Enhancing Image-Tabular Data Classifiers under Distribution Shift via Multimodal Augmentation Strategies

Master Thesis

Master of Science in Applied Computer Science

Kevin Gutjahr

January 29, 2026

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Francesco Di Salvo, M.Sc.

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

In literature, machine learning models are often trained and tested with datasets that share similar distributions among their features. The rationale behind this approach is that these algorithms perform most effectively under these conditions. However, this approach does not reflect the real world, as data encountered after deployment in real life hardly follows the distributions of the training data. In particular, in medical settings, factors such as patient demographics, varying scanner settings, and missing data can produce previously unseen data distributions, which may cause incorrect predictions by the machine learning model. These mistakes can result in catastrophic consequences for patients.

Therefore, we investigate the effectiveness of various augmentation strategies in mitigating the adverse effects of such distribution shifts. We analyze both state-of-the-art latent embedding augmentation techniques and the integration of the Mixture-of-Experts (MoE) framework into the STiL architecture. The STiL architecture extracts and leverages both modality-specific and modality-shared information. Building on this, we integrated the Mixture-of-Experts framework in two variants: STiL-MoFE, in which classifiers receive a fixed set of inputs, and STiL-Switch, in which input tokens are dynamically routed by the gating network. In particular, our findings indicate that integrating the Mixture-of-Experts framework can lead to noticeable improvements in performance under distribution shifts, although the model may exhibit instabilities.

The code associated with this thesis is available on GitHub [1].

[1] https://github.com/kgutjahr/STiL-MoE

Abstract

In der Fachliteratur werden Machine-Learning Modelle häufig an Datensätzen trainiert und getestet, deren Attribute ähnliche Verteilungen aufweisen. Der Grund für dieses Vorgehen liegt darin, dass diese Algorithmen unter solchen Bedingungen am effektivsten arbeiten. Dieses Vorgehen spiegelt jedoch nicht die reale Welt wider, da die Daten, die in der Praxis anfallen, kaum den Verteilungen der Trainingsdaten entsprechen. Besonders im medizinischen Bereich können Faktoren wie Patientendemografie, unterschiedliche Scanner-Einstellungen und fehlende Daten zu bisher unbekannten Datenverteilungen führen, die wiederum zu falschen Vorhersagen des Machine-Learning-Modells führen können. Solche Fehler können katastrophale Konsequenzen für Patient:innen haben.

Daher untersuchen wir die Wirksamkeit verschiedener Augmentierungsstrategien, um die negativen Auswirkungen solcher Verteilungsverschiebungen abzumildern. Wir analysieren sowohl Methoden zur Augmentierung latenter Embeddings als auch die Integration des Mixture-of-Experts (MoE) Frameworks in die STiL-Architektur. Die STiL-Architektur extrahiert und nutzt sowohl modalspezifische als auch modalübergreifende Informationen. Darauf aufbauend haben wir das Mixture-of-Experts-Framework in zwei Varianten integriert: STiL-MoFE, bei der die Klassifikatoren einen festen Satz von Eingaben erhalten, und STiL-Switch, bei der die Eingabetoken dynamisch durch das Gating-Network weitergeleitet werden. Insbesondere zeigen unsere Ergebnisse, dass die Integration des Mixture-of-Experts-Frameworks zu spürbaren Verbesserungen der Leistung bei Verteilungsverschiebungen führen kann, auch wenn das Modell dabei Instabilitäten aufweist.

Der zu dieser Arbeit gehörende Code ist auf GitHub[2] verfügbar.

[2]https://github.com/kgutjahr/STiL-MoE

Acknowledgements

I would like to sincerely thank my supervisor, Francesco, for his guidance and support throughout this thesis. I greatly enjoyed our discussions about potential next steps.

I would also like to thank all my friends and family, who supported me during the most stressful moments of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

ACC	Accuracy
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
BACC	Balanced Accuracy
AE	Auto-Encoders
CDR	Clinical Dementia Rating
CGPL	Consensus-Guided Pseudo-Labeling
CLUB	Contrastive Log-ratio Upper Bound
DCC	Disentangled Contrastive Consistency
DVM-Car	Deep Visual Marketing Car
FAQ	Functional Activities Questionnaire
GDS-Short	Short form of the Geriatric Depression Scale
GAN	Generative Adversarial Networks
MCI	Mild Cognitive Impairment
MMS	Mini Mental State
MoE	Mixture of Experts
MLP	Multi Layer Perceptron
MMFM	Multimodal Fusion Model
MRI	Magnetic Resonance Imaging
NPI-Q	Neuropsychiatric Inventory Questionnaire
PGLS	Prototype-guided Label Smoothing
Semi-SL	Semi-supervised Learning
STiL	Semi-supervised Tabular-Image Learning
STiL-MoFE	STiL-Mixture-of-Fixed-Experts
STiL-Switch	STiL with Switch Transformer MoE Implementation

Notation

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix

Sets and Graphs

$\mathbb{R}$	The set of real numbers
$\{0, 1, \ldots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b

Linear Algebra Operations

$\boldsymbol{A} \odot \boldsymbol{B}$	Element-wise (Hadamard) product of $\boldsymbol{A}$ and $\boldsymbol{B}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$p(\mathrm{y}\|\mathrm{x})$	The conditional distribution of y given x
$\mathrm{a} \sim P$	Random variable a has distribution P
$H(\mathrm{a}, \mathrm{b})$	The cross-entropy between a and b
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{U}(\mathrm{a}, \mathrm{b})$	Uniform distribution with boundaries a and b

Functions

$f(\boldsymbol{x};\boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log(x)$	Natural logarithm of x
$\mu(x)$	Sample mean of x, $\frac{1}{N}\sum_{i=0}^{N-1} x_i$
$\sigma(x)$	standard deviation of x, $\sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(x_i-\mu(x))^2}$
$\zeta(x)$	Softplus, $\log(1+\exp(x))$
$\mathbb{1}\{\text{condition}\}$	Indicator function, returning 1 if the condition is met, 0 otherwise
$CV(x)$	Coefficient of variation, $\frac{\sigma(x)}{\mu(x)}$
$\text{argmax}(x)$	Index of the maximum value of x
$kth\_excluding(x,k,i)$	the top-k elements of x, excluding x_i itself
$\Pr(X=x)$	Probability that the random variable X equals x
$\max(x)$	Maximum value of x
$\Psi(\boldsymbol{a},\boldsymbol{b})$	exponential of the cosine similarity between $\boldsymbol{a}$ and $\boldsymbol{b}$, $\exp\left(\cos\left(\boldsymbol{a},\boldsymbol{b}\right)/\kappa\right)$ with temperature parameter κ
$\text{sim}\left(\boldsymbol{a},\boldsymbol{b}\right)$	log-softmax of the cosine similarities between $\boldsymbol{a}$ and $\boldsymbol{b}$, $\log\left(\frac{\Psi\left(\boldsymbol{a},\boldsymbol{b}\right)}{\sum_{k=1}^{B}\Psi\left(\boldsymbol{a},\boldsymbol{b}_k\right)}\right)$

Datasets

$\mathcal{X}$	A batch of input features
$\boldsymbol{x}$	A sample of input features
$\boldsymbol{x}^i$	A sample of image features
$\boldsymbol{x}^t$	A sample of tabular features
$\boldsymbol{y}$	The target associated with x for supervised learning
$\mathcal{Y}$	The set of all unique labels present in the dataset
B_l	Amount of labeled data samples in the batch $\mathcal{X}$
B_u	Amount of unlabeled data samples in the batch $\mathcal{X}$
B	Total size of the batch $\mathcal{X}$, $B = B_u + B_l$

Mixture of Experts

f_i	expert i
K	The total number of experts in the framework
k	The amount of activated experts
G	The gating Network
$G(x)$	Gating weights
$g(x)$	The gating function's output given x
$\mathcal{R}_{noise}$	Noise added to the gating function's input
N_T	Total amount of tokens in the batch $\mathcal{X}$

STiL

f	A Classifier inside CGPL
ϕ	An Encoder
$\boldsymbol{I}$	A raw latent embedding of the image encoder
$\boldsymbol{T}$	A raw latent embedding of the tabular encoder
$\mathbf{z}$	A latent embedding
$\mathbf{z}_c$	A latent embedding containing modality-specific information
$\mathbf{z}_s$	A latent embedding containing modality-shared information
$\hat{\mathbf{z}}$	A refined latent embedding
$\boldsymbol{p}$	A logit
$\bar{\boldsymbol{p}}$	A refined logit

1 Introduction

Humans perceive the world through a diverse set of signals, such as visual sensory, sound, scent, touch, and taste. Information obtained from such heterogeneous sources or senses can be organized into categories referred to as *modalities*. These signals give us complementary cues about different aspects of our environment. As living beings, humans utilize multiple modalities simultaneously to derive a more holistic view of the world. For instance, when purchasing a new car, its appearance alone is usually not sufficient to make an informed decision. Other aspects, such as mileage and fuel efficiency, can also highly influence such decisions, as they give insight into the car's condition and upcoming costs. Considering such differing aspects gives us a *multimodal* perception and it guides us in various scenarios, such as decision-making. Consequently, multiple data modalities are also used in medical settings, including diagnosis. In this context, multimodal data usually includes medical imaging, such as magnetic resonance imaging (MRI) and computed tomography (CT), tabular data, such as laboratory results and questionnaires, or semi-structured or unstructured texts, such as notes Kubben (2019). For instance, both imaging data and tabular data are used in the diagnosis of Alzheimer's disease (AD) (Bhagwat et al., 2018). In this case, utilizing both modalities is crucial as examining one modality alone would not provide sufficient information necessary to create a reliable diagnosis (Huang et al., 2020). MRI images may hold specific features that are also present in scans of healthy, aging patients (Huang, 2023). Thus, relying solely on imaging data may result in an inaccurate diagnosis here. Moreover, tabular data in particular is becoming increasingly popular in the deep learning literature (Borisov et al., 2024). In research areas such as medicine, finance and marketing, it is one of the most prevalent modalities (Shwartz-Ziv and Armon, 2022; Sahakyan et al., 2021). This is due to its ability to hold heterogeneous features, such as categorical and numerical ones (Arik and Pfister, 2021). For instance, in a medical setting, age and body weight are continuous features, whereas sex and genome type are categorical. In a marketing context, such as car advertising, attributes like color and fuel type are categorical, while variables such as mileage and price are continuous. However, this characteristic, along with weaker inter-feature correlations compared to the strong spatial correlations in image features, makes processing this modality a challenging task (Borisov et al., 2024).

Multimodal fusion models (MMFM) are deep learning models designed to leverage the complementary characteristics of different data sources (Baltrušaitis et al., 2019). Their basic structure in the context of image-tabular data is shown in Fig. 1. In recent years, these models have gained rising interest in different areas, such as biomedical research (Stahlschmidt et al., 2022), marketing (Wang et al., 2025a), and clinical diagnostics, such as Parkinson's disease (Lv et al., 2024) and Alzheimer's disease (Bhagwat et al., 2018). Particularly, image-tabular MMFMs are becoming an increasingly important subfield in deep learning literature (Hager et al., 2023). According to Du et al. (2024), together with large medical image-tabular datasets, such models could help us to better understand human health.

However, the performance of deep learning models relies heavily on the assumption that data encountered during deployment is independent and identically distributed from the data used for training (Shen et al., 2021). In other words, the model expects the test data to share the same statistical characteristics as the training data. Distribution shifts can manifest in various forms, such as covariate shift (Quiñonero-Candela et al., 2009). In covariate shift, the marginal distributions of input features shifts between the training and test datasets, i.e. $p(x_{train}) \neq p(x_{test})$, while the conditional distribution of labels given the input features, $p(y|x)$, remains the same. For example, in a medical setting, the age distribution of two patient groups may differ, while the relationship between the age and the diagnosis remains the same. Consequently, the marginal distribution of the diagnosis, $p(y)$, will shift according to the impact of the shift of $p(x)$. In reality, however, this can happen due to a variety of reasons. For instance, in a medical setting, differences in patient populations can affect both tabular and image data. Tabular data usually contains demographic information, while in image data, anatomical structures, such as organs, may vary in shape and size (Yoon et al., 2024). Additionally, if the scanners are not configured uniformly, the resulting scans may exhibit different pixel intensity levels, leading to variations in brightness and contrast (Guan and Liu, 2022). Furthermore, tabular data may be affected by errors arising from both human and machine sources. For example, improperly configured electronic healthcare systems may not accurately capture relevant information or may even confuse the data of different patients (Bowman, 2013).

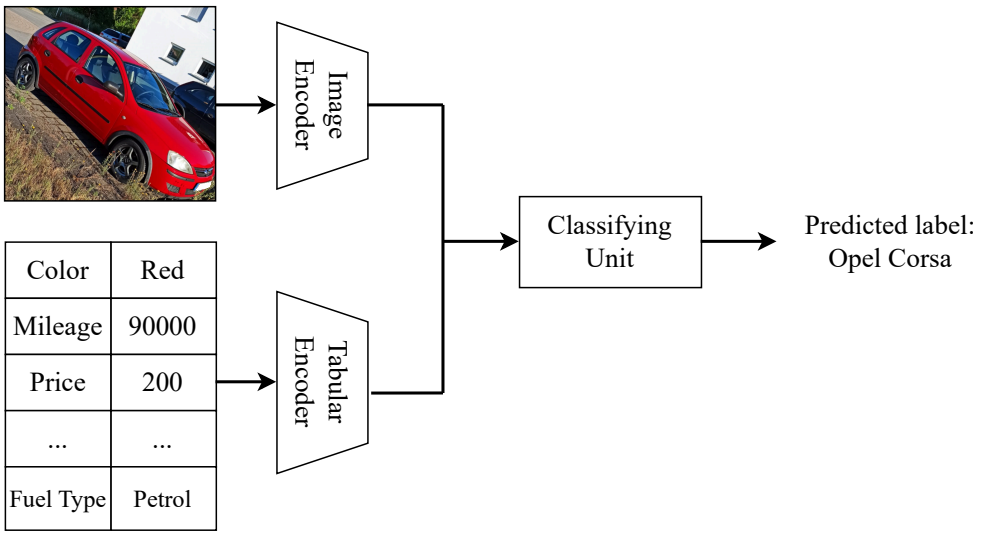


Figure 1: Basic structure of a classifying MMFM using image-tabular data. The incoming modalities are converted to latent embeddings using modality-specific encoders. The resulting embeddings are then used by a classifying unit to create a prediction.

If such a dataset shift occurs, even modern classifiers can become biased toward the training data, leading to confident but erroneous predictions (Tamang et al., 2025). In a medical setting, such an inaccurate prediction can lead to misdiagnosis with serious consequences for the patient. For instance, an inaccurate prediction of whether a patient suffering from mild cognitive impairment (MCI) will be stable or develop AD will lead to the delay of essential treatment (Moraveji and Mansouri, 2025). Therefore, this work aims to investigate various augmentation strategies to enhance the robustness of a multimodal image-tabular model against a set of handcrafted datasets exhibiting distribution shifts.

The examined model in this thesis is the *STiL* model developed by Du et al. (2025). STiL is short for Semi-supervised Tabular-Image Learning, and it employs semi-supervised learning (Semi-SL) to classify image-tabular data. The objective of this model is to achieve high performance despite the absence of labeled training data. In a supervised setting, deep learning models require large amounts of high-quality labeled data (Jang and Cho, 2019). In the context of medical data, this task is particularly challenging, since labeling often relies on qualified experts performing manual annotations, which can be both time-consuming and error-prone (Aljuaid and Anwar, 2022). Thus, it applies Semi-SL. However, the reason why this model was chosen is its unique approach of utilizing information from multiple modalities. The majority of multimodal Semi-SL methods can be classified as either Co-Training or Co-Regularization methods (Yang et al., 2019). Co-Training (Blum and Mitchell, 1998) utilizes one training algorithm per modality, and each classifier labels unlabeled samples. The most confident labels are then added to the other classifier's training set. On the other hand, Co-Regularization (Brefeld et al., 2006) trains one training algorithm per modality and aims to minimize their disagreement as much as possible. However, according to Du et al. (2025), such methods do not fully utilize the potential of the information given by the input data.

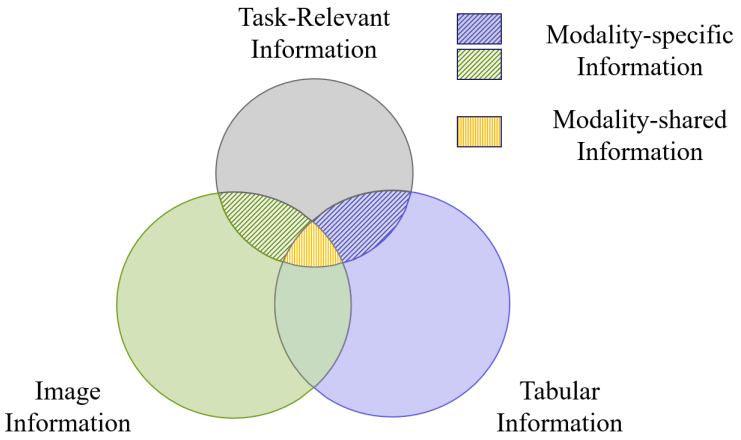


Figure 2: Visualization of the Information Modality Gap (Du et al., 2025). STiL focuses on the information gained by image and tabular data.

Co-Training methods operate under the assumption that each unimodal view is sufficient to make proper predictions. Conversely, Co-Regularization methods operate under the assumption that task-relevant information is common between the modalities. This means that these methods only utilize either the modality-specific or modality-shared information. Du et al. (2025) identified this insufficient usage of information as the *Information Modality Gap* (illustrated in Fig. 2) and thus integrated the usage of both types of information into their model. A detailed explanation of STiL's architecture and training process is provided in Section 3.2.

The datasets used in this thesis are subsets of the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2006) and the Deep Visual Marketing Car (DVM-Car) (Huang et al., 2022) datasets, which are presented in Section 4.1.1. The use of these datasets enables the investigation of the impact of the examined augmentation strategies on both medical and natural data. From these datasets, subsets have been extracted that exhibit data distribution shifts either in both or in individual modalities. This approach enables the evaluation of the effectiveness of the augmentation strategies against distribution shifts.

First, we investigate augmentation strategies applied in both the input space and the latent embedding space. Input space data augmentation is a standard method for improving robustness to unseen data (Mumuni and Mumuni, 2022). In contrast, latent embedding space augmentation techniques are an emerging, yet sparsely researched area in the literature (Liu et al., 2023). Thus, the effect of popular state-of-the-art methods in this field is examined. However, the results of these experiments showed that the individual modalities contribute differently to the current classification task, leading to higher uncertainty in the respective classifying units. Thus, the second strategy followed aims to augment the model's architecture. Here, we integrate the Mixture of Experts (MoE) framework (Jacobs et al., 1991). It allows to dynamically weight the contributions of classifying units inside model architectures, reflecting their relevance to the current task. This framework is gaining prominence in literature and is commonly utilized inside transformer-based architectures (Mu and Lin, 2025). Nevertheless, it has also been applied inside multimodal architectures in literature, such as the Gemini 2.X family (Comanici et al., 2025) and Uni-MoE by Li et al. (2025). These works report promising performance due to improved classifier collaboration. Thus, this thesis examines the impact of MoE on data distribution shifts.

This thesis is structured as follows: Section 2 conveys theoretical background knowledge about MMFMs, data augmentation strategies, and the MoE framework. In addition to STiL's architecture, Section 3 also describes the integration of latent embedding augmentation techniques and the MoE framework into STiL. Section 4 first presents the created subsets of the datasets that exhibit distribution shifts and then showcases the results of the experiments. In Section 5, the key findings and the limitations of this work are presented. Finally, Section 6 concludes this thesis.

2 Theoretical Background

2.1 Multimodal Fusion Models

Learning from multiple modalities is not easy. Data is created by complex systems, which are controlled by many latent variables that are not directly accessible (Lahat et al., 2015). An overly simplistic approach to data fusion could possibly result in the loss of important, task-relevant information. Therefore, MMFMs aim to replicate our multimodal ability to understand the world by transforming data from multiple modalities into a fused encoding that captures inter-modal relationships (Zhao et al., 2024). These encodings, also known as embeddings or representations, are numerical, n-dimensional vectors that act as a concise representation of the task-relevant information in the data samples (Baltrušaitis et al., 2019). The resulting set of vectors corresponds to distinct points within an embedding space, as illustrated in Fig. 3. This representation empowers the model to perform essential mathematical operations, such as cosine similarity, enabling it to interpret the semantics of the input data and discern instances of shared characteristics. Therefore, multimodal models create a shared embedding space from incoming data to learn both similarities and discrepancies between the differing modalities (Bayoudh et al., 2022). MMFMs generate these embeddings using dedicated, often pretrained, encoders. These encoders are usually smaller deep learning models tailored to each modality.

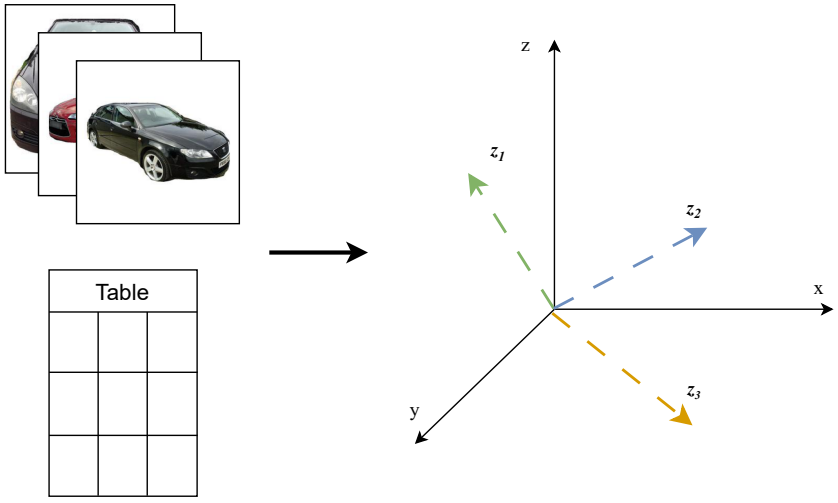


Figure 3: Simplified illustration of multimodal data samples represented by the embeddings z_1, z_2, and z_3, which are pointing to different coordinates within a three-dimensional vector space. Vectors representing similar semantic meanings point towards the same direction, while vectors representing contrary meanings point towards opposite directions. The car images are taken from the DVM-Car dataset (Huang et al., 2022).

For example, imaging data can be encoded using ResNet-50 (He et al., 2016), while tabular data can be encoded using multi layer perceptrons (MLPs) or transformer-based (Vaswani et al., 2017) encoders. Multiple multimodal models use these architectures to generate their embeddings (Hager et al., 2023; Du et al., 2025, 2024; Radford et al., 2021). Later on, these representations can be fused in three ways: Early-Fusion, Late-Fusion and Hybrid-Fusion (Yang et al., 2022). They are presented in Fig. 4. Early-Fusion combines the embeddings right after they have been extracted via operations like vector concatenation. Early-Fusion enables learning from multimodal embeddings from the beginning (Snoek et al., 2005). However, Martínez and Yannakakis (2014) argued that this approach does not fully capture low-level interactions between the modalities, leading to an insufficient understanding of the data. Especially with an increasing amount of modalities, Early-Fusion struggles to capture these interactions (Atrey et al., 2010). In the Late-Fusion-paradigm, each modality-embedding will be processed by its own dedicated, task-performing model and the resulting outputs will be combined. This can be done using, for example, rule-based methods or linear weighting of the outputs, allowing a deeper understanding of the individual modalities (Atrey et al., 2010). However, it does not capture the synergies that arise from multiple modalities interacting with one another (Yang et al., 2022). Hybrid-Fusion is a combination of both Early- and Late-Fusion. Similar to Early-Fusion, the embeddings of selected modalities are merged right before they are passed to a task-performing model, while other modalities are directly processed. At the end, the outputs of all models are combined like it is done in the Late-Fusion paradigm. This approach allows to exploit the benefits of both paradigms, although it results in a more complex model (Atrey et al., 2010; Yang et al., 2022). Many recent methods build on multimodal training paradigms such as Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), which aims to align visual and textual representations through contrastive learning.

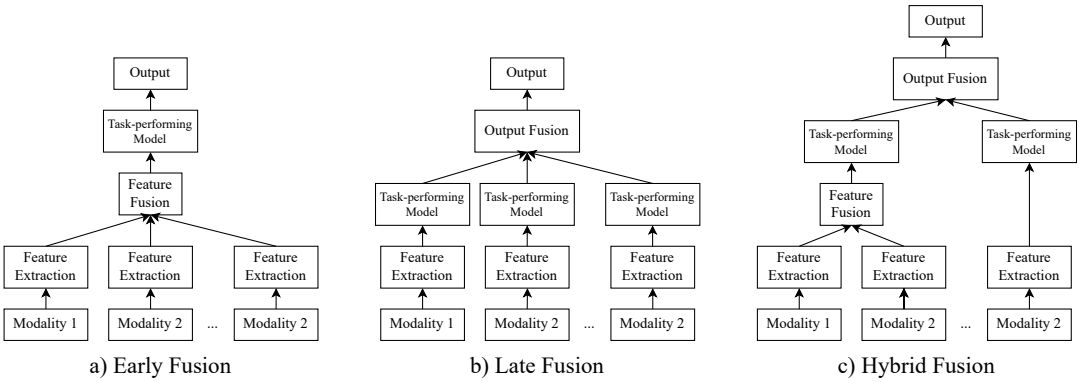


Figure 4: Multimodal Fusion Paradigms (Yang et al., 2022). (a) In Early-Fusion, the model combines the created embeddings immediately after their extraction. In the Late-Fusion approach (b), each modality embedding is assigned its own task-performing deep learning model. Subsequently, the results are integrated. Hybrid-Fusion (c) is a combination of both approaches.

Contrastive learning utilizes both similar and dissimilar instances to capture their differences and similarities (Hu et al., 2024). This has inspired medical adaptations that pair imaging data with clinical reports for tasks like image retrieval and zero-shot classification (Zhao et al., 2025). Likewise, BERT-based models (Devlin et al., 2019) are used to encode text-based modalities such as EHR documentation and questionnaire responses (AlSaad et al., 2024). However, most MMFMs are data-hungry, relying on large, well-annotated datasets and assuming that all modalities are present, which is often unrealistic in practice (Zhan et al., 2025). Recent works like TIP (Du et al., 2024) aim to address the issue of missing data by randomly masking the tabular data and subsequently pre-training the MMFM to reconstruct the missing parts, thereby learning the correlation between table entries. In contrast, STiL (Du et al., 2025) is designed specifically to tackle the challenge posed by sparsely labeled data by employing semi-supervised learning and extracting both shared and modality-specific representations.

2.2 Data Augmentation

Data augmentation is widely regarded as one of the most essential and broadly adopted strategies in modern machine learning (Feng et al., 2021; Zha et al., 2025). It performs various transformations on the training data to generate synthetic samples, thereby producing a larger and more diverse dataset without the need for labor-intensive collection of additional real data. This facilitates the ability of machine learning models to identify the features that are truly relevant, thereby improving their robustness to minor variations in the data.

Wang et al. (2025b) defined three levels of granularity on which data augmentation techniques can operate, namely the single-instance level, the multi-instance level, and the dataset level. Single-instance level augmentation produces additional data samples by introducing perturbations, such as noise addition, to individual instances. As a result, the augmented samples typically occupy positions in the feature space that are proximate to their corresponding original samples. In contrast, Multi-instance level augmentation takes multiple data samples into account. This may be beneficial, since the resulting samples typically lie between the original instances and thus have the potential to occupy portions of the feature space that remain relatively unexplored by the original training dataset. Dataset-level augmentations learn the data distribution of the entire training dataset and then create new data samples using algorithms like Auto-Encoders (AE) (Li et al., 2023) or Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). These approaches are especially effective for mitigating class imbalance, as they enable the generation of additional samples for underrepresented classes. However, these samples will also lie inside the training data distribution, since the generative models are trained to replicate samples from it.

Beyond this taxonomy, data augmentation techniques can also be categorized according to the point at which they act, resulting in two main groups: input space augmentation and latent embedding space augmentation.

2.2.1 Input Space Augmentation

Input space augmentation techniques transform incoming training data before it is processed by the model. These techniques are very popular as they are the most intuitive approach of augmentation (Mumuni and Mumuni, 2022). It allows to transform the data to directly mimic domain-specific variations that happen in real life. Therefore, libraries such as Albumentation (Buslaev et al., 2020) offer a variety of methodologies, in this case tailored for images. Each modality has its own set of augmentation techniques. This section focuses on methods for image and tabular data. For image data, the spatial relationship between pixels is important. Image augmentation techniques can be classified as traditional or advanced (Khalifa et al., 2022; Mumuni and Mumuni, 2022). Traditional methods are the easiest to implement and therefore the most commonly used. They can be further divided into geometric transformations and photometric transformations. A selection of such methods is shown in Fig. 5. Geometric augmentation techniques apply transformations that translate the image without altering the original pixel values. This includes methods such as rotation, flipping, scaling, and translation. These augmentations allow the model to learn task-relevant features, independent of their initial positioning and perspective in the original training data. However, these methods could also potentially remove relevant features in the process. In contrast, photometric image augmentation techniques generate new data samples by modifying only the pixel values of the original image.

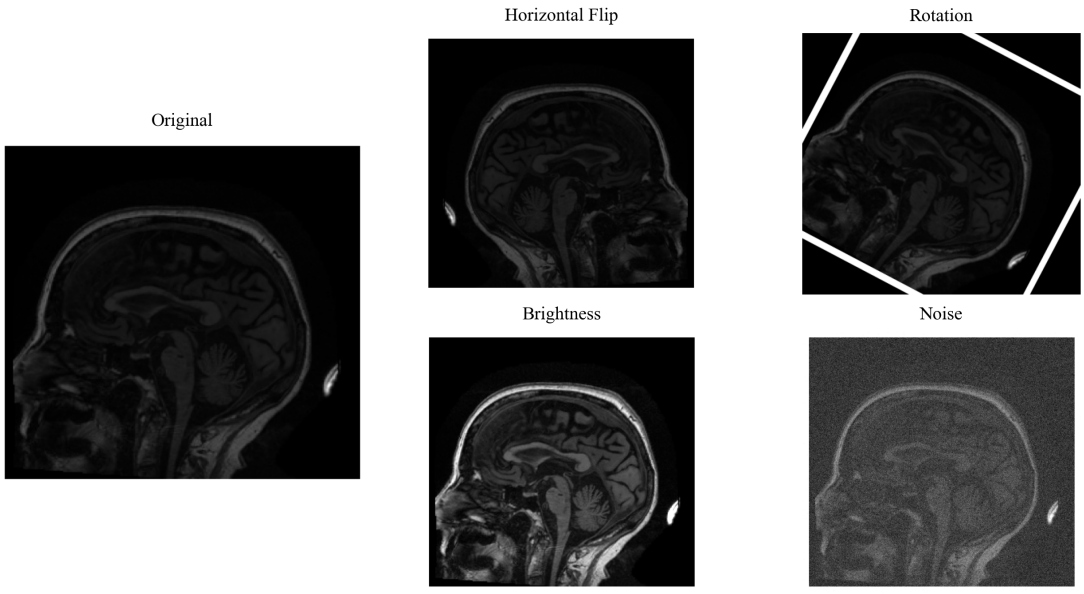


Figure 5: Visualization of different image augmentation strategies. On the left, you can see the original image (taken from the ADNI Dataset (Mueller et al., 2006)). Horizontal Flip and Rotation are typical examples of geometric image augmentations, while adjusting brightness and adding noise represent photometric augmentations.

These methods aim to create diversity in visual characteristics like brightness, saturation, hue and contrast. The main goal here is to mimic variations originating from differing sensors or environmental conditions. Nevertheless, these transformations could also render task-relevant features unrecognizable if they are applied excessively. Advanced techniques encompass more sophisticated approaches, such as region-based methods like Cutout (Devries and Taylor, 2017), Image-Mixing (Lewy and Mańdziuk, 2022), and GAN based methods. Region-based methods only alter certain regions, such as arbitrarily selected squares in the image. Image-mixing techniques involve merging multiple images by applying patches from one image to another, or blending features from one image to create a composite. These methods can strengthen a model's robustness to domain shifts, as they can combine images originating from different sources. However, this also introduces a risk of label misalignment.

Tabular data, on the other hand, has a different structure than images. It is constructed by rows and columns. Each row in the table represents an individual data sample, while the columns represent data attributes. Thus, the order in which rows or columns are listed is inconsequential for the data samples contained within the table. Additionally, tables can hold heterogeneous data such as continuous and categorical attributes. These characteristics complicate the development of data augmentation techniques for tabular data (Cui et al., 2024). Consequently, augmenting tabular data by applying it directly to the source is mostly limited to corrupting individual cells or generating entirely new rows. For instance, Bahri et al. (2022) introduced corruption by randomly replacing cells with values drawn from its features' marginal distribution. In other words, a value from the same column but from another row with an uniform distribution is selected. Yoon et al. (2020) followed a similar approach. First, they created multiple views of their tabular data by applying different masks to it. Subsequently, a trained encoder processes these views and attempts to reconstruct the missing values. The recently generated data samples will then act as the augmented data samples. While the approach is used to generate individual cells, it is also possible to create entire artificial rows by utilizing architectures like GANs or AEs (Sauber-Cole and Khoshgoftaar, 2022). However, the unique structure of tabular data poses a challenge in this regard. For instance, categorical variables cannot be recreated as easily as continuous variables by GANs because their components return continuous values. This is because both types of components must be differentiable for training (Engelmann and Lessmann, 2021). Similarly, variational autoencoders encounter challenges with categorical data due to the same reason. Additionally, these architectures tend to generate implausible data samples for underrepresented classes as their embedding spaces tend to collapse to a limited range (Somvanshi et al., 2024).

Input augmentation techniques are a powerful tool for increasing dataset diversity. However, these methods are mainly focused on singular modalities (Liu et al., 2023). As a result, applying them independently to different modalities in an MMFM may disrupt the semantic alignment of a data sample, as, for instance, an augmented

image may no longer reflect the features present in its corresponding (un-)augmented tabular data sample (Hao et al., 2023).

2.2.2 Latent Embedding Space Augmentation

Since MMFMs operate in latent embedding space, augmenting embeddings provides an alternative to input-level data augmentation (Kumar et al., 2019). According to Bengio et al. (2013) and Ozair and Bengio (2014), this approach may be more beneficial than input space augmentation techniques, as it may provide a higher degree of freedom. They claim that augmentations at this stage are less likely to generate data samples with incoherent features that would severely deteriorate model training. In their work, they analyzed the manifolds of the data at the input and latent embedding levels. They observed that the manifold at the latent embedding level was flattened and thus wider than the manifold at the input level. This is illustrated in Fig. 6. Mathematically, a manifold describes a set of points, with each point having a neighborhood around it (Goodfellow et al., 2016). This means that one can apply transformations on a data point to move inside the manifold. The manifold assumption (Chapelle et al., 2010) posits that high-dimensional data of interest lies near a low-dimensional manifold embedded in the higher-dimensional space. As a result, meaningful variations in the data would occur along this manifold, and nearby points on the manifold correspond to similar data instances. Consequently, points far from the manifold would be unlikely to correspond to realistic data. Thus, a broader manifold would offer augmentation techniques additional space for new data points, which would correspond to realistic input data.

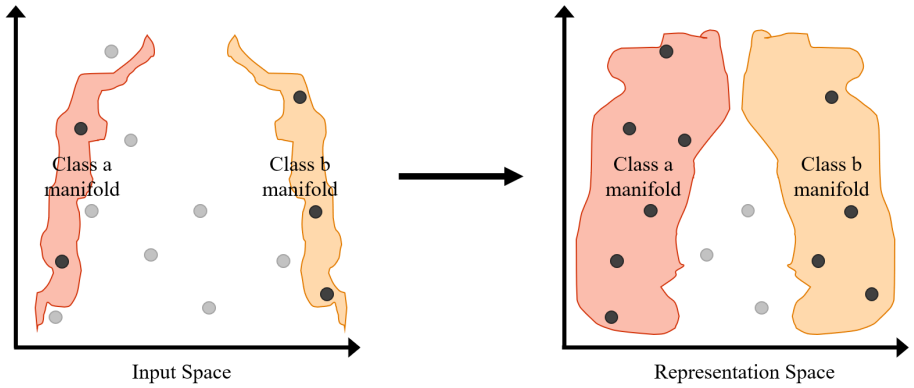


Figure 6: Visualization of the manifold unfolding effect, documented by Bengio et al. (2013). The shapes represent the manifold of each class, meaning they represent the space of coherent samples of that class. The black dots represent augmented samples containing coherent features, and the gray dots represent samples with unrealistic features. Their results indicate that the manifolds become less twisted in the embedding space, occupying more volume and thereby allowing for additional data samples without unrealistic properties.

By performing augmentation in the latent embedding space, it becomes possible to apply transformations that would be inappropriate or semantically destructive at the input level, yet still generate new samples that remain aligned with their corresponding target labels. An illustration of this would be interpolation between images (Ozair and Bengio, 2014). At the input level, this would result in an image that resembles an overlap of the two original samples, creating an artificial and unnatural appearance. For instance, interpolating between medical images would result in an image that looks like two patients have been scanned at once. However, at the latent embedding level, this is a widely known technique. One prominent example is SMOTE (Chawla et al., 2002), which is designed to address class imbalance by generating synthetic samples for underrepresented classes. It achieves this by interpolating between the latent embeddings of two samples belonging to the same minority class, thereby producing new embeddings and resulting in a more balanced training dataset. Other approaches use affine transformations, such as rotation, scaling, and translation, and have been applied to augment the feature maps of CNNs (Shen et al., 2016). Similar to input space augmentation, noise perturbation is a common technique for augmenting latent embeddings. Kurata et al. (2016) applied both additive and multiplicative noise, sampled from an uniform distribution, to the hidden vectors and memory cells of an encoder-decoder LSTM architecture. Similarly, DeVries and Taylor (2017) also augmented the latent embeddings of an LSTM-based sequence AE by applying additive noise, sampled from a Gaussian distribution. They also tested the effects of interpolation and extrapolation and found that extrapolation produced better results for datasets with more complex decision boundaries. More sophisticated methods, such as Neural Style Transfer (Gatys et al., 2016), aim to generate new data samples by mixing the defining features of latent embeddings. A state-of-the-art example in this regard is MixStyle (Zhou et al., 2021). It combines the feature statistics of two data samples.

However, these latent embedding space augmentation techniques are not based on domain knowledge. Consequently, these techniques may alter or introduce attributes that are not beneficial for the current task (Mumuni and Mumuni, 2022). Furthermore, relative to input space augmentation, this remains a sparsely researched area, despite its potential (Liu et al., 2023).

2.3 Mixture of Experts

Today, large deep learning models are trained on extensive datasets that include diverse data structures, such as multiple modalities, and complex relationships among them. Learning these correlations is a difficult task. As datasets grow larger and models become more complex, the training process becomes more demanding, resulting in rising computational costs that can exceed improvements in hardware performance (Thompson et al., 2020). Furthermore, the hypothesis space of a given task may exceed the capacity of a single model, potentially resulting in sub-optimal performance (Dietterich, 2000). One possible approach to address both of these challenges simultaneously is the Mixture of Experts (MoE) framework. Introduced

by Jacobs et al. (1991), this approach trains classifiers specialized in different regions of the input space or sub-tasks, activating only relevant parameters for each input token and thereby increasing model capacity. This framework has been attracting growing interest due to its integration into the transformer architecture (e.g. GLaM (Du et al., 2022) and Switch Transformers (Fedus et al., 2022)), and this approach has been adopted in several recent large language models, including DeepSeek-V2 (DeepSeek-AI, 2024), Mixtral 8×7B (Jiang et al., 2024), the GPT-oss family (OpenAI, 2025), and the Gemini 2.X family (Comanici et al., 2025). The latter of which supports text, vision, and audio.

2.3.1 MoE Architecture

The MoE framework is comprised of two key components: namely, the experts and the gating network (Mu and Lin, 2025). The interaction between these components is illustrated in Fig. 7. (1) The model includes multiple classification units $\{f_1, f_2, \ldots, f_K\}$, also referred to as experts, rather than a single classifier. K denotes the total number of experts present in the framework. Although MoE supports heterogeneous experts, in practice, experts often use a shared architecture and training algorithm, naturally specializing in distinct data clusters through exposure to different inputs (Chen et al., 2022). Typically, they are implemented as small MLPs comprising two linear layers with a ReLU-activation layer between them (Cai et al., 2025). The activation of an individual expert is controlled by the gating network G (Mu and Lin, 2025). (2) The gating network is crucial as it controls the experts' utilization. Usually, it is implemented as a simple linear layer, followed by a softmax activation function (Cai et al., 2025). It learns to produce weights $G(\boldsymbol{x})$ that are used to compute a weighted sum of the experts' outputs.

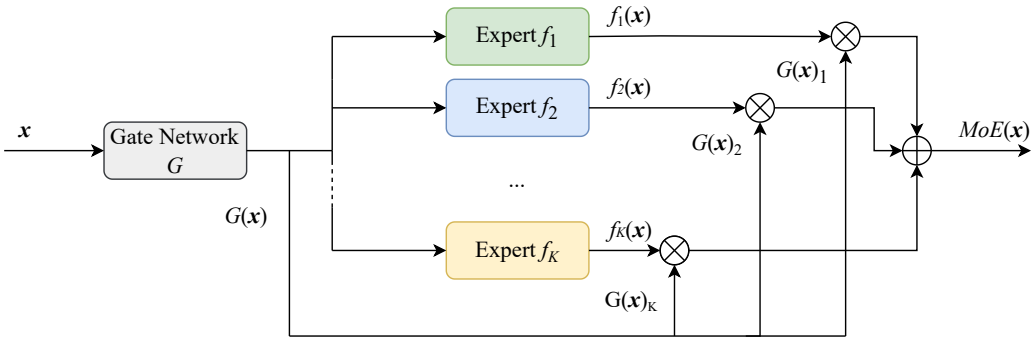


Figure 7: Basic Mixture-of-Experts Architecture (Mu and Lin, 2025). Here, the input tokens x are received by the gate G, which then determines the experts f_i responsible for each token. Additionally, the outputs of the experts are weighted and then combined using weights $G(x)$ learned by the gating network.

Therefore, the MoE framework can be mathematically described as follows:

$$\mathrm{MoE}\left(\boldsymbol{x}\right) = \sum_{i=1}^{K} G\left(\boldsymbol{x}\right)_i f_i\left(\boldsymbol{x}\right). \tag{1}$$

When a gating weight is zero, the corresponding expert does not need to compute an output, saving computational resources. Depending on the deployed routing mechanism, it also controls how input data is directed to the experts. The gating network can decide how many experts may be simultaneously activated to process a given input token. While each input token is processed by all available experts in the original setting of Jacobs et al. (1991), Shazeer et al. (2017) introduced a sparsely gated version of the MoE framework. Here, the gate routes each input token to only the top-k experts from a total of K experts. Thus, their sparse gating function is defined as:

$$G\left(\boldsymbol{x}\right) = \mathrm{softmax}\left(\mathrm{TopK}\left(g\left(\boldsymbol{x}\right) + \mathcal{R}_{noise}, k\right)\right), \tag{2}$$

$$\mathrm{TopK}\left(\boldsymbol{v}, k\right)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } \boldsymbol{v}, \\ -\infty & \text{otherwise.} \end{cases} \tag{3}$$

$g\left(x\right)$ denotes the gate's original output before softmax and top-k is applied. If $k = K$, the gating function can be considered as dense. All values that are not selected by the top-k operation are set to $-\infty$, so that they return 0 in the softmax operation. Additionally, noise, denoted by $\mathcal{R}_{noise}$, is added to the gate's original output. The type of noise sampled varies in literature. For instance, Shazeer et al. (2017) applied Gaussian noise $\mathcal{R}_{noise} \sim \mathcal{N}\left(0, 1\right)$ controlled by the term $\zeta\left(x \cdot W_{noise}\right)$, with W_{noise} being a trainable parameter. Meanwhile, Fedus et al. (2022) applied a noise sampled from an uniform distribution $\mathcal{R}_{noise} \sim \mathcal{U}\left(1 - \epsilon, 1 + \epsilon\right)$ with $\epsilon = 0.01$. This is applied because the MoE framework may collapse to using only a single expert during training, as the gating function may be optimized to favor a single expert. Consequently, the noise encourages the model to explore other experts as well. Furthermore, it is also possible to switch the order of the softmax and the top-k operation. Applying top-k before softmax breaks the normalization effect of the softmax output, resulting in an invalid probability distribution and potentially disproportionate scaling of expert contributions (Mu and Lin, 2025). Nevertheless, Shazeer et al. (2017) have not observed any deterioration in their results using their original order.

The MoE framework introduces additional hyperparameters, namely k and K, which have to be considered when designing a MoE-based model. For the Switch Transformer, Fedus et al. (2022) reported that the performance of their model was enhanced as the number of experts K increased. However, the way of input routing also has an effect. Fan et al. (2024) observed that, for small-scale GPT-2 models,

performance improves with increasing k when input tokens are routed at the sequence level, whereas higher values of K are more beneficial when tokens are routed individually.

2.3.2 MoE Loss

As mentioned in Section 2.3.1, noise is added to the sparse-gating function's original output, before it is processed by top-k and softmax. The goal of this perturbation is to mitigate the favoritism of a small subset of experts during training. Otherwise, the excluded experts are not properly trained and the model cannot fully utilize their contributions. An additional way to prevent this is to introduce auxiliary loss terms alongside the model's overall loss function. Shazeer et al. (2017) incorporated two auxiliary loss terms in their training, namely $\mathcal{L}_I$ and $\mathcal{L}_L$. $\mathcal{L}_I$ aims to keep the individual importance of the experts balanced. It is calculated as:

$$I(\mathcal{X}) = \sum_{\boldsymbol{x} \in \mathcal{X}} G(\boldsymbol{x}), \tag{4}$$

$$\mathcal{L}_I(\mathcal{X}) = w_I \cdot CV(I(\mathcal{X}))^2. \tag{5}$$

$\mathcal{X}$ refers to the current batch and CV denotes the coefficient of variation (Brown, 1998), which is defined as:

$$CV(\boldsymbol{x}) = \frac{\sigma(\boldsymbol{x})}{\mu(\boldsymbol{x})}. \tag{6}$$

As the CV is the ratio of the standard deviation and the mean value, it equals zero if all inputs have the same value. Thus, minimizing $\mathcal{L}_I$ encourages the gate to return similar weights for each expert. This ensures that each expert contributes approximately equally to the final result. w_I controls the contribution of $\mathcal{L}_I$ to the overall loss. $\mathcal{L}_L$ aims to encourage the gate to dispatch equal amounts of training input to each expert. It is calculated as:

$$L(\mathcal{X})_i = \sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{x}, i), \tag{7}$$

$$\mathcal{L}_L(\mathcal{X}) = w_L \cdot CV(L(\mathcal{X}))^2. \tag{8}$$

w_L denotes the hyperparameter controlling the influence of $\mathcal{L}_L$. $P(\boldsymbol{x}, i)$ denotes the probability that the weight of the i-th expert $G(\mathcal{X})_i$ is nonzero, given new noise has been sampled for the i-th expert, while the noise sampled for all other experts remains the same.

It is calculated as:

$$g_{Noise}(\boldsymbol{x}) = g(\boldsymbol{x}) + \mathcal{R}_{noise}, \tag{9}$$

$$P(\boldsymbol{x}, i) = \Pr\left(g_{Noise}(\boldsymbol{x})_i > kth\_excluding\left(g_{Noise}(\boldsymbol{x}), k, i\right)\right). \tag{10}$$

$G(\mathcal{X})_i$ is considered nonzero only if $g_{Noise}(\boldsymbol{x})_i$ exceeds the top-k elements of $g_{Noise}(\boldsymbol{x})$ excluding itself, given by $kth\_excluding(g_{Noise}(\boldsymbol{x}), k, i)$. Thus, it is only considered if the corresponding expert would have actually been selected in the top-k setting.

Here, $\mathcal{L}_I$ and $\mathcal{L}_L$ influence load and importance balancing separately. Fedus et al. (2022) introduced a simplified combination of both terms in the Switch Transformer. However, this simplification is explicitly tailored for sparse gating functions with $k = 1$. The loss term, which will be called $\mathcal{L}_{\text{Switch}}$ in this work, is defined as:

$$\ell_i = \frac{1}{N_T} \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{1}\left\{\text{argmax}\left(p(\boldsymbol{x})\right) = i\right\}, \tag{11}$$

$$P_i = \frac{1}{N_T} \sum_{\boldsymbol{x} \in \mathcal{X}} p_i(\boldsymbol{x}), \tag{12}$$

$$\mathcal{L}_{\text{Switch}} = w_{Switch} \cdot K \cdot \sum_{i=1}^{K} \ell_i \cdot P_i. \tag{13}$$

In this loss function, the load of the expert i is represented by ℓ_i and its importance by P_i. In ℓ_i, the function first counts how many input tokens x are assigned to expert i, with $\mathbb{1}$ denoting the indicator function, which returns 1 when the condition is met and 0 otherwise. Then, it normalizes this value by dividing by the total amount of tokens N_T in the batch $\mathcal{X}$. In P_i, it first sums the probabilities of input token $\boldsymbol{x}$ being routed to expert i, denoted by $p_i(x)$, and then normalizes the sum in the same manner as in ℓ_i. The overall loss $\mathcal{L}_{\text{Switch}}$ is weighted by w_{Switch}.

However, according to Mu and Lin (2025), while this approach is more efficient, it achieves this efficiency at the cost of a minor drop in accuracy.

3 Methods

3.1 Overview and Notation

This work investigates the impact of latent embedding space augmentations and MoE on the performance of STiL under data distribution shift. Augmentation techniques are applied at the internal classifiers after disentangling modality-specific and shared representations. The integration of MoE is examined in two variants: classifier output weighting and embedding-level expert routing.

Let $\mathcal{X} = \{(\boldsymbol{x}^i, \boldsymbol{x}^t), \boldsymbol{y}\}^{B_l} + \{(\boldsymbol{x}^i, \boldsymbol{x}^t)\}^{B_u}$ be a batch consisting of labeled and unlabeled data. $\boldsymbol{y}$ denotes a one-hot ground-truth label. B_l denotes the amount of labeled data samples and B_u the amount of unlabeled samples. Thus, the batches have the size $B = B_u + B_l$. Images are represented as pixel matrices $\boldsymbol{x}^i \in \mathbb{R}^{H \times W \times C}$, with H, W, and C denoting height, width, and number of channels, respectively. The tabular data sample $\boldsymbol{x}^t = (x_i^t)_{i=1}^n$ consists of n attributes, with n_{cat} categorical and $n_{con} = n - n_{cat}$ continuous attributes. Covariate shift can arise in either a single modality, i.e. $p(x^t_{train}) \neq p(x^t_{test})$ or $p(x^i_{train}) \neq p(x^i_{test})$, or in both. The datasets consist of a set of $\mathcal{C}$ unique labels, namely $\mathcal{Y}$.

3.2 STiL

The STiL model follows the hybrid fusion paradigm and consists of four modules: modality-specific encoders, the disentangled contrastive consistency (DCC) module, the consensus-guided pseudo-labeling (CGPL) module, and the prototype-guided label smoothing (PGLS) module. Its architecture is shown in Fig. 8.

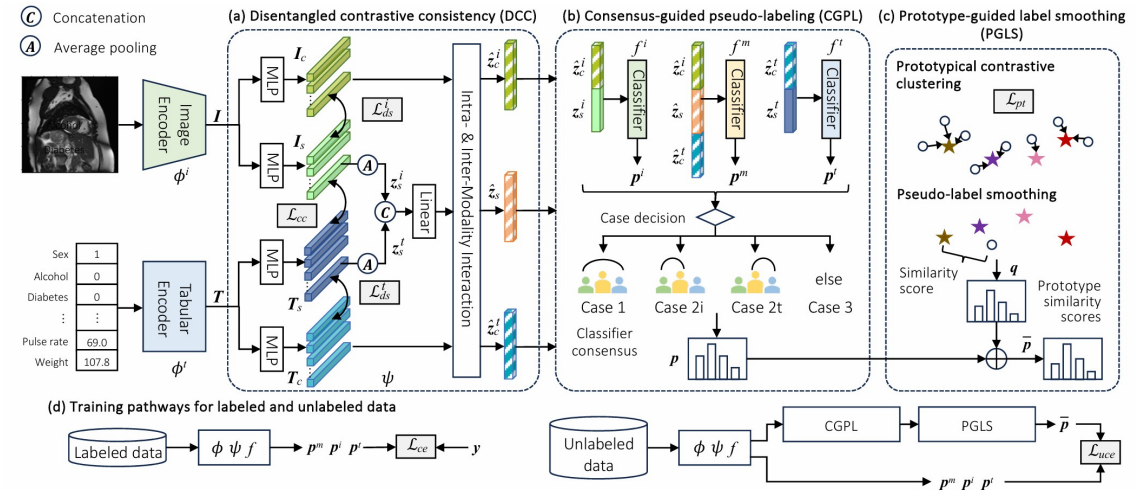


Figure 8: The STiL model architecture (Du et al., 2025). The DCC module (a) disentangles modality-specific and -shared information. The CGPL module (b) generates pseudo-labels for unlabeled samples, which are refined by the PGLS module (c). The separate training pipelines for labeled and unlabeled data are shown in (d).

3.2.1 Disentangled Contrastive Consistency

The first module of the STiL architecture is designed to extract and separate the modality information parts from the input data. At first, it creates latent embeddings of the input data using the encoders ϕ^i and ϕ^t. ϕ^i consists of a pre-trained ResNet-50 (He et al., 2016), while ϕ^t is a transformer-based encoder, which is able to handle both categorical and continuous data (Du et al., 2024). The corresponding Embeddings $\boldsymbol{I}$ and $\boldsymbol{T}$ are then processed by two MLPs each, which will output modality-specific embeddings ($\boldsymbol{z}_c^i$ and $\boldsymbol{z}_c^t$) and modality-shared embeddings ($\boldsymbol{z}_s^i$ and $\boldsymbol{z}_s^i$). These MLPs are trained using the Contrastive Log-ratio Upper Bound (CLUB) loss (Cheng et al., 2020) ($\mathcal{L}_{ds}^t$ and $\mathcal{L}_{ds}^t$), which is designed to minimize the mutual information between the resulting modality-specific ($\boldsymbol{I}_C$, $\boldsymbol{T}_C$) and modality-shared ($\boldsymbol{I}_S$, $\boldsymbol{T}_S$) embeddings. Additionally, the loss $\mathcal{L}_{cc}$, which is based on InfoNCE (van den Oord et al., 2018), is used to learn the modality-shared embeddings. Average pooling over the sequence dimension is applied to standardize the resulting embedding representations. The emerging embeddings $\boldsymbol{z}_s^i$ and $\boldsymbol{z}_s^t$ are then processed by the projection heads $proj^i$ and $proj^t$. The projection heads facilitate the InfoNCE loss objective by mapping the embeddings to a subspace where contrastive learning works more effectively (Gupta et al., 2022). Finally, the *Intra- & Inter Modality Interaction* component is used to purify the modality-specific and the projected modality-shared embeddings to $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}_c^t$, and $\hat{\boldsymbol{z}}_s$. This component consists of a transformer layer that employs self-attention on $\boldsymbol{z}_c^i$ and $\boldsymbol{z}_c^t$. To extract $\hat{\boldsymbol{z}}_s$, it applies cross-attention on $\boldsymbol{z}_s^i$ and $\boldsymbol{z}_s^t$. Overall, the loss of the DCC module is calculated as:

$$\mathcal{L}_{dcc} = \beta \mathcal{L}_{cc} + \gamma \left(\mathcal{L}_{ds}^i + \mathcal{L}_{ds}^t \right). \tag{14}$$

where β and γ weight the contribution of the respective loss terms. These terms either maximize the lower bound (for $\boldsymbol{I}_S$–$\boldsymbol{T}_S$) or minimize the upper bound (for $\boldsymbol{I}_C$–$\boldsymbol{I}_S$ and $\boldsymbol{T}_S$–$\boldsymbol{T}_C$) on the mutual information between the embeddings. The mutual information cannot be calculated precisely here, as it is intractable (Zhang et al., 2024). $\mathcal{L}_{cc}$ is defined as:

$$\mathcal{L}_{cc} = -\frac{1}{2B} \sum_{b=1}^{B} \left(\text{sim}\left(\boldsymbol{z}_{s_b}^i, \boldsymbol{z}_{s_b}^t \right) + \text{sim}\left(\boldsymbol{z}_{s_b}^t, \boldsymbol{z}_{s_b}^i \right) \right), \tag{15}$$

$$\text{sim}\left(\boldsymbol{z}_{s_b}^i, \boldsymbol{z}_{s_b}^t \right) = \log \left(\frac{\Psi\left(proj^i\left(\boldsymbol{z}_{s_b}^i \right), proj^t\left(\boldsymbol{z}_{s_b}^t \right) \right)}{\sum_{k=1}^{B} \Psi\left(proj^i\left(\boldsymbol{z}_{s_b}^i \right), proj^t\left(\boldsymbol{z}_{s_k}^t \right) \right)} \right). \tag{16}$$

With $\Psi(\cdot,\cdot) = \exp\left(\cos(\cdot,\cdot)/\kappa\right)$ and κ as the temperature parameter, $\text{sim}(\cdot,\cdot)$ is the log-softmax of the cosine similarities between the vectors $\boldsymbol{z}_{s_b}^i$ and $\boldsymbol{z}_{s_b}^t$. With this, $\mathcal{L}_{cc}$ represents the contrastive cross-entropy of the cosine similarities. Minimizing this loss maximizes a lower bound on mutual information, leading to embeddings that encode shared information (van den Oord et al., 2018). For the modality-specific embeddings, the losses $\mathcal{L}_{ds}^i$ and $\mathcal{L}_{ds}^t$ are defined as:

$$\mathcal{L}_{ds}^{i} = I_{vCLUB}\left(\boldsymbol{z}_c^i, \boldsymbol{z}_s^i\right) - \mathcal{L}_{q_\theta}\left(\boldsymbol{z}_c^i, \boldsymbol{z}_s^i\right), \tag{17}$$

$$\mathcal{L}_{ds}^{t} = I_{vCLUB}\left(\boldsymbol{z}_c^t, \boldsymbol{z}_s^t\right) - \mathcal{L}_{q_\theta}\left(\boldsymbol{z}_c^t, \boldsymbol{z}_s^t\right). \tag{18}$$

These losses aim to minimize the higher bound of mutual information between their corresponding inputs. With

$$I_{vCLUB}\left(\boldsymbol{a}, \boldsymbol{b}\right) = \frac{1}{B^2}\sum_{j=1}^{B}\sum_{k=1}^{B}\left[\log\left(q_\theta\left(\boldsymbol{b}_j|\boldsymbol{a}_j\right)\right) - \log\left(q_\theta\left(\boldsymbol{b}_k|\boldsymbol{a}_j\right)\right)\right], \tag{19}$$

they estimate the upper bound by calculating the difference between the log-likelihood of matching sample pairs and the log-likelihood of mismatching sample pairs (Cheng et al., 2020). However, this calculation would require the conditional distribution between the samples to be known, which is unobtainable in STiL (Du et al., 2025). Therefore, it is estimated with the MLP $q_\theta\left(\boldsymbol{b}|\boldsymbol{a}\right)$, which is trained using:

$$\mathcal{L}_{q_\theta}\left(\boldsymbol{a}, \boldsymbol{b}\right) = \frac{1}{B}\sum_{j}^{B}\log\left(q_\theta\left(\boldsymbol{b}_j|\boldsymbol{a}_j\right)\right). \tag{20}$$

3.2.2 Consensus-guided Pseudo-Labeling

The CGPL module aims to derive a prediction from the embeddings created in DCC (Du et al., 2025). It consists of three classifiers. f^m receives the concatenation of $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}_s$, and $\hat{\boldsymbol{z}}_c^t$ as input and acts therefore as a multimodal classifier. f^i classifies $\hat{\boldsymbol{z}}_c^i$ concatenated with $\boldsymbol{z}_s^i$, which makes it the image-data classifier. Finally, f^t receives the concatenation of $\hat{\boldsymbol{z}}_c^t$ and $\boldsymbol{z}_c^t$ and acts as the tabular-data classifier. These classifiers return the logits $\boldsymbol{p}_m$, $\boldsymbol{p}_i$, and $\boldsymbol{p}_t$. During testing and at labeled training samples, performance evaluation only considers the output of f^m. However, if the training sample is unlabeled, it will create a pseudo-label $\boldsymbol{p}$ following a rule-based approach with four cases:

1. Case 1: All classifiers predict the same class.
2. Case 2_i: f^m and f^i predict the same class.
3. Case 2_t: f^m and f^t predict the same class.
4. Case 3: None of the above, only consider f^m

The classifiers are trained using the cross-entropy loss. For labeled training samples, the loss is calculated as:

$$\mathcal{L}_{ce} = H\left(\boldsymbol{p}^m, \boldsymbol{y}\right) + H\left(\boldsymbol{p}^i, \boldsymbol{y}\right) + H\left(\boldsymbol{p}^t, \boldsymbol{y}\right), \tag{21}$$

where $H(\cdot,\cdot)$ denotes the cross-entropy of the respective classifier. However, for unlabeled training samples, a selective update process is used, as this approach serves to reduce the probability of the classifiers erroneously agreeing on an incorrect label (Du et al., 2025). Depending on the current case, the cross-entropy of only a subset of the classifiers will be calculated. Since no ground-truth label $\boldsymbol{y}$ is available, the cross-entropy will be calculated using the refined logits $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{p}}^m$ obtained from the PGLS module (see Section 3.2.3). Additionally, only predictions with a confidence level above a threshold τ are considered in the calculation, ensuring that the model does not learn from unreliable guesses. Therefore, the loss for unlabeled data samples is defined as:

$$\mathcal{L}_{uce} = \frac{1}{B_U} \sum_{b=1}^{B_U} \mathbb{1}\left\{\max\left(\bar{\boldsymbol{p}}_b^m \geq \tau\right)\right\} \mathcal{L}\left(\boldsymbol{p}_b^m, \boldsymbol{p}_b^i, \boldsymbol{p}_b^t, \bar{\boldsymbol{p}}_b\right), \tag{22}$$

$$\mathcal{L}\left(\boldsymbol{p}_b^m, \boldsymbol{p}_b^i, \boldsymbol{p}_b^t, \bar{\boldsymbol{p}}_b\right) = \begin{cases} \text{Case 1}, & H\left(\boldsymbol{p}^m, \bar{\boldsymbol{p}}\right) + H\left(\boldsymbol{p}^i, \bar{\boldsymbol{p}}\right) + H\left(\boldsymbol{p}^t, \bar{\boldsymbol{p}}\right) \\ \text{Case } 2_i, & H\left(\boldsymbol{p}^t, \bar{\boldsymbol{p}}\right) \\ \text{Case } 2_t, & H\left(\boldsymbol{p}^i, \bar{\boldsymbol{p}}\right) \\ \text{Case 3}, & \text{choose randomly } H\left(\boldsymbol{p}^i, \bar{\boldsymbol{p}}\right) \text{ or } H\left(\boldsymbol{p}^t, \bar{\boldsymbol{p}}\right) \end{cases} \tag{23}$$

3.2.3 Prototype-guided Label Smoothing

The pseudo-labels generated in Section 3.2.2 are refined by the PGLS module. Therefore, it is only utilized during the training process when unlabeled data is present in the training dataset. The purpose of this module is to make the generated pseudo-labels as reliable as possible (Du et al., 2025). The prototypes reflect the entire training dataset and are computed at the end of each training epoch as the average of all embeddings belonging to each class $c \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of all unique classes in the dataset. The prototypes are defined as:

$$\boldsymbol{v}_c = \frac{1}{n_c} \left(\sum_{y_j^l = c}^{N_l} \boldsymbol{v}_j^l + \sum_{\tilde{y}_k^u = c}^{N_u} \mathbb{1}\left\{\max \bar{\boldsymbol{p}}_k^m \geq \tau\right\} \boldsymbol{v}_k^u \right), \tag{24}$$

$$n_c = \sum_{y_j^l = c}^{N_l} 1 + \sum_{\tilde{y}_k^u = c}^{N_u} \mathbb{1}\left(\max \bar{\boldsymbol{p}}_k^m \geq \tau\right), \tag{25}$$

with N_l and N_u denoting the total amount of labeled and unlabeled data samples in the training dataset. Again, to enhance reliability, the predictions for unlabeled data must exceed the threshold τ for the corresponding embeddings to be considered. These calculations use a projection of the DCC output embeddings into a low-dimension embedding space $\boldsymbol{v} = h\left([\hat{\boldsymbol{z}}_c^i, \hat{\boldsymbol{z}}_s, \hat{\boldsymbol{z}}_c^t]\right)$ as input. Following the manifold

assumption, $\boldsymbol{p}$ and $\boldsymbol{p}^m$ will be refined by adding a prototype similarity score $\boldsymbol{q} = \text{softmax}\left(\left[\boldsymbol{v}_1, ..., \boldsymbol{v}_C\right]^T \boldsymbol{v}\right)$ to them:

$$\bar{\boldsymbol{p}}, \bar{\boldsymbol{p}}^m = r\bar{\boldsymbol{p}} + (1-r)\,\bar{\boldsymbol{q}}, r\bar{\boldsymbol{p}}^m + (1-r)\,\bar{\boldsymbol{q}}. \tag{26}$$

At the end, this module is trained with the loss:

$$\begin{aligned}\mathcal{L}_{pt} = &- \frac{1}{B_l}\sum_{b=1}^{B_l}\sum_{c\in\mathcal{Y}} \mathbb{1}\left\{y_b^l = c\right\} \log \frac{\Psi\left(\boldsymbol{v}_b^l, \boldsymbol{v}_c\right)}{\sum_{c'\in\mathcal{Y}} \Psi\left(\boldsymbol{v}_b^l, \boldsymbol{v}_{c'}\right)} \\ &- \frac{1}{B_u}\sum_{b=1}^{B_u} \mathbb{1}\left\{\max\left(\bar{p}_b^m \geq \tau\right)\right\} \sum_{c\in\mathcal{Y}} \mathbb{1}\left\{\tilde{y}_b^u = c\right\} \log\left(\frac{\Psi\left(\boldsymbol{v}_b^u, \boldsymbol{v}_c\right)}{\sum_{c'\in\mathcal{Y}} \Psi\left(\boldsymbol{v}_b^u, \boldsymbol{v}_{c'}\right)}\right).\end{aligned} \tag{27}$$

Similar to Eq. (16), the log-softmax of the cosine similarities between the prototypes of both labeled and unlabeled data is calculated. This forces them to become more similar to their respective data classes, while also enhancing the distinction between the different classes.

3.2.4 Training

Combining all modules, the complete loss is calculated as:

$$\mathcal{L} = \alpha\mathcal{L}_{ce} + \mathcal{L}_{dcc} + \lambda_p\mathcal{L}_{pt} + \lambda_u\mathcal{L}_{uce}. \tag{28}$$

α, λ_p, and λ_{uce} denote the respective loss contribution weights. As a final measure for training stabilization during unlabeled data, the model is trained as a teacher-student-framework. In such a framework, the actual model of interest (the student) is trained by another model, which can create better predictions (the teacher) (Meng et al., 2019). Here, both models are built with the STiL architecture and the teacher model is updated via exponential moving average (He et al., 2020), which is calculated as:

$$\theta' = m\theta' + (1-m)\,\theta, \tag{29}$$

with θ denoting the model parameters and m denoting the momentum coefficient. Thus, the teacher is updated faster than the student. The model is optimized using the Adam algorithm (Kingma and Ba, 2015) without weight decay. Du et al. (2025) trained the model under two settings: with 10% of the training data labeled and with 1% labeled.

3.3 Latent Embedding Augmentation Techniques

The latent embedding techniques examined in this thesis include linear extrapolation, noise perturbation, linear delta, and MixStyle. In this work, the impact of these methods is examined by augmenting the inputs of the classifiers f^i, f^t, and f^m inside the CGPL module, both individually and all at once.

Linear Extrapolation

Linear extrapolation is one of the simplest and therefore easiest to implement latent embedding augmentation strategies (DeVries and Taylor, 2017). This method simply creates a new embedding in the feature space by calculating:

$$\boldsymbol{z}' = (\boldsymbol{z}_j - \boldsymbol{z}_k)\,\lambda + \boldsymbol{z}_j. \tag{30}$$

λ is a weight that controls the level of extrapolation. λ is in the range of $[0, \infty]$. In order to keep the new embedding aligned with the target label, this approach is used only on data points from the same class. Here, the newly generated sample $\boldsymbol{z}'$ replaces $\boldsymbol{z}_j$.

Noise Perturbation

Noise Perturbation is another simple and widely used method of augmentation (Huang et al., 2025; DeVries and Taylor, 2017; Kumar et al., 2019). Adding noise to the latent embedding allows the model to see more diverse data samples. It is calculated as:

$$\boldsymbol{z}' = \boldsymbol{z} + \delta\mathcal{R}_{noise}, \mathcal{R}_{noise} \sim \mathcal{N}\left\{0, \sigma_{\boldsymbol{z}}^2\right\}. \tag{31}$$

$\mathcal{R}_{noise}$ is drawn from a Gaussian distribution $\mathcal{N}$ with $\mu = 0$ and and per-element variance σ_z^2, estimated from the context vectors. δ controls the magnitude of the added noise. $\boldsymbol{z}$ is replaced by $\boldsymbol{z}'$ inside the batch.

Linear Delta

Similar to linear extrapolation, linear delta (Kumar et al., 2019) aims to create new samples by adding the difference between two data samples to another data sample:

$$\boldsymbol{z}' = (\boldsymbol{z}_i - \boldsymbol{z}_j) + \boldsymbol{z}_k. \tag{32}$$

However, the difference is that linear delta may explore the latent embedding space in a different direction than linear extrapolation, as it is not restricted to the space between two samples. Similarly to linear extrapolation, it is also applied only on data samples of the same class. $\boldsymbol{z}'$ takes the place of $\boldsymbol{z}_k$ in the batch.

MixStyle

MixStyle has been developed by Zhou et al. (2021) and has become a state-of-the-art method in latent embedding augmentation. Its main idea is to create new data instances by mixing the styles of different images. The style of an image is encapsulated in the statistics, namely the mean and standard deviation, of its feature map (Huang and Belongie, 2017). Thus, the original embedding is first normalized to eliminate its statistical properties $\mu(\boldsymbol{z}_i)$ and $\sigma(\boldsymbol{z}_i)$. Subsequently, the blends of the statistical attributes from both embeddings, namely γ_{mix} and β_{mix}, are applied to replace them:

$$\boldsymbol{z}' = \gamma_{mix} \odot \frac{\boldsymbol{z}_i - \mu(\boldsymbol{z}_i)}{\sigma(\boldsymbol{z}_i)} + \beta_{mix} \tag{33}$$

$$\gamma_{mix} = \lambda\, \sigma(\boldsymbol{z}_i) + (1 - \lambda)\, \sigma(\boldsymbol{z}_j) \tag{34}$$

$$\beta_{mix} = \lambda\, \mu(\boldsymbol{z}_i) + (1 - \lambda)\, \mu(\boldsymbol{z}_j). \tag{35}$$

Here, λ is drawn from a beta distribution $\lambda \sim Beta\,(\alpha_{MixStyle}, \alpha_{MixStyle})$ with $\alpha_{MixStyle} \in \{0, \infty\}$ for each data instance specifically. $\boldsymbol{z}_j$ is selected randomly. Finally, $\boldsymbol{z}'$ replaces $\boldsymbol{z}_i$ inside the batch.

3.4 STiL-MoE

In the original setting, only the output of the multimodal classifier f^m is considered at test time, with the unimodal classifiers f^i and f^t being ignored, despite the fact that they are trained. f^m receives the refined embeddings $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}_s$, and $\hat{\boldsymbol{z}}_c^t$, containing the shared and specific information of the modalities. However, if a distribution shift occurs in at least one modality, f^m will exhibit uncertainty. Since $\hat{\boldsymbol{z}}_s$ contains the refined shared information from both modalities, at least two parts of its input are affected. Thus, the idea behind incorporating MoE into STiL is to rely not solely on the multimodal classifier, but also on the other classifiers. This is tested with two variants. In the first variant, the classifiers remain unchanged, with only their outputs being weighted by the gate network. In this work, this variant will be referred to as *STiL-Mixture-of-Fixed-Experts* (STiL-MoFE). In the second variant, the gate determines which classifier processes the individual embedding parts. As it is trained with the training procedure of the Switch-Transformer (Fedus et al., 2022), this approach will be referred to as *STiL-Switch*. These modules replace the CGPL module in the original architecture. However, due to the complexity of the STiL model, both variants are tested only with fully labeled training datasets and the training procedure is switched from Semi-SL to fully supervised learning. Thus, the PGLS module is removed from STiL here. This approach ensures more reliable results by increasing training stability.

3.4.1 STiL-MoFE

In the first variant, the logits $\boldsymbol{p}^m$, $\boldsymbol{p}^i$, and $\boldsymbol{p}^t$ are combined to the final logit $\boldsymbol{p}^{\text{STiL-MoFE}}$ using weights learned from a gate network. An illustration of the STiL-MoFE architecture is shown in Fig. 9. In other implementations of MoE in literature, the gate determines which expert processes which embedding token (Mu and Lin, 2025). However, in this case, the inputs of the classifiers f^i, f^m, and f^t remain fixed and consist of the original embedding components. The gate receives the input of the classifiers. Thus, it ingests the concatenation of $\boldsymbol{z}_s^i$, $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}^s$, $\hat{\boldsymbol{z}}_c^t$, and $\boldsymbol{z}_s^t$. With this, it learns to produce weights for each classifier. The idea behind this approach is to preserve the existing specialization of the classifiers. Here, f^i acts as the expert for image data, f^t as the expert for tabular data, and f^m as an expert that is knowledgeable about both modalities. This approach would prove beneficial when the gate learns which modality contributes more valuable information for the classification task. For instance, if the tabular data does not contribute as much as the image data, the gate could learn to assign a smaller weight to f^t and a higher weight to f^i, thus relying more on the image data. This way, the impact of distribution shifts in the less important modality could be mitigated. The gate is tested both as a simple linear layer and as a MLP.

The STiL-MoFE model is trained using the following loss function:

$$\mathcal{L} = \alpha \mathcal{L}_{ce} + \mathcal{L}_{dcc} + \epsilon \mathcal{L}_{\text{STiL-MoFE}}. \tag{36}$$

Since this model is only tested using a fully labeled training dataset, the loss terms responsible for unlabeled data and prototypes ($\mathcal{L}_{uce}$ and $\mathcal{L}_{pt}$ respectively), are removed.

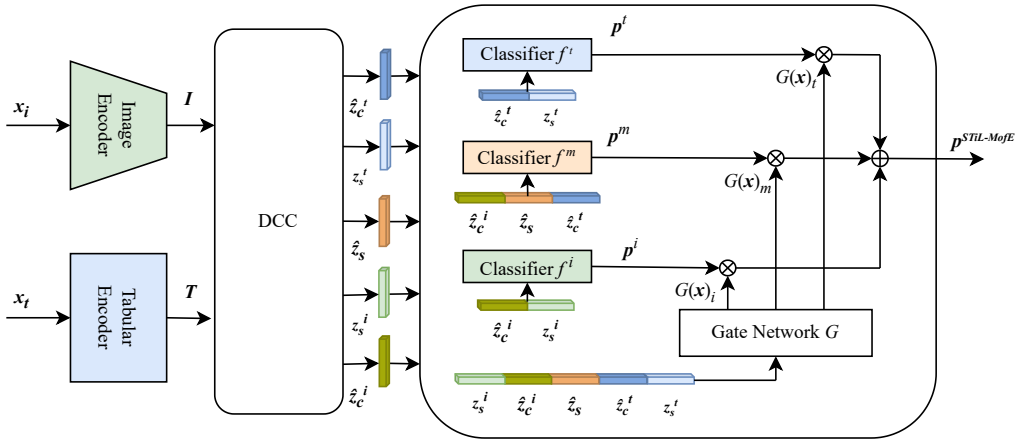


Figure 9: Architecture of the STiL-MoFE variant. Here, the classifiers f^i, f^m, and f^t are unchanged. Their outputs are then combined using weights learned by the gate network. The gate network receives the concatenation of all embedding parts processed by the classifiers.

Instead, the loss term $\mathcal{L}_{\text{STiL-MoFE}}$, which is controlled by the weight ϵ, is introduced. It is calculated as

$$\mathcal{L}_{\text{STiL-MoFE}} = \mathcal{L}_I + H\left(\boldsymbol{p}^{\text{STiL-MoFE}}, \boldsymbol{y}\right). \tag{37}$$

It regulates the weights produced by the gate network as it consists of the importance-loss (Eq. (5)) and the cross-entropy-loss between $\boldsymbol{p}^{\text{STiL-MoFE}}$ and the ground-truth y. This way, the gate is encouraged to learn weights that represent a balance between equal expert utilization and high training accuracy.

The cross-entropy encourages the gate to learn weights that lead to a high training accuracy. However, since the goal of this work is to enhance the robustness against data distribution shifts, this term alone may lead to additional overfitting. It could encourage the model to focus on classifiers that process task-relevant modalities, which can then be shifted. In this case, the performance may be adversely affected rather than improved. Therefore, the importance-loss explained in Section 2.3.2 is also employed. It encourages the gate to produce equal weights for the classifiers, thereby training it not to rely solely on the presumably most important classifier. Since the classifiers always receive a fixed, respective set of inputs, the load-loss term (Eq. (8)) is unnecessary and therefore not used here. $\mathcal{L}_{ce}$ is kept to preserve the classifiers' specialization.

3.4.2 STiL Switch

In the second variant, the implementation follows the Switch-Transformer training procedure. Its architecture is illustrated in Fig. 10. Since the Switch-Transformer is commonly used within transformer-based architectures, the gate usually receives language tokens (Mu and Lin, 2025). In this setting, however, the output embeddings of DCC act as the tokens and the classifiers do not receive a fixed input set. Instead, the gate receives all embedding parts coming from the DCC module and distributes them among the classifiers. Therefore, they are no longer hardcoded to act as fixed experts for specific modalities. Instead, they become experts implicitly through the MoE training process. When the classifier processed their assigned tokens, the results are summarized using average pooling. This way, the contributions of all experts are weighted equally. This approach has two benefits. First, the model gains greater flexibility, as it can now dynamically choose which expert processes the current token. This allows the model to send tokens to more suitable experts individually rather than as a concatenation. Under a distribution shift, the model can route affected tokens to experts who may exhibit greater certainty. Second, since the experts may receive embeddings from both modalities, they can develop a broader understanding of the task, giving it more robustness against distribution shifts. The STiL-Switch module is trained using the following loss function:

$$\mathcal{L} = \alpha H\left(\boldsymbol{p}^{\text{STiL-Switch}}, \boldsymbol{y}\right) + \mathcal{L}_{dcc} + \xi \mathcal{L}_{\text{Switch}}. \tag{38}$$

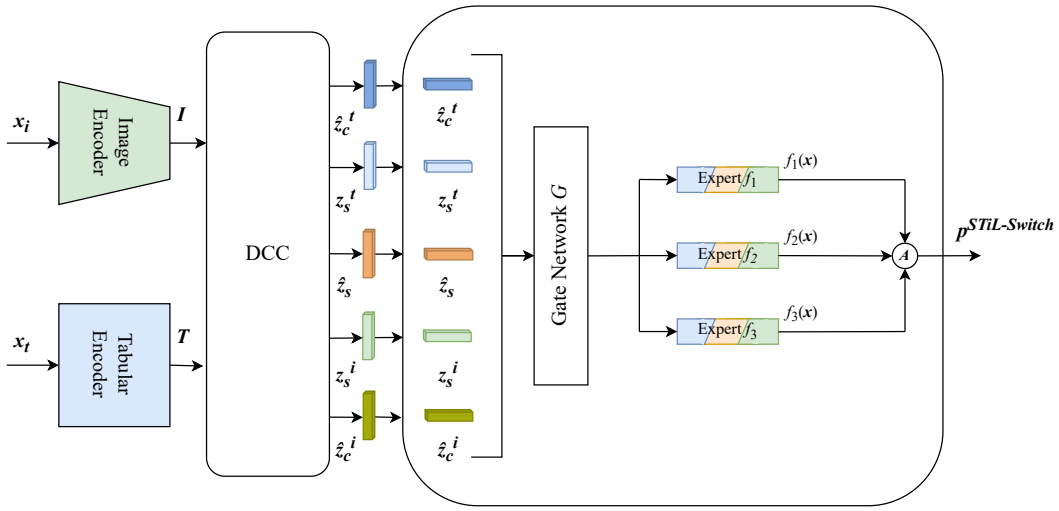


Figure 10: Architecture of the STiL-Switch variant. Here, the experts are no longer constrained to fixed input sets. For each embedding part, the gate determines which expert will process it. As multiple embedding tokens correspond to the same data sample, the results from the individual experts are summarized using average pooling.

$\mathcal{L}_{ce}$ is replaced by $H\left(\boldsymbol{p}^{\text{STiL-Switch}}, \boldsymbol{y}\right)$, since the classifiers no longer receive their fixed input sets. Instead, the loss depends on $\boldsymbol{p}^{\text{STiL-Switch}}$ as the sole logit. The auxiliary loss term $\mathcal{L}_{\text{Switch}}$ (Eq. (13)) is incorporated. This loss term is controlled by the hyperparameter ξ. Similar to the STiL-MoFE loss, this also helps to prevent overfitting as it balances both expert importance and load.

3.5 Evaluation Metrics

The metrics used for evaluation in this thesis are presented here. As the goal of this thesis is to enhance the performance of the STiL model architecture under distribution shifts, standard accuracy (ACC) alone is not sufficient for a proper evaluation. Therefore, balanced accuracy (BACC) is also considered. In order to evaluate the certainty of the classifiers, Shannon entropy is also evaluated.

Accuracy

The accuracy score is one of the most commonly used and most intuitive metrics in machine learning literature. It measures the proportion of correct predictions generated by the model over a set of data samples (Goodfellow et al., 2016). It is calculated as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{39}$$

Here, TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Simplified, it measures task correctness. However, the standard accuracy score tends to return inflated scores when it is employed to measure model performance on imbalanced datasets (Goodfellow et al., 2016). For instance, if the majority of data samples in an AD dataset correspond to healthy patients, it is trivial to achieve a high accuracy score by assigning every single data sample to a healthy person, regardless of its features.

Balanced Accuracy

For class-imbalanced datasets, the balanced accuracy (BACC) is more reliable than the standard accuracy (Brodersen et al., 2010). It is the average of the model's per-class recall scores (Bulavas et al., 2021). It is calculated as:

$$BACC = \frac{1}{\mathcal{C}} \sum_{i=1}^{\mathcal{C}} \frac{TP_i}{TP_i + FN_i}. \tag{40}$$

$\mathcal{C}$ denotes the number of unique classes in the dataset. If the dataset exhibits no class imbalance, it converges to the same value as the standard accuracy metric.

Entropy

The model's uncertainty can be measured by evaluating the Shannon entropy metric (Shannon, 1948). It is calculated as:

$$H(\boldsymbol{p}) = -\sum_{i=1}^{\mathcal{C}} \boldsymbol{p} \log(\boldsymbol{p}). \tag{41}$$

Here, it takes the logits $\boldsymbol{p}$ generated by the model. After applying a softmax function, these logits define a probability distribution over the possible labels for the current data sample. Here, the entropy reaches its maximum value when the probabilities for each class are evenly distributed. For example, if the dataset exhibits three unique labels, it reaches its maximum value, which is equal to $\log(\mathcal{C})$, when all labels are assigned a probability of $\frac{1}{3}$. If the model assigns a probability of 1.0 to a single class and 0 to all other classes, the entropy is equal to 0.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

4.1.1.1 ADNI The ADNI dataset (Mueller et al., 2006) contains medical scans as image data and subject-related information as tabular data. ADNI is designed as a longitudinal study and it continuously updates its clinical protocols to keep up-to-date with current Alzheimer diagnosis protocols. Thus, it consists of multiple datasets. Due to accessibility, the dataset used in this thesis is the first one created, namely ADNI1 (ADNI, 2005). Here, the subjects recruited come from the United States and Canada. The initiative also aims to gain an understanding of the disease's progression. Therefore, multiple data samples have been acquired of the same subjects in the span of 2 to 3 years. In total, the available dataset comprises 645 subjects, with an average of approximately 5 data samples per subject. Thus, it consists of 3412 data samples. The subjects are divided into three research groups, depending on their cognitive state. Thus, these groups consist of cognitively normal subjects, subjects showing signs of MCI and subjects suffering from AD.

The tabular data consists of subject demographics, questionnaire results, and genomic test results. Specifically, the demographic information consists of the subjects' sex, age, and weight, with age and weight being considered as continuous features and sex considered as a categorical feature in this work. The questionnaire data consists of questionnaires commonly used in practice, namely the Neuropsychiatric Inventory Questionnaire (NPI-Q) (Kaufer et al., 2000), the Functional Activities Questionnaire (FAQ) (Pfeffer et al., 1982), the short form of the Geriatric Depression Scale (GDS-Short) (Yesavage and Sheikh, 1986), the Mini Mental State (MMS) questionnaire (Folstein et al., 1975), and the Clinical Dementia Rating (CDR) (Berg, 1988). These questionnaires assess different aspects of the subjects' cognitive states and score them using individual rating systems. NPI-Q assesses neuropsychiatric aspects such as apathy, depression, and psychosis, based on their severity. FAQ is designed to assess the subjects' degree of dementia by evaluating the subject's ability to perform everyday activities, such as writing checks and playing card games. Here, the individual activities are scored based on the subject's dependency on help. GDS-Short evaluates the severity of the subject's depression using questions with binary responses, such as "Are you basically satisfied with your life?" (Yesavage and Sheikh, 1986). The MMS questionnaire examines the subject's mental state by evaluating cognitive abilities such as orientation, attention and recall. Finally, CDR also evaluates the subject's degree of dementia by assessing aspects such as memory loss, orientation, and problem solving. NPI-Q, FAQ, GDS-Short, and MMS do not define clear boundaries for their final assessments and are therefore considered as continuous variables in this work. However, CDR is considered a categorical variable, as it provides explicit ratings (0 = "None", 0.5 = "Questionable", 1 = "Mild", 2 = "Moderate", 3 = "Severe") (Berg, 1988). The genomic test consists of apolipoprotein genotyping. Humans inherit two copies of this gene from their parents, and it is

considered a genetic risk factor indicator for AD (Roses and Saunders, 1994). Since it has three different alleles, the two copies are considered as categorical variables. In total, the tabular data of this dataset consists of four categorical ($n_{cat} = 4$) and six continuous variables ($n_{con} = 6$). The scans consist of volumetric MRI scans of the subjects' brain. To create an image dataset, the middle slice of each volume is extracted. The resulting images have the dimensions $240 \times 256 \times 1$.

However, this dataset has been curated by excluding data samples with missing data entries. Initially, 1612 data samples remained. Given that the dataset must be divided into a training, a validation, and a test set, this is a relatively low number. To further increase it, the GDS-Short results have been omitted, given that it is the attribute with the highest missing entry rate (35.4%). Thus, the dataset used in this work contains 5 continuous features ($n_{con} = 5$). Thus, the resulting curated dataset has a size of 2620 samples. The assigned research groups are used as the ground-truth label.

4.1.1.2 DVM The DVM-Car dataset (Huang et al., 2022) consists of images of cars alongside advertisement data and car specifications as tabular data. In total, it contains over 1.4 million data samples of 916 car models. The model name will be used as the ground-truth label. The images present the cars from 8 different angles, ranging from 0° to 315° in 45° increments. Additionally, the background has been removed to focus solely on the cars. The images have the dimensions $300 \times 300 \times 3$. The tabular data contains the following specification data: seat number, door number, engine size, color, body type, gearbox, fuel type, and mileage measured in miles. Furthermore, the table includes data concerning the cars' advertisements, such as the year and month of creation, both the selling and the entry prices, and the year of first registration. Out of these features, the color, the body type, the gearbox, and the fuel type are considered as categorical features, while the rest are continuous ($n_{cat} = 4$ and $n_{con} = 13$). However, the tabular dataset contains only 247236 unique samples, limiting the total amount of data samples available in this work.

The dataset was prepared following the procedure used by Du et al. (2024). All date samples exhibiting missing entries were removed. Additionally, all samples corresponding to models that are represented less than 100 times are also excluded. Therefore, the remaining dataset contains 176414 data samples with 286 unique car models.

4.1.2 Shifted Subsets

The objective of this study is to examine how distribution shifts occurring in specific modalities influence model performance and whether the examined strategies can mitigate their effects. Thus, multiple subsets have been extracted from both the ADNI and the DVM datasets, based on modality-specific features. Additionally, we intended to create subsets exhibiting shifts with varying degrees of severity. Therefore, features were selected based on their influence on the classification task.

To ensure comparability between results, all subsets were reduced to the size of the smallest subset obtained after extraction. Otherwise, results could be unfairly influenced by differences in training set size, which may either improve the learning of task-relevant features or cause overfitting. Furthermore, this thesis only investigates the effects of covariate shift. Therefore, label distributions between the respective training and test sets were kept as similar as possible. However, since the labels aren't evenly distributed, this approach has led to minor differences between respective subsets. After extraction, the subsets have been further split into training, validation, and test sets. Following Du et al. (2025), categorical features were encoded as ordinal values, and continuous features were z-score normalized using training set statistics. An overview of the final subsets is given in Table 1.

Table 1: Overview of the created subsets. "Train Size" denotes the number of samples in the training set, "Val Size" the number in the validation set, and "Test Size" the number in the test set. "Num. of Classes" shows the number of labels left in the respective subset. "Shifted modality" shows which modality feature has been shifted.

Subset	Train. Size	Val. Size	Test. Size	Num. of classes	Shifted Modality
DVM-Car Subsets					
No-Shift	29000	5000	5000	286	-
B-S-B-G-W	29000	5000	5000	286	Image
B-S-G	29000	5000	5000	286	Image
Black	29000	4999	4992	285	Image
Adv-Year	28894	4977	5000	281	Tabular
Miles	29000	4977	5000	281	Tabular
Color-Miles	29000	4997	4938	276	Both
ADNI Subsets					
No-Shift	1460	375	375	3	-
TE	1460	375	375	3	Image
Weight	1460	375	375	3	Tabular
Age	1460	375	375	3	Both

4.1.2.1 ADNI Subsets From the ADNI dataset, four subsets were extracted. Two of these subsets exhibit distribution shifts based on tabular features, while one exhibits a distribution shift based on an image feature. For all subsets, patient-level separation between the training and test split is ensured.

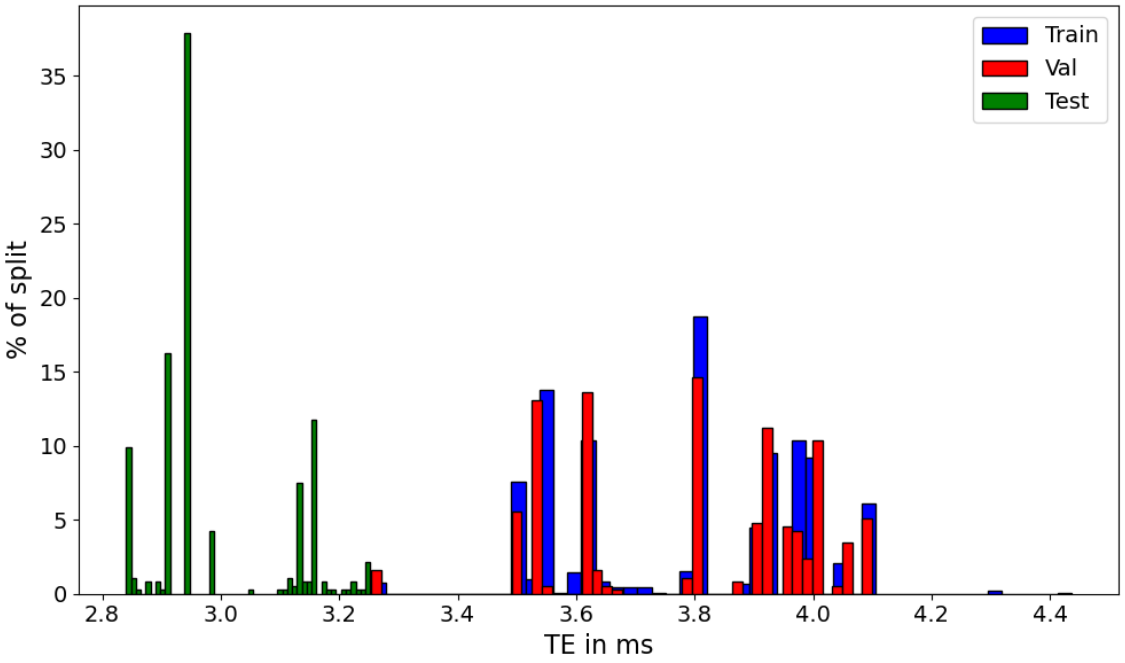


Figure 11: TE distributions of the TE subset.

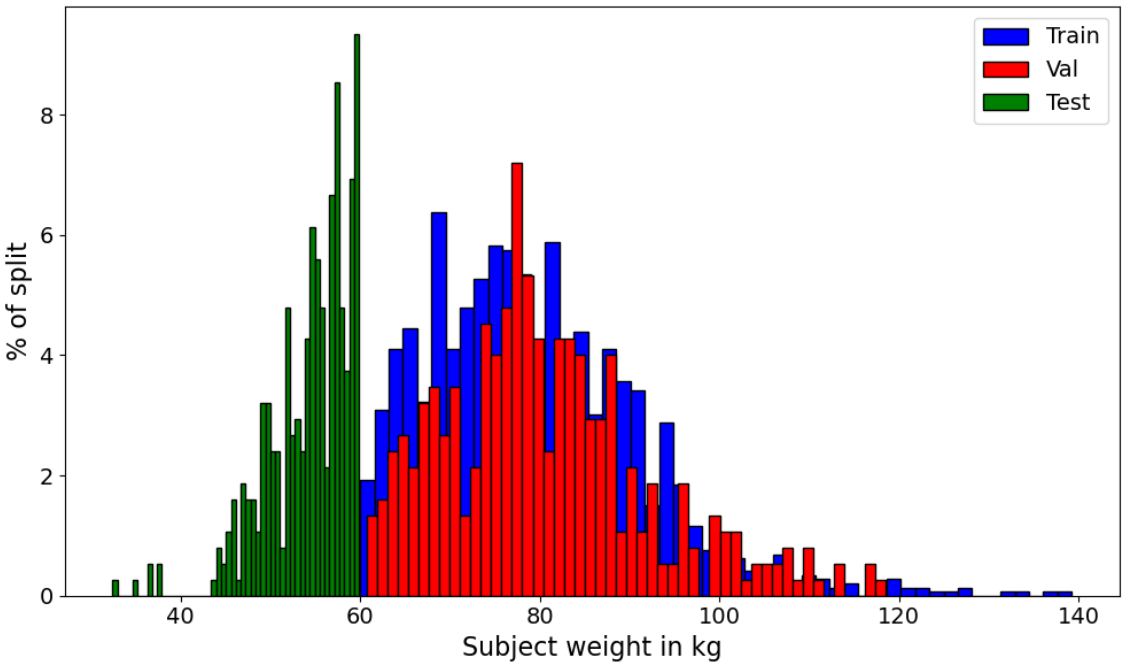


Figure 12: Distribution of the subjects' weight in the Weight subset.

(1) First, a smaller subset that exhibits no distribution shift is extracted. This subset is used to evaluate whether the applied augmentation strategies can further improve model performance in a non-shifted setting or whether they instead degrade it. It is called the *No-Shift* subset. (2) The feature selected to induce a distribution shift in the image modality in this subset is the MRI echo time (TE). This parameter measures the time between the magnetic pulse and its echo and influences the scan's contrast: a longer TE causes liquids to appear brighter than softer tissues (Jung and Weigel, 2013). Here, samples with a TE greater than 3.25 ms are assigned to the training and validation sets, while all samples with a lower TE are placed in the test set (see Fig. 11). This subset will be referred to as the *TE* subset. (3) In this subset, the distribution shift is created based on the subjects' weight. Data samples from patients weighing less than 60 kg are placed in the test dataset, whereas all other samples are included in the training and validation datasets (see Fig. 12). Since AD can lead not only to cognitive but also to behavioral and eating disorders, it is also associated with weight loss (Gillette-Guyonnet et al., 2000). This subset will be referred to as the *Weight* subset. (4) Finally, the last subset of this study aims to induce a distribution shift based on the subjects' age. It is called the *Age* subset. All data samples from patients aged between 68 and 87 were assigned to the training and validation dataset. All other data samples were assigned to the test dataset (see Fig. 13). The incidence rate of AD is positively correlated with age (Launer et al., 1999). However, MRI studies indicate that younger AD patients experience faster rates of brain volume loss (Dukart et al., 2013; Fiford et al., 2018). Therefore, the intuition behind this split is to deteriorate the model's performance by exposing it to these extremes during testing.

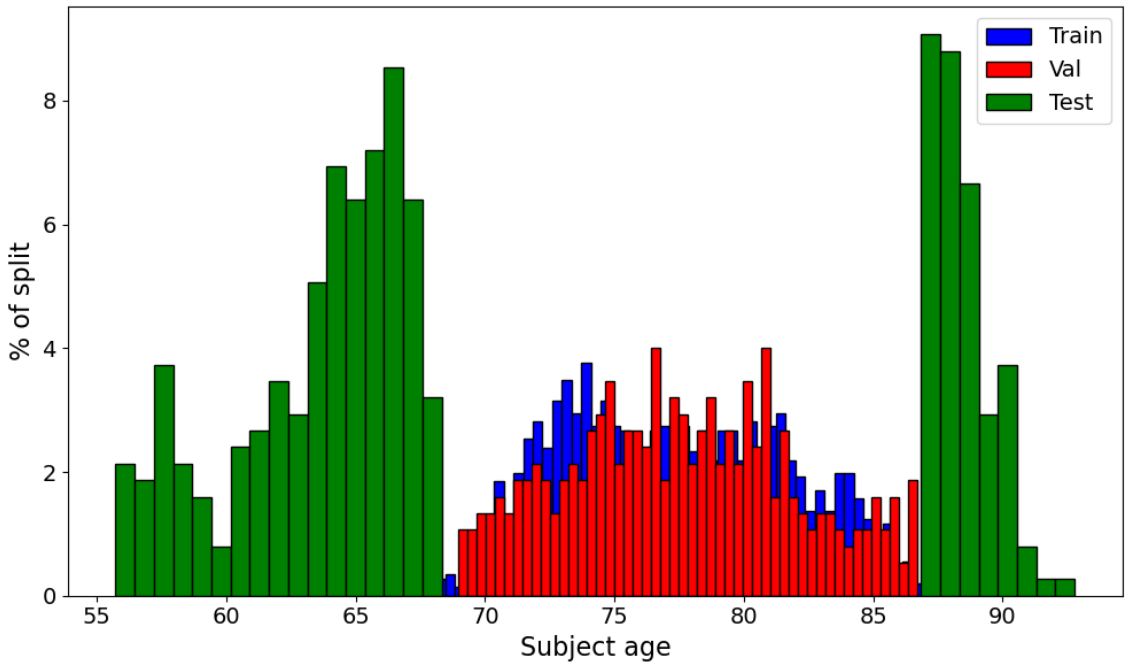


Figure 13: Subject age distributions in the Age subset.

It will encounter a higher AD incidence rate among older patients and an increased brain loss rate among younger patients, which it hasn't seen during training. Since the MRI scans are also affected here, this subset is shifted in both modalities.

4.1.2.2 DVM Subsets From the DVM dataset, seven subsets were extracted. Three subsets focus on image features, two on tabular features, one on both modalities, and a smaller version of the original subset. The selected image feature is the color of the car. The subsets here exhibit increasing levels of difficulty.

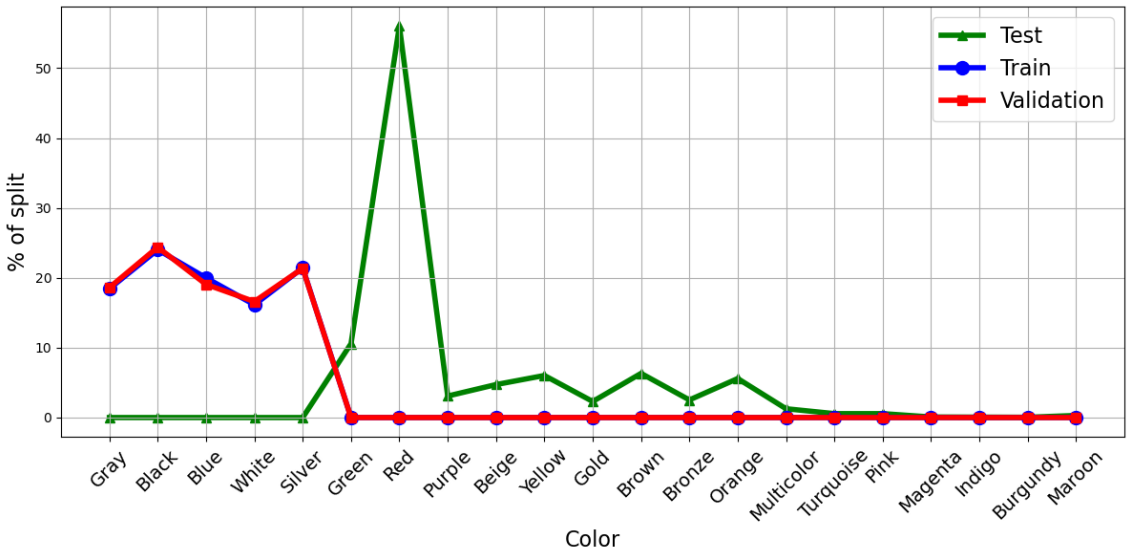


Figure 14: Color distributions of the Black-Silver-Blue-Grey-White (B-S-B-G-W) Subset.

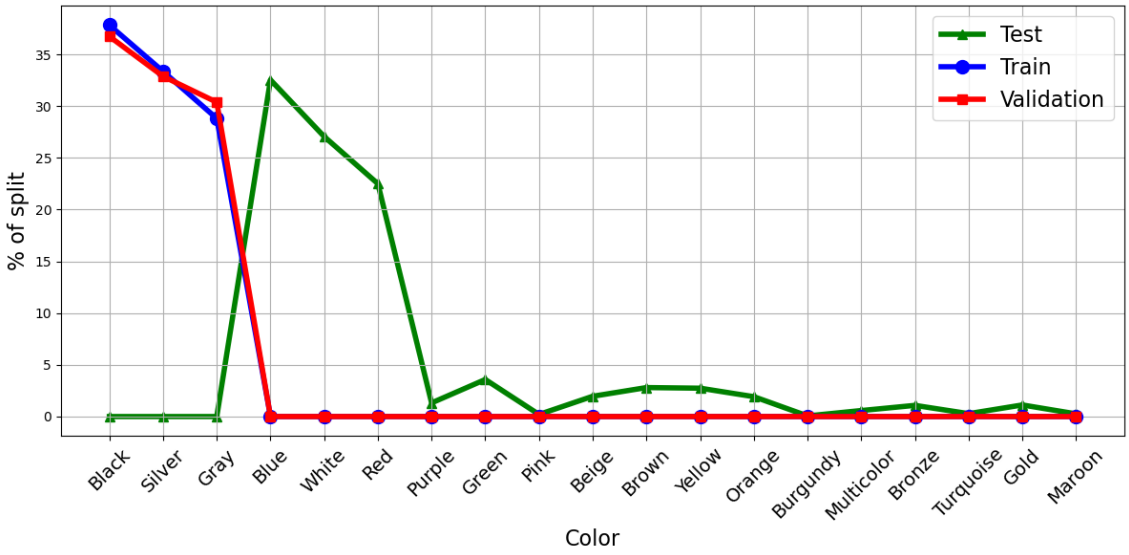


Figure 15: Color distributions of the Black-Silver-Grey (B-S-G) Subset.

(1) The first subset preserves the most prominent colors in the training and validation sets, while all remaining colors are assigned to the test set (see Fig. 14). The training dataset consists of black, silver, blue, gray, and white cars. Therefore, this subset will be called the *B-S-B-G-W* subset. The majority of this subset's test samples are red cars. The difficulty is expected to be relatively low, as the training colors make up the majority of the dataset. (2) The second subset only contains cars with black, silver, and gray colors (see Fig. 15) in the training and validation dataset. It will be referred to as the *B-S-G* subset. The test dataset consists mostly of blue, white, and red cars. This subset is designed to test whether the model's performance declines when it sees only grayscale colors.

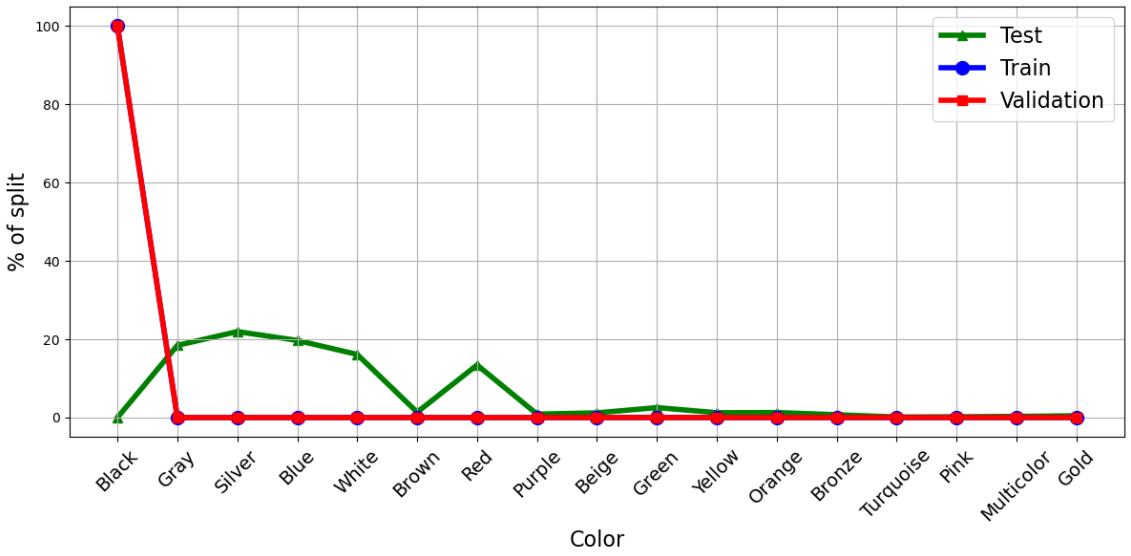


Figure 16: Color distributions of the Black Subset.

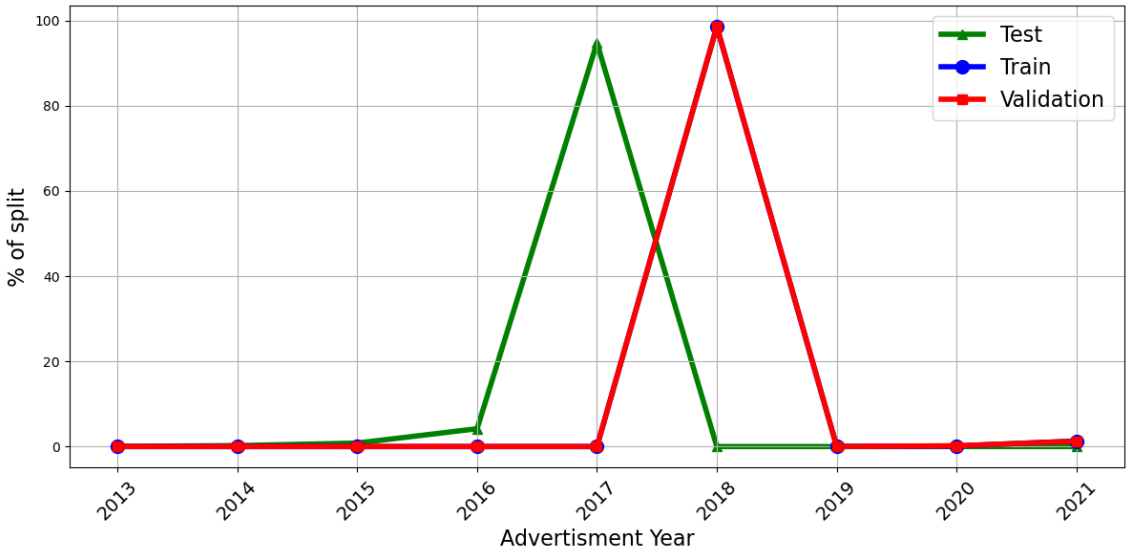


Figure 17: Year distributions of the Advertisement Year (Adv-Year) Subset.

Additionally, since the images are set to grayscale during input space augmentation, this subset will appear as mostly gray cars to the model. (3) In the third subset, only black cars are present in the training and validation dataset, while all other colors are present in the test dataset. Thus, this subset will be referred to as the *Black* subset. As the training dataset consists of just one color, the model will lack diversity to a considerable extent. Consequently, this subset exhibits a high level of difficulty for the model. (4) The first subset focusing on a tabular feature exhibits a shift based on the advertisement's year of creation (see Fig. 17). The training and validation dataset consists of all data samples of advertisements not older than 2018, while the test dataset's newest advertisement was created in the year 2017. As these years make up the majority of the entire dataset, they are also the most prominent years in the respective splits. Nevertheless, this subset is not expected to pose a major challenge for the model, as the year of the advertisement does not provide any information about the car model. This subset will be referred to as the *Adv-Year* subset. (5) The next subset exhibits a distribution shift based on the cars' mileage. The training and validation sets consist of samples with mileage ranging from 17,038 to 79,000 miles (see Fig. 18). Therefore, the test dataset consists of all samples below and above that range. Unlike the Adv-Year subset, this subset is expected to pose a challenge for the model. Responsible for this is that the mileage is correlated with the car's age. Older cars tend to have been driven more than younger cars. With a limited training dataset, this might confuse the model. This subset will be referred to as the *Miles* subset. (6) The multimodal subset exhibits distribution shifts on both modalities (see Fig. 19).

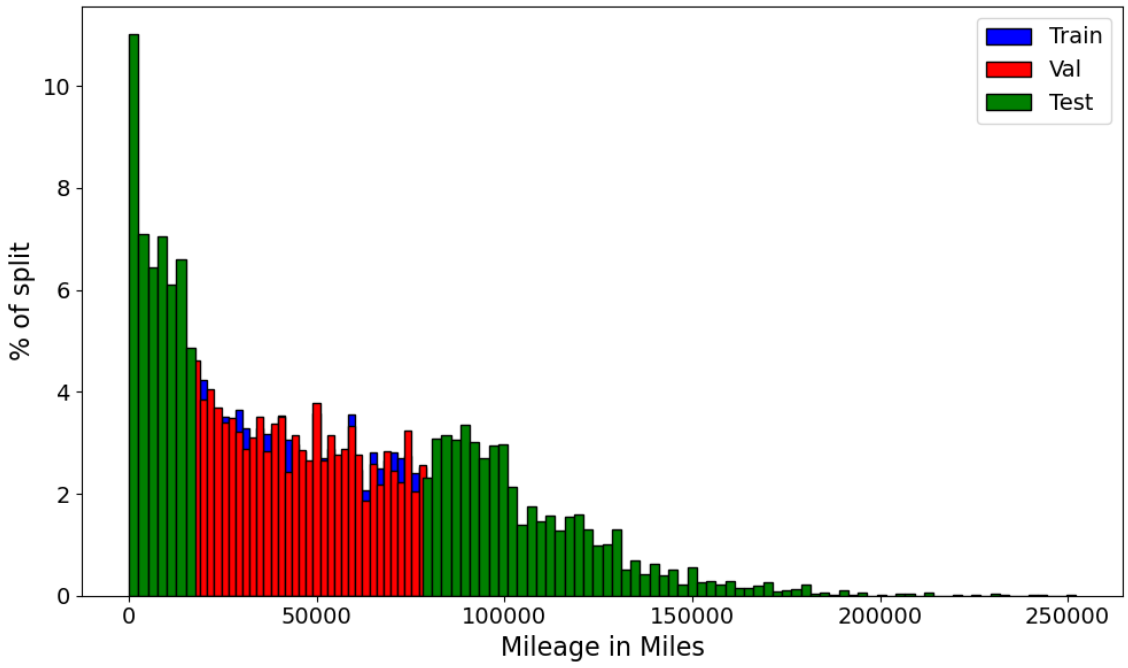
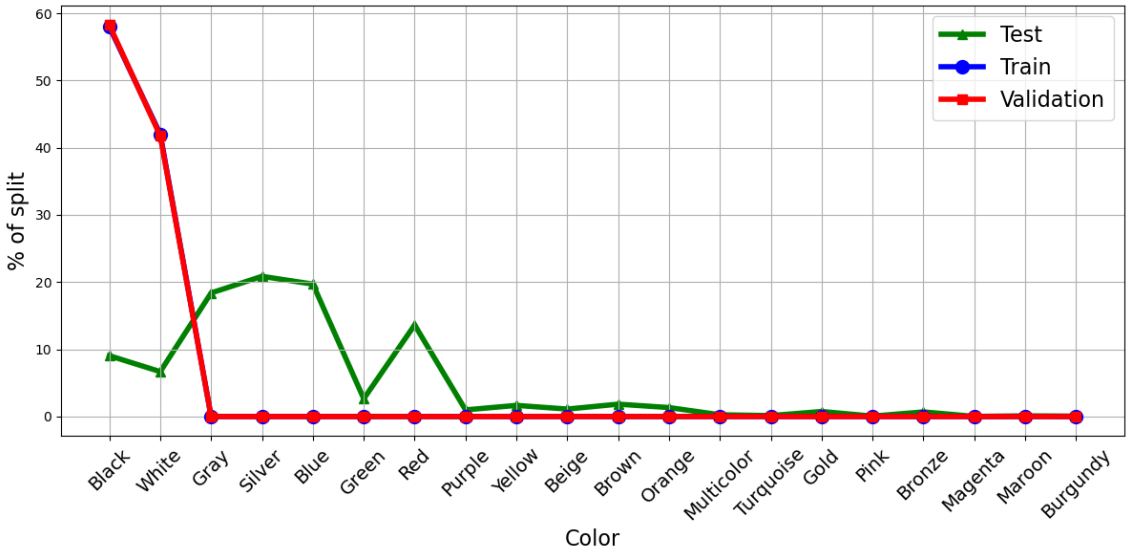
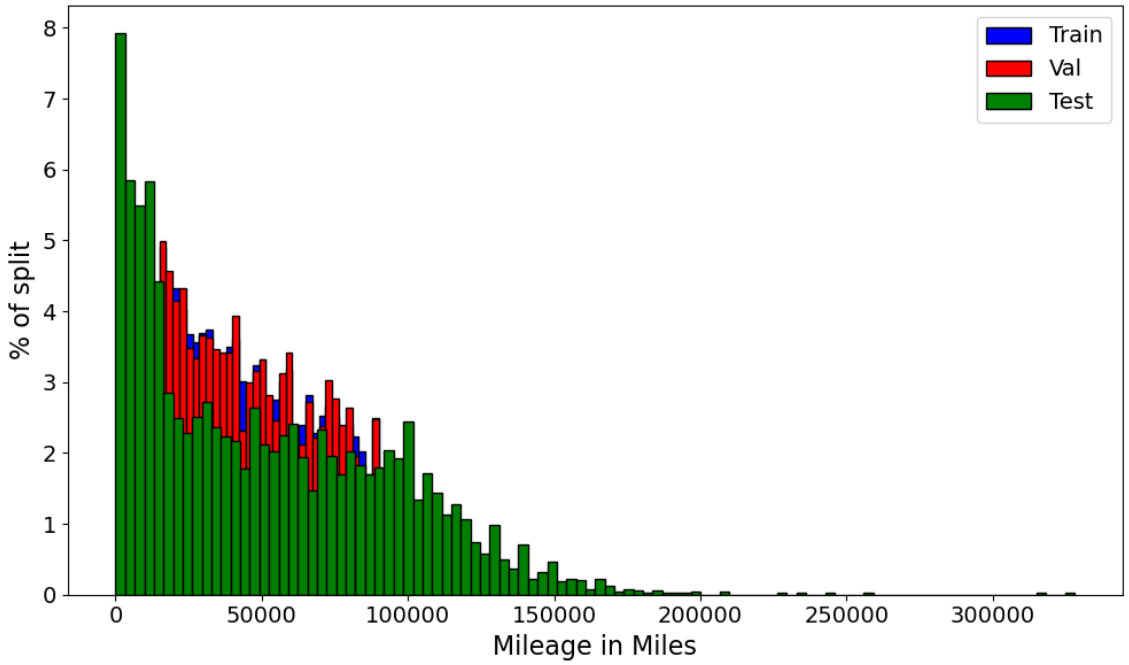


Figure 18: Mileage distributions of the Miles Subset.

First, the training and validation datasets consist only of black and white cars. Second, these datasets also contain only data samples with a mileage ranging from 15,000 to 90,000 miles. Thus, this subset is referred to as the *Color-Miles* subset. In contrast to the previously introduced subsets, these value ranges are not exclusive to the training and validation sets and are also present in the test dataset.



(a) Color-Miles: Color distributions.



(b) Color-Miles: Miles distributions.

Figure 19: The distributions of the shifted features in the Color–Miles Subset. (a) presents the color distribution, whereas (b) shows the mileage distribution.

However, as this subset exhibits distribution shifts in both modalities, it is still expected that it poses a considerable challenge for the model. (7) Finally, analogous to the ADNI dataset, a smaller non-shifted subset was extracted from the DVM-Car dataset.

4.1.3 Implementation Details

All experiments were conducted on a system outfitted with 125 GiB of RAM and an NVIDIA RTX A5000 GPU with 24564 MiB of available VRAM. The system's CPU is an Intel Xeon W-2265 processor, featuring a base frequency of 3.50 GHz and 12 available cores. The code is implemented using PyTorch (Paszke et al., 2019) version 1.11.0. For GPU computation, CUDA version 13.0 is used.

The configuration of the input space augmentation techniques follows Du et al. (2025). Image space augmentations are listed in Table 2 and applied with a probability of 95 %. In tabular space, randomly selected cells are corrupted by replacing values with samples from their marginal distributions. This procedure is applied to 30% of all cells. When evaluating the latent embedding augmentation methods, the application probability was set to 0.5. The remaining hyperparameters were set in accordance with the settings in the respective works of the techniques: DeVries and Taylor (2017) reported optimal performance with $\lambda = 0.5$ for linear extrapolation, while Zhou et al. (2021, 2024) used $\alpha_{MixStyle} = 0.1$ for MixStyle. For noise perturbation, noise is sampled from $\mathcal{R}_{noise} \sim \mathcal{N}\{0, \sigma_{\boldsymbol{z}}^2\}$ following DeVries and Taylor (2017), where $\sigma_{\boldsymbol{z}}$ is computed across embeddings in the current batch instead of the full dataset. The input space augmentation techniques were disabled during the evaluation of the latent embedding methods.

When examining the two STiL-MoE variants, only input space augmentation techniques are applied. For the ADNI subsets, STiL-MoFE is tested with both a simple linear layer and an MLP as the gate network. As the incoming embeddings $\boldsymbol{z}_s^i$, $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}^s$, $\hat{\boldsymbol{z}}_c^t$, and $\boldsymbol{z}_s^t$ have the shape $B \times 512$, the input dimension of the linear gate is 2560, while its output has the dimension $\mathcal{C}$. The MLP has two linear layers with a ReLU-layer in between. Its input dimension is 2560, its hidden dimension is 1280, and its output dimension is $\mathcal{C}$. For DVM-Car subsets, STiL-MoFE is evaluated only with the MLP gate. The STiL-Switch variant uses only the linear gate. However, in both variants, the gating function's output is perturbed with additive noise during training to increase expert exploration. The noise is sampled from a standard normal distribution $\mathcal{R}_{noise} \sim \mathcal{N}(0, 1)$. Furthermore, K is set to 3 in all experiments. For STiL-MoFE, k is varied from 1 to 3 for the ADNI subsets and set to 3 for the DVM-Car subsets. Since STiL-Switch follows the Switch Transformer training procedure, it is only tested with $k = 1$ for all subsets.

Most of the remaining training hyperparameters have been adopted from the DVM-Car settings of Du et al. (2025). Here, they are set as follows: $\alpha = 0.2$, $\beta = 3.0$, $\gamma = 0.5$, $\lambda_p = 1.0$, $\lambda_u = 0.2$, $r = 0.9$, $\tau = 0.9$, and $m = 0.996$. The learning rate is set to 0.0001. Additionally, early stopping is employed to stop training prematurely to reduce overfitting. Its minimum delta is set to 0.0001, while the maximum number

of epochs is reduced to 350 and the patience to 50 due to the smaller size of the subsets. For the DVM-Car subsets, the batch size is also adopted, thus it is set to $B = 512$. Here, when experimenting in a sparsely labeled setting, the ratio between unlabeled and labeled training samples is also adopted, resulting in $B_l = 64$ and $B_u = 448$, with 10% of the total training dataset being labeled.

Table 2: Image space augmentation configuration for the natural and medical data subsets. Floating-point numbers indicate the likelihood of the technique being applied. Tuples control more advanced parameters: "Scale" in "Random Resized Crop" controls the lower and upper bounds for the crop area, respective to the original image size, while "Ratio" sets the bounds of the aspect ratio of the new image. For "Gaussian Blur", "Blur Limit" defines the range of possible kernel sizes, while "Sigma Limit" controls the range of sigma values.

Augmentation Method	Parameter	Value used for DVM-Car subsets	Value used for ADNI subsets
Color-Jitter	Probability	0.8	0.5
	Brightness	0.8	0.5
	Contrast	0.8	0.5
	Saturation	0.8	0.5
Horizontal-Flip	Probability	0.5	0.5
Random-Resized-Crop	Probability	1.0	1.0
	Height	128	128
	Width	128	128
	Scale	(0.6, 1.0)	(0.6, 1.0)
	Ratio	(0.75, 1.333)	(0.75, 1.333)
Gray-Scale	Probability	0.2	Not used
Gaussian-Blur	Probability	0.5	Not used
	Blur Limit	(29, 29)	Not used
	Sigma Limit	(0.1, 2.0)	Not used
Rotation	Probability	Not used	0.5
	Limit	Not used	45

For simplicity, this setting is also referred to as the *unlabeled* setting. For the ADNI subsets, all values of $B \in \{32, 64, 128, 256\}$ have been tested to find the optimal setting. Additionally, a range of values is tested to determine the optimal settings for the respective loss terms of the STiL-MoE variants. For STiL-MoFE, $\epsilon \in \{0.1, 0.5, 1, 2, 3\}$ is tested, whereas $\xi \in \{0.001, 0.01, 0.1, 1, 2\}$ is tested for STiL-Switch. The value range of STiL-Switch differs from the one tested in STiL-MoFE, as the value range used by Fedus et al. (2022) is adopted, who examined a range from 10^{-1} to 10^{-5} in power of 10 steps. However, it was adapted in relation to the already complex loss function of the STiL-Switch variant.

To ensure reproducibility, all experiments except those in the unlabeled setting were run with three different seeds: 2022, 2023, and 2024. These experiments were only run once with the seed 2022 due to time constraints. The scores shown are the mean of the three runs. At first, the experiments have been run in the unlabeled setting to investigate the behavior of the original STiL architecture under distribution shifts. However, the setting was later changed to fully labeled training, as the unlabeled setup substantially reduced the effectiveness of the augmentation strategies.

4.2 Results

4.2.1 Baselines

The first experiments aimed to find baseline values for the remaining experiments conducted in this thesis. The baseline results for the DVM-Car subsets in the unlabeled setting are shown in Table 3. As expected, Black, Miles, and Color-Miles provide the lowest accuracy (ACC) scores. Meanwhile, the effect isn't as severe for the other subsets. Here, the model struggles to generalize to new data, as it is trained with very limited training datasets.

Table 3: Baseline accuracy scores (ACC) of the DVM-Car subsets in the unlabeled setting.

Subset	ACC in %
No-Shift	83.22
B-S-B-G-W	80.48
B-S-G	82.10
Black	76.46
Adv-Year	80.14
Miles	75.59
Color-Miles	72.64

Table 4: Baseline results for the DVM-Car subsets in the fully labeled setting. The scores show the accuracy (ACC) scores across all seeds (mean ± std).

Subset	ACC in %	BACC in %
Black	90.09 ± 1.14	88.42 ± 1.22
Miles	91.77 ± 0.59	89.14 ± 0.62
Color-Miles	91.58 ± 0.51	90.19 ± 0.66
No Shift	96.65 ± 0.17	96.04 ± 0.35

Furthermore, this experiment was also meant to reduce the number of DVM-Car subsets in order to limit computation time throughout this thesis. The subsets were selected based on their test accuracy in the unlabeled setting. The worst-performing subsets were retained for this work, with one subset selected per modality. Black and Miles have the lowest scores and are thus used in subsequent analyses in this thesis. Furthermore, the DVM-Car subsets were also evaluated in a fully labeled setting in this work. The baseline results of this setting are presented in Table 4.

The ADNI subsets were only examined in a fully labeled setting. Since the ADNI dataset is substantially smaller than the DVM dataset, we conducted a small hyperparameter tuning experiment to determine the optimal batch size first. Therefore, four test runs using the No-Shift subset were conducted, with the batch sizes $B = \{32, 64, 128, 256\}$. The results are shown in Table 5. As the test accuracy is almost identical across all batch sizes, we also considered the balanced accuracy (BACC). Additionally, we also evaluated the results stemming from the validation set.

Table 5: Accuracy (ACC) and balanced accuracy (BACC) scores (mean ± std) of the validation and test dataset of the No-Shift subset using different batch sizes.

	Validation Set		Test Set	
Batch Size	ACC in %	BACC in %	ACC in %	BACC in %
32	83.11 ± 1.01	79.39 ± 1.71	84.09 ± 0.15	75.74 ± 1.29
64	82.75 ± 0.56	77.45 ± 2.15	84.18 ± 0.56	75.36 ± 1.20
128	83.64 ± 0.81	78.55 ± 0.93	84.27 ± 0.27	75.14 ± 0.69
256	82.23 ± 0.19	76.41 ± 0.00	84.53 ± 0.38	75.69 ± 0.94

Table 6: Baseline results of the ADNI subsets.

Subset	ACC in %	BACC in %
TE	85.24 ± 0.94	81.93 ± 1.26
Age	81.87 ± 1.60	81.84 ± 0.95
Weight	85.42 ± 0.77	84.13 ± 2.22
No Shift	84.09 ± 0.15	75.74 ± 1.29

While the resulting scores are highly similar, they indicate that $B = 32$ is the best option, as it provides the highest BACC scores in both the validation and the test set. Thus, this value was used throughout the remainder of this work.

Afterwards, the baseline scores of the remaining ADNI subsets were determined. They are presented in Table 6. Surprisingly, the No-Shift subset yielded the worst BACC score, with a margin of over 6% to the second lowest score. In terms of the accuracy metric, it performs better than the Age subset. The Weight subset achieves the best performance across both metrics. Such an effect is not seen in the baseline results of the DVM-Car subsets.

4.2.2 Modality-specific Input Space Augmentation Ablation Study

The next experiment is a modality-specific ablation study of the input space augmentation techniques. The goal is to examine the contribution of the specific modalities to the classification task. Therefore, the model was trained using the Black, Miles, and Color-Miles subsets with the input space augmentation techniques disabled either for each modality individually or for both modalities simultaneously. With this approach, we can observe from which modality the model benefits most and how increased diversity in each modality affects model performance. The experiment settings were also run using the No-Shift subset to observe its effect in an evenly distributed setting. This experiment was run in the unlabeled setting. The results are presented in Fig. 20. Interestingly, it shows that the augmentation technique applied to the tabular data actually deteriorates model performance rather than improving it. For the Black and Color-Miles subsets, model accuracy increases by 3-4% when only the image modality is augmented, whereas No-Shift exhibits only a negligible improvement of 0.1%. However, the performance drops by 2.2% at the Miles subset in this setting. Additionally, subsets whose primarily shifted feature is tabular (Miles and Color-Miles) generally exhibit lower accuracy scores than the other subsets. Furthermore, the results also show that the experiments suffer severely across all subsets if the tabular data is the only modality being augmented. Compared to image-only augmentation, the model's accuracy decreases by at least 38% (Miles) and by up to 46.1% (Black). Moreover, the results of the settings with no augmentation technique enabled demonstrate that this effect is severe.

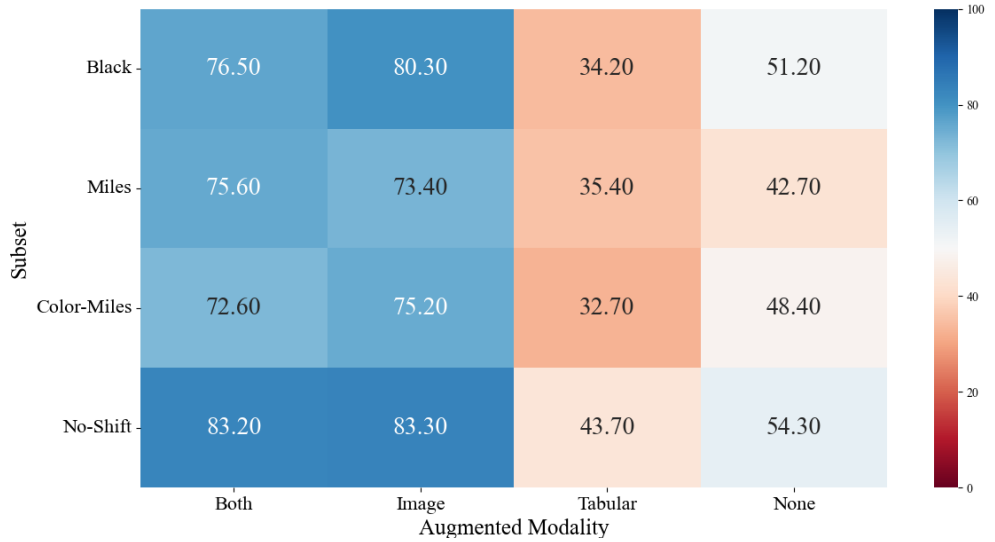


Figure 20: Accuracy scores (in %) of the DVM-Car subsets with different input space augmentation settings. The augmentation methods have been disabled for the two different modalities, either individually, both, or not at all. The column "Both" presents the baseline values taken from Section 4.2.1. The remaining columns indicate the modalities for which augmentation techniques are enabled.

In fact, the model benefits more when no modality is augmented than when only the tabular data is augmented. Here, the model performs better by an average of 12.7%. The model's behavior can be explained by the applied augmentation technique, which leads to training samples that are highly improbable in the context of the dataset used. Here, the tabular features are corrupted by replacing them with values from their marginal distribution. This can result in confusing car specifications, which disrupts the semantics between the data sample and its label. For instance, it may augment a data sample originally describing a Smart Fortwo and replace its engine size with the value of a Ford Mustang, which has an engine approximately five times that size. Such a misalignment confuses the model. Furthermore, the results demonstrate that diversity in the image modality is the main contributor to the success of the classification task when using the DVM-Car dataset. The model benefits from the color-altering techniques such as Color-Jitter and Gray-Scale as these methods increase the limited selection of colors presented in the training datasets in the Black and Color-Miles subsets.

4.2.3 Latent Embedding Augmentation Experiments

In the next experiments, the impact of the previously introduced latent embedding space augmentation techniques was evaluated. This was done by applying these methods to the input embeddings batches of the unimodal and multimodal classifiers within the CGPL module. To determine which of these inputs provides the greatest benefit when augmented, we applied the selected methods on each input

individually. However, since literature indicates that isolated augmentation may disrupt the semantic relationships between modalities, we also augmented the input batches of all classifiers simultaneously. With all embeddings corresponding to the same data sample being augmented, we can observe if the STiL model benefits from consistent latent embedding space augmentation across modalities. All input space augmentation methods were disabled during these experiments. This approach not only allows us to observe the impact of the latent embedding space techniques more directly, but also enables comparison with the results in Section 4.2.2 to determine in which data space augmentation is most effective. At first, this experiment was run with the DVM-Car subsets in the unlabeled setting. Its results are presented in Fig. 21. First, we observe that no method is able to compensate for the absence of input space augmentation techniques. However, in most settings, they still perform better in comparison to the scenario where no augmentation is applied at all. The highest accuracy score (57.5%) is achieved on the Black subset using extrapolation, whereas the lowest score (42.1%) is obtained on the Miles subset using noise perturbation. Nevertheless, the same pattern regarding the primarily shifted feature is observed, consistent with the input space augmentation ablation study. Here, the Miles and Color-Miles subsets also perform the worst, while Black achieves similar scores as No-Shift. Additionally, the results show that the best scores are achieved when transforming the classifiers' inputs consistently.

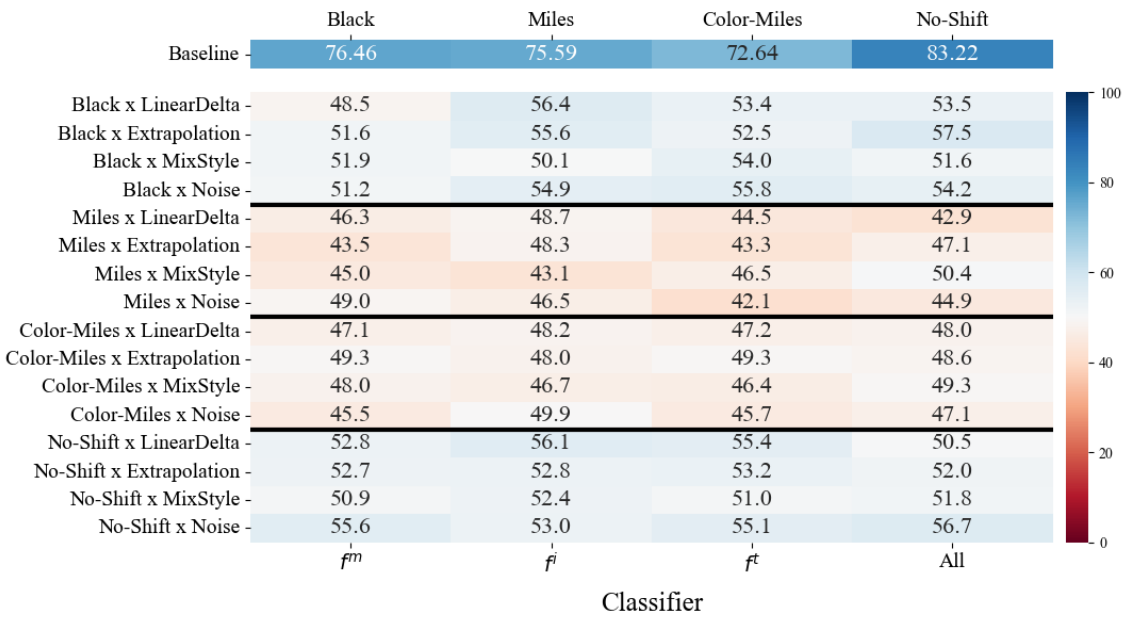


Figure 21: Accuracy scores (in %) of the DVM-Car subsets using different latent embedding space augmentation techniques in the unlabeled setting. The inputs of the classifiers f^m, f^i, and f^t are augmented individually or jointly, yet none of these methods fully compensate for the absence of input-space augmentation. The top row shows the baseline values for each subset, respectively. The y-axis ticks indicate the subset and the augmentation method used (Subset × Method).

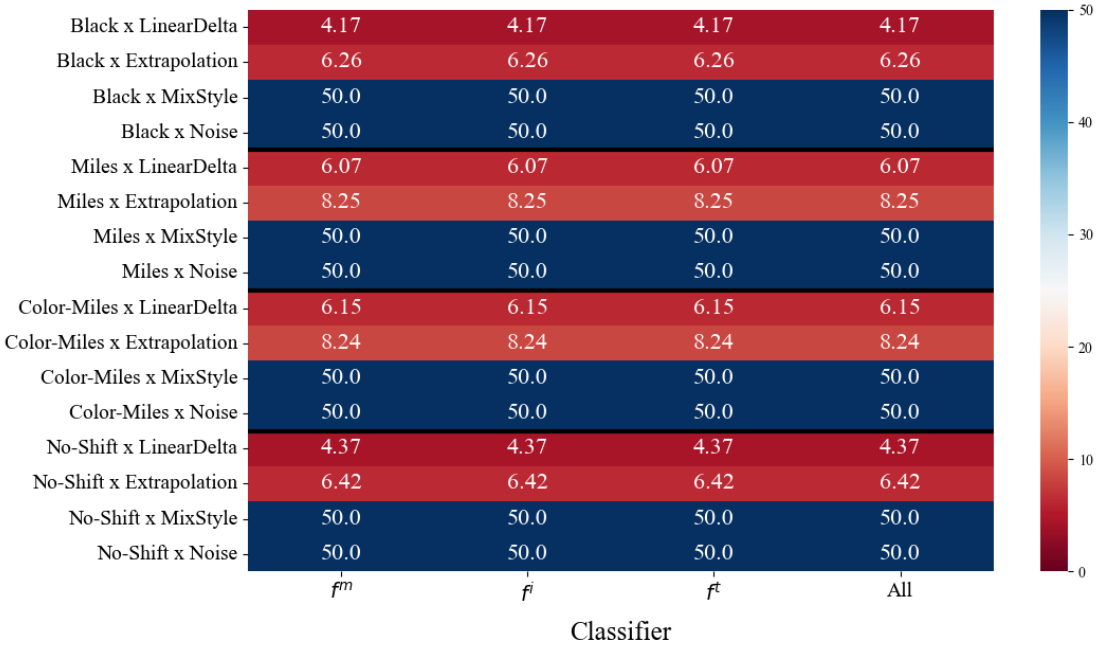


Figure 22: Augmentation rate (in %) relative to the training set size for DVM-Car subsets using latent embedding space augmentation methods. Class-agnostic methods (noise perturbation and MixStyle) transform 50%, while class-dependent methods (linear delta and extrapolation) operate at substantially lower rates.

Here, the best accuracy score of the Miles subset is obtained when applying the MixStyle, whereas the best score of the No-Shift subset is achieved when the input embeddings are transformed using noise perturbation. Only Color-Miles benefits most from an isolated setting, as it achieves its highest score of 49.9% when noise perturbation is applied on the input of the image classifier f^i. However, it is closely followed by MixStyle being applied to all input embedding batches. Here, it achieved an accuracy score of 49.3%. Thus, the results are in alignment with observations in literature in this regard.

However, the rate of transformed data samples shows that the unlabeled setting inhibits the effectiveness of the applied augmentation techniques, especially of the class-dependent methods linear delta and extrapolation. The rates are illustrated in Fig. 22 and are calculated relative to the total amount of data samples present in the respective training datasets. While the rates of the class-agnostic methods, MixStyle and noise perturbation, are fixed at 50%, the rates for the other methods drop below 10%. Extrapolation and linear delta are applied only to data samples with the same label, ensuring that the resulting data sample remains consistent with its label. Because this experiment was conducted in the unlabeled setting, each batch contained only $B_l = 64$ labeled data samples. Additionally, the subsets contain at least $\mathcal{C} = 276$ unique labels, resulting in a low amount of multiple data samples sharing the same label. This leads to a very limited number of data samples being transformed, particularly when using linear delta, which requires three samples with the same label to generate a new one (see Eq. (32)). Nevertheless, the class-agnostic

methods show no clear advantage over the class-dependent methods, indicating that the limited number of data samples sharing the same label per batch is not the primary issue.

To further illustrate the negative effect of the unlabeled setting, we also applied the latent embedding augmentation techniques within the model while using the ADNI subsets in the fully labeled setting. Again, all input space augmentation methods were disabled during this experiment. The resulting accuracy scores are presented in Fig. 23, while the balanced accuracy scores are shown in Fig. 24. Here, the applied methods improve the model's performance to the point where it matches that of having all input-space augmentation techniques enabled. Considering the accuracy score, the augmentation methods also outperform the baseline, albeit by only small margins. These are the highest improvements per subset: extrapolation increased the model's mean score on the TE subset by 0.48% when applied to the input of all classifiers. For the Age subset, it incremented it by 0.18% when applied to the input of f^t. For the other subsets, the best-performing method is MixStyle. Here, it provides an improvement of 0.34% when applied on the input of f^i. For the No-Shift subset, it increments model performance by 0.31%. Considering balanced accuracy however, only the performances using the Weight and No-Shift subsets could be enhanced. Here, linear delta improved the mean score by a negligible margin of 0.05% when applied on the input of f^i, while noise perturbation enhanced it by 0.16% when applied to the input of f^t. Nevertheless, the fact that these methods are able to reach the baseline indicates that the unlabeled setting is mitigating the effect of the applied augmentation techniques.

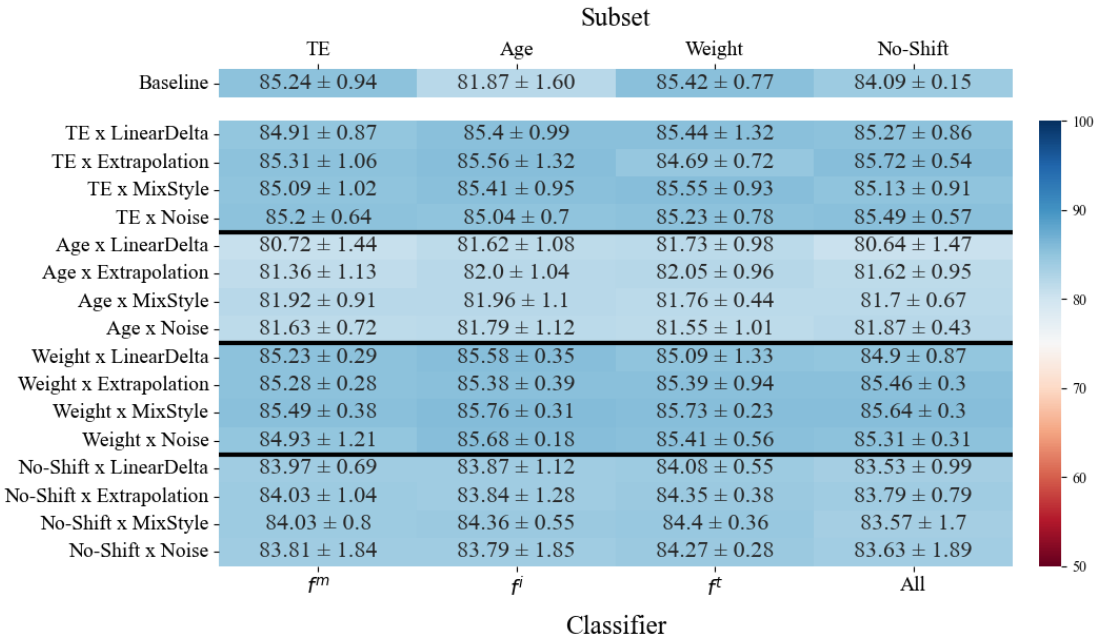


Figure 23: Accuracy scores (in %) of the ADNI subsets with the latent embedding augmentation techniques being applied in the fully labeled setting.

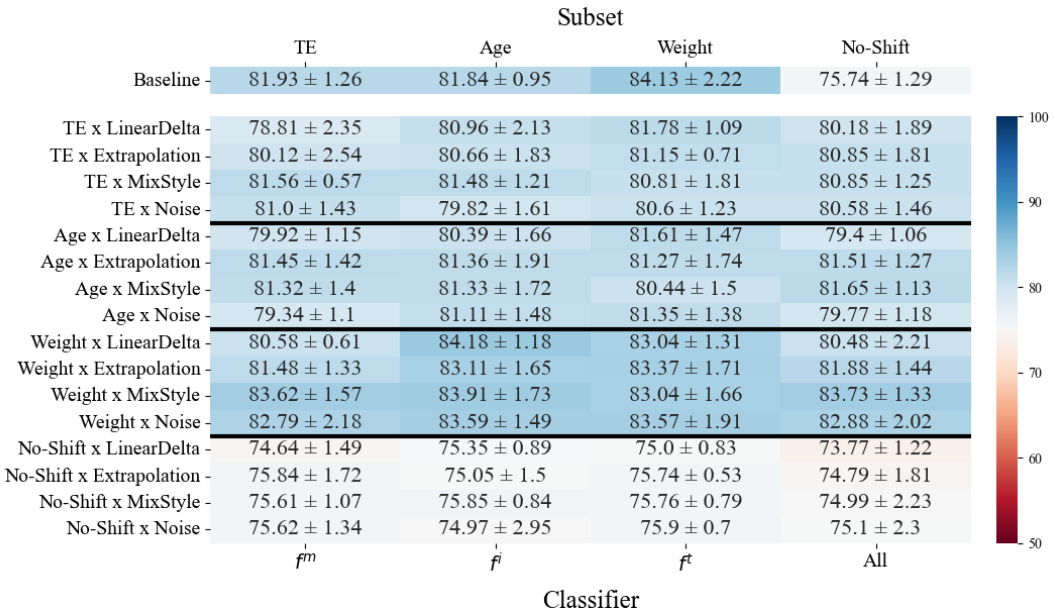


Figure 24: Balanced accuracy scores (in %) of the ADNI subsets with the latent embedding augmentation techniques being applied in the fully labeled setting.

Furthermore, although the ADNI subsets contain substantially fewer classes ($\mathcal{C} = 3$) compared to the DVM-Car subsets ($286 \geq \mathcal{C} \geq 276$, see Table 1), no differences can be observed between the resulting scores of the class-agnostic and the class-dependent methods. This further indicates that the applied augmentation methods are indeed affected by the unlabeled setting, rather than by a low amount of data samples sharing the same label per batch. Thus, the remaining experiments of this work were conducted in a fully labeled setting.

4.2.4 STiL-MoE Experiments

In the original architecture, only the multimodal classifier f^m is used at test time. As f^m receives the concatenation of $\hat{\boldsymbol{z}}_c^i$, $\hat{\boldsymbol{z}}^s$, and $\hat{\boldsymbol{z}}_c^t$ as input, it is affected regardless of which modality is shifted. On the other hand, the remaining classifiers, f^i and f^t, are only affected when their respective modality is shifted. Consequently, if only the other modality is shifted, they are expected to perform similarly to their behavior during training. To validate this, we evaluated the accuracy (ACC) and balanced accuracy (BACC) of each classifier individually on all subsets, applying only the original augmentation strategy to establish baseline behavior. For the ADNI subsets, the scores are presented in Table 7. While f^i shows severe performance issues on all subsets, with both ACC and BACC scores significantly lower than those of the other classifiers, the results still confirm our hypothesis, as they show that f^t is able to provide higher scores than f^m. With regard to the ADNI subsets, it attained higher balanced accuracy scores on the TE, the Age, and the No-Shift subsets, while its standard accuracy score is also superior on the latter subset. Particularly, the biggest margin can be observed on the No-Shift subset, which has the lowest

Table 7: Accuracy (ACC), balanced accuracy (BACC), and entropy ($H(f^j)$) scores of the individual classifiers during test time across the ADNI subsets. Values of f_i or f_t that exceed the corresponding scores of f_m are highlighted in bold. Considering BACC, f^t surpasses f^m on almost all subsets. ACC and BACC are presented in %.

Subset	Metric	f^m	f^i	f^t
TE	ACC	85.24 ± 0.94	50.67 ± 0.00	84.98 ± 08.57
	BACC	81.93 ± 1.26	55.56 ± 38.49	**82.86 ± 0.86**
	$H(f^j)$	0.523 ± 0.008	0.992 ± 0.005	0.531 ± 0.013
Age	ACC	81.87 ± 01.60	66.13 ± 0.00	81.68 ± 1.52
	BACC	81.84 ± 0.95	55.56 ± 38.49	80.57 ± 0.89
	$H(f^j)$	0.503 ± 0.017	1.005 ± 0.008	0.510 ± 0.009
Weight	ACC	85.42 ± 0.76	47.73 ± 0.00	85.24 ± 0.31
	BACC	84.13 ± 2.22	55.56 ± 38.49	**84.66 ± 0.23**
	$H(f^j)$	0.511 ± 0.009	1.000 ± 0.006	0.515 ± 0.006
No-Shift	ACC	84.00 ± 0.27	49.33 ± 0.00	**84.53 ± 0.92**
	BACC	75.23 ± 0.43	55.56 ± 38.49	**77.24 ± 0.44**
	$H(f^j)$	0.502 ± 0.009	0.996 ± 0.005	0.506 ± 0.013

score in the baseline. Here, the balanced accuracy score of f^t surpasses the one of f^m by 2.01%. Its entropy matches f^m, indicating that f^m makes more confident predictions, even when wrong. A similar behavior can be seen on the Black subset of the DVM-Car dataset, which is shown in Table 8. Here, the BACC score of f^t is 2.58% higher than the one of f^m. Considering the standard accuracy, the margin between the two classifiers is 2.61%. However, in this case, f^t exhibits a higher entropy score, indicating higher uncertainty. Additionally, for all other DVM-Car subsets and the Age subset, f^m remains the best-performing classifier. This highlights that the individual classifiers need to be included dynamically based on the current situation during test time. The classifiers' performance depends not only on the current distribution shift but also on the nature of the dataset itself.

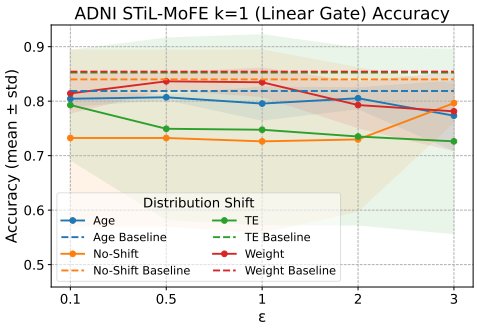
However, the contributions of f^t to the classification task are currently ignored. Therefore, we integrated the MoE framework into the STiL architecture. With this approach, the results of all classifiers are dynamically integrated into a final prediction. The goal is to make the model aware of the distribution shift it is currently encountering so that it can rely more on the classifier that is likely to provide better results.

Table 8: Accuracy (ACC), balanced accuracy (BACC), and entropy ($H(f^j)$) scores of the individual classifiers during test time across the DVM-Car subsets. Values of f_i or f_t that exceed the corresponding scores of f_m are highlighted in bold. ACC and BACC are presented in %.

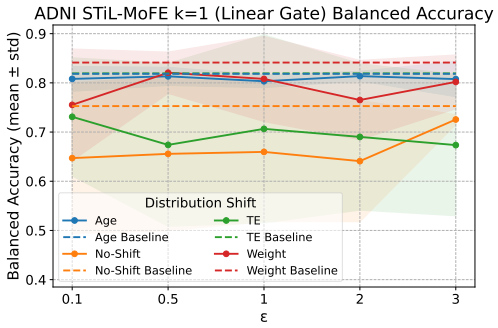
Subset	Metric	f^m	f^i	f^t
Black	ACC	90.09 ± 1.14	36.28 ± 2.43	**92.70 ± 0.38**
	BACC	88.42 ± 1.22	30.47 ± 2.19	**91.60 ± 0.49**
	$H(f^j)$	0.235 ± 0.014	1.476 ± 0.098	0.428 ± 0.009
Miles	ACC	91.77 ± 0.59	51.24 ± 1.81	89.16 ± 0.93
	BACC	89.14 ± 0.62	43.51 ± 1.76	87.19 ± 1.28
	$H(f^j)$	0.181 ± 0.009	1.095 ± 0.083	0.503 ± 0.019
Color-Miles	ACC	91.58 ± 0.51	59.28 ± 1.81	88.29 ± 0.92
	BACC	90.19 ± 0.66	54.42 ± 2.44	87.19 ± 0.87
	$H(f^j)$	0.159 ± 0.014	0.868 ± 0.123	0.543 ± 0.045
No-Shift	ACC	96.65 ± 0.17	66.14 ± 1.07	94.97 ± 0.14
	BACC	96.04 ± 0.35	62.20 ± 1.43	94.01 ± 0.11
	$H(f^j)$	0.089 ± 0.008	0.843 ± 0.043	0.346 ± 0.009

4.2.4.1 STiL-MoFE On the ADNI subsets, the first STiL-MoE variant was tested extensively with all $k \in \{1, 2, 3\}$. The goal behind this approach is to evaluate if the model benefits from all experts contributing to the task or only from a subset of them. As f^t outperforms f^m on three out of four subsets, it is expected that the model will benefit from at least 2 activated experts per data sample. On the other hand, the model could also learn to dynamically select the most fitting expert without being made uncertain by unreliable experts if k is set to 1. Furthermore, we also tested each setting with both a simple linear layer and an MLP network as the gate network. With this experiment setting, we can observe whether the simple architecture design that is commonly used in literature is sufficient for our setting, or whether a more complex approach would be preferable. Since the gate network has to process the information of all modalities at once, a network structure with a higher capacity may prove beneficial, as it could learn the complex relationships within its input. To determine the effect of the introduced loss terms, we evaluated each experiment with the values $\epsilon \in \{0.1, 0.5, 1, 2, 3\}$. Both the ACC and BACC scores of the linear layer and the MLP implementation with k set to 1 are illustrated in Fig. 25. The results demonstrate that the model severely suffers from this approach.

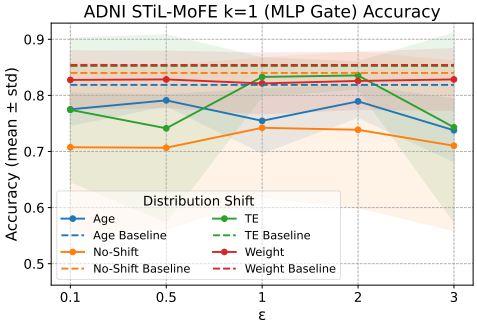
The mean scores of all subsets drop by high margins. For instance, the highest ACC score achieved with the No-Shift subset using the linear layer gate network is 4.35% lower (79.64%) than its baseline (84.00%) at $\epsilon = 3$, whereas the lowest score obtained is 11.38% (72.62%) lower at $\epsilon = 1$. The Age subset remains closest to its baseline value (81.86%) as it achieves an ACC score of 80.53% at $\epsilon = 2$, which corresponds to a margin of 1.33%. As for the BACC score, the model suffers as well. The highest performance degradation can be observed on the TE subset. Here, model performance drops by 8.83% (73.10%) compared to its baseline (81.93%) at the best setting of ϵ, which is $\epsilon = 0.1$ here. Its lowest score is 67.39%, which corresponds to a degradation of 14.54%, and it is obtained at $\epsilon = 0.5$. However, the Age subset remains relatively close to its baseline score across all settings of ϵ. Moreover, the MLP gate network implementation also performs poorly. Here, the No-Shift subset suffers the most. Its highest ACC score is obtained at $\epsilon = 1$ (74.22%) and it is 9.78% lower than its baseline value (84.00%). Meanwhile, the highest BACC score it achieves is still 12.4% (62.87%) lower than its baseline (75.27%). Furthermore, across all settings, the results not only exhibit great variability across all values of ϵ,



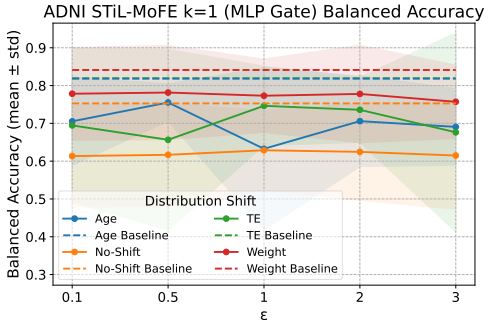
(a) ACC scores using the linear gate network.



(b) BACC scores using the linear gate network.



(c) ACC scores using the MLP gate network.

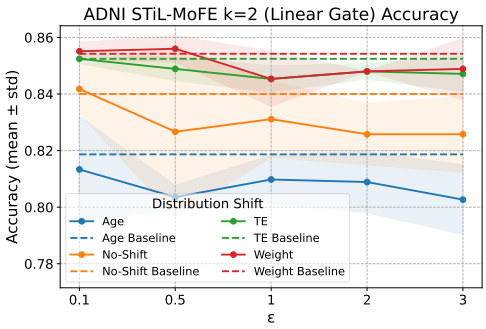


(d) BACC scores using the MLP gate network.

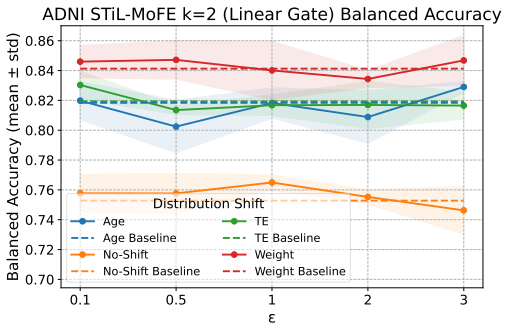
Figure 25: Performance metrics of the STiL-MoFE variant on the ADNI subsets across the different hyperparameter settings of ϵ, using both the linear layer and the MLP as the gate network. k is set to 1. The shades represent the standard deviation of the respective scores across the three different seeds.

they also showcase high standard deviations across the seed settings, indicating high instability with respect to network weight initialization. Therefore, in the context of STiL-MoFE, $k = 1$ is highly insufficient and destabilizes the model severely.

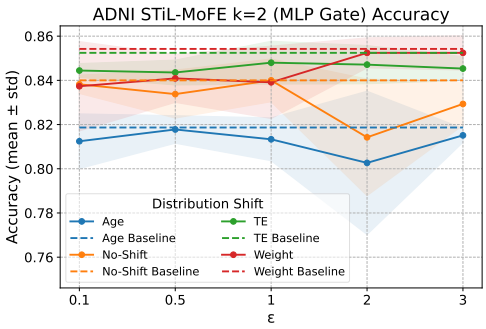
With k set to 2, the model performs noticeably better. The results are presented in Fig. 26. Not only are the per-seed standard deviations across the experiment settings substantially narrower than in the $k = 1$ results, but the mean scores also do not drop as severely as in the previous setting. In fact, the results even surpass the baseline by small margins at certain settings. For instance, the linear gating network provides an improvement of 1.1% in the BACC score (from 81.93% to 83.03%) at $\epsilon = 0.1$, whereas its ACC score remains at baseline level (85.24%). Conversely, the linear leads to a higher BACC value but also a lower ACC score on the No-Shift subset. Here, with ϵ being set to 1, the highest BACC score obtained (76.49%) is 1.22% higher than the baseline (75.27%), while its ACC score achieved a value of 83.11%, which is 0.89% lower than the baseline. However, as for the ACC score, the model exhibits the highest per-seed standard deviation on the No-Shift subset ($\sigma = 3.01\%$). The Weight subset benefits only by negligible margins from the linear layer. At $\epsilon = 0.5$, model accuracy increases by 0.18% (from 85.42% to



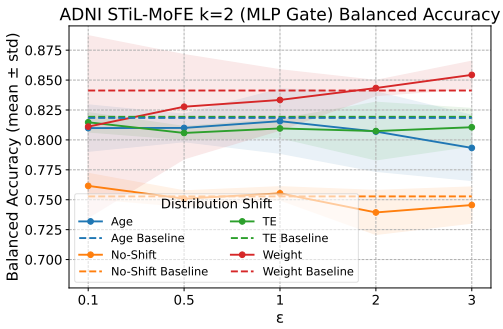
(a) ACC scores using the linear gate network.



(b) BACC scores using the linear gate network.



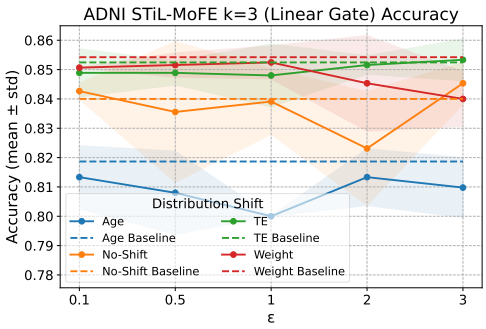
(c) ACC scores using the MLP gate network.



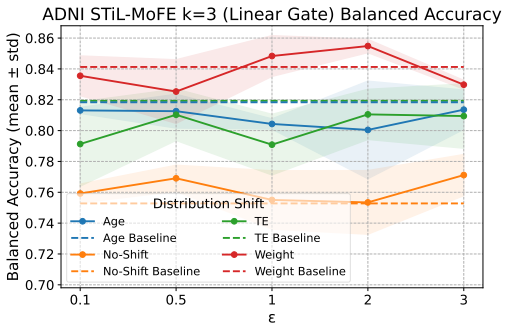
(d) BACC scores using the MLP gate network.

Figure 26: Performance metrics of the STiL-MoFE variant on the ADNI subsets across the different hyperparameter settings of ϵ, using both the linear layer and the MLP as the gate network. k is set to 2.

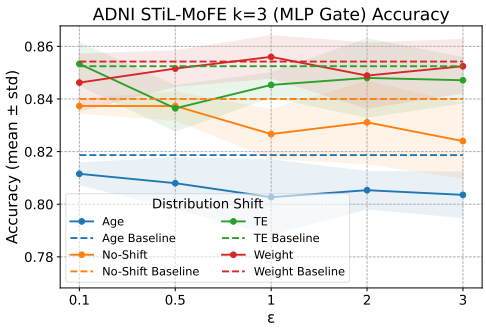
85.60%), whereas BACC increases by 0.59% (from 84.12% to 84.71%). The Age subset suffers across all settings of ϵ. Here, the best results are achieved at $\epsilon = 3$, where the ACC score is 1.60% below the baseline (from 82.90% to 81.83%) and BACC is 1.07% below (from 81.87% to 80.27%). The MLP gating network returns similar results. Considering the BACC score, the STiL-MoFE variant exceeds the baselines on two subsets. For the No-Shift subset, it exceeds the baseline by 0.88% with a score of 76.15% ($\epsilon = 0.1$), whereas it exceeds it by 1.31% on the Weight subset with a score of 85.43% ($\epsilon = 3$). Moreover, while the results of the other subsets remain mostly consistent across ϵ, the BACC score of the Weight subset increases with the value of ϵ, from 81.11% to 85.43%. The results also show that the per-seed standard deviation also shrinks (from $\sigma = 7.58\%$ to $\sigma = 1.13\%$). The ACC scores do not exceed the baseline values but approach them closely. For instance, the highest score, obtained at $\epsilon = 0.1$, using the No-Shift subset, is 83.82%, which is just 0.18% below its baseline (84.00%). With the Weight subset, the model also performs similarly to the baseline. At $\epsilon = 3$, it reaches 85.24%, which is 0.22% below its baseline (85.42%).



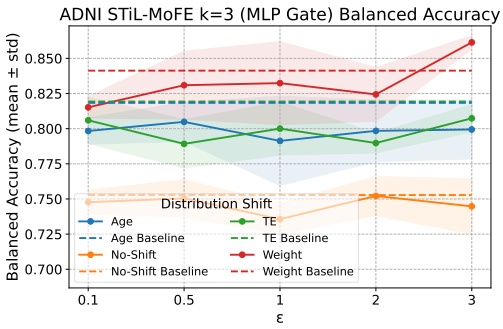
(a) ACC scores using the linear gate network.



(b) BACC scores using the linear gate network.



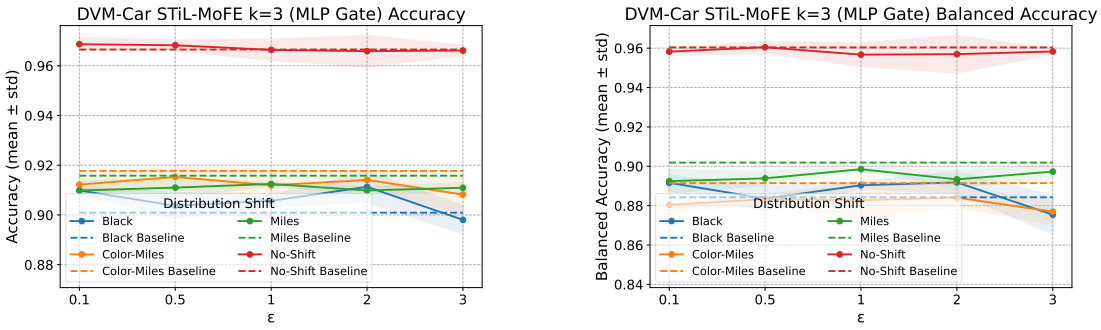
(c) ACC scores using the MLP gate network.



(d) BACC scores using the MLP gate network.

Figure 27: Performance metrics of the STiL-MoFE variant on the ADNI subsets across the different hyperparameter settings of ϵ, using both the linear layer and the MLP as the gate network. k is set to 3.

However, with respect to balanced accuracy, the model achieves its largest improvement margins when k is set to 3. The obtained scores of the performance metrics are illustrated in Fig. 27. With the linear gating network, the model reported better BACC scores when processing both the No-Shift and Weight subsets. On the Weight subset with ϵ set to 2, it achieves a BACC score of 85.48%, which is 1.35% higher than the respective baseline value of 84.13%. However, its ACC score also drops by 0.9% to 84.53%. While using the No-Shift subset, on the other hand, the model improves on both metrics. With $\epsilon = 3$, its ACC score climbs up by 0.53% (84.53%), while its BACC score reaches 77.11%, which corresponds to an improvement of 1.84%. While using the Age and TE subsets, however, the model does not benefit from architectural augmentation. With the Age subset, the model's highest BACC score is obtained at $\epsilon = 3$ and it is 0.48% (81.36%) lower than the baseline (81.84%). At this setting, its ACC score is 80.97%, which is 0.89% lower (81.86%). The highest BACC score using the TE subset (81.05%) is achieved with $\epsilon = 2$. Here, the model's ACC score remains close to the baseline with a value of 85.15%. For these two subsets, the MLP gating network does not lead to better results. Here, the model achieves a maximum BACC score of 80.48% (Age subset, $\epsilon = 0.5$), which corresponds to a 1.35% drop. The maximum score is 80.73% when using the TE subset. When assessing the performance of the model with the MLP gating network when processing the No-Shift subset, it exhibits a drop of 0.89% in accuracy (from 84.00% to 83.11%). Its BACC score remains consistent with the baseline. Nevertheless, the largest improvement margin can be observed on the Weight subset with ϵ being set to 3. Here, the BACC score achieves 86.14%. This is an improvement of 2.01% with respect to its baseline, which is 84.13%. Additionally, in this setting, the model's accuracy remains close to the baseline (85.24%), with a minor difference of 0.18% compared to the baseline (85.42%). Furthermore, the per-seed standard deviation of the balanced accuracy in this setting is also narrow, with a value of $\sigma = 0.46\%$, which implies robustness against weight initialization for this setting. Overall, these results indicate that $k = 3$ is the most promising setting when processing the ADNI subsets.



(a) ACC scores.

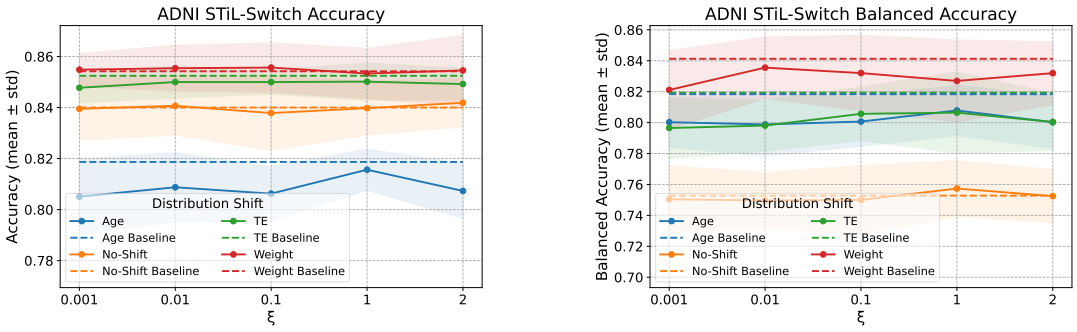
(b) BACC scores.

Figure 28: Performance metrics of the STiL-MoFE variant on the DVM-Car subsets across the different hyperparameter settings of ϵ. k is set to 3.

On the DVM-Car subsets, however, this effect cannot be observed. Here, the performance scores remain consistently close to the baseline values. The results are presented in Fig. 28. Furthermore, the model produces consistent performance values across all settings of ϵ, indicating robustness against this hyperparameter in this experiment. The largest margin between values of ϵ can be observed on the Black subset, from $\epsilon = 2$ to $\epsilon = 3$. Here, the model's ACC score drops from 91.13% to 89.80%. The BACC score is reduced by 1.64%, from 89.18% to 87.54%. Nevertheless, the model exceeds the baselines at $\epsilon = 2$ by 1.04% on the ACC score and by 0.66% on the BACC scores. These are the highest improvement margins relative to the baseline observed in this experiment. Moreover, the model only exhibits improvements when processing the Black subset. On the other subsets, it suffers regardless of the setting of ϵ. Overall, the results indicate that the extent to which the STiL-MoFE variant is beneficial depends on the underlying characteristics of the dataset. Also, the differing behaviors regarding the settings of ϵ between the datasets also indicate that the loss term $\mathcal{L}_{\text{STiL-MoFE}}$ further influences the model's stability.

4.2.4.2 STiL-Switch The second variant adopts the Switch Transformer architecture, allowing us to compare our first variant with a widely used approach in the literature. In this architecture, only a single expert is activated per embedding part, thus k is set to 1 across all settings. The model generates predictions for each embedding component individually, leading to greater granularity, which could result in better results since the embedding components are not evaluated jointly. However, the classifiers are not explicitly trained to become experts dedicated to handling the respective unimodal or multimodal data, as in the first variant, which could deteriorate performance in our setting. The experiments were run with the values $\xi \in \{0.001, 0.01, 0.1, 1, 2\}$.

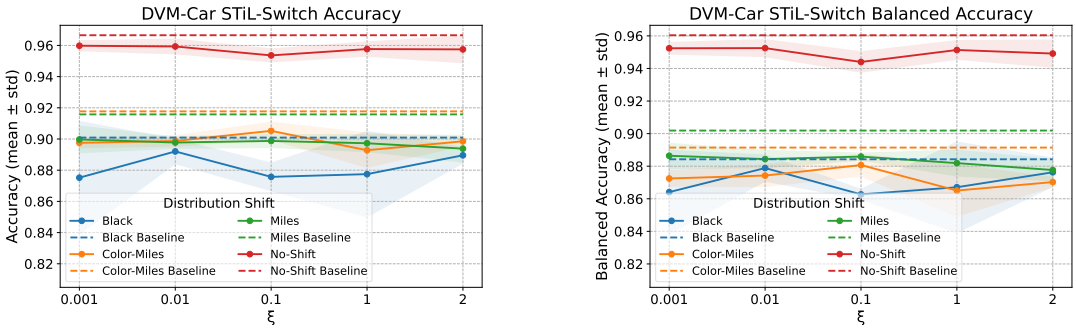
Compared to the results obtained from the STiL-MoFE variant with $k = 1$, the performance of STiL-Switch is substantially more stable.



(a) ACC scores.

(b) BACC scores.

Figure 29: Performance metrics of the STiL-Switch variant on the ADNI subsets across the different hyperparameter settings of ξ.



(a) ACC scores.

(b) BACC scores.

Figure 30: Performance metrics of the STiL-Switch variant on the DVM-Car subsets across the different hyperparameter settings of ξ.

The achieved scores while processing the ADNI subsets are illustrated in Fig. 29. The per-seed mean scores remain close to the baseline values, while the per-seed standard deviations also indicate an improved robustness against weight initialization. Furthermore, the model exhibits robustness against the value of ξ, as its performance metrics remain consistent throughout the hyperparameter's settings. However, compared to STiL-MoFE with $k = 3$, it provides only negligible improvement margins. The highest improvement margin can be observed at $\xi = 1$ while processing the No-Shift subset. Here, the BACC score is improved by 0.36% (from 75.28% to 75.74%), while the ACC score remains at baseline level.

Finally, the STiL-Switch variant was also evaluated using the DVM-Car subsets. The results of the final experiment are presented in Fig. 30. Here, the model exhibits a similar level of robustness to the choice of the hyperparameter ξ as observed when processing the ADNI subsets. However, the per-seed standard deviations on the Black subset on $\xi = 0.001$ and $\xi = 1$ indicate a higher instability with respect to weight initialization than in the STiL-MoFE setting. Additionally, none of the evaluated settings in this experiment surpass the baseline performance metrics. Therefore, the results indicate that while the STiL-Switch variant may be more stable than the STiL-MoFE variant, it does not achieve the same level of improvement.

5 Discussion

The results have shown that the effectiveness of the augmentation strategies examined in this thesis is highly dependent on multiple factors. Nevertheless, these outcomes still deliver valuable insights. First, the experiments using the latent embedding augmentation techniques show that the applied methods only lead to improvements when they are applied in a fully supervised setting. In the original Semi-SL setting, they result in serious deterioration in model performance. On the other hand, in the fully supervised setting while processing the ADNI subsets, the results remain close to the respective baselines throughout the experiments, which shows that augmentation applied in the latent embedding space can lead to equivalent results as input space augmentation. Here, the outcomes also indicate no difference between the class-agnostic and class-dependent methods. Nevertheless, the applied methods fail to close the gap between the different subsets and surpass the baseline scores only by negligible margins, indicating that they are not sufficient for handling data distribution shifts. The integration of the MoE framework, on the other hand, proves to be a more promising approach. Considering the balanced accuracy, the STiL-MoFE variant improves the model performance by up to 1-2% on various settings of ϵ when processing the ADNI subsets. However, when processing the ADNI subsets, the model demonstrates greater instability across different settings of both the MoE loss term hyperparameter and the weight initialization seed than when processing the DVM-Car subsets. This indicates that dataset characteristics, such as dataset size and data type, strongly influence the stability of the STiL-MoFE variant. In contrast, the STiL-Switch variant does not exhibit instability at a comparable scale. Here, the performance scores remain largely consistent across all settings of ϵ for both subsets. However, this variant does not result in improvements of the same magnitude, with both ACC and BACC remaining close to the baseline in the best-case scenarios. Nevertheless, the differing behavior across the evaluated performance metrics provides insight into the model's class-wise performance. The balanced accuracy calculates the average of the per-class recall (see Eq. (40)), whereas the standard accuracy calculates the overall proportion of correct predictions (see Eq. (39)). Thereby, the balanced accuracy gives all classes the same weighting regardless of their frequency. Thus, in cases where the STiL-MoE variants exhibit higher BACC scores while ACC remains at similar levels, this indicates that the proposed model architectures can become more sensitive to underrepresented classes when appropriately tuned. In cases where ACC decreases, this suggests that sensitivity to minority classes is traded off for increased sensitivity to majority classes.

Considering the results on the ADNI subsets, model performance is increased on three out of four subsets, namely the No-Shift, the Weight, and the TE subsets, which exhibit data distribution shifts in both image and tabular features. However, to this extent, these improvements are only observed when processing the ADNI subsets. When processing the DVM-Car subsets, we observe improvements only on the Black subset. Here, this is the only subset from this dataset that is purely shifted

within an image feature. This indicates that the augmentation strategy cannot be applied to real-world data without first considering the origin of the underlying data distribution shift. For medical data, the results imply that the shifted feature should not be positively correlated with the disease incidence rate, such as age in the case of AD. However, together with the findings presented in Table 7 and Table 8, they also imply that the performance of the unimodal classifiers such as f^t could be utilized as an indicator of successful integration. When f^t exhibits stronger performance than f^m, the STiL-MoE variants tend to yield better results.

However, this thesis also has limitations. First, all experiments concentrate on enhancing the performance of the STiL architecture. No other model has been evaluated. Thus, it is unclear whether the proposed augmentation strategies would lead to similar results if they were applied to other models. In particular, the benefit of integrating MoE into the model remains uncertain, as its effectiveness depends on the model successfully bridging the information modality gap by leveraging both modality-specific and modality-shared information. Thus, it remains an open question for future works to examine whether MoE can lead to benefits when applied to MMFMs following either the Co-Training or the Co-Regularization paradigms. Second, regarding the STiL-MoFE variant, the variability in performance across different MoE loss-term settings suggests that the proposed architecture is sensitive to hyperparameter selection. Since no extensive hyperparameter tuning was performed, it remains uncertain whether optimal hyperparameter settings could lead to more stable performance improvements. However, given the size of the model and the complexity of its loss function, searching for optimal hyperparameters is a challenging task. Another possible explanation for the observed instability is the relatively small size of the subsets. For instance, the DVM-Car subsets are substantially smaller than the dataset used by Du et al. (2025) to train the original model. Additionally, we only acquired access to the first dataset of ADNI, ADNI1, which consists of a relatively small number of data samples. Consequently, we argue that the examined performances and model stability in this work could also be enhanced by larger dataset sizes. The examination of the impacts of these properties is left for future works. Moreover, given that the STiL-MoE variants have only been examined in a fully supervised setting, it remains unclear whether the architectural changes would yield any benefits if applied in a Semi-SL setting.

Finally, other approaches could lead to improvements as well. For example, since data augmentation in both the input space and the latent embedding space leads to similar results, future work could explore whether combining these approaches further enhances model performance. Moreover, since the integration of the MoE framework has led to improved performance metric scores, model performance could also be improved by other model-merging paradigms. For instance, methodologies such as optimization merging and pruning merging have been reported to improve model generalization (Ruan et al., 2025). Here, the parameters of multiple, pretrained models sharing the same architecture are combined. Under this approach, multiple models could be trained on predefined subsets that exhibit distribution shifts across different features and subsequently merged to produce a final model

that is more robust to such shifts. Another promising approach is test-time adaption (Liang et al., 2025), which adapts the weights of a pre-trained model directly at test time. These methods have already shown promising results when evaluated under data distribution shifts.

6 Conclusion

In this thesis, we examined the effectiveness of different augmentation strategies to mitigate the deteriorating impact of data distribution shifts between training and test data on the STiL model architecture. While we analyzed the efficacy of state-of-the-art latent embedding augmentation techniques, we also proposed two novel model architectures, STiL-MoFE and STiL-Switch, that adapt the Mixture-of-Experts framework and demonstrate its potential to mitigate distribution shifts. With this approach, we were able to exploit both modality-shared and modality-specific information more efficiently by dynamically weighting the multimodal and unimodal classifiers within the CGPL module of the STiL architecture. The results of this work indicate that while latent embedding augmentation techniques can achieve performance comparable to transformations applied at the input space level, they do not surpass their effectiveness when used in isolation. As for the STiL-MoE variants, both have their own advantages and disadvantages: the STiL-MoFE can achieve noticeable improvements, but its performance is unstable with respect to the introduced hyperparameters. The STiL-Switch variant, on the other hand, exhibits greater stability in this regard, but at the cost of reduced performance gains. Future work could therefore investigate the influence of hyperparameters in greater detail, as optimal configurations may help mitigate the limitations of both variants.

Bibliography

ADNI. Alzheimer's disease neuroimaging protocol (adni). Technical report, 2005. URL `https://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI-1_Protocol.pdf`. Accessed: 08 January 2026.

Abeer Aljuaid and Mohd Anwar. Survey of Supervised Learning for Medical Image Processing. *SN Computer Science*, 3(4):292, May 2022. ISSN 2661-8907. doi: 10.1007/s42979-022-01166-1. URL `https://doi.org/10.1007/s42979-022-01166-1`.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: Applications, challenges, and future outlook. *Journal of Medical Internet Research*, 26:e59505, 2024. doi: 10.2196/59505. URL `https://doi.org/10.2196/59505`.

Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6679–6687. AAAI Press, 2021. doi: 10.1609/AAAI.V35I8.16826. URL `https://doi.org/10.1609/aaai.v35i8.16826`.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, November 2010. ISSN 1432-1882. doi: 10.1007/s00530-010-0182-0. URL `https://doi.org/10.1007/s00530-010-0182-0`.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=CuV_qYkmKb3`.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, August 2022. ISSN 1432-2315. doi: 10.1007/s00371-021-02166-7. URL `https://doi.org/10.1007/s00371-021-02166-7`.

Yoshua Bengio, Gregoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 552–560, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/bengio13.html.

L. Berg. Clinical Dementia Rating (CDR). *Psychopharmacology bulletin*, 24(4): 637–639, 1988. ISSN 0048-5764. Place: United States.

Nikhil Bhagwat, Joseph D. Viviano, Aristotle N. Voineskos, and M. Mallar Chakravarty. Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLoS Computational Biology*, 14(9): e1006376, 2018. doi: 10.1371/journal.pcbi.1006376. URL https://doi.org/10.1371/journal.pcbi.1006376.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279962. URL https://doi.org/10.1145/279943.279962.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, June 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3229161.

Sue Bowman. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information managemen*, 10(Fall):1c, 2013. ISSN 1559-4122. Place: United States.

Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 137–144, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143862. URL https://doi.org/10.1145/1143844.1143862.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124, Aug 2010. doi: 10.1109/ICPR.2010.764.

Charles E. Brown. *Coefficient of Variation*, pages 155–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-642-80328-4. doi: 10.1007/978-3-642-80328-4_13. URL https://doi.org/10.1007/978-3-642-80328-4_13.

Viktoras Bulavas, Virginijus Marcinkevičius, and Jacek Rumiński. Study of multi-class classification algorithms' performance on highly imbalanced network intrusion datasets. *Informatica*, 32(3):441–475, 2021. doi: 10.15388/21-INFOR457.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL `https://www.mdpi.com/2078-2489/11/2/125`.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):3896–3915, July 2025. ISSN 1558-2191. doi: 10.1109/TKDE.2025.3554028.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 2010.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1): 321–357, June 2002. ISSN 1076-9757.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23049–23062. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf`.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1779–1788. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/cheng20b.html`.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek

Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, and Nicole Brichtova. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv e-prints*, art. arXiv:2507.06261, July 2025. doi: 10.48550/arXiv.2507.06261.

Lingxi Cui, Huan Li, Ke Chen, Lidan Shou, and Gang Chen. Tabular data augmentation for machine learning: Progress and prospects of embracing generative AI. *CoRR*, abs/2407.21523, 2024. doi: 10.48550/ARXIV.2407.21523. URL `https://doi.org/10.48550/arXiv.2407.21523`.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL http://arxiv.org/abs/1708.04552.

Terrance DeVries and Graham W. Taylor. Dataset augmentation in feature space. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HyaF53XYx.

Thomas G. Dietterich. Ensemble methods in machine learning. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1. URL https://doi.org/10.1007/3-540-45014-9_1.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/du22c.html.

Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P O'Regan, and Chen Qin. Tip: Tabular-image pre-training for multimodal classification with incomplete data. In *European Conference on Computer Vision*, pages 478–496. Springer, 2024.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Stil: Semi-supervised tabular-image learning for comprehensive task-relevant information exploration in multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15549–15559, June 2025.

Juergen Dukart, Karsten Mueller, Arno Villringer, Ferath Kherif, Bogdan Draganski, Richard Frackowiak, and Matthias L. Schroeter. Relationship between imaging biomarkers, age, progression and symptom severity in alzheimer's disease. *NeuroImage: Clinical*, 3:84–94, 2013. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2013.07.005. URL https://www.sciencedirect.com/science/article/pii/S2213158213000922.

Justin Engelmann and Stefan Lessmann. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.114582. URL https://www.sciencedirect.com/science/article/pii/S0957417421000233.

Dongyang Fan, Bettina Messmer, and Martin Jaggi. Towards an empirical understanding of moe design choices. *CoRR*, abs/2402.13089, 2024. doi: 10.48550/ARXIV.2402.13089. URL https://doi.org/10.48550/arXiv.2402.13089.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, aug 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.84. URL https://aclanthology.org/2021.findings-acl.84/.

Cassidy M. Fiford, Gerard R. Ridgway, David M. Cash, Marc Modat, Jennifer Nicholas, Emily N. Manning, Ian B. Malone, Geert Jan Biessels, Sebastien Ourselin, Owen T. Carmichael, M. Jorge Cardoso, and Josephine Barnes. Patterns of progressive atrophy vary with age in alzheimer's disease patients. *Neurobiology of Aging*, 63:22–32, 2018. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2017.11.002. URL https://www.sciencedirect.com/science/article/pii/S0197458017303718.

Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975. ISSN 0022-3956. doi: https://doi.org/10.1016/0022-3956(75)90026-6. URL https://www.sciencedirect.com/science/article/pii/0022395675900266.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, June 2016. doi: 10.1109/CVPR.2016.265.

S. Gillette-Guyonnet, F. Nourhashemi, S. Andrieu, I. de Glisezinski, P. J. Ousset, D. Riviere, J. L. Albarede, and B. Vellas. Weight loss in alzheimer disease. *The American Journal of Clinical Nutrition*, 71(2):637S–642S, feb 2000. doi: 10.1093/ajcn/71.2.637s. URL https://doi.org/10.1093/ajcn/71.2.637s.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, March 2022. ISSN 1558-2531. doi: 10.1109/TBME.2021.3117407.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *CoRR*, abs/2212.11491, 2022. doi: 10.48550/ARXIV.2212.11491. URL https://doi.org/10.48550/arXiv.2212.11491.

Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23924–23935, June 2023.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 379–389. IEEE, 2023. doi: 10.1109/WACVW58289.2023.00042. URL https://doi.org/10.1109/WACVW58289.2023.00042.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. A comprehensive survey on contrastive learning. *Neurocomputing*, 610:128645, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.128645. URL https://www.sciencedirect.com/science/article/pii/S0925231224014164.

Jingming Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. DVM-CAR: A large-scale automotive dataset for visual marketing research and applications. In *Proceedings of the IEEE International Conference on Big Data*, pages 4130–4137, 2022.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3 (1):136, October 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z. URL `https://doi.org/10.1038/s41746-020-00341-z`.

Weichen Huang. Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2465–2474, Oct 2023. doi: 10.1109/ICCVW60793.2023.00261.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Generalizable audio deepfake detection via latent space refinement and augmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888328.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. doi: 10.1109/ICCV.2017.167.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, March 1991. ISSN 0899-7667. doi: 10.1162/neco.1991.3.1.79.

Hyun-Jong Jang and Kyung-Ok Cho. Applications of deep learning for the analysis of medical data. *Archives of Pharmacal Research*, 42(6):492–504, June 2019. ISSN 1976-3786. doi: 10.1007/s12272-019-01162-9. URL `https://doi.org/10.1007/s12272-019-01162-9`.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *CoRR*, abs/2401.04088, 2024. doi: 10.48550/ARXIV.2401.04088. URL `https://doi.org/10.48550/arXiv.2401.04088`.

Bernd André Jung and Matthias Weigel. Spin echo magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, 37(4):805–817, 2013. doi: https://doi.org/10.1002/jmri.24068. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.24068`.

Daniel I. Kaufer, Jeffrey L. Cummings, Patrick Ketchel, Vanessa Smith, Audrey MacMillan, Timothy Shelley, Oscar L. Lopez, and Steven T. DeKosky. Validation of the npi-q, a brief clinical form of the neuropsychiatric inventory. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 12(2):233–239, 2000.

doi: 10.1176/jnp.12.2.233. URL https://psychiatryonline.org/doi/abs/10.1176/jnp.12.2.233.

Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, 55(3):2351–2377, March 2022. ISSN 1573-7462. doi: 10.1007/s10462-021-10066-4. URL https://doi.org/10.1007/s10462-021-10066-4.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Pieter Kubben. Data sources. In *Fundamentals of Clinical Data Science*, pages 3–9. Springer, Cham (CH), 2019.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 1–10, 2019. doi: 10.18653/v1/D19-6101. URL https://aclanthology.org/D19-6101/.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. Labeled data generation with encoder-decoder LSTM for semantic slot filling. In Nelson Morgan, editor, *17th Annual Conference of the International Speech Communication Association, Interspeech 2016, San Francisco, CA, USA, September 8-12, 2016*, pages 725–729. ISCA, 2016. doi: 10.21437/INTERSPEECH.2016-727. URL https://doi.org/10.21437/Interspeech.2016-727.

Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103 (9):1449–1477, Sep. 2015. ISSN 1558-2256. doi: 10.1109/JPROC.2015.2460697.

Lenore Launer, Kjeld Andersen, M Dewey, Luc Letenneur, Alewijn Ott, L Amaducci, Carol Brayne, John Copeland, J Dartigues, P Kragh-Sorensen, Antonio Lobo, Jose Martinez-Lage, Theo Stijnen, and A Hofman. Rates and risk factors for dementia and alzheimer's disease: Results from eurodem pooled analyses. *Neurology*, 52:78–84, 02 1999.

Dominik Lewy and Jacek Mańdziuk. An overview of mixing augmentation methods and augmentation strategies. *Artif. Intell. Rev.*, 56(3):2111–2169, June 2022. ISSN 0269-2821. doi: 10.1007/s10462-022-10227-z. URL https://doi.org/10.1007/s10462-022-10227-z.

Pengzhi Li, Yan Pei, and Jianqiang Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, 138:110176, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110176. URL https://www.sciencedirect.com/science/article/pii/S1568494623001941.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439, May 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3532688.

Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision*, 133 (1):31–64, jan 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02181-w. URL https://doi.org/10.1007/s11263-024-02181-w.

Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. Learning multimodal data augmentation in feature space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=6SRDbbvU8s.

Cuihua Lv, Lizhou Fan, Haiyun Li, Jun Ma, Wenjing Jiang, and Xin Ma. Leveraging multimodal deep learning framework and a comprehensive audio-visual dataset to advance parkinson's detection. *Biomedical Signal Processing and Control*, 95:106480, 2024. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2024.106480. URL https://www.sciencedirect.com/science/article/pii/S174680942400538X.

Héctor P. Martínez and Georgios N. Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 34–41, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328852. doi: 10.1145/2663204.2663236. URL https://doi.org/10.1145/2663204.2663236.

Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6445–6449, 2019. doi: 10.1109/ICASSP.2019.8683438.

Mahsa Moraveji and Najme Mansouri. Recent Advances in Computational and Machine-Learning Approaches for Alzheimer's Disease Classification: A Comprehensive Review. *Archives of Computational Methods in Engineering*, December 2025. ISSN 1886-1784. doi: 10.1007/s11831-025-10476-5. URL https://doi.org/10.1007/s11831-025-10476-5.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *CoRR*, abs/2503.07137, 2025. doi: 10.48550/ARXIV.2503.07137. URL https://doi.org/10.48550/arXiv.2503.07137.

Susanne Mueller, M.W. Weiner, L.J. Thal, Ronald Petersen, Clifford Jack, and W. Jagust. Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative. *Cogn Dement*, 5:56–62, 01 2006.

Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022. ISSN 2590-0056. doi: https://doi.org/10.1016/j.array.2022.100258. URL https://www.sciencedirect.com/science/article/pii/S2590005622000911.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025. doi: 10.48550/ARXIV.2508.10925. URL https://doi.org/10.48550/arXiv.2508.10925.

Sherjil Ozair and Yoshua Bengio. Deep directed generative autoencoders. *CoRR*, abs/1410.0630, 2014. URL http://arxiv.org/abs/1410.0630.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

R. I. Pfeffer, T. T. Kurosaki, Jr. Harrah, C. H., J. M. Chance, and S. Filos. Measurement of functional activities in older adults in the community1. *Journal of Gerontology*, 37(3):323–329, 05 1982. ISSN 0022-1422. doi: 10.1093/geronj/37.3.323. URL https://doi.org/10.1093/geronj/37.3.323.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in MAchine Learning*. MIT Press, Cambridge, MA, USA, 2009. ISBN 978-0-262-17005-5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Allen D. Roses and Ann M. Saunders. Apoe is a major susceptibility gene for alzheimer's disease. *Current Opinion in Biotechnology*, 5(6):663–667, 1994. ISSN 0958-1669. doi: 10.1016/0958-1669(94)90091-4. URL https://www.sciencedirect.com/science/article/pii/0958166994900914.

Wei Ruan, Tianze Yang, Yifan Zhou, Tianming Liu, and Jin Lu. From task-specific models to unified systems: A review of model merging approaches. *CoRR*, abs/2503.08998, 2025. doi: 10.48550/ARXIV.2503.08998. URL https://doi.org/10.48550/arXiv.2503.08998.

Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE Access*, 9:135392–135422, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3116481.

Rick Sauber-Cole and Taghi M. Khoshgoftaar. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1):98, August 2022. ISSN 2196-1115. doi: 10.1186/s40537-022-00648-6. URL `https://doi.org/10.1186/s40537-022-00648-6`.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=B1ckMDqlg`.

Xu Shen, Xinmei Tian, Anfeng He, Shaoyan Sun, and Dacheng Tao. Transform-invariant convolutional neural networks for image classification and search. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1345–1354, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336031. doi: 10.1145/2964284.2964316. URL `https://doi.org/10.1145/2964284.2964316`.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021. URL `https://arxiv.org/abs/2108.13624`.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Inf. Fusion*, 81:84–90, 2022. doi: 10.1016/J.INFFUS.2021.11.011. URL `https://doi.org/10.1016/j.inffus.2021.11.011`.

Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 399–402, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930442. doi: 10.1145/1101149.1101236. URL `https://doi.org/10.1145/1101149.1101236`.

Shriyank Somvanshi, Subasish Das, Syed Aaqib Javed, Gian Antariksa, and Ahmed Hossain. A survey on deep tabular learning. *CoRR*, abs/2410.12034, 2024. doi: 10.48550/ARXIV.2410.12034. URL `https://doi.org/10.48550/arXiv.2410.12034`.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, jan 2022. ISSN 1477-4054. doi: 10.1093/bib/

bbab569. URL https://doi.org/10.1093/bib/bbab569. _eprint: https://academic.oup.com/bib/article-pdf/23/2/bbab569/42805085/bbab569.pdf.

Lakpa Tamang, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. Handling out-of-distribution data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):5948–5966, Oct 2025. ISSN 1558-2191. doi: 10.1109/TKDE.2025.3592614.

Neil C. Thompson, Kristjan H. Greenewald, Keeheon Lee, and Gabriel F. Manso. The computational limits of deep learning. *CoRR*, abs/2007.05558, 2020. URL https://arxiv.org/abs/2007.05558.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Haoran Wang, Zhen-Song Chen, Mingjie Fang, Yilong Wang, and Feng Liu. Panoramic sales insight: Using multimodal fusion to improve the effectiveness of flash sales. *Decision Support Systems*, 190:114401, 2025a. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2025.114401. URL https://www.sciencedirect.com/science/article/pii/S0167923625000028.

Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2025b. doi: 10.1109/TKDE.2025.3622600.

Fan Yang, Bo Ning, and Huaiqing Li. An overview of multimodal fusion learning. In Yan Chenggang, Wang Honggang, and Lin Yun, editors, *Mobile Multimedia Communications*, pages 259–268, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-23902-1.

Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 4092–4098. AAAI Press, 2019. ISBN 9780999241141.

Jerome A. Yesavage and Javaid I. Sheikh. 9/geriatric depression scale (gds). *Clinical Gerontologist*, 5(1-2):165–173, 1986. doi: 10.1300/J018v05n01_09. URL https://doi.org/10.1300/J018v05n01_09.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, Oct 2024. ISSN 1558-2256. doi: 10.1109/JPROC.2024.3507831.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self- and semi-supervised learning to tabular domain. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11033–11043. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/7d97667a3e056acab9aaf653807b4a03-Paper.pdf.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5), January 2025. ISSN 0360-0300. doi: 10.1145/3711118. URL https://doi.org/10.1145/3711118.

Yifan Zhan, Rui Yang, Junxian You, Mengjie Huang, Weibo Liu, and Xiaohui Liu. A systematic literature review on incomplete multimodal learning: techniques and challenges. *Systems Science & Control Engineering*, 13(1):2467083, 2025. doi: 10.1080/21642583.2025.2467083.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=otHZ8JAIgh.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9), April 2024. ISSN 0360-0300. doi: 10.1145/3649447. URL https://doi.org/10.1145/3649447.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. CLIP in medical imaging: A survey. *Medical Image Analysis*, 102:103551, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103551. URL https://www.sciencedirect.com/science/article/pii/S1361841525000982.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6xHJ37MVxxp.

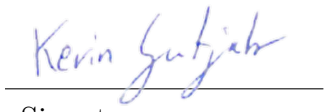
Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg, 29.01.2026

Place, Date



Signature