues and needs to be taken into account when training the models.

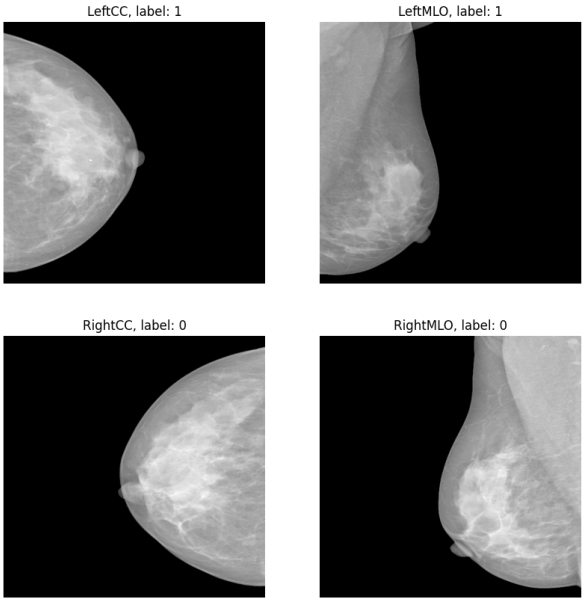

Figure 1. Example of a four-view observation from test split.

We are using pytorch lightning tooolkit to train the models, in line with its encapsulation practices we implemented three distinct `pl.LightningDataModule` classes to handle the dataset based on the needs of current type of model. `View_Cancer_Dataloader` loads images and their labels completely individually. `Breast_Cancer_Dataloader` loads both CC and MLO view images for each breast and its label. We tackle the imbalance in these two dataloaders in a relatively standard way, by sampling the less common classes with higher probability in the training dataset, so that in each epoch the class occurences were balanced. The third implemented class `Patient_Cancer_Dataloader` loads all four images from a single examination, along with two labels corresponding to the left and right breast. Imbalance handling is more difficult here, as each observation has two labels, we solved it by balancing for the maximum of the two labels, which does not seem to be ideal since the total number of BI-RADS 1 images in the dataset still remains much higher than for other labels in training split as can be seen in figure 3.

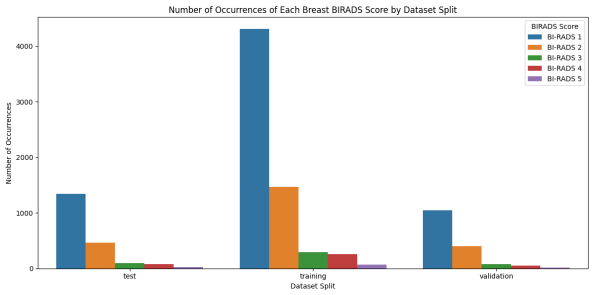

Figure 2. Histogram of label distribution in the original dataset.

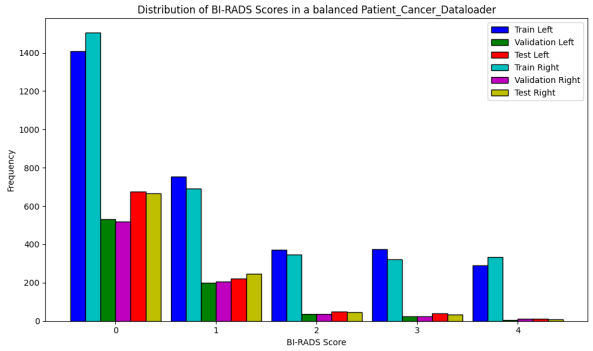

Figure 3. Distribution of labels in a balanced `Patient_Cancer_Dataloader`.

Classification networks As mentioned before, we based our classifiers on ResNet18 architecture and used its ImageNet-1K weights as initialization rather than random values. Our goal was initially to directly train a multiview model right away, using all the available images to inspect the interactions between different views. On figure 4, there is a diagram of a `4_View_2_Branch` model, that we proposed. When taking all the four views as the input we have different labels for left and right breasts, therefore we keep the two branches separate, let them predict labels separately and then sum up the two

losses before passing them to the optimizer. Here is a wandb project with training runs and logged metrics. The performance of this setup was initially very unsatisfying as we experienced significant overfitting. Gradient accumulation across batches improved this problem as the maximum batch size to fit into memory without issues is 16. Also, to improve the initialisation of individual ResNet featurizers we decided to train a single view model on all image views.

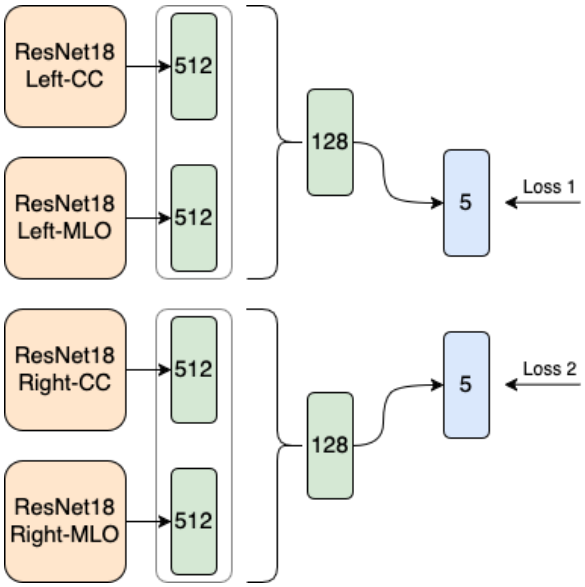

Figure 4. Diagram of a `4_View_2_Branch` model architecture.

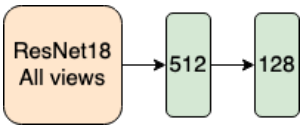

Figure 5. Diagram of `Single_View` model architecture.

This `Single_View` model is a simple wrapper around the ResNet18 model as shown on figure 5. It is trained using `View_Cancer_Dataloader` on a mix of all image views, and even though its results, available here, are relatively tragic, using these weights as initialisation improved the performance of multi-view models.

Due to complications caused by the presence of two labels in the case of `4_View_2_Branch` model and high time demands in finding the optimal architecture and optimization strategy for that setting we decided to chose a simpler approach to get the proof of concept and to test the whole intended pipeline of the project. For this, we chose a dual-view setting (CC and MLO of a single breast), where we only have one label per observation. The model `2_View_1_Branch` follows the design choices of others as is apparent from figure 6. Train runs for this model are logged here. The confusion matrix of final classification of test dataset is displayed on figure 7. The overall quality of predictions of the model is not very good, especially for categories 2,3 (BI-RADS scores 3,4), however relatively good accuracy for BI-RADS 1 and 5 scores gives hope that the model did pick up on some relevant structure in the images.

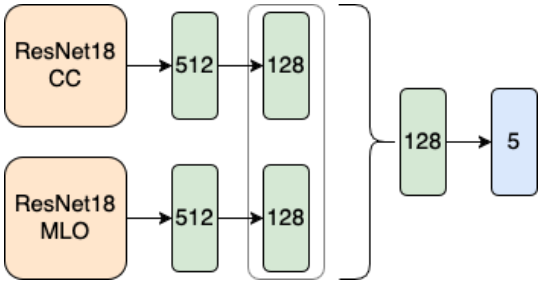

Figure 6. Diagram of `2_View_1_Branch` model architecture.

Saliency maps In this part, we only work with the trained `2_View_1_Branch` model mentioned above. Before attempting to visualize decomposed modes using saliency maps it was natural to just inspect the image-specific saliency map for predicted class as presented in [9]. To produce these maps, a single backpropagation pass is done with respect to the predicted class for a given image, then the absolute value of gradients corresponding to individual pixels is taken as the map. We display processed images from the test dataset, showing both view of the original image, then the saliency map on its own and then an overlay of the given saliency map over the original image to try to highlight the regions important for classifiers decision.

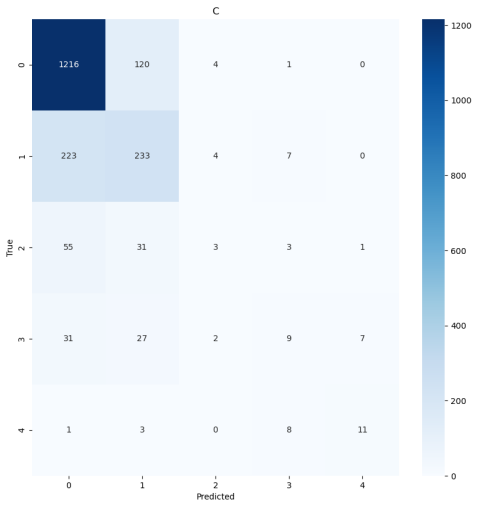

Figure 7. Confusion Matrix for classification results of `2_View_1_Branch` on test dataset.

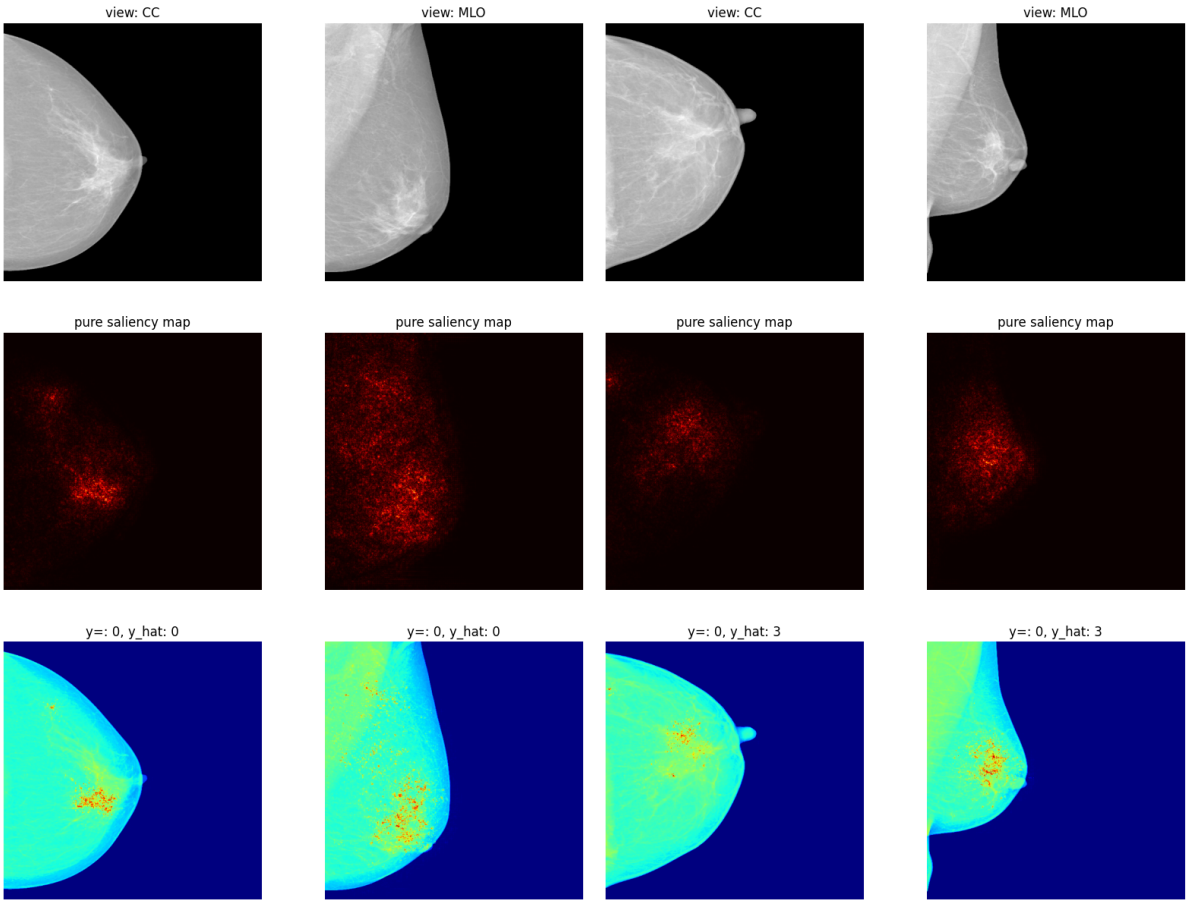

Figure 8. True labe: 0, Predicted label: 0

Figure 9. True labe: 0, Predicted label: 3

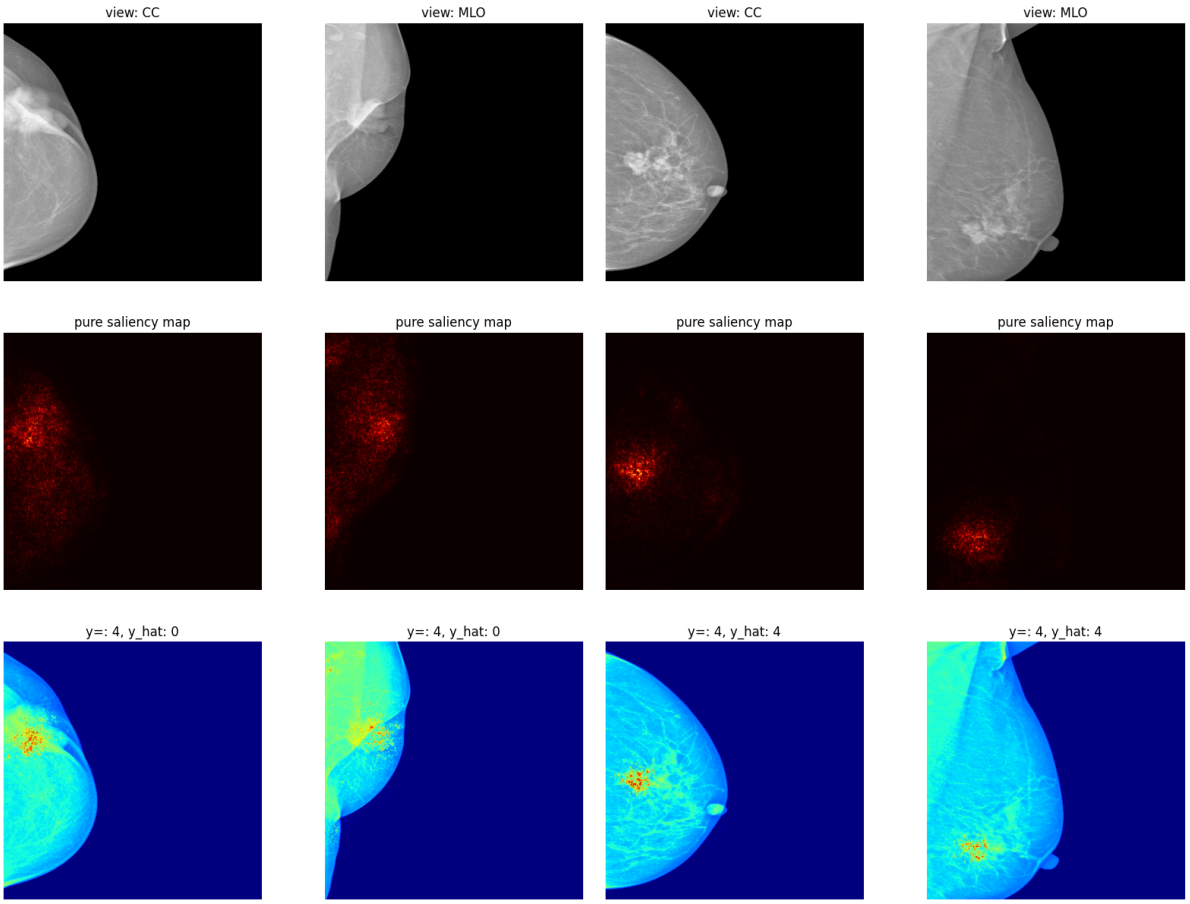

Figure 10. True labe: 4, Predicted label: 0

Figure 11. True labe: 4, Predicted label: 4

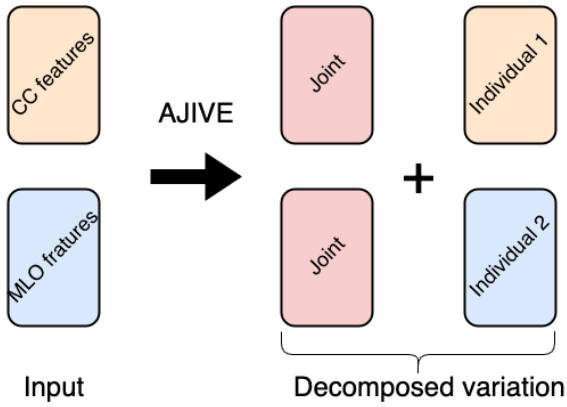

Figure 12. Visualisation of AJIVE method.

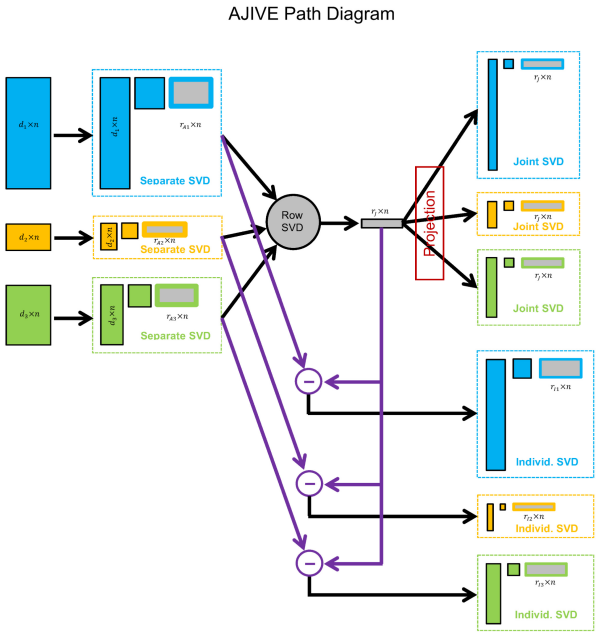

Figure 13. Chart demonstrating AJIVE from [2].

In figures 8, 9, 10 and 11 we are showing the results for four samples from test dataset, each contains both CC an MLO view of the same breast. Ground truth BI-RADS score as well as the prediction of `2_View_1_Branch` model are mentioned in the visualisation, we have chosen samples corresponding to the four corners of a confusion matrix in the sense of a true vs predicted class. It is difficult to draw conclusions just from these limited visualisation, but after inspecting other correctly classified samples from classes 0 and 4 (BI-RADS 1,5) it seems that the region of the image affecting the value of the output neuron corresponding to class 4 is much more centralised around a small region, which could be the around the tumor. In contrast, the region responsible for correctly classifying an image of a helthy breast from class 0 is much larger and the saliency map covers the whole foreground of the image. Further research is necessary to give meaning to the saliency maps corresponding to incorrectly classified images.

This approach to visualisation of the decision process of our CNN is just the first iteration using a simple and relatively old method. Much deeper study into the importance of different regions in the original image meeds to be conducted in future research, leveraging more complex methods such as Grad-CAM or Integrated gradients. It also seems obvious that we need to improve the performance of our classification network, before we can attribute some weight to the conclusions built on saliency map analysis.

Ajive decomposition In this section we present the current stage of involving the AJIVE method, presented in [2], in our project and the difficulties encountered. The adoption of this method turned out to be difficult both technically and methodologically. We remain in the context of a dual-view classification and `2_View_1_Branch` model. In our setting, we extract features from the input images using the frozen, trained model. We work with inner representations provided by the network just before concatenation of features from the two views (see figure 6), so we have two blocks of 128 features each.

The idea of the AJIVE algorithm is to estimate both joint and individual structure by providing factorisations that partition the input data blocks into joint and individual components as displayed on figure 12 for our setting. It does so by firstly low-rank approximating each data block, this is visualised on the left part of diagram from the original paper 13. Then in the middle joint structure between the low-rank approximations is extracted using SVD of the stacked row basis matrices. Finally, on the right, the joint components (upper) are obtained by projection of each data block onto the joint basis (middle) and the individual components (lower) come from orthonormal basis subtraction.

In our work we used the AJIVE implementation in python from [7] that also uses code from [8]. Unfortunately, these repositories are around five years old, the documentation is limited and they are no longer maintained, so it was not straightforward to put them to use and it was necessary to adjust multiple segments of the code to make it compatible with other packages used in our environment. For ease of adjustment the packages were cloned and form part of our codebase now. As input to the AJIVE algorithm it is also necessary to specify initial rank estimations of the individual feature blocks. A deeper look into how to choose these values is necessary in the future research, the idea of estimating intrinsic dimension of each input data block is to be investigated further. At this point, we present results obtained when choosing the initial values as DANCo [12] estimates of intrinsic dimension - 10 for

Feature set (dim)	Averaged f1 score
Original feat. (256)	0.339
CC original feat. (128)	0.310
MLO original feat. (128)	0.320
CC+MLO on Joint (6)	0.341
CC+MLO on Individual (15)	0.337
CC+MLO on Joint+Individual (21)	0.339

Table 1. Results of 5-fold k-nn classification on different feature sets of test dataset.

both blocks. The AJIVE algorithm then estimates the joint rank to be 3 and the individual ranks to be 7 and 8 for the CC and MLO feature blocks respectively. The projections of feature blocks onto the joint and individual spaces are depicted in the figure 14, with noise being the remainder of observed data after the joint and individual structure has been subtracted. As usually, features are ordered along the horizontal axis and data points, sorted by their class, are ordered along the vertical axis, with class 0 on top and class 4 on the bottom. It is worth to mention that in the joint projection the features of class 4 observations follow a significantly different pattern than for the other classes.

To try to asses the meaningfullness of the factorization we implemented a 5-fold cross-validation k-nn classifier to be employed on differently sets of features obtained using the AJIVE algorithm. Results of this experiment are listed in table 1. Dimension reduction using AJIVE does not seem to affect the performance of the validation classifier, however the scores are so low that we do not dare to draw serious conclusions from this, further investigation of the "quality" of factorized spaces is necessary.

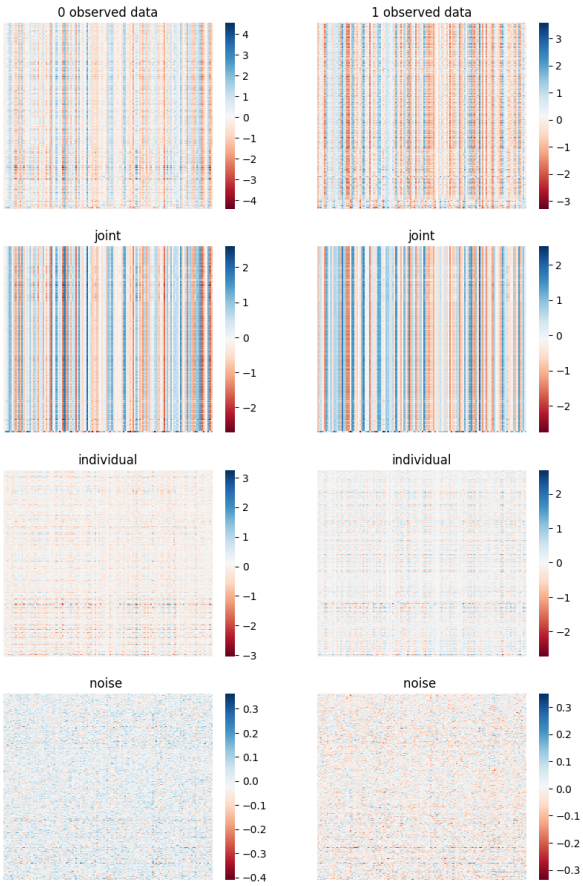

Figure 14. Projection of feature blocks onto the joint and individual variation subspaces.

5 Conclusion

This project is still in early stages and presented results are of an exploratory nature as the methodology and data processing pipelines are still in development. Due to limited time and technical difficulties encountered both with training on a cluster and when reusing unmaintained implementation of AJIVE, this report does not contain all the foreseen results mentioned in project proposal, however we have made progress in building a multi-view mammography classifier and investigating its decision process through explainability techniques.

Our experiments demonstrate the feasibility of a multi-view classification approach, where mammography images are treated as separate modalities. While the initial classifier encountered issues such as overfitting and limited classification performance for intermediate BI-RADS scores, the model showed relatively strong performance in distinguishing extreme cases (BI-RADS 1 and 5), which gives hope that with further improvements the model might also be able to capture more fine-grained information and distinguish the intermediate classes as well. In future work we intend to focus on image preprocessing, enhancement and augmentation. Also, since the X-ray images are grayscale, it might be beneficial to abandon the ResNet18 as the main featurizer and implement our own CNN to be trained from scratch specifically for this task. The incorporation of saliency maps provided early-stage insights into the classifier's focus regions. We have observed possible differences in how BI-RADS extremal cases are learned by the network. As mentioned before, we intend to use more sophisticated techniques in the future and it remains our goal to try to visualize the AJIVE acquired joint and individual structure in the original images, even though the methodology for this remains unclear. It might be beneficial to implement our own version of the core AJIVE functionality to ensure compatibility and flexibility in its usage.

6 Recommendations

It could be nice to have a more practical workshop related to high performance computing and cutting

edge deep learning frameworks. As deep learning models continue to grow in size, computational efficiency becomes increasingly critical and, in my experience, many machine learning researchers may lack the necessary computer science background to effectively optimize performance. However, NLDL might not be the ideal format for such a workshop.

References

[1] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types". In: *The Annals of Applied Statistics* 7.1 (Mar. 2013). ISSN: 1932-6157. DOI: 10.1214/12-aoas597. URL: http://dx.doi.org/10.1214/12-AOAS597.

[2] Q. Feng, M. Jiang, J. Hannig, and J. Marron. "Angle-based joint and individual variation explained". In: *Journal of Multivariate Analysis* 166 (2018), pp. 241–265. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2018.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0047259X1730204X.

[3] J. Prothero, M. Jiang, J. Hannig, Q. Tran-Dinh, A. Ackerman, and J. Marron. "Data Integration Via Analysis of Subspaces (DIVAS)". In: (Dec. 2022). DOI: 10.48550/arXiv.2212.00703.

[4] I. Carmichael, B. C. Calhoun, K. A. Hoadley, M. A. Troester, J. Geradts, H. D. Couture, L. Olsson, C. M. Perou, M. Niethammer, J. Hannig, and J. S. Marron. *Joint and individual analysis of breast cancer histologic images and genomic covariates*. 2020. arXiv: 1912.00434 [q-bio.QM]. URL: https://arxiv.org/abs/1912.00434.

[5] H. Nguyen Trung, H. Q. Nguyen, H. Pham, K. Lam, L. Linh, M. Dao, and V. Vu. "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography". In: *Scientific Data* 10 (May 2023). DOI: 10.1038/s41597-023-02100-7.

[6] E. Sickles, D. CJ, and B. L. et al. "ACR BI-RADS® Mammography". In: *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology;* (2013).

[7] I. Carmichael. *idc9/mvdr: First release*. Version 0.0.0. Oct. 2020. DOI: 10.5281/zenodo.4091757. URL: https://doi.org/10.5281/zenodo.4091757.

[8] I. Carmichael. *idc9/ya_pca: First release of the code!* Version 0.0.0. Oct. 2020. DOI: 10.5281/zenodo.4091759. URL: https://doi.org/10.5281/zenodo.4091759.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV]. URL: https://arxiv.org/abs/1312.6034.

[10] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

[11] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV]. URL: https://arxiv.org/abs/1409.1556.

[12] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli. *DANCo: Dimensionality from Angle and Norm Concentration*. 2012. arXiv: 1206.3881 [cs.LG]. URL: https://arxiv.org/abs/1206.3881.