DAIsy – Developing AI ecosystems improving diagnosis and care of mental diseases

ITEA 4 – 21016

Work package 4 (WP4) : AI Technology Development

Deliverable 4.3 : Limited data availability, transfer learning (doc & software)

Document type	: Deliverable
Document version	: No. 1
Document Preparation Date	: September 2025
Classification	: Confidential
Due Date	: October 2025

---

25 years ITEA 4 Community 1998–2023

Contents

I. Introduction

AI models achieve the best accuracy when trained on very large data sets and they perform much worse when trained with small datasets. The study of rare mental illnesses, unusual pathological conditions, comorbidities, and tailored therapy inevitably reduces the amount of statistical information available. This lack of statistical evidence makes AI-based diagnosis, prognosis, and treatment less reliable and poor in efficacy. Additionally, some datasets are highly rich in features, which may need careful consideration with regards to fitting a model to the data.

This deliverable document describes the tools that were developed in work-package-4, for task 4.3, which is to deal with these common challenges of lack of enough data. Although different modalities and different data types call for tailored approaches, here we have divided the possible ways to handle limited data in DAIsy in three sections: dimensionality reduction, data augmentation, and ML models tailored to limited data. It should be noted that techniques and models discussed throughout this deliverable document are limited to only those developed or applied by the partners of the DAIsy consortium.

II. Dimensionality reduction

Dimensionality reduction techniques play a crucial role in enhancing the analysis of small datasets by extracting and emphasizing the most relevant features, thereby improving the efficiency of machine learning models. Examples include feature engineering and regularization.

1. Feature selection

Feature selection can be via filtering, wrapper methods, or embedded techniques. When statistical tests such as Chi-square tests, (m)ANOVA, Wilcoxon tests, t-tests are used for modelling a dataset to a lower dimension, while preserving the distribution and characteristics of the original high(er) dimension of the dataset, the feature selection method is called filtering(Ghosh et al. 2024; James et al. 2023; Nikolaou 2016; Rainio, Teuho, and Klén 2024). Feature selection via recursive feature elimination or sequential feature selection constitute wrapper methods, whereas that via methods such as logistic regression with LASSO, decision trees (and random forests) are known as embedded methods. In this section we discuss the different techniques that were used in DAIsy, and how they were used.

a. **Chi-Square based** feature selection is a well-established Statistical technique for interpretable dimensionality reduction.
   In DAIsy, this has been applied by AMC within MRI radiomics features was performed on MRI radiomics features, as represented in the work published in (Poirot et al. 2024) and PhD thesis (Poirot n.d.).
b. **Minimum redundancy maximum relevance (mRMR)** was used as an iterative feature selection tool by AMC, within a MSc. Thesis project on classifying recurrent MDD in a population study (UK Biobank).
c. **Forward selection and recursive feature elimination (RFE)**, along with Random Forest (RF)-based feature importance scores can also be used for feature selection.
   ARD Group used feature selection to reduce dimensionality of normative modelling techniques, avoid its overfitting, and retain the model using the most informative features, for classifying bipolar vs. unipolar depression. To model the course of depression, Ascora used patient-related mood, sensor and input data, whose value ranges were defined using clinical scales and empirical distributions and whose

measurement error probabilities were taken into account through factor analysis and reliability tests. This information formed the basis for the creation of realistic synthetic data using generative models.

d. **Prior domain knowledge**: To cut back on the overflow of radiomics features available in the UK Biobank and prevent overfitting, AMC used previous literature reviews to find overlap of radiomics features between MDD and ED within a BSc. Thesis project.

2. Feature engineering

While feature selection constitutes selection of a subset of features from existing input feature set, feature engineering refers to creation of new feature set from original input features, often requiring domain-knowledge. Power spectral density (PSD) is a feature engineering method of measuring the energy of any signals, such as brain waves (EEG), characterise such signals based on their different constituent frequencies, and distinguish the EEG signal into alpha, beta, and gamma waves. While PSD transforms signal data in terms of the distribution of signal power across different frequencies, applying Fast Fourier Transform (FFT) on a signal, its frequency spectrum can be extracted. FFT illustrates how strongly each frequency is represented in a signal.

For ARD group, feature engineering included averaging gray matter across anatomical brain regions, encoding clinical variables, and removing low-variance features, with all data standardized for consistent input to classification models.

OFFIS used feature engineering for finding relevant features for fNIRS-, EEG, and fNIRS-EEG-based neurofeedback. MATERNA used PSD evaluation tool that is embedded in the NeurositySDK toolbox(Brainwaves | Neurosity SDK n.d.) they used for processing EEG signals. They applied FFT on EEG data to decompose it into distinct frequency bands, such as alpha, beta, etc., and enabling analysis of the different mental states.

3. Feature extraction

Feature extraction can be regarded as a type of feature engineering, that is more data-driven.

a. **Factor Analysis (FA):** FA assumes the existence of Q independent latent variables, which are assumed to follow a Gaussian distribution, denoted as (t ~N(0,I), where I is the identify matrix serving as the covariance matrix, indicating that these latent variables are uncorrelated. Alongside these latent variables, FA incorporates a Gaussian noise model. This noise model represents the randomness or errors in the data not explained by the Q latent variables. The relationship between the observable data ($X_{D\times N}$) and the latent variables ($t_{Q\times N}$) is given by equation: $\vec{X}=W\vec{t}+\mu+\epsilon$, where $\epsilon \sim N(0, \Psi)$. Here W is a matrix that details how each of the D observable variables contributes to each of the Q latent variables; it is referred to as the factor loadings. The covariance matrix Σ of the observed data X is related to the matrix W and the noise covariance Ψ, as in equation $\Sigma = WW^T + \Psi$. This indicates that the total variance of the observed data is the sum of the variance explained by the latent factors (WW^T) and the variance due to noise(Bishop 2006; Ghosh et al. 2024; Tipping and Bishop n.d.) .

b. **Principal Component Analysis (PCA):** Probabilistic PCA (PPCA) is a special case of FA, where the covariance Σ is expressed by a simpler term $\sigma^2 I$ (assumed to be spherical), instead of the diagonal matrix Ψ. Both FA and PPCA are capable of handling high dimensional data with missingness which is common in healthcare. The classical PCA is yet a special case of PPCA which assumed that the data is centered (μ=0), and the covariance σ=0. PCs are arranged such that the first PC contains the maximum information, followed by the second, and so on. Loadings of PCA illustrate the contributions

of each feature towards a PC. The higher the contribution of a feature (from input space) towards the lower numbered PC, greater is its importance in explaining the variance of the dataset(Ghosh et al. 2024; Tipping and Bishop n.d.).

c. **Independent Component Analysis (ICA)** is similar to PCA in terms of existence of a linear relationship between the observed and latent variables, except in ICA the latent distribution is non-Gaussian, given mathematically as$p(z) = \prod_{j=1}^{M} p(z_j)$(Bishop 2006). This is especially useful in preprocessing of signal data(Ghosh et al. 2024).

TU/e used FA and PCA for dimensionality reduction, preceding both classification of Eating Disorder (ED) patients of GGz OB into their respective ED types, and clustering to find inherent groups of ED patients, based on their baseline responses after intake. These dimensionality reduction techniques proved useful in improving classification performance and cluster quality, given the low sample sizes in the dataset, even at baseline. Similarly, given the low sample sizes relative to the high dimensionality, ARD group used PCA and FA prior to normative modelling for classifying bipolar vs. unipolar depression, to circumvent overfitting, while retaining the most informative dimensions. MATERNA used PCA for reducing dimensionality of EEG signals, while capturing the most variance in the data. They used ICA to separate the EEG signals into independent components, which can help in identifying the sources of brain activity related to motor imagery.

III. Data augmentation

Data augmentation is a key tactic for addressing limited data in machine learning. By applying transformations like rotation and scaling to existing data, it expands the dataset, improving model generalization and performance. This technique is especially beneficial in tasks like image and natural language processing, where it enhances model robustness with a sparse dataset. Therefore, in this section we discuss the data augmentation strategies used by DAIsy partners for their respective use-cases.

1. BLENDS

BLENDS is an Augmentation model of functional MRI, using Anatomically Constrained Warping(Nguyen et al. 2023). AMC applied this to fMRI for the prediction of treatment response in a clinical trial of Major Depressive Disorder patients, as illustrated in the PhD thesis (Poirot n.d.) and in the MSc thesis work of another student.

2. Audiomentations

Audiomentations is a Python library to augment audio, allowing customizable audio transformations and generating a variety of realistic audio samples(Jordal, Tamazian, and Dhyani n.d.). By augmenting segments of the mockup intake interviews at GGz OB (e.g. shift pitch, change volume, add noise). Semlab expanded the training dataset and tried to make the trained ASR-model more robust, by preventing overfitting on specific speakers.

3. GenAI based data generation of clinical notes

Philips used these tools for question-answering and summarization of conversations between physicians and patients at GGz OB, customer and service agents.

4. Windowing

OFFIS used this technique to create overlapping windows from EEG and fNIRS signals.

IV. ML models for limited data

Effective machine learning modelling relies on strategic techniques. Ensemble methods, such as Random Forests, aggregate predictions from diverse models, providing robust results with minimal data. Transfer learning leverages knowledge from a related domain with ample data, fine-tuning models for tasks with limited data. Embracing simplicity is key, where straightforward models like linear regression and decision trees prove advantageous. These simpler models mitigate overfitting and offer interpretability, making them valuable in data-scarce scenarios. We discuss here the different data-frugal models that were applied on the data from our clinical partners, without the need for data augmentation.

1. Data-frugal classification and regression models

a. **Logistic regression with LASSO regularization (Log-LASSO)** introduces L1-regularization, which penalizes the absolute size of regression coefficients. By imposing a constraint on the sum of the absolute values of the coefficients, LASSO effectively shrinks less informative coefficients to zero. This not only aids in feature selection and reducing high dimensionality, by identifying the most significant variables, but also enhances model interpretability and prevents overfitting (Alpaydin 2020; Bishop 2006; James et al. 2023; Nikolaou 2016; Tan et al. n.d.).

b. **Learning Vector Quantization (LVQ)** models are a family of dissimilarity- and prototype-based-classifiers (PBCs), which perform a discrimination task by learning representative examples of the ground-truth classes it trained on; these examples are called prototypes. Generalized Matrix LVQ (GMLVQ), in addition to learning the prototypes, also learn a relevance matrix, which contain information about which of the input features are relevant by themselves, or in combination with others, for the discrimination task at hand. Local GMLVQ are variants of LVQ where a relevance matrix is local to each prototype(Ghosh et al. 2025; Schneider et al. 2008).

c. **Linear Discriminant Analysis (LDA)** performs classification by transforming the higher dimensional input data into a lower dimensional space, such that the separation between the target classes are maximized. Thus, it can be interpreted as the supervised equivalent of PCA(Alpaydin 2020; Duda, Hart, and Stork 2015; Tan et al. n.d.).

d. **Support Vector Machine (SVM)** uses a transformation function (kernel trick), which can be linear (LSVM Classifier), Gaussian or radial basis (RSVM Classifier), or polynomial for instance, to map the original data into a higher dimensional space (hyperspace). Thus, for data that is not linearly separable into the target classes while in their original input space, SVM can find an optimal plane in the hyperspace to achieve such separability(Alpaydin 2020; Bishop 2006; Duda et al. 2015; James et al. 2023).

e. **Random Forest:** Decision Trees (DTs) are grown in a recursive fashion by partitioning the training records successively into ‘purer’ or more homogeneous subsets, thus giving these models an inverted tree structure. However, their performance depends on the selected feature at each decision node and thus the subsequent subtree, and can be unstable and unreliable on new data, as they tend to overfit(Bishop 2006; Tan et al. n.d.). Breiman's Random Forest (RF) method addresses this by aggregating multiple DTs, each trained on distinct but random subsets of instances and features. The collective decision-making process in RF involves majority voting for categorical

outcomes (classification) and averaging for continuous predictions (regression)(Alpaydin 2020; Breiman 2001; Tan et al. n.d.) .

f. **XGBoost:** Developed by Tianqi Chen et al, eXtreme Gradient Boosting or XGBoost is a decision tree-boosting based algorithm, where base learners that are weak learners are trained into 'strong' learners by reintroduction of the challenging instances, with higher weights attached to these being correctly identified(Bishop 2006; Chen and Guestrin 2016).

g. **Few-shot learning and transfer learning:** Few-shot learning(Wang et al. 2020) is used by MEDrecord to contextualize earlier prompts and data to train new models. When new models are deployed, transfer learning is used to distil data from a more advanced model to newly deployed models to achieve better results.

For classification of Eating disorder (ED) patients of GGz OB, based on their intake data, that included demographic information (age, biological sex, education), BMI, and the responses to ED relevant questionnaires such as SQ48, LAV, and MHC-SF, and EDEQ, into ED types Anorexia nervosa, Bulimia nervosa, Binge-eating disorder, and Other ED, TU/e used the classifiers RF, global and local variants of LVQ, SVM, LDA along with baseline log-LASSO. All mentioned classifiers, except LVQ were from the standard scikit-learn library of Python. LVQ model was from the 2021-2024 version of van Veen's SKLVQ toolbox.

OFFIS used RF, SVM and LDA in the context of Neurofeedback in real-time, such as for classification of task vs. resting states, using EEG and fNIRS data. ARD group used XGBoost, Log-LASSO, SVM, and RF for classification of unipolar vs bipolar depression, using structural MRI of the brain.

2. Data-frugal clustering techniques

a. **Gaussian mixture models (GMM)** are density-estimation (computing means and variances of each cluster) based probabilistic clustering strategy. The centroid of each of the clusters obtained from a GMM can be interpreted as a prototypical representative of the respective cluster. Therefore, all the cluster centroids from a GMM represent not only the underlying, distinctive profiles within a dataset, it essentially summarizes the dataset in terms of fewer representatives (Bishop 2006).

b. **Agglomerative clustering (AggC)** is a dissimilarity-based bottom-up hierarchical clustering which begins with the assumption that all data points are individual clusters, and at each step two closest clusters are coalesced into one. The pairwise proximity of clusters at every step is determined by the dissimilarity measure of choice and the type of linkages selected for defining a cluster (Tan et al. n.d.).

For clinical profiling of these patients on aforementioned dataset of GGz OB, based on their respective intake (baseline), TU/e used GMM and AggC from Scikit-learn library of Python. The clustering techniques were preceded by PCA and FA (also from scikit-learn).

3. Anomaly detection techniques

Anomaly detection is the process of identifying data points, events, or patterns that deviate significantly from expected behaviour.

a. **Isolation Forest** is an unsupervised machine learning model designed specifically for anomaly detection and is particularly effective in high-dimensional datasets. It operates on the principle that anomalies are data points that are more susceptible to isolation than normal observations, meaning they can be separated with fewer splits in a recursive partitioning process. The model builds an ensemble of binary trees, known as

isolation trees, anaolgous to DTs, thus making IFs analogous to RFs(Liu, Ting, and Zhou 2008).

b. **One-Class Support Vector Machine** is an unsupervised anomaly detection method based on the principles of support vector learning. The algorithm was introduced in(Schölkopf et al. 2001), and is designed to identify outliers by learning the boundary that encompasses the majority of data points in a high-dimensional space. In this formulation, the training data is assumed to represent only one class: the "normal" class. The algorithm maps the input data into a higher-dimensional feature space using a kernel function. In this space, it attempts to find a hyperplane that best separates the data points from the origin, such that most points lie on one side of the hyperplane (normal) and anomalies lie on the other side (closer to the origin).
c. **LOF algorithm**, as first introduced in (Breuniq et al. 2000) is a density-based anomaly detection method that identifies data points with substantially lower local density compared to their neighbours. The key idea is that normal data points are expected to lie in regions of similar density to their neighbours, while outliers will have noticeably lower density.

A BSc. Student of TU/e carried out their final thesis on laboratory data of hospitalized AN patients at GGz Oost-Brabant, focussing on detection of treatment non-response (anomalies), and especially investigate if those non-response patients coincided with a severe condition called re-feeding syndrome. The student used the aforementioned anomaly detection models from scikit-learn library for this objective.

4. Fine-tuning model for transcription

a. By fine-tuning state-of-the-art Whisper-large-v3 model for ASR Semlab improved its performance on in-domain (intake interviews, eating disorders, Dutch) audio transcription with limited available data.
b. **Voice activity detection and speaker segmentation models:** By fine-tuning state-of-the-art Voice Activity Detection (Pyannote) and Speaker Segmentation (NeMo) models, Semlab improved the speaker diarization pipeline on Dutch, 2-person intake interviews with limited available data.

REFERENCES

Alpaydin, Ethem. 2020. *Introduction to Machine Learning*. MIT press.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Brainwaves | Neurosity SDK. n.d. Retrieved 28 July 2025. https://docs.neurosity.co/docs/api/brainhttps://github.com/neurosity/neurosity-sdk-jswaves/.

Breiman, Leo. 2001. 'Random Forests'. *Machine Learning* 45(1):5–32.

Breuniq, Markus M., Hans Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. 'LOF'. *ACM SIGMOD Record* 29(2):93–104. doi:10.1145/335191.335388.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016:785–94. doi:10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2015. *Pattern Classification*. Wiley.

Ghosh, S., E. S. Baranowski, M. Biehl, W. Arlt, P. Tino, and K. Bunte. 2025. 'Interpretable Modelling and Visualization of Biomedical Data'. *Neurocomputing* 626:129405. doi:10.1016/J.NEUCOM.2025.129405.

Ghosh, Sreejita, Pia Burger, Mladena Simeunovic-Ostojic, Joyce Maas, and Milan Petković. 2024. 'Review of Machine Learning Solutions for Eating Disorders'. *International Journal of Medical Informatics* 189:105526. doi:10.1016/J.IJMEDINF.2024.105526.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. 'An Introduction to Statistical Learning'. doi:10.1007/978-3-031-38747-0.

Jordal, Iver, Araik Tamazian, and Tushar Dhyani. n.d. 'Audiomentations'.

Liu, Fei Tony, Kai Ming Ting, and Zhi Hua Zhou. 2008. 'Isolation Forest'. *Proceedings - IEEE International Conference on Data Mining, ICDM* 413–22. doi:10.1109/ICDM.2008.17.

Nguyen, Kevin P., Vyom Raval, Abu Minhajuddin, Thomas Carmody, Madhukar H. Trivedi, Richard B. Dewey, and Albert A. Montillo. 2023. 'BLENDS: Augmentation of Functional Magnetic Resonance Images for Machine Learning Using Anatomically Constrained Warping'. *Brain Connectivity* 13(2):80–88. doi:10.1089/BRAIN.2021.0186.

Nikolaou, Vasilis. 2016. 'Statistical Analysis: A Practical Guide for Psychiatrists'. *BJPsych Advances* 22(4):251–59. doi:10.1192/APT.BP.115.014696.

Poirot, M. G. n.d. 'Artificial Intelligence and MRI for the Prediction of Treatment Outcome in Depression'.

Poirot, Maarten G., Henricus G. Ruhe, Henk Jan M. M. Mutsaerts, Ivan I. Maximov, Inge R. Groote, Atle Bjørnerud, Henk A. Marquering, Liesbeth Reneman, and Matthan W. A. Caan. 2024. 'Treatment Response Prediction in Major Depressive Disorder Using Multimodal MRI and Clinical Data: Secondary Analysis of a Randomized Clinical Trial'. *The American Journal of Psychiatry* 181(3):223–33. doi:10.1176/APPI.AJP.20230206.

Rainio, Oona, Jarmo Teuho, and Riku Klén. 2024. ‘Evaluation Metrics and Statistical Tests for Machine Learning’. *Scientific Reports 2024 14:1* 14(1):1–14. doi:10.1038/s41598-024-56706-x.

Schneider, P. ;., F. M. ;. Schleif, T. ;. Villmann, and M. Biehl. 2008. ‘Generalized Matrix Learning Vector Quantizer for the Analysis of Spectral Data’. http://www.rug.nl/research/portal.

Schölkopf, Bernhard, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. ‘Estimating the Support of a High-Dimensional Distribution’. *Neural Computation* 13(7):1443–71. doi:10.1162/089976601750264965.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. n.d. *Introduction to Data Mining*. Pearson Education.

Tipping, Michael E., and Christopher M. Bishop. n.d. ‘Mixtures of Probabilistic Principal Component Analysers’. *Neural Computation* 11(2):443–82. http://www.miketipping.com/papers.htm.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. ‘Generalizing from a Few Examples’. *ACM Computing Surveys (CSUR)* 53(3). doi:10.1145/3386252.

APPENDIX

Here we share excerpts from public and internal dissemination of DAIsy partners that are relevant to this deliverable, and provided building blocks for this document.

A. Relevant excerpts from AMC

Artificial intelligence and MRI for the prediction of treatment outcome in depression
(Poirot n.d.)

From Chapter 2, page 50

We assumed that manual feature selection outperforms data-driven feature selection for two reasons: First, the number of samples per feature is extremely low number and second imaging biomarker features of sleep deprivation have been identified in previous studies.73–77 To test this assumption, we compared the performance of manual feature selection with the 500 best Chi-squared selected features. Chi-squared feature selection was implemented in Scikit-learn101 (v. 1.0.2).

From Chapter 5, page 179

Each machine learning pipeline consisted of three components. First, feature selection which selects the K best features based on the univariate linear correlation for regression, or χ2-test for classification. The hyperparameter for this pipeline component K ranged from 1 to N/5, with N the number of analyzed participants. Second, is scaling, which consists of subtraction of the mean and scaling to unit variance. Finally, a machine learning estimator. For regression we trained linear regression models, and for classification logistic regression classifiers. Both were regularized using elastic net regularization with hyperparameter L1/L2-ratio α between 0 and 1. The criterion for classification was accuracy, whilst for regression if was the negative root mean square error (RMSE). Machine learning components were implemented in the Scikit-Learn (v. 1.3.2)[62] package for Python (v. 3.9.1).

From thesis summary, page 236

This thesis addresses several novel methodologies for predicting antidepressant treatment response based on brain MRI imaging in depression. This work highlights the challenges posed by the heterogeneity of depressive symptoms and emphasizes the need for a multimodal approach for improved performance. Our findings are promising for better predicting treatment outcomes, albeit with some nuances. Our predictive models performed better than chance, but further validation in external cohorts is needed to translate these findings into clinical practice. Moreover, validation in prospective cohorts is considered ideal for robust clinical implementation. Our work underscores the value of research on developing predictive models and subsequent validation for clinical applicability in individualized treatment planning for depression.

Novel Models for Depressive Symptom Prediction using Functional Connectivity fMRI in the UK Biobank Cohort
(MSc. Thesis, L. De Vries)

This thesis investigates novel neural network models for predicting Major Depressive Disorder (MDD) using Functional Connectivity (FC) data derived from resting-state fMRI in the UK Biobank cohort. Four classifier types are compared: Multilayer Perceptron (MLP), Graph Convolution Network (GCN), Kolmogorov-Arnold Network (KAN), and Support Vector Machine (SVM). The study compares these models based on classifier performance, architectural complexity and investigates the interpretability of the KAN. Furthermore, the effect of feature reduction using minimum Redundancy Maximum Relevance

(mRMR) is investigated, and different definitions of MDD are tested comparing long-term depression with and without recent symptoms. It was found that complex neural models such as the KAN generally outperform simpler SVMs without feature reduction. However, this gap is narrowed down with mRMR applied. Performance is improved across all models, with the MLP achieving the highest accuracy of 81.3% for predicting severe recurrent MDD with recent symptoms. Predicting long-term depression without recent symptoms proved to be a more difficult task, with lower accuracies across all models despite a larger dataset, with some models achieving accuracies as low as 49%. The novel KAN architecture showed promising results, outperforming traditional MLPs and GCNs in some scenarios while having fewer parameters and more interpretable activations. However, most of these interpretable activations are actually similar to the activations found in an MLP.

Predicting Treatment Outcome in Major Depressive Disorder Using a Siamese Graph Isomorphism Neural Network with Augmented Longitudinal Emotional Conflict fMRI

(MSc. Thesis of L. Krook, AMC)

The wide variability of treatment outcome in Major Depressive Disorder (MDD), highlights the need for predictive models that can capture this treatment outcome as early as possible. With the few studies available that capture longitudinal changes in brain activation and functional connectivity from both pre-treatment and early-treatment emotional conflict functional Magnetic Resonance Imaging (fMRI) scans, the current study aims to address this gap by developing a Siamese Graph Isomorphism Neural Network to predict treatment outcome in MDD. Two sub-aims are addressed: (1) examining the effect of data augmentation using the Brain Library Enrichment through Nonlinear Deformation Synthesis (BLENDS) framework, to enrich dataset size and support effective deep learning, and (2) investigating whether incorporating radiomic features, which capture complex spatial characteristics from activation maps beyond mean activation values, improves prediction of treatment outcome. These sub-aims are investigated through ablation studies comparing models with and without augmentation and radiomics, based on their significance of the area under the curve (AUC) using the DeLong comparison test. All remission models showed statistically significant performance above chance. Furthermore, results demonstrate a detectable signal in predicting sertraline response, with the best significant model achieving an AUC of 0.64 and F1 score of 0.59 on augmented emotional conflict fMRI with radiomic features included. These findings underscore the potential of Siamese Graph Isomorphism Neural Networks as an approach for modelling longitudinal emotional conflict fMRI data to predict treatment outcome in the early stages of treatment.

B. Relevant excerpts from TU/e

Detecting Anomalies in Clinical Data Using Unsupervised Learning: A Study on Anorexia Nervosa Patients

(BSc. End project report, P. Kwaspen, TU/e)

From Chapter 2

2.2 Anomaly Detection Models

Supervised learning methods rely on the availability of accurately labelled positive and negative examples, but in many real-world healthcare scenarios, ground truth labels are either unavailable, unreliable, or only emerge retrospectively. A fundamental challenge in the application of machine learning to clinical domains is the scarcity or complete absence of labelled data, particularly for rare or underdiagnosed conditions such as RFS. This renders supervised models vulnerable to class

imbalance, label noise, and limited generalisability (Johnson and Khoshgoftaar, 2019; Wei et al., 2024; Ktena et al., 2024).

This study evaluates three unsupervised anomaly detection algorithms, Isolation Forest (IForest), One-Class Support Vector Machine (OCSVM), and Local Outlier Factor (LOF), to identify deviations in data that may signal the onset of anomalies like RFS. Each method captures abnormality from a distinct theoretical perspective: partitioning-based isolation (IForest), boundary-based separation (OCSVM), and density-based local deviation (LOF). The following subsections describe each of the models and their applications in more detail.

2.2.1 Isolation Forest

The Isolation Forest is an unsupervised machine learning model designed specifically for anomaly detection and is particularly effective in high-dimensional datasets. The original algorithm was introduced by Liu et al. (2008). It operates on the principle that anomalies are data points that are more susceptible to isolation than normal observations, meaning they can be separated with fewer splits in a recursive partitioning process.

The model builds an ensemble of binary trees, known as isolation trees. For each tree, a feature is selected at random, and a split value is chosen uniformly between the minimum and maximum values of that feature. This process divides the data and is repeated recursively in each subset until the instance is isolated or a maximum depth is reached (Figure 2.4).

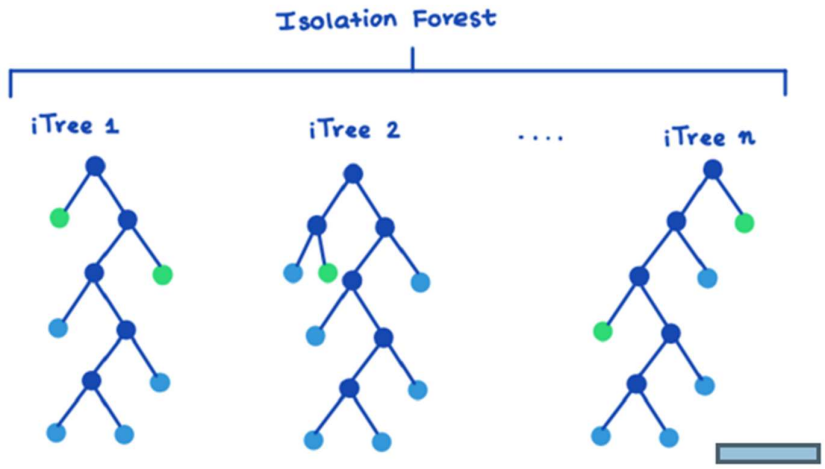

Figure 2.4: Isolation Forest

... ..

2.2.2 One-Class Support Vector Machine

The One-Class Support Vector Machine is an unsupervised anomaly detection method based on the principles of support vector learning. The algorithm was introduced by Schölkopf et al. (2001), and is designed to identify outliers by learning the boundary that encompasses the majority of data points in a high-dimensional space.

In this formulation, the training data is assumed to represent only one class: the“normal” class. The algorithm maps the input data into a higher-dimensional feature space using a kernel function. In this space, it attempts to find a hyperplane that best separates the data points from the origin, such that most points lie on one side of the hyperplane (normal) and anomalies lie on the other side (closer to the origin). An example of this can be found in Figure 2.5. This approach relies on the idea that anomalies differ significantly in structure or density from the normal data.

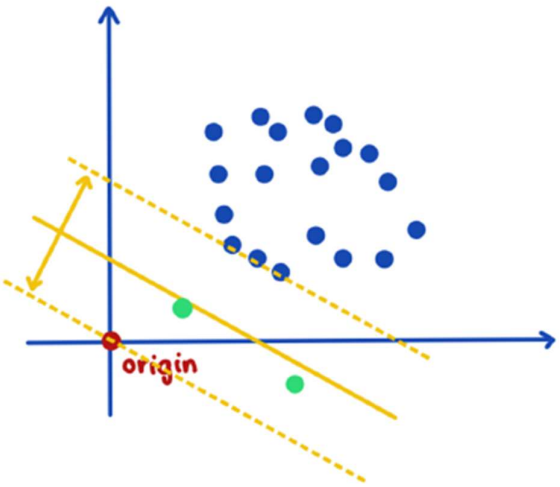

Figure 2.5: One-Class Support Vector Machine

A decision function f(x) is learned, and points for which f(x) < 0 are classified as anomalies, while those for which f(x) ≥ 0 are considered normal.

... ..

2.2.3 Local Outlier Factor

The LOF algorithm, as first introduced by Breunig et al. (2000), is a density-based anomaly detection method that identifies data points with substantially lower local density compared to their neighbours. Unlike global approaches such as IForest or OCSVM, LOF focuses on the local neighbourhood of each data point and computes how isolated it is with respect to the density of surrounding observations.

The key idea is that normal data points are expected to lie in regions of similar density to their neighbours, while outliers will have noticeably lower density.

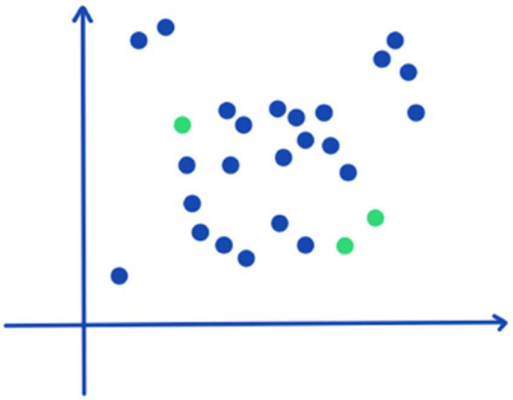

Figure 2.7: Local Outlier Factor Identifies Local instead of Global Outliers

2.3.1 Feature Contribution Analysis using Factor Analysis

To address sub-question 2, which focuses on understanding which features contribute most to anomaly detection, Factor Analysis (FA) is employed. FA is a statistical method used to simplify a dataset by taking a larger number of observed variables, and reducing them to a smaller set of unobserved factors, also known as factors (Spearman, 1961). If a latent factor has a strong relationship with observed variables, it accounts for a large portion of their variance.

Selection of Factor Extraction Method

To investigate which factor extraction method would be most appropriate for the dataset, a Spearman correlation matrix was created (Figure 2.8). Spearman's rank-order correlation was selected over Pearson's correlation because it is more robust to non-normality and skewed distributions, which is the case for the data (Spearman, 1904).

The matrix shows that several features are notably interrelated. In particular, ALT and AST exhibit a strong positive monotonic correlation ($\rho = 0.77$), indicating they may capture overlapping physiological processes. Furthermore, phosphate levels are moderately correlated with BMI ($\rho = 0.34$), suggesting possible associations between nutritional status and electrolyte imbalance.

These results support the use of dimensionality reduction techniques that can handle correlated inputs, while also reinforcing the need to verify normality assumptions due to observed skewness in features like AST (see Figure 2.9).

Given these characteristics, Principal Axis Factoring (PAF) as introduced by Thurstone (1931) and Harman (1976) was selected as the extraction method. PAF is well-suited for this type of medical data, as it is robust to violations of normality and less sensitive to outliers. PAF focuses specifically on common variance, making it appropriate for identifying latent physiological factors driving correlations between variables. This is particularly important in the context of anomaly detection, where subtle, multivariate changes in biochemical markers may indicate emerging risk. By reducing dimensionality while preserving the underlying shared structure among variables, PAF aids in isolating meaningful patterns that could distinguish normal from anomalous clinical profiles.

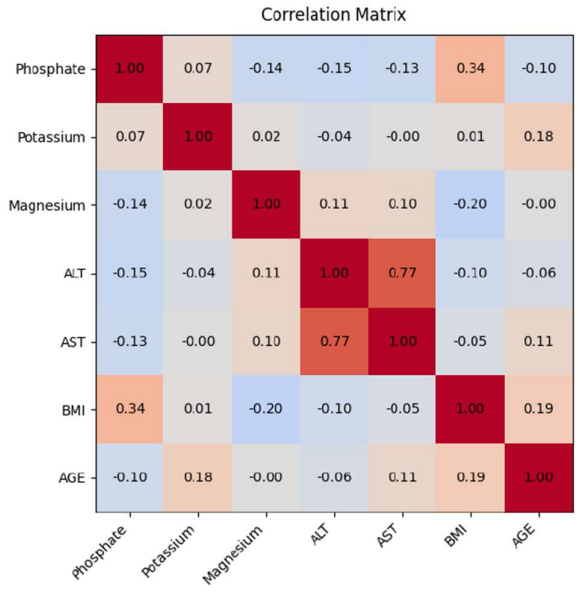

Figure 2.8: Feature Correlation Matrix