University of Sheffield

Multimodal Bayesian fusion for detection and classification of Alzheimer's disease

Charalambos Georgiades

Supervisor: Mr. Areeb Sherwani

A report submitted in fulfilment of the requirements
for the degree of MSc in Cybersecurity and Artificial Intelligence

in the

Department of Computer Science

September 11, 2024

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Charalambos Georgiades

Signature: Charalambos Georgiades

Date: 07 June 2024

Abstract

Alzheimer's disease is a prevalent neurodegenerative condition accounting for the majority of dementia cases worldwide. In recent years, there has been a surge of machine learning models being developed for its diagnosis, as early detection is critical for effective intervention. While models that make use of multiple biomarkers have shown great promise in aiding the diagnostic process, their adoption in clinical practice remains heavily limited in part due to challenges in implementation and lack of reliability in their predictions. The question our work aims to answer is: could uncertainty quantification, an approach commonly used to heighten trust in machine learning models, serve as a viable alternative to complex multimodal fusion? In this dissertation project, we developed a simple, distinct model for each modality and quantified the uncertainty in its predictions using MC Dropout. Through uncertainty-based model selection, we achieved accuracies upwards of 93%, highlighting the potential of this approach. Although the uncertainties themselves did not enchance the interpretability of the models as much as we initially hoped, our findings provide a compelling foundation for future research, with several promising directions discussed.

Acknowledgements

I would like express my gratitude to Mr. Areeb Sherwani, my supervisor, for his invaluable assistance and direction throughout my dissertation project. In addition, I want to thank my parents, brother, and other family members for their unwavering support over the past year. I would also like to express my gratitude to my friends, George and Stefanos, for supporting me no matter what and for helping me unwind and enjoy my time at university. Finally, I acknowledge that all data used in this dissertation project was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and has been crucial to the completion of my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disease which causes memory loss and cognitive decline due to degeneration and death of brain cells. It is the most common form of dementia, a general term for diseases which affect memory, thinking, and ability to perform daily tasks, accounting for 60%-70% of cases (World Health Organization, 2023). The gradual loss of bodily functions eventually leads to death, with the average life expectancy post diagnosis being 1 to 12 years (Todd et al., 2013; Schaffert et al., 2022).

Despite substantial research, the cause of AD remains mostly unknown with no current treatments able to stop or reverse its progression, though medication may temporarily help cope with the symptoms or slow down the disease in some patients. Research suggests that by 2050, 1 in 85 persons worldwide will be living with the disease (Brookmeyer et al., 2007), highlighting the urgent need for effective early diagnostic tools.

Early detection is critical, especially during the disease's early stage known as amnestic mild cognitive impairment (MCI) where interventions may have a greater chance of delaying progression to AD. Current diagnostic approaches mostly rely on techniques such as neuroimaging and cerebrospinal fluid analysis. Despite evidence indicating that combining multiple biomarkers can enhance diagnostic accuracy by providing complementary information (Foster et al., 2007; de Leon et al., 2007; Fjell et al., 2010; Apostolova et al., 2010), the majority of existing pattern classification techniques for AD diagnosis only employ a single modality.

Recent works such as (Zhang et al., 2011; Venugopalan et al., 2021; Qiu et al., 2022) have made use of multimodal models, models which integrate biomarkers from different modalities, aiming to improve diagnostic accuracy. The proportions of studies using either single-modality or multi-modality approaches in AD research are illustrated in *Figure 1.1*. Despite showing great potential, these multimodal models still face significant challenges in their widespread adoption. One major issue is the difficulty in implementation. These models

require the integration of diverse data types which necessitates sophisticated architectures and considerable computational resources. This complexity can pose a barrier for organizations that lack the necessary infrastructure or expertise to effectively create and deploy such systems. Moreover, the lack of transparency in how multimodal models process and interpret different modalities raises concerns. These models usually function as black boxes, making it difficult to understand how they combine and weigh various inputs, leading to potential issues with trust, accountability, and bias in decision-making.

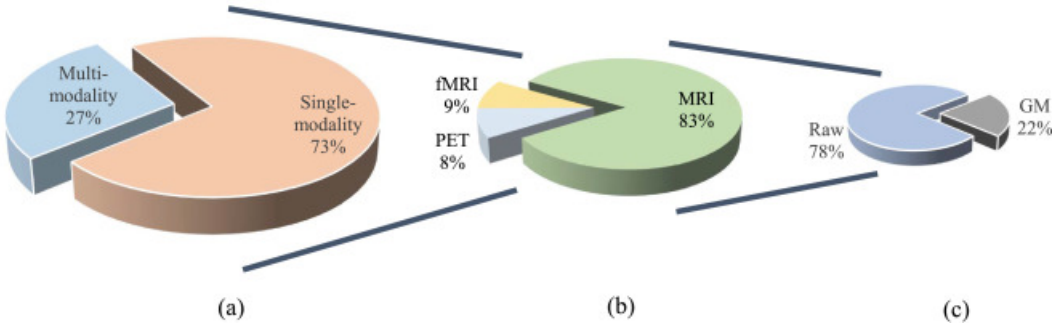

Figure 1.1. (a) The proportion of studies using either a single modality or multiple modalities; (b) among the single-modality studies, the specific neuroimaging technique utilized; and (c) the frequency with which grey matter measures (GM) are employed in MRI-based studies. (Ebrahimighahnavieh et al., 2020)

1.2 Aims and Objectives

Uncertainty quantification (UQ), a technique that measures the confidence associated with a prediction, is an approach often used to improve the interpretability of machine learning models. But can it also be used as a means of fusing multiple modalities? In this dissertation project we aim to investigate whether UQ, when applied to individual biomarkers, is sufficient for accurate prediction, reducing the need for the more resource- and computationally intensive multimodal fusion process. To achieve this aim, the following objectives have been outlined:

1. **Review and analyze existing literature**: Conduct a comprehensive review of current methodologies in both uncertainty quantification and multimodal fusion, exploring relevant works in the process.

2. **Develop a framework leveraging uncertainty for predictions**: Develop a framework that leverages uncertainty when combining modalities based on the research done in the literature review.

3. **Implement the proposed framework**: Implement the proposed framework using state-of-the-art machine learning techniques, ensuring it is robust and scalable.

4. **Evaluate and interpret the framework's performance**: Evaluate the performance of the proposed method using popular metrics and conduct further testing to better understand the observed behavior.

5. **Suggest future research directions**: Based on the findings, propose potential avenues for future research in the field.

1.3 Overview of the Report

The remainder of the dissertation project is structured as follows: Chapter 2 covers necessary and relevant concepts in machine learning, uncertainty quantification, and multimodal fusion in the form of a literature survey. Chapter 3 analyzes the project and breaks it down into smaller steps, Chapter 4 details the implemented design, Chapter 5 lists the experiments conducted and corresponding results. Finally, a brief conclusion along with possible future research directions can be found in Chapter 6.

Chapter 2

Literature Survey

2.1 Overview

Achieving this dissertation's goal requires knowledge of multiple areas of machine learning and Bayesian statistics. The fundamental concepts of neural networks and their Bayesian variants are covered first, providing the necessary background. Following this, uncertainty quantification techniques are explored, with an emphasis on their relevance and suitability for this project. The survey subsequently moves to an analysis of popular multimodal fusion approaches. Finally, relevant works are discussed in the context of the dissertation's objectives.

2.2 Machine Learning Prerequisites

This section serves as a brief introduction to the types of neural networks needed for this project.

2.2.1 Feed-Forward Neural Network

A Feed-Forward Neural Network (FFNN) (Rumelhart et al., 1986) is a type of artificial neural network characterized by the direction of the flow of information between its layers. They are unidirectional, with information flowing forward from the input nodes, through the hidden nodes (if any), and to the output nodes without any cycles or loops. In contrast, recurrent neural networks have a bidirectional flow as data can flow in cycles over time. Modern FFNNs are trained using backpropagation and are often referred to as basic or "vanilla" neural networks.

In the forward pass, the weights $\mathbf{W}$ and bias $\mathbf{b}$ are used to transform the input vector $\mathbf{x}$ into a vector of elements using $\mathbf{W}\mathbf{x} + \mathbf{b}$. An activation function σ such as Rectified Linear Unit (ReLU) or Hyperbolic Tangent (Tanh) is subsequently applied to the resulting vector to introduce non-linearity, allowing the model to learn more complex patterns, and obtain the output of the layer. If multiple hidden layers are used, this process is repeated for each one

in order where the input of one layer is the ouput of the previous one. The output weight is then used to map the last hidden layer to the output. For a FFNN with a single hidden layer: $\hat{y} = \mathbf{W_2}\sigma(\mathbf{W_1}\mathbf{x} + \mathbf{b_1}) + \mathbf{b_2}$. In a binary classification problem, the sigmoid function computes the probability p of input vector $\mathbf{x}$ belonging to class 1 as follows:

$$p(\text{class=1}) = \frac{1}{1 + e^{-\hat{y}}}$$

where $\hat{y}$ is the model's output score. The probability for class 0 is $1 - p(\text{class=1})$. For multiclass classification problems on the other hand, the softmax function normalizes the output of the model to a probability distribution over the predicted output classes, essentially calculating the probability p_i for each class i using:

$$p_i = \frac{e^{\hat{y}_i}}{\sum_{j=1}^{K} e^{\hat{y}_j}}$$

where $\hat{y}_i$ is the output score for class i and K represents the number of classes.

As previously mentioned, backpropagation is used to train the network after the forward pass is completed. It involves a backward pass through the network, computing the gradients of the loss function with respect to each weight using the chain rule, enabling their efficient computation. The algorithm begins by computing the gradient of the loss with respect to the output and subsequently propagates these gradients backwards through the network layers while adjusting the weights in each one to minimize the loss. The weight updates are typically performed using optimization algorithms such as Stochastic Gradient Descent (SGD) or Adaptive Moment Estimation (Adam). For a given layer, the weight update rule for a weight $\mathbf{W}$ is:

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$$

where η is the learning rate. The gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ is computed as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathbf{W}}$$

where $\mathcal{L}$ is the loss function and $\frac{\partial \hat{y}}{\partial \mathbf{W}}$ is derived from the chain rule through each subsequent layer of the network.

2.2.2 Convolutional Neural Network

Convolutional Neural Networks, or CNNs (LeCun et al., 1989), are a type of FFNN that use filter (or kernel) optimization to learn spatial hierarchies of features. Due to their ability to effectively capture spatial and temporal dependencies, they have become the de-facto standard when it comes to computer vision and image processing tasks.

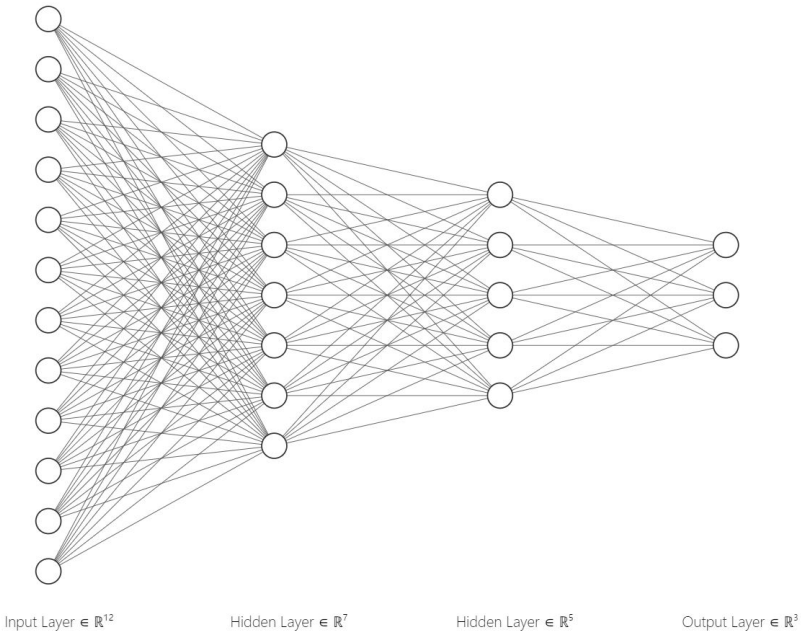

Figure 2.1. An example Feed Forward Neural Network (LeNail, 2019).

A typical CNN architecture is comprised of multiple types of layers: convolutional layers, activation layers, pooling layers, and fully connected layers. Convolutional layers, are the primary components of a CNN, consisting of learnable filters that glide over the input data to produce output feature maps, detecting specific features such as edges and textures. The operation is mathematically expressed as:

$$F_l = \sigma\left(\mathbf{W_l}\mathbf{x_{l-1}} + \mathbf{b_l}\right)$$

where F_l is the output feature map of the l-th layer, $\mathbf{W_l}$ are the weights (filters), $\mathbf{x_{l-1}}$ is the input from the previous layer, and $\mathbf{b_l}$ is the bias. Activation functions introduce non-linearity and allow a CNN to learn more complex patterns. Pooling layers reduce the dimensions of the feature maps, reducing the number of learnable parameters and amount of computation performed. Lastly, fully connected or dense layers are almost always used at the end of the network to aggregate the features extracted from the convolutional layers and produce class probabilities (such as in the case of classification).

CNNs are trained using a backpropagation algorithm similar to that used in FFNNs. During the forward pass, input data is passed through the network, and feature maps are computed. The final output is compared to the true labels using a loss function and during the backward pass, gradients of the loss with respect to the weights and biases are computed using the chain rule. The gradients are then used to adjust the weights through optimization algorithms.

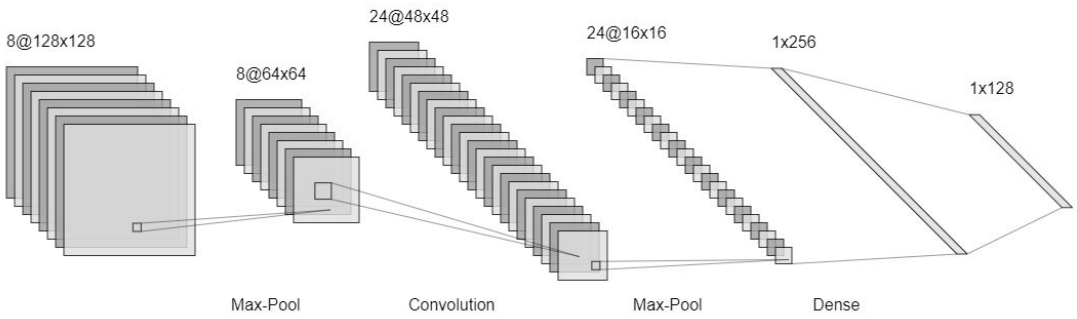

Figure 2.2. An example Convolutional Neural Network (LeNail, 2019).

2.3 Uncertainty

Uncertainty refers to the state of being uncertain; not being able to predict an outcome with complete confidence. It can arise due to lack of information, inadequate observed quantity, varying procedure of measurement, and the measuring device itself. Uncertainty can affect planning, risk assessment, and the ability to make informed choices, often necessitating strategies to manage or mitigate its impact. In the context of machine learning, uncertainty refers to the lack of confidence in a model's prediction.

There are two main types of uncertainty: epistemic and aleatoric uncertainties (Hüllermeier and Waegeman, 2019). Epistemic, also known as knowledge uncertainty, occurs due to inadequate knowledge or gaps in the data. On the other hand, aleatoric or data uncertainty is the irreducible uncertainty inherent from the data distribution and not the property of a model.

In machine learning, uncertainty can be modeled in various ways depending on its type-epistemc or aleatoric. Epistemic uncertainty can be model using Bayesian approaches such as Bayesian Neural Networks or variational inference, which allow for capturing the uncertainty in the model's parameters. Aleatoric uncertainty, on the other hand, is modeled by estimating the noise or variability inherent in the data. This can be captured by augmenting the output of the model with a probability distribution, where the variance reflects the degree of data uncertainty.

2.3.1 Bayesian Neural Network

An extension of FFNNs, BNNs use probability distributions for weights and biases instead of single values (Mullachery et al., 2018). This allows them to incorporate probabilistic reasoning, expressing uncertainty and continuously updating their beliefs. Consequently, BNNs can assess the probability of each prediction and offer insights into the accuracy of the output.

Bayesian probabilistic modeling captures uncertainty by estimating the posterior distribution over the weights:

$$P(\mathbf{W}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{D})} = \frac{P(\mathbf{D}|\mathbf{W})P(\mathbf{W})}{\int P(\mathbf{D}|\mathbf{W})P(\mathbf{W})\, d\mathbf{W}}$$

where $\mathbf{W}$ are the weights, $\mathbf{D}$ represents the data, $P(\mathbf{W})$ is the prior probability based on $\mathbf{W}$, $P(\mathbf{D}|\mathbf{W})$ is the likelihood of the observations, and $P(\mathbf{W}|\mathbf{D})$ is the posterior probability.

When new data is observed, its likelihood is used to update the prior, forming the posterior distribution which reflects the updated belief about the parameter. Since computing the exact posterior can be complex or impractical for large datasets and complicated models, approximation methods such as Markov Chain Monte Carlo and variational inference are often used.

Due to their architecture, BNNs are particularly useful when predictions require an understanding of confidence levels, when dealing with small or noisy datasets, or in safety-critical applications such as healthcare and medicine. Although computationally intensive, difficult to scale, and complex, BNNs require less training data and can handle uncertainty well while simultaniously preventing overfitting.

2.4 Uncertainty quantification using Bayesian techniques

Bayesian deep learning in general is a substitute to standard deep learning methods utilized primarily when information about the reliability of predictions is necessary. Several Bayesian deep learning techniques used for quantifying uncertainty are described below.

2.4.1 Monte Carlo Dropout

Dropout, and more specifically Bernoulli Dropout, is an extremely popular and effective regularization technique employed to combat overfitting in deep neural networks (Srivastava et al., 2014). During training, each unit in a layer is retained with a probability p_i, and a binary mask $\mathbf{z}$ is sampled for each unit from a Bernoulli distribution:

$$z_i \sim \text{Bernoulli}(p_i)$$

The forward pass for layer l with dropout is:

$$y_l = \mathbf{z} \odot (\mathbf{W_l x_{l-1}} + \mathbf{b_l})$$

where $\odot$ denotes element-wise multiplication. At test time, dropout is not applied, but activations are scaled by p_i:

$$h_l = p_i \cdot (\mathbf{W_l x_{l-1}} + \mathbf{b_l})$$

Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) approximates the prediction uncertainty by applying dropout at test time and performing multiple stochastic forward passes. In BNNs, we would ideally integrate over all possible network weights to fully capture uncertainty. However, this is computationally intractable for large models. MC Dropout approximates this process by using dropout as a variational approximation to the posterior distribution of the model weights. Thus, it doesn't exactly estimate the true uncertainty but provides a computationally efficient approximation. For a test input $\mathbf{x}$, predictive mean $\hat{y}$ and variance σ^2 are estimated as:

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^{T} y_t(\mathbf{x})$$

$$\sigma^2(x) = \frac{1}{T} \sum_{t=1}^{T} (y_t(\mathbf{x}) - \hat{y}(\mathbf{x}))^2$$

where $\{y_t(\mathbf{x})\}_{t=1}^{T}$ are the predictions from T stochastic forward passes.

Different variants of dropout techniques exist and can be found here (McClure and Kriegeskorte, 2017).

2.4.2 Markov Chain Monte Carlo

Monte Carlo algorithms are computational techniques that use repeated random sampling to approximate numerical results. The basic idea involves generating random samples, evaluating a function or process at these points, and aggregating the results to produce an approximation. For example, the integral of a function $f(x)$ over an interval $[a, b]$ can be approximated by:

$$I \approx \frac{b-a}{N} \sum_{i=1}^{N} f(x_i)$$

where x_i are random points uniformly distributed in $[a, b]$. Similarly, the expected value of a random variable X with probability density function $p(x)$ can be estimated by:

$$E[X] \approx \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_i are samples drawn from $p(x)$.

A Markov Chain is a stochastic model that describes a sequence of events where the probability of each event depends only on the state attained in the preceding event. Mathematically, for a sequence $x_1, x_2, \ldots, x_n$, the probability of transitioning to x_i given the previous states $x_{i-1}, \ldots, x_1$ is determined solely by the previous state x_{i-1}:

$$p\left(x_i \mid x_{i-1}, \ldots, x_1\right) = T\left(x_i \mid x_{i-1}\right)$$

Here, T represents the transition probabilities between states, indicating the likelihood of moving from state x_{i-1} to state x_i.

Combining these two methods, Markov Chain Monte Carlo or MCMC (Brooks, 1998) obtains approximate solutions to complex problems by sampling from probability distributions. These methods rely on constructing a Markov Chain where the stationary distribution of the chain corresponds to the target distribution of interest. In MCMC, a sequence of states $x_1, x_2, \ldots, x_n$ is generated such that the probability of transitioning to x_i from x_{i-1} depends on a proposal distribution and an acceptance criterion, often guided by detailed balance conditions. The transition probability in MCMC is typically defined as:

$$p\left(x_i \mid x_{i-1}\right) = T\left(x_i \mid x_{i-1}\right)$$

where T represents the transition kernel, which governs how the chain moves between states.

2.4.3 Variational inference

Classical methods for approximating Bayesian inference such as the aforementioned MCMC are notoriously slow (Andrieu and Thoms, 2008). Variational inference (VI) is a technique used to approximate compound posterior densities for Bayesian models by reframing it as an optimization problem (Blei et al., 2017). Assuming a family of approximate densities $\mathscr{Q}$, we approximate the posterior with the optimized member of the family $q^*(\cdot)$ that minimizes the Kullback-Leibler (KL) divergence to the exact posterior.

$$q^*(z) = \arg\min_{q(z)\in\mathscr{Q}} \mathrm{KL}(q(z) \parallel p(z \mid x)).$$

Mathematically, the KL divergence from distribution Q to distribution P is defined as:

$$\mathrm{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

for discrete probability distributions, or

$$\mathrm{KL}(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

for continuous probability distributions.

VI and MCMC solve the same problem using different approaches. MCMC algorithms approximate the posterior with samples from the Markov Chain while VI algorithms approximate the posterior using the result of the optimization. MCMC techniques yield (asymptotically) accurate samples from the target density, although being computationally more demanding than VI (Robert and Casella, 2000). While Variational inference tends to be faster than MCMC, it does not enjoy such guarantees and can only identify a density close to the target. Variational inference readily leverages techniques such as stochastic and distributed optimization since it is based on optimization.

2.4.4 Bayes By Backprop

Bayes by Backprop is a VI method used to train Bayesian Neural Networks, allowing uncertainty quantification of the weights (Blundell et al., 2015). The method uses variational inference to approximate the intractable posterior distribution of the weights, $p(\theta|D)$, with a simpler distribution $q(\theta|\phi)$. Once again, the objective is to minimize the KL divergence between these distributions. The loss function is defined as follows:

$$F(\mathcal{D}, \theta) = \mathrm{KL}[q(\mathbf{w}|\theta) \parallel P(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w}|\theta)}[\log P(\mathcal{D}|\mathbf{w})]$$

Stochastic gradient descent (SGD) and backpropagation are used to minimize the compression cost, known as variational free energy (VFE) or expected lower bound of the marginal likelihood, using the above loss function. Gradients are computed by backpropagating through the network with weights sampled from $q(\mathbf{w}|\theta)$, allowing the model to balance data fitting and uncertainty estimation effectively.

2.4.5 Laplace Approximations

Laplace approximations (LA) is one of the simplest family of approximations used for calculating the intractable posteriors of Deep Neural Networks (Laplace, 1774; MacKay, 1992). Despite its simplicity, the LA family is not as popular as alternative UQ methods such as variational inference or deep ensembles. LA approximates a posterior distribution $p(\theta|y)$ with a Gaussian distribution centered at the mode of the posterior. Given a posterior distribution from Bayes' theorem:

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

where θ represents the parameters and y the observed data, the Laplace approximation centers the Gaussian at the mode $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y).$$

The second-order Taylor expansion of the log-posterior around $\hat{\theta}$ is:

$$\log p(\theta|y) \approx \log p(\hat{\theta}|y) - \frac{1}{2}(\theta - \hat{\theta})^T H(\theta - \hat{\theta}),$$

where H is the Hessian matrix of the negative log-posterior evaluated at $\hat{\theta}$:

$$H = -\nabla^2 \log p(\theta|y)\big|_{\theta=\hat{\theta}}.$$

This yields the Gaussian approximation:

$$p(\theta|y) \approx \mathcal{N}(\hat{\theta}, H^{-1}),$$

where $\mathcal{N}(\hat{\theta}, H^{-1})$ is a normal distribution with mean $\hat{\theta}$ and covariance H^{-1}. This

approximation simplifies the integration in Bayesian inference, making it computationally efficient to estimate the posterior mean and variance.

2.5 Uncertainty quantification using Ensemble methods

Ensemble methods are techniques that create and combine multiple machine learning models to get better predictive performance than a single model can. They have recently gained traction due to their ability to improve predictive performance and quantify uncertainty. By combining multiple models, ensemble methods can provide more reliable predictions and better capture the uncertainty inherent in the data.

2.5.1 Deep Ensemble

This method involves training multiple instances of the same neural network architecture, each initialized with different random weights and potentially trained with different subsets of the data (through techniques like bagging) (Lakshminarayanan et al., 2017). The final prediction is obtained by averaging the predictions of all individual models in the ensemble.

Deep ensembles offer several advantages. Firstly, the varying initial weights of the models lead to different convergence paths even if trained on the same dataset, which introduces diversity in the models. Secondly, by averaging the predictions of multiple models, deep ensembles mitigate overfitting, leading to more robust and generalized predictions. Thirdly, the variance among the predictions of individual models provides an estimate of the epistemic uncertainty. If the models disagree significantly, it indicates high uncertainty in the prediction. Mathematically, if $f_i(x)$ represents the prediction of the i-th model in the ensemble for input x, the ensemble prediction $\hat{y}$ is given by:

$$\hat{y}(x) = \frac{1}{M} \sum_{i=1}^{M} f_i(x)$$

where M is the total number of models in the ensemble. The uncertainty can be quantified by the variance of these predictions:

$$\sigma^2(x) = \frac{1}{M} \sum_{i=1}^{M} \left(f_i(x) - \hat{y}(x)\right)^2$$

Empirical studies have demonstrated that deep ensembles often outperform single models in terms of both predictive performance and uncertainty estimation (Wilson and Izmailov, 2019; Brigato and Iocchi, 2021).

2.5.2 Bayesian Deep Ensemble

Bayesian deep ensembles extend the concept of deep ensembles by incorporating Bayesian principles to provide a more rigorous approach to uncertainty quantification. This approach

combines the strengths of deep ensembles and Bayesian inference, aiming to capture both epistemic and aleatoric uncertainties. Each model in the ensemble is interpreted as a sample from the posterior distribution over the model parameters given the data.

Implementing Bayesian deep ensembles involves training multiple models, similarly to the deep ensemble approach. However, each model is trained with techniques that encourage sampling from the posterior distribution, such as stochastic gradient descent with warm restarts, dropout as approximate Bayesian inference and MC Dropout.

The ensemble prediction and uncertainty estimation in Bayesian deep ensembles are similar to those in deep ensembles, with the added Bayesian interpretation providing a richer understanding of the uncertainty. Mathematically, if θ_i represents the parameters of the i-th model sampled from the posterior, the ensemble prediction $\hat{y}$ and uncertainty can be formulated as:

$$\hat{y}(x) = \frac{1}{M} \sum_{i=1}^{M} f(x; \theta_i)$$

$$\sigma^2(x) = \frac{1}{M} \sum_{i=1}^{M} (f(x; \theta_i) - \hat{y})^2$$

By leveraging Bayesian inference, Bayesian deep ensembles provide a more principled way to quantify uncertainty, capturing both epistemic and aleatoric uncertainties. Additionally, predictions from Bayesian deep ensembles tend to be better calibrated, meaning the predicted probabilities better reflect the true likelihood of events.

2.6 Other uncertainty quantification techniques

Below, other UQ techniques not suitable for this project are briefly mentioned.

2.6.1 Bayesian Active Learning

Active Learning (AL) methods learn from unlabeled samples by querying an oracle for the most informative data points (Settles, 2011). As semi-supervised learning techniques, they are primarily used when labels are scarce or expensive to obtain. The main challenge they face is defining a suitable acquisition function. In this regard, Bayesian approaches can be combined with the AL structure to represent uncertainty and probe the oracle for uncertain samples.

AL is generally suitable for Alzheimer's disease classification as the labeled data can be limited, has high dimensionality and is imbalanced regarding class. However, AL methods are not suitable for our project as an oracle or expert is not available and thus, will not be explored further.

2.6.2 Variational Autoencoders

Variational Autoencoders (VAEs) are a type of generative model that combine neural networks with VI to learn latent representations of data (Pinheiro Cinelli et al., 2021). Unlike traditional autoencoders, VAEs impose a probabilistic structure on the latent space, enabling them to generate new, similar data points by sampling from this learned distribution. This is achieved by encoding input data into a latent space that follows a prior distribution, usually a Gaussian, and then decoding samples from this space back into the data space. VAEs are widely used in tasks such as image generation, data compression, and anomaly detection due to their ability to model complex data distributions.

VAEs are not ideal for the task at hand as they focus on unsupervised learning and generative tasks rather than optimizing discriminative features necessary for accurate diagnosis. They often capture broad data variations, including irrelevant features, which can hinder classification performance. Additionally, they may struggle with the complexity of medical imaging data, leading to poor reconstruction quality and potential loss of critical diagnostic information.

2.6.3 Deep Gaussian processes

Deep Gaussian Processes (DGPs) extend Gaussian Processes (GPs) by stacking multiple GP layers, creating a hierarchical model that captures complex, non-linear relationships in data. DGPs introduce uncertainty in predictions through their probabilistic nature, inheriting the uncertainty quantification from GPs at each layer. This results in a richer expression of epistemic uncertainty, representing the model's uncertainty about the underlying function, and allowing for more robust and interpretable predictions in the presence of limited data.

They are well-suited for Alzheimer's disease classification due to their ability to model complex, non-linear relationships and quantify uncertainty in predictions. By stacking multiple Gaussian Process layers, DGPs can handle high-dimensional and heterogeneous data such as neuroimaging and genetic information, which are crucial for accurate Alzheimer's diagnosis. However, the computational complexity and data requirements of DGPs are practical considerations that need to be managed for effective implementation. DGPs require careful tuning of hyperparameters, which can be a time-consuming process. Due to computational and time limitations, they are not viable options for this project.

2.7 Multimodal fusion

Multimodal fusion is the process of combining data from several modalities to predict a deisred outcome, such as a class through classification or a continuous value through regression. (Baltrusaitis et al., 2017) have separated the key approaches into two categories: model-agnostic and model-based methods.

2.7.1 Model-agnostic methods

Model-agnostic methods for multimodal fusion are not directly dependent on specific machine learning methods and can be implemented using almost any unimodal classifier or regressor, the main motive behind their popularity (D'mello and Kory, 2015). These approaches are commonly split into early, late, intermediate, and hybrid fusion (Atrey et al., 2010; Stahlschmidt et al., 2022).

In early multimodal fusion, raw or minimally processed data from different sources are merged into a single, unified representation before being fed into a machine learning model. The notion behind early fusion is that it allows the model to learn and leverage correlations and interactions between low level features of each modality. It is the easiest to implement out of the four as it only requires training a single model (Barnum et al., 2020).

On the other hand, late fusion makes use of unimodal decision values and fuses them using a learnt model, channel noise and signal variance weighting, voting systems, or averaging. Different models can be used for each modality allowing for more flexibility and better individual performance. It also allows for training when no parallel data is available and predictions when one or more modalities are absent is made easier. Unfortunately, any low level correlations or interactions between the modalities are ignored.

Hybrid fusion aims to combine the benefits of early and late fusion by performing fusion at both the feature and decision levels.

Finally, intermediate fusion, which is the most widely used fusion strategy, fuses and processes the latent representations of each individual modality to obtain the final output scores. Examples of how this fusion is performed includes element-wise addition, concatenation, averaging, and attention mechanisms. Intermediate fusion allows the model to learn rich interactions between modalities by converting them into machine-understandable representations, but it requires individual processing for each modality, impacting inference and processing speed.

2.7.2 Model-based methods

The unimodal machine learning methods used by model-agnostic strategies are not intended for multimodal data whereas model-based methods are specifically designed to perform multimodal fusion. Three such methods are Multiple Kernel Learning (MKL), graphical models, and neural networks.

MKL methods are an extension to kernel Support Vector Machines (KSVMs) which allow for better fusion of heterogeneous data by using different kernels for different modalities (Gönen and Alpaydın, 2011). An advantage of MKLs is their flexibility in kernel selection and the convexity of their loss function, which enables the use of standard optimization packages and guarantees global optimum solutions. MKLs are versatile, applicable to both regression

and classification tasks. However, a significant disadvantage is their dependency on training data during test time, resulting in slow inference times and a large memory footprint. (Liu et al., 2014) have utilized MKL for multimodal fusion in Alzheimer's disease classification with promising results.

Graphical models can be split into two categories: generative models, which model joint probability, and discriminative models, which model conditional probability (Getoor and Taskar, 2007). Early generative models for multimodal fusion include coupled and factorial hidden Markov models, as well as dynamic Bayesian networks. However, generative models have largely been replaced by discriminative models such as conditional random fields, which prioritize predictive power. The main advantages of graphical models are their ability to exploit spatial and temporal data structures, incorporate expert knowledge, and produce interpretable results.

Neural networks have been extensively applied to multimodal fusion tasks (Akkus et al., 2023). Both shallow and deep neural models have been explored for multimodal fusion, leveraging architectures like Recurrent Neural Networks and Long Short-Term Memory networks for temporal information fusion. Deep neural networks excel in data fusion due to their ability to learn from large datasets and perform end-to-end training of multimodal representations and fusion components. However, their major drawbacks lie in interpretability and training requirements. It is challenging to discern which features or modalities contribute most to predictions, and they require substantial amounts of training data to achieve optimal performance.

2.8 Relevant works

Deep learning has made notable advancements in the field of medical imaging and diagnosis in the last few years, with CNNs having attracted the most attention out of all architectures due to their high performance in image classification tasks. Numerous works have utilized such models to identify signs of AD or MCI in brain scans, predominantly using MRIs.

(Gunawardena et al., 2017) used 2D coronal MRI scans converted from 3D ones for conducting two experiments aiming to evaluate the best pre-detection method for AD, with the experiment using CNN outperforming the SVM one. (Puente-Castro et al., 2020) successfully used transfer learning techniques to detect AD in saggital MRI images, an imaging view not typically used for such tasks. (Bae et al., 2020)'s model used coronal slices from the MRI scans of two diverse populations to achieve high within-dataset and between-dataset AUCs with a mean processing time of 23–24 seconds per person.

Many studies have shifted their focus towards 3D convolutional networks for medical imaging tasks due to their ability to capture spatial and temporal features in scans by processing volumetric data albeit at an increased computational cost. (Aaraji and Abbas, 2022)

constructed and evaluated several deep learning architectures on brain MRI images and segmented images in an attempt to investigate the influence of image segmentation on deep learning classification. (Cuingnet et al., 2011)'s paper aimed to evaluate ten classification methods on the same subset of data so that their performance can be properly compared.

Regarding multimodal learning, works such as (Zhang et al., 2011), (Venugopalan et al., 2021), and (Qiu et al., 2022) have concluded that both, deep and shallow models integrating multiple biomarkers typically outperform single modality models.

In recent years, there has been an insufficient number of papers which have attempted to quantify uncertainty in medical imaging tasks despite it being a field requiring informed decision making. One such paper, (Yang and Fevens, 2021), proposed an UQ system for general medical imaging classification tasks utilizing MC Dropout and Deep Ensembles, evaluating it on two distinct datasets.

Chapter 3

Analysis

3.1 Project Analysis

AD diagnosis presents unique challenges, particularly in its early stages like amnestic mild cognitive impairment, where accurate detection can significantly impact disease progression. Traditional machine learning diagnostic methods often rely on single-modality approaches, typically neuroimaging or cerebrospinal fluid (CSF) analysis, which limit the scope of analysis by focusing on just one type of data. While effective to a degree, these methods may fail to capture the complex, multi-faced nature of AD, resulting in lower diagnostic accuracy.

Multimodal models, models that integrate various biomarker modalities-such as magnetic resonance imaging (MRI), positron emission tomography (PET), cerebrospinal fluid (CSF) analysis, and electronic health records (EHR)—aim to improve diagnostic performance by providing complementary information. However, the integration of diverse data sources introduces new challenges, including increased model complexity and computational requirements. One example is how and at what point should the data be fused? Furthermore, the interaction between different biomarkers in multimodal models can be difficult to decipher, further reinforcing the black box architecture of machine learning models. This lack of transparency hinders clinical adoption as healthcare professionals require clear explanations of how diagnoses are reached. (Lahat et al., 2015; Zhang et al., 2019)

Is the increased complexity of multimodal models justified by their potential for greater diagnostic accuracy? While multimodal fusion holds promise, its complex and opaque nature raise concerns. This project will explore whether simplifying the diagnostic process by relying on individual biomarkers and their associated uncertainty can offer a more interpretable and practical alternative to full multimodal fusion.

3.2 Project breakdown

Given the scope of this project, it will be broken down into smaller, more manageable steps.

3.2.1 Dataset acquisition and preprocessing

The first step involves identifying, acquiring, and preprocessing the relevant dataset. Due to the sensitive and multimodal nature of the data required, sources are limited and often require permission to access and use. Some of the most widely used databases for Alzheimer's research include the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Australian Imaging, Biomarkers & Lifestyle Study of Ageing (AIBL), the Alzheimer's Disease Data Initiative (ADDI), and the European Prevention of Alzheimer's Dementia (EPAD). Once an appropriate dataset has been identified, the data must be retrieved in compliance with all necessary data protection regulations. Finally, preprocessing is an essential step needed to ensure that the data is consistent and ready for analysis. This step may involve converting various medical imaging formats (e.g., NIfTI, MINC, HDF5, DICOM) into a standardized format, as well as performing any necessary data cleaning or applying augmentations. Although many works limit themselves to one dataset, we believe that using multiple could reveal more regarding the generalization abilities of our models.

3.2.2 Model design, implementation, and optimization

The second step focuses on the design, implementation, and subsequent optimization of machine learning models for each modality. Examples of hyperparameters that should be tuned include, but are not limited to, the input image dimensions, number of layers, number of neurons, and dropout rate. Given the inherent differences between modalities, it is possible that the performance of a single model may vary across different inputs. Therefore, modality-specific models may be required.

3.2.3 Uncertainty quantification

For the third step, we will focus on quantifying the uncertainty in the predictions generated by our implemented models. Among the various UQ methods discussed in the literature review, we will prioritize Monte Carlo Dropout due to its simplicity, computational efficiency, and proven effectiveness. If time permits, we aim to explore additional UQ techniques and compare their performance in terms of the quality of the uncertainty estimates.

3.2.4 Bayesian fusion

The final step is developing a framework that uses the calculated uncertainty to fuse the modalities. This step is crucial for evaluating whether our approach can serve as a valid alternative to typical multimodal fusion.

3.3 Evaluation

The evaluation of the framework will focus on both performance metrics and interpretability. Traditional performance metrics such as accuracy, precision, recall, and F-score will be used to

assess its predictive power. Additionally, the area under the receiver operating characteristic (ROC) curve or AUC score will be considered to evaluate the its ability to distinguish between different classes. The same evaluation will also be conducted on the single-modality models for a comparison to be conducted and assess the added value of integrating different data types.

The impact of uncertainty quantification on the reliability and trustworthiness of the predictions will be analyzed, primarily through figures, with a focus on how uncertainty estimates correlate with model errors.

3.4 Ethical, professional, and legal issues

Given that this project involves the handling of sensitive medical data, strict adherence to ethical, professional, and legal standards is crucial. The project will fully comply with the UK's General Data Protection Regulation (UK GDPR) and other relevant data protection laws such as the Data Protection Act 2018 to ensure the privacy and security of all personal data. To protect patient confidentiality, all data will be anonymized to guarantee confidentiality and ensure that individuals cannot be directly identified. The project will adhere to the principles of data minimization, ensuring that only the data strictly necessary for the analysis is collected, processed, and stored. The data will only be used for the purposes of this project and not for any other unrelated activities.

Chapter 4

Design and Implementation

4.1 Description of dataset

The data to be used for this project has been taken from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database due to its suitability for our cause. ADNI is a longitudinal multicenter study which began in 2004 with the goal of developing clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease.

The data consists of 1,320 3D MRI and PET scans in total, with 660 scans from each modality. With 220 scans per class in each group, the three categories were Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN). The scans were obtained from 191 distinct individuals. Due to the lack of subjects available, multiple scans were taken from individual subjects with a 1-1 ratio, meaning that an equal number of MRI and PET scans were taken from each subject.

Figure 4.1 reveals age imbalance within the classes as all groups are skewed towards the 70-79 and 80-89 age ranges, with fewer participants in the younger and older brackets. In regards to gender, *Figure 4.2* indicates that the MCI group is notably imbalanced, with almost double the number of males than females, while the CN and AD groups show a more balanced distribution, though the AD group still has slightly more males than females.

The scans taken from the ADNI dataset had already undergone slight preprocessing. The MRI scans had been processed using the GradWarp technique to correct geometric distortions caused by gradient field non-linearities which can introduce minor spatial distortion artifacts and blurring at the extreme margins of MR images. On the other hand, the PET scans underwent longer processing. First, the scans were smoothed to reduce noise and enhance clarity. Next, coregistration was calculated to align the PET scans with a reference image, and this alignment was then applied to the scans to correct any misalignments. Finally, multiple frames were averaged to create a single image.

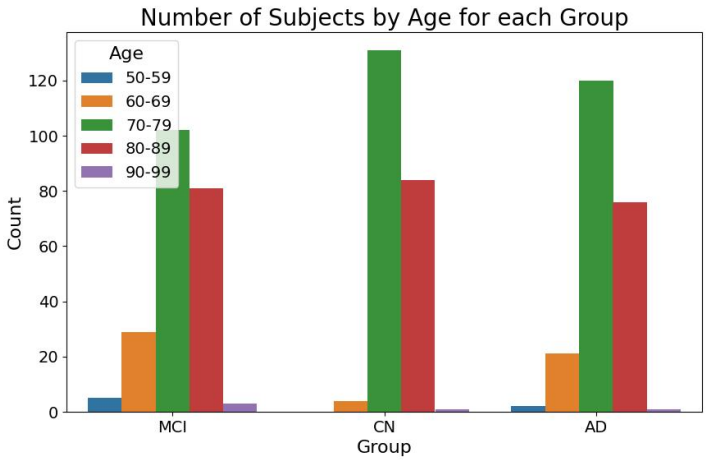

Figure 4.1. Age demographics for a single modality

4.2 Data preprocessing

Data preprocessing is the essential task of transforming raw data into a clean dataset, guaranteeing its accuracy and usefulness for modeling and analysis (Singh et al., 2021). Noise, missing numbers, and inconsistencies are common in raw data, which can result in incorrect outcomes and subpar model performance. Preprocessing solves these problems, boosts the accuracy and efficiency of machine learning algorithms, and improves the trustworthiness of insights by cleansing, normalizing, and augmenting the data. *Figure 4.5* visualizes the preprocessing methodology implemented in this work while *Figure 4.6* showcases its effects on a sample MRI scan.

4.2.1 Converting to unified data format

Converting the multimodal imaging data to a unified format is a necessary preprocessing step that ensures consistency, compatibility, and data quality. It simplifies data handling, making it more efficient and reliable, which results in fewer errors and enhanced interoperability.

The ADNI dataset uses different formats to store the various modalities. MRIs are stored using the Neuroimaging Informatics Technology Initiative (NIfTI) format while PET scans are stored using the Digital Imaging and Communications in Medicine (DICOM) imaging format. We have chosen to convert all images to NIfTI, for our implementation. The reasoning behind this decision is that DICOM image files are made up of multiple 2D layers while NIfTI images are stored in a 3D format which was specifically designed to overcome spatial orientation challenges arising in other medical file formats.

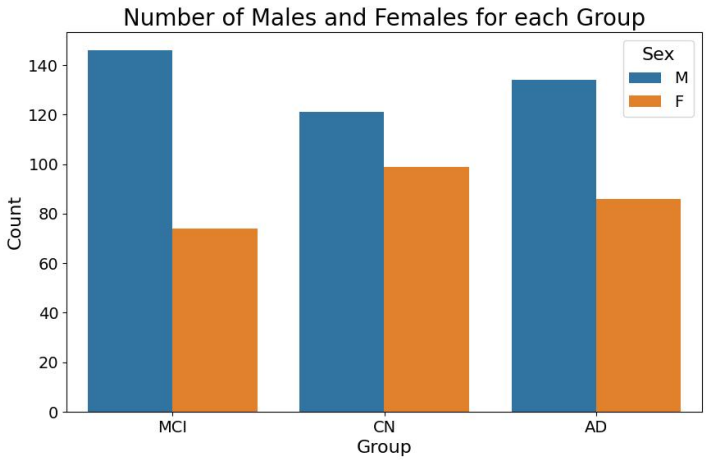

Figure 4.2. Sex demographics for a single modality

4.2.2 Data scaling

Anatomical differences among patients, variability in imaging protocols, and the adaptability of medical imaging technology lead to variations in the dimensions and number of slices in medical images, even in those of the same modality. Therefore, the data must be standardized to common dimensions. While altering the dimensions of 2D images is a fairly straightforward task, dealing with 3D images of varying depth is a more complicated one.

In our approach, both MRI and PET scans are resized to 128 x 128 x 60 voxels. To do so, we first create a binary mask identifying voxels in the original image that fall below a set threshold, marking them as background. The image is then resized using cubic spline interpolation, which provides a smooth transformation. The background mask is resized using nearest-neighbor interpolation to maintain its binary nature. Finally, the resized image's voxels corresponding to the background mask are set to zero, ensuring that the background regions remain unchanged in the output image.

4.2.3 Data splitting

At this point, the data was randomly split into two subsets, training and testing, with 80% for training and 20% for testing. This was done early in the preprocessing pipeline to prevent data leakage and ensure that the model evaluation remains unbiased. The split resulted in 1056 total scans for training and 264 for testing.

4.2.4 Data augmentation

Data augmentation is a particularly popular method for expanding a dataset in the medical imaging field due to the lack of labelled photos and limited availability of expert knowledge. By definition, data augmentation is the process of artificially enriching the training set

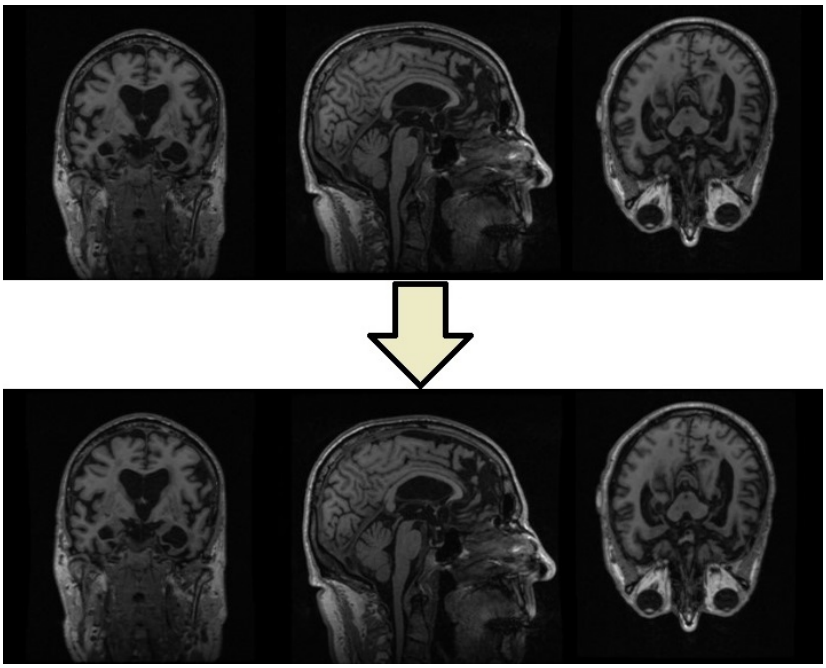

Figure 4.3. An MRI scan before and after preprocessing done by the ADNI group.

by creating altered copies of existing data without changing their semantic meaning. The variability introduced by augmented data tends to improve a model's generalization ability and reduce overfitting, the driving force behind the approach's popularity (Shorten and Khoshgoftaar, 2019).

In our work, an augmented copy of an image is created by applying augmentations based on probabilities. This process is repeated 10 times per scan in the original training set increasing its size from 528 scans per modality to 5808 or 11,616 scans in total. A list of the possible augmentations along with their corresponding probabilities can be found in *Table 4.1*.

Augmentation	**Probability**
Cropping	0.3
Scaling	0.2
Flipping	0.3
Blur	0.2
Elastic transform (Simard et al., 2003)	0.2
Brightness	0.2
Contrast	0.2
Random erasing	0.2
Pixel dropout	0.2

Table 4.1. The list of augmentations and their probabilities

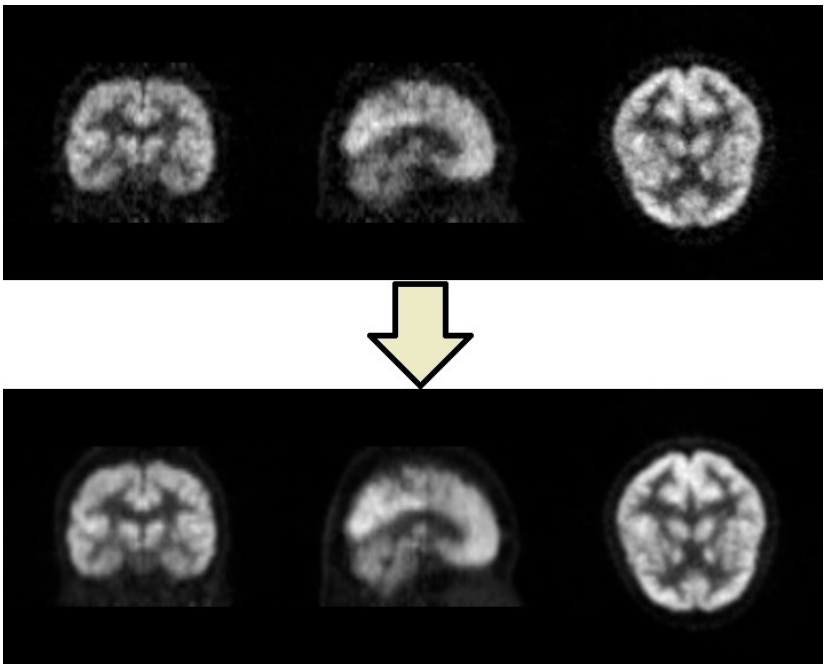

Figure 4.4. A PET scan before and after preprocessing done by the ADNI group.

4.2.5 Data Normalisation

Data normalization refers to the process of adjusting the data values to a standard scale or distribution. In the case of medical imaging, it reduces variability due to different scanners, settings, or acquisition protocols, ensuring more consistent and comparable data. It enhances image quality by reducing noise and artifacts, improves contrast for better visualization, and prepares the images for effective processing and analysis by algorithms that assume standardized input. Normalisation takes place after augmentation to ensure that the normalised features are have not been affected due to the applied augmentation operators.

Inspired by (Kociołek et al., 2020)'s study on the impact of image normalization on textures, we apply percentile normalization to our scans. In percentile normalization, upper and lower bounds, defining the range within which the data will be clipped, are specified. After clipping, the data is scaled between 0 and 1, ensuring that the entire dataset is uniformly distributed relative to the selected percentiles. The training and test sets are normalized separately to, once again, ensure that model evaluation remains as fair as possible.

4.2.6 Transposing the data and adding channel dimension

The models are build using the PyTorch machine learning library in Python which requires 3D CNN inputs to have shape $(N_{in}, C_{in}, D_{in}, H_{in}, W_{in})$ where N is the batch size, C is the number of channels, D is the image depth, H is the image height, and W is the image width.

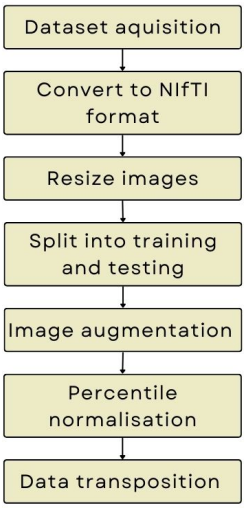

Figure 4.5. The preprocessing methodology implemented

NIfTI format, in which the scans are stored, has shape (X, Y, Z), where X, Y, and Z are the number of voxels in the three spatial dimensions. The axial plane is often preferred for such tasks due to its common usage and intuitive interpretation. Therefore, the data needs to be transposed to ensure that the axial view serves as the image depth. Adding the channel dimension is straightforward in this case as both types of scans used are greyscaled and require only 1 channel.

4.3 Model implementation

The two CNNs developed for this classification task can be seen in *Figure 4.7*. Both models share a common structure but differ in specific details such as the number of convolutional filters and features used at each stage.

The MRI Model passes the 3D input images through a series of 3D convolutional layers with increasing filters (16, 32, 64, 128), all using 3x3x3 kernels with padding of 1. Every two convolutional layers are followed by a 3D max-pooling layer with a kernel size of 2 that downsamples the feature maps. After eight convolutional layers, batch normalization with 128 features is applied, followed by a dropout layer to prevent overfitting and allow for the implementation MC Dropout. The 3D feature maps are flattened into a 1D vector and passed through three fully connected layers (512, 256, 3) to produce the final output.

The PET Model follows a similar architecture but uses more filters (32, 64, 128, 256) in each convolutional layer. Like the MRI Model, it uses 3D convolutions and max-pooling but applies batch normalization with 256 features to match the output of the final convolutional layer. The flattened feature vector is larger due to the higher number of filters, but dtill passes through the same sequence of fully connected layers (512, 256, 3) for classification.

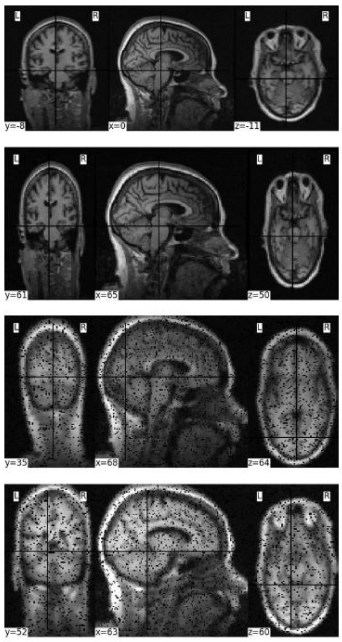

Figure 4.6. Sample MRI scan through the stages of preprocessing. The above figure illustrates the original scan being resized, the augmented copy being created and the resulting normalized augmented scan.

Overall, both models operate extremely similarly, differing only in the number of filters and batch normalization features. These architectures were entirely based on manual experiments so their performance should be not be considered final.

Although we were not able to make extensive use of optimization techniques, we did implement one to help us select the optimal dropout rate, learning rate, and weight decay. We employed another probabilistic method, Bayesian optimization (Snoek et al., 2012). Bayesian optimization is a global optimization technique typically used to optimize functions that are too expensive to evaluate. Since the objective function is unknown, it is treated as a random function with a prior distribution placed over it to capture beliefs about its behavior. As function evaluations are collected, a posterior distribution is formed over the objective function based on the updated prior. This posterior is subsequently used to construct an acquisition function, guiding where to query next. The most common methods for defining the prior and posterior distribution use Guassian processes. Our implementation uses the optimal hyperparameters obtained after 30 trials. *Table 4.2* contains a complete list of the final hyperparameter values used.

4.4 Measuring uncertainty

The choice between using logits (raw model outputs) and softmax probabilities for estimating uncertainty when using MC Dropout involves a trade-off, with each approach offering its

Hyperparameter	MRI model	PET model
Batch size	4	4
Learning rate	1.658670526543198e-05	1.025435302692212e-06
Weight decay	0.00892559113070601	0.006289134887218049
Num epochs	50 with early stopping	50 with early stopping
Activation function	RELU	RELU
Dropout rate	0.15452504447184434	0.3209841104276942
Optimizer	Adam	Adam
Classifier	Softmax	Softmax
Loss function	Categorical Cross-entropy	Categorical Cross-entropy

Table 4.2. The developed CNNs' hyperparameters

own advantages and disadvantages. Logits are more closely tied to the model's underlying parameters which makes them potentially more informative from a Bayesian perspective, as they provide a more direct and unbounded measure of uncertainty. The variance in logits can reflect model uncertainty in a more unprocessed way before being compressed into probabilities, which can be useful when analyzing how uncertain the model is about its internal decision-making process.

On the other hand, softmax probabilities transform logits into a normalized probability distribution (as per the literature survey), which aligns directly with the decision-making process of classification tasks. While this makes uncertainty more interpretable, the softmax function can often mask uncertainty, especially when probabilities are near 0 or 1 due to its squashing effect. Nevertheless, this method provides a clearer representation of the uncertainty that affects the final predictions, making it a practical choice for applications where model interpretability and calibrated confidence are key.

We decided to experiment with both logits and softmax probabilities, comparing their performance in quantifying uncertainty. To calculate the uncertainty in our models' predictions, we calculated the variance of the logits or softmax probabilities for each class across all forward passes. We then averaged these variances across the classes to produce a single uncertainty value for each prediction. While averaging the uncertainty across the classes is not a standard approach, we adopted this method to simplify model comparisons and provide a single value that reflects the overall reliability of a prediction.

4.5 Uncertainty-based model selection

A model's final prediction is obtained by averaging the logits across all forward passes for each class and selecting the one with the highest value (the softmax function is not needed since the one with the highest logit value will always have the highest softmax probability).

Having obtained the predictions and their associated uncertainties from each model, we use a variation of late fusion called uncertainty-based model selection (UBMS) to decide which modality to trust. As the name implies, the chosen prediction is always the one with the least uncertainty, under the assumption that it is more reliable.

4.6 Evaluation

The performance and robustness of the framework have been accurately assessed using multiple metrics. In the following equations, TP, FP, TN, and FN represent True Positive, False Positive, True Negative, and False Negative values, respectively.

- **Accuracy:** Accuracy measures how often the outcome is predicted correctly.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Recall:** Recall (or sensitivity) measures the number of correctly predicted positive instances from all actual positive instances.

$$recall = \frac{TP}{TP + FN}$$

- **Precision:** Precision measures the number of correctly predicted positive instances to all instances predicted as positive.

$$precision = \frac{TP}{TP + FP}$$

- **F1-score:** The F1 score is the harmonic mean of the precision and recall.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

- **Confusion matrix:** A table that compares the predicted labels to the true labels for each class.
- **Area under the ROC Curve (AUC):** The ROC (receiver operating characteristic) curve displays the performance of a classification model by plotting the TP and FP rates for each class at different classification thresholds. The metric AUC measures the two-dimensional area underneath this curve.

4.7 Software and hardware employed

All code was run on the Stanage High Performance Computing cluster owned by the University of Sheffield. Specifically, the experiments were conducted on a single node

equipped with an NVIDIA A100 GPU, featuring 80GB of GPU memory, alongside 120GB of CPU memory. The software stack included Python 3.9.7 and made use of libraries such as NumPy, Scikit-learn, Matplotlib, and Pandas. The implemented models were developed using the PyTorch machine learning library and Torchvision package.

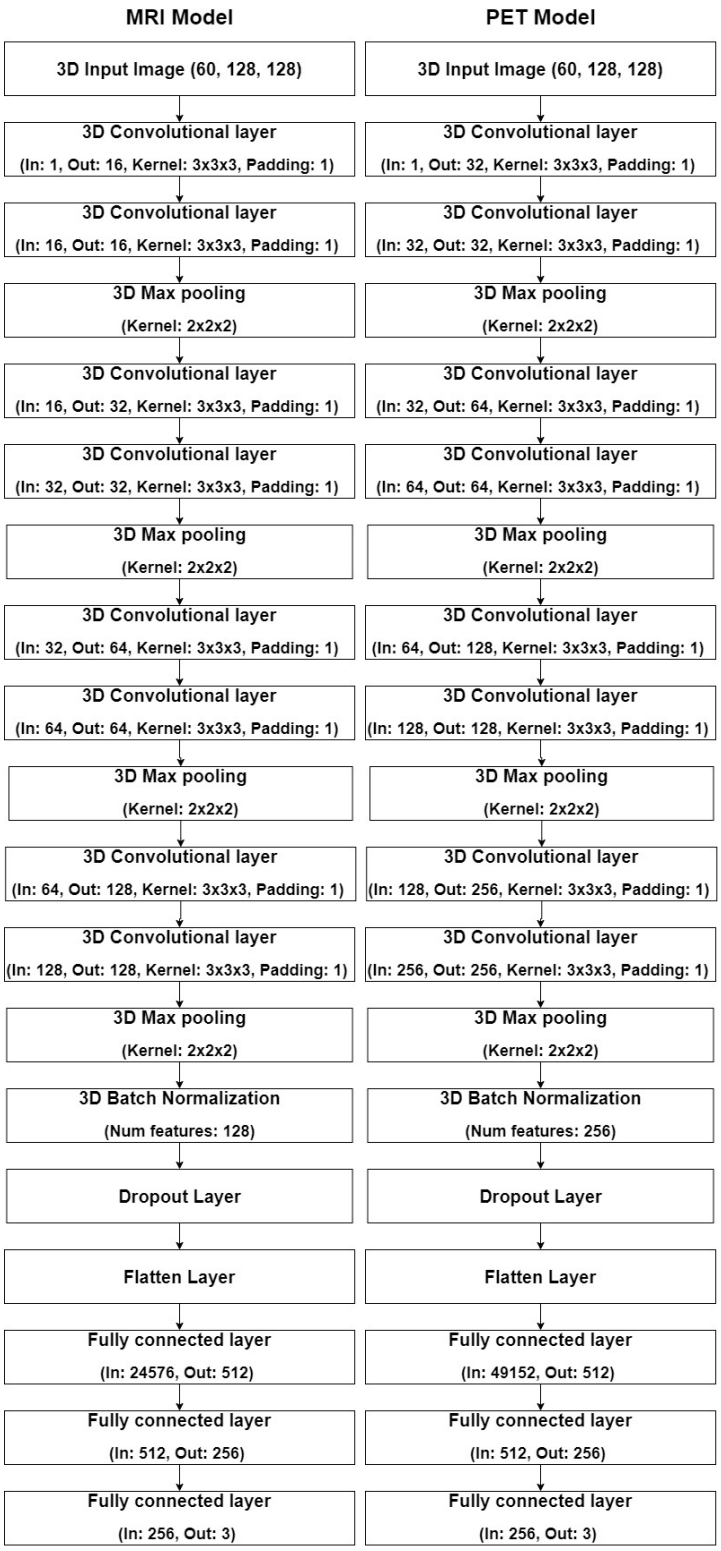

Figure 4.7. The architectures of the implemented models.

Chapter 5

Experimentation and Results

5.1 Individual model performance

The graphs in *Figure 5.1* portray the accuracy and loss metrics over training epochs for the individual modality models. In both cases, the training accuracy increases steadily before plateauing at around 40 epochs for MRI and 50 epochs for PET, achieving close to 90% accuracy. The validation accuracy on the other hand, levels off earlier, indicating possible overfitting. The loss graphs show a rapid decline in training loss in both models, while the validation loss decreases less sharply, stabilizing at higher loss values compared to the training loss, further suggesting some degree of overfitting, especially in the MRI model. Although time constrains prevented us from combating the overfitting issue, we attempted to mitigate it as much as possible by saving and loading the model state which results in the lowest validation loss. Additionally, as the validation loss does not start increasing again, the generalization of the model does not suffer any negative effects. These models perform adequately by themselves with the MRI and PET models achieving an accuracy score of 84.09% and 87.88% on the test set respectively.

5.2 Logits versus softmax probabilities for uncertainty calculations

The histograms in *Figure 5.2* depict the distribution of prediction uncertainties for the two models when using logits versus softmax probabilities at 1000 samples. Uncertainties derived from logits for both models show a broader spread, almost forming distributions that approximate a normal curve. For the MRI model, uncertainties range from about 0.10 to 0.80, peaking at around 0.45, while the PET model's uncertainties range from 0.10 to 0.70, peaking near 0.40. In contrast, the uncertainties derived from the softmax probabilities are tremendously skewed towards near-zero values, with most uncertainties clustering below 0.005 for both models. Using logits yields a wider range of uncertainty, indicating varied confidence levels in the models' predictions while softmax probabilities lead to much smaller

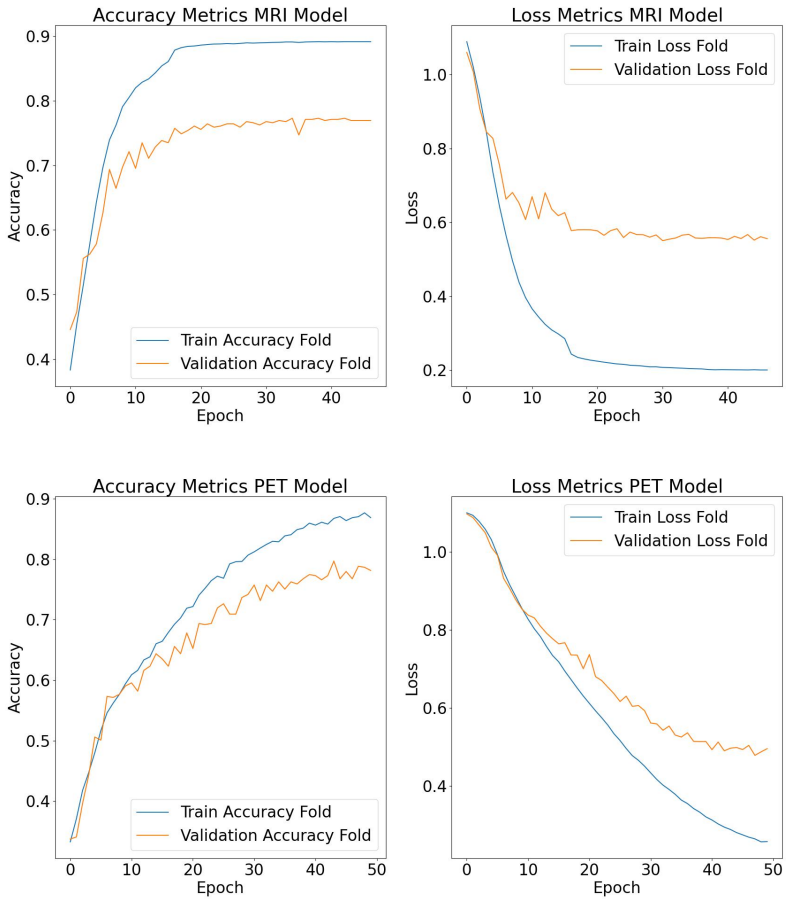

Figure 5.1. Training/Validation accuracy and loss for the MRI (top) and PET (bottom) models.

uncertainties, expressing an extremely high overall confidence in the models' predictions.

At first glance, the uncertainties generated using logits appear more trustworthy due to their distribution and values. The broader spread suggests that the models are capable of expressing a wider range of confidence levels in their predictions, which might indicate better calibration and reliability in the uncertainty estimates. The tight clustering which occurs when using softmax probabilities suggests overconfidence leading to less meaningful uncertainty measures, undermining trust in the predictions.

Our experiments with varying number of samples consistently show that fusing the modalities using the variance (or uncertainty) calculated on the logits performs significantly worse in practice when compared to the variance calculated on the softmax probabilities, contradicting the aforementioned observation. The metrics for the individual models along with the proposed framework's are shown in *Table 5.1*. When using only 1 sample per prediction, no uncertainty is calculated thus the fused model always chooses the prediction of the most accurate model. For the rest of the sample sizes, where uncertainty-based model selection

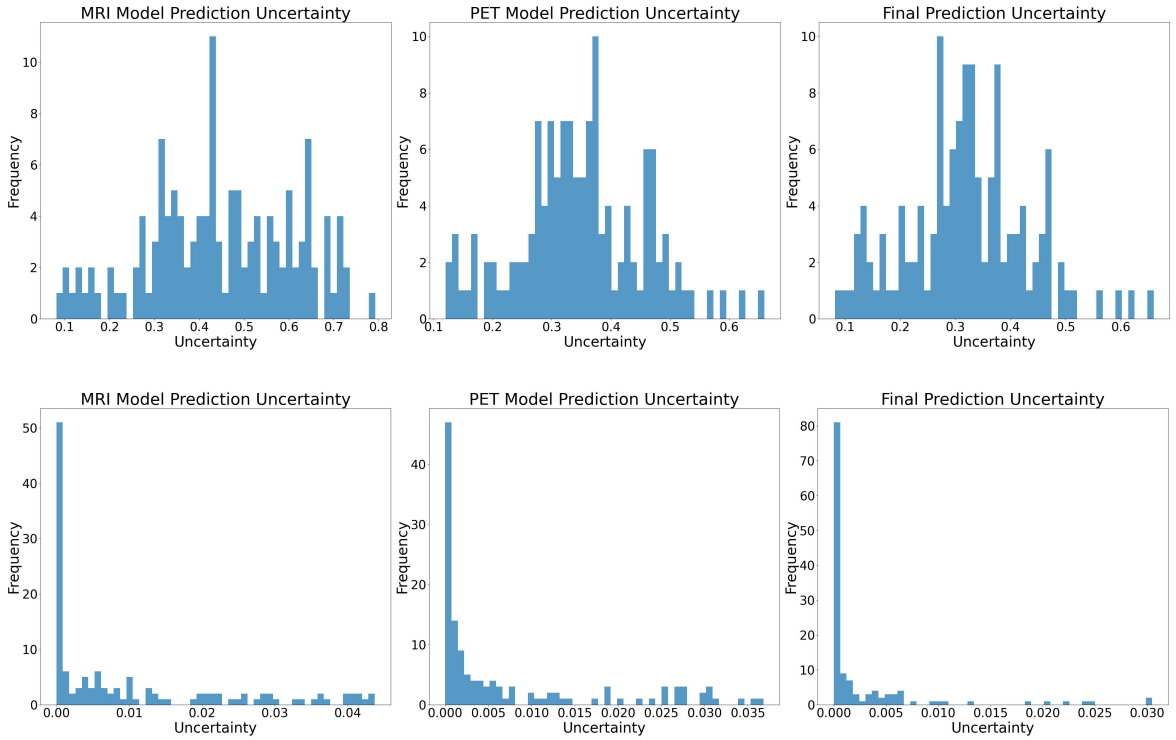

Figure 5.2. Distribution of uncertainties when using logits (top) and softmax probabilities (bottom).

is used (50, 500, 1000), the fused model consistently outperforms both individual models in all metrics. The accuracy is noticeably higher when using softmax probabilities but stable across all samples sizes. Although lower, the accuracy of the fused model when logits are used increases when going from 500 samples to 1000 which hints that the issue may lie in the inadequate number of samples used. The performance of the MRI and PET models fluctuates as the number of samples size increases, most likely attributed to the law of large numbers which states that as the number of samples increases, the average of the results converges to the true value.

The confusion matrices in *Figure 5.3* and ROC curves in *Figure 5.4* are those obtained from using 1000 samples, as they are regarded as the most accurate, again due to the law of large numbers. Regarding the confusion matrices, the individual models perform identically, as expected. Both fused models indicate strong, and almost identical, performance when classifying AD and MCI but the model using softmax probabilities is notably better at correctly identifying the CN class. The ROC curves indicate strong performance, with high AUC values for all classes. The individual models, once again, perform quite similarly with the MRI models achieving an AUC score in the 0.92-0.94 range for all three classes while the PET models perform marginally better with an AUC score of 0.96-0.99. The difference between the logits and softmax probabilities is minimal, though the softmax curves are slightly smoother and provide a more consistent performance across the classes.

Samples	Metric	Logits			Softmax Probabilities		
		MRI	PET	Fused	MRI	PET	Fused
1	Accuracy	0.8409	0.8788	0.8788	0.8409	0.8788	0.8788
	Precision	0.8521	0.8823	0.8823	0.8521	0.8823	0.8823
	Recall	0.8409	0.8788	0.8788	0.8409	0.8788	0.8788
	F1-score	0.8425	0.8793	0.8793	0.8425	0.8793	0.8793
50	Accuracy	0.8561	0.8788	0.8864	0.8485	0.8788	0.9318
	Precision	0.8647	0.8823	0.8903	0.8594	0.8823	0.9323
	Recall	0.8561	0.8788	0.8864	0.8485	0.8788	0.9318
	F1-score	0.8573	0.8793	0.8861	0.8497	0.8793	0.9320
500	Accuracy	0.8409	0.8788	0.8864	0.8333	0.8788	0.9318
	Precision	0.8521	0.8823	0.8914	0.8466	0.8823	0.9323
	Recall	0.8409	0.8788	0.8864	0.8333	0.8788	0.9318
	F1-score	0.8425	0.8793	0.8870	0.8348	0.8793	0.9320
1000	Accuracy	0.8333	0.8788	0.8939	0.8333	0.8788	0.9318
	Precision	0.8466	0.8823	0.9003	0.8466	0.8823	0.9323
	Recall	0.8333	0.8788	0.8939	0.8333	0.8788	0.9318
	F1-score	0.8348	0.8793	0.8939	0.8348	0.8793	0.9320

Table 5.1. Combined performance of implemented models at different number of samples for MC Dropout when working with logits and softmax probabilities

Figure 5.5 seems to explain how the gap in performance between the two approaches arises. Normally, we would expect the accuracy of a model to decrease as the uncertainty increases, similarly to what we observe when using softmax probabilities. When using logits however, the accuracy of both models paradoxically increases alongside the uncertainty, meaning that the models are becoming overly confident about their incorrect predictions. This indicated that the logits fail to provide a reliable measure of uncertainty, leading to poor calibration and misleading predictions.

Despite the aforementioned issues regarding the representation of uncertainty, the fused models display extremely promising performance results. *Table 5.2* compares the implemented framework against other known multimodal models, highlighting its potential if developed further.

5.3 Importance of adequate number of forward passes

The number of forward passes performed in MC Dropout directly impacts the performance of uncertainty estimation and prediction quality. More samples generally lead to better uncertainty estimates and improved accuracy by averaging out the noise from dropout, but this comes at the cost of increased computational overhead. Fewer samples reduce computational costs but result in noisier uncertainty estimates and less stable predictions. Typically, there is a diminishing return on performance as the number of samples increases,

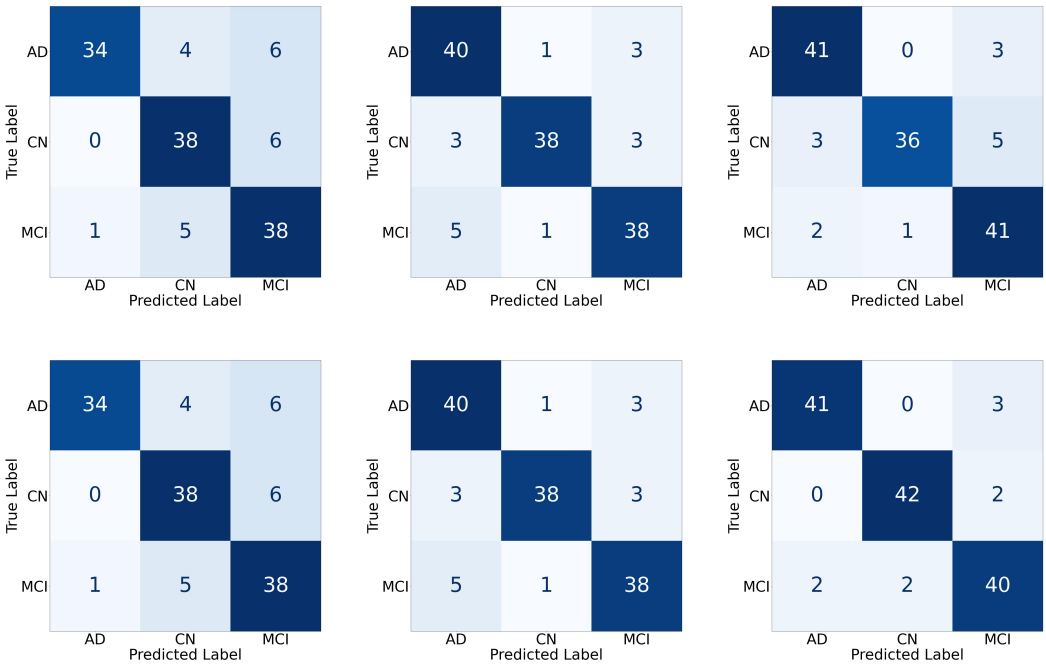

Figure 5.3. Confusion matrices for models at different samples sizes (logits at the top and softmax probabilities at the bottom).

so the optimal number balances accuracy and computational efficiency based on the specific application. *Figure 5.6* visualizes the relationship between the number of forward passes used for MC Dropout and the accuracy of the fused modalities.

The accuracy in both graphs fluctuates as the number of samples increases, showcasing the framework's sensitivity to sample size. In the top graph, where logits are used, the accuracy fluctuates between values close to 0.880 and 0.900. While we expected more samples to smooth out the noise introduced by dropout, the performance does not fully stabilize even with 2000 samples. This suggests that, despite the larger number of forward passes, uncertainty in the model itself continues to cause variability in predictions.

The bottom graph, which uses softmax probabilities, shows more noticeable but narrower fluctuations, ranging between 0.932 and 0.939. Although the oscillations are sharper in this case, the range of variability is significantly smaller compared to the top graph. This can be attributed to the softmax's squashing effect, which compresses the logits into a more compact range of values, reducing the overall variability in accuracy. Thus, while the softmax-based model also exhibits sensitivity to uncertainty, the fluctuations are more controlled.

Both graphs reinforce the importance of selecting an adequate number of forward passes in MC Dropout. More samples help average out the randomness from dropout, improving the stability of predictions. However, there's a balance to be struck between computational cost

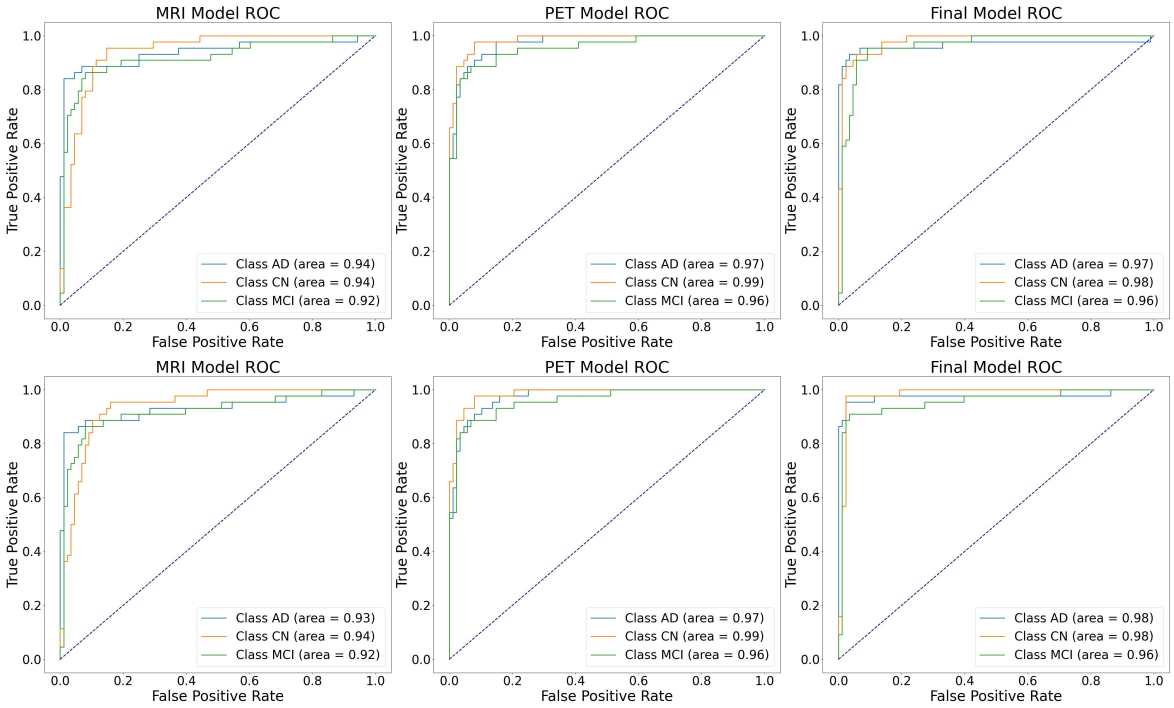

Figure 5.4. ROC curves and corresponding AUC scores for models at different samples sizes (logits at the top and softmax probabilities at the bottom).

and performance, as increasing the number of forward passes doesn't completely eliminate the model's inherent uncertainty.

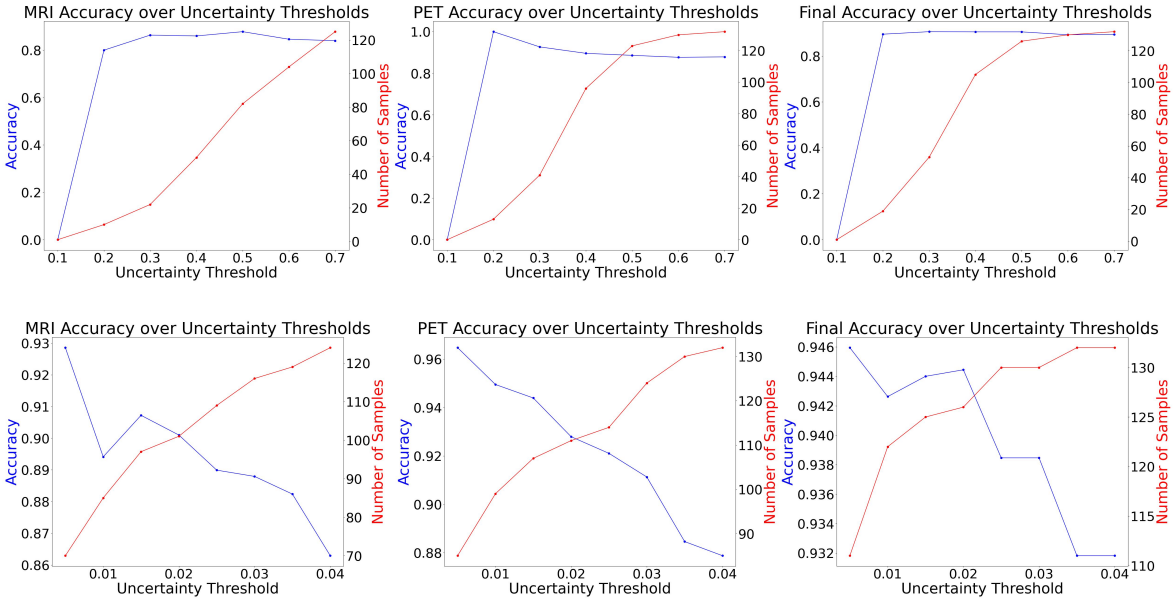

Figure 5.5. Accuracy of models at different uncertainty thresholds when using logits (top) and softmax probabilities (bottom).

Paper	Modality	Accuracy
(Venugopalan et al., 2021)	MRI, SNPs, EHR	88.00%
(Walhovd et al., 2010)	MRI, PET, CSF	88.80%
UBMS using logits	MRI, PET	**89.39%**
(Huang et al., 2019)	MRI, FDG-PET	90.10%
(Liu et al., 2014)	MRI, CSF	90.56%
(Westman et al., 2012)	MRI, CSF	91.80%
UBMS using softmax probabilities	MRI, PET	**93.18%**
(Zhang et al., 2011)	MRI, FDG-PET, CSF	93.20%

Table 5.2. Performance of multimodal models against implemented one

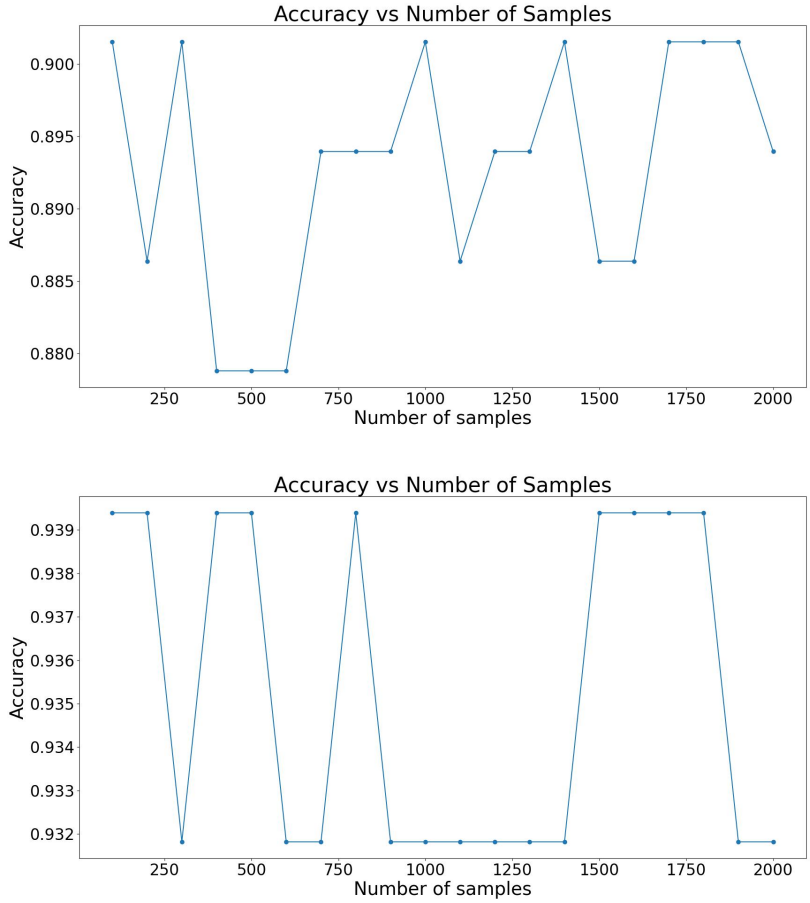

Figure 5.6. Accuracy of fused model at different sample sizes (logits at the top and softmax probabilities at the bottom)

Chapter 6

Conclusions and discussion

6.1 Overview

In this dissertation project, we have successfully proven that utilizing the predictions' uncertainty, obtained through Bayesian techniques, to decide between conflicting predictions from different biomarkers can be a simpler but still effective approach of modality fusion. By developing individual models trained to each modality and selecting the prediction with the least associated uncertainty, we were able to improve the decision-making process in scenarios where data from multiple sources is available. When prioritizing the most confident predictions, the risks associated with inaccurate or misleading outputs from any single modality are mitigated. Our methodology has demonstrated the potential of UQ to refine model performance, offering a more robust framework for multimodal integration in complex tasks. Although using the quantified uncertainty is an effective approach for accurate fusion of the modalities, it is still not as informative to humans as we had hoped and cannot be used reliably in a clinical setting as of yet.

6.2 Limitations

Despite the promising results, the project faced several limitations primarily due to time and computational constraints. The scope of our implementation was narrower than what we had hoped, limiting the exploration of more advanced or alternative techniques. For instance, the dataset and modality diversity used in our experiments were constrained, potentially affecting the outcomes of the project. Additionally, while the project focused on Monte Carlo Dropout, we were hoping to explore other available UQ techniques, which might have offered different insights or allowed for performance comparisons. The computational intensity or time required to implement or optimize some methods was beyond our available resources, further constraining our analysis. These limitations suggest that while our findings are valuable, further research should be conducted to produce definite results.

6.3 Future Work

Several avenues for future work are evident from the limitations and findings of this study. Firstly, addressing the issue of overfitting would undoubtedly be beneficial. Further hyperparameter tuning should also be conducted to optimize the performance of the individual models, perhaps continuing with Bayesian optimization or switching to other automated techniques such as grid search.

Furthermore, there is significant potential in experimenting with different UQ techniques, such as Bayesian neural networks, MCMC, or ensemble methods, to assess their impact on model performance and the quality of uncertainty estimates. This could also involve comparative studies to identify which UQ techniques are best suited for specific types of multimodal data or tasks. As previously discussed, the uncertainty values produced by our models are not informative enough to be used in clinical settings despite aiding performance. Another promising direction could be the integration of UQ methods into the training process itself, enabling models to learn more robust representations that inherently account for uncertainty. Additionally, modelling and making use of aleatoric uncertainty would also be an interesting addition.

Finally, future work could also explore the scalability of our approach, particularly in the context of real-time systems where computational efficiency is as critical as accuracy. This could involve optimizing the model architecture or leveraging parallel processing and hardware acceleration to handle larger and more complex datasets efficiently.

Bibliography

Aaraji, Z. S. and Abbas, H. H. (2022). Automatic classification of alzheimer's disease using brain mri data and deep convolutional neural networks.

Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., and Aßenmacher, M. (2023). Multimodal deep learning.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and Computing*, 18(4):343–373.

Apostolova, L. G., Hwang, K. S., Andrawis, J. P., Green, A. E., Babakchanian, S., Morra, J. H., Cummings, J. L., Toga, A. W., Trojanowski, J. Q., Shaw, L. M., Jack, C. R., Petersen, R. C., Aisen, P. S., Jagust, W. J., Koeppe, R. A., Mathis, C. A., Weiner, M. W., and Thompson, P. M. (2010). 3d pib and csf biomarker associations with hippocampal atrophy in adni subjects. *Neurobiology of Aging*, 31(8):1284–1303. Alzheimer's Disease Neuroimaging Initiative (ADNI) Studies.

Atrey, P. K., Hossain, M. A., Saddik, A. E., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379.

Bae, J. B., Lee, S., Jung, W., Park, S., Kim, W., Oh, H., Han, J. W., Kim, G. E., Kim, J. S., Kim, J. H., and Kim, K. W. (2020). Identification of alzheimer's disease using a convolutional neural network model based on t1-weighted magnetic resonance imaging. *Scientific Reports*, 10(1):22252.

Baltrusaitis, T., Ahuja, C., and Morency, L. (2017). Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Barnum, G., Talukder, S., and Yue, Y. (2020). On the benefits of early fusion in multimodal representation learning.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks.

Brigato, L. and Iocchi, L. (2021). On the effectiveness of neural ensembles for image classification with small datasets.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 3(3):186–191.

Brooks, S. (1998). Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):69–100.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., and Colliot, O. (2011). Automatic classification of patients with alzheimer's disease from structural mri: A comparison of ten methods using the adni database. *NeuroImage*, 56(2):766–781. Multivariate Decoding and Brain Reading.

de Leon, M. J., Mosconi, L., Li, J., De Santi, S., Yao, Y., Tsui, W. H., Pirraglia, E., Rich, K., Javier, E., Brys, M., Glodzik, L., Switalski, R., Saint Louis, L. A., and Pratico, D. (2007). Longitudinal csf isoprostane and mri atrophy in the progression to ad. *Journal of Neurology*, 254(12):1666–1675.

D'mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3).

Ebrahimighahnavieh, M. A., Luo, S., and Chiong, R. (2020). Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 187:105242.

Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Holland, D., Brewer, J. B., Dale, A. M., Initiative, A. D. N., et al. (2010). Csf biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and alzheimer's disease. *Journal of Neuroscience*, 30(6):2088–2101.

Foster, N. L., Heidebrink, J. L., Clark, C. M., Jagust, W. J., Arnold, S. E., Barbas, N. R., DeCarli, C. S., Turner, R. S., Koeppe, R. A., Higdon, R., and Minoshima, S. (2007). FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. *Brain*, 130(10):2616–2635.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Getoor, L. and Taskar, B. (2007). An introduction to conditional random fields for relational learning.

Gunawardena, K. A. N. N. P., Rajapakse, R. N., and Kodikara, N. D. (2017). Applying convolutional neural networks for pre-detection of alzheimer's disease from structural mri data. In *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 1–7.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2019). Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. *Frontiers in Neuroscience*, 13:509.

Hüllermeier, E. and Waegeman, W. (2019). Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *CoRR*, abs/1910.09457.

Kociołek, M., Strzelecki, M., and Obuchowicz, R. (2020). Does image normalization and intensity resolution impact texture classification? *Computerized Medical Imaging and Graphics*, 81:101716.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles.

Laplace, P.-S. (1774). *Mémoires de Mathématique et de Physique, Tome Sixième*. Académie Royale des Sciences, Paris.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 2.

LeNail, A. (2019). Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33):747.

Liu, F., Zhou, L., Shen, C., and Yin, J. (2014). Multiple kernel learning in the primal for multimodal alzheimer's disease classification. *IEEE Journal of Biomedical and Health Informatics*, 18(3):984–990.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

McClure, P. and Kriegeskorte, N. (2017). Representing inferential uncertainty in deep neural networks through sampling.

Mullachery, V., Khera, A., and Husain, A. (2018). Bayesian neural networks.

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A., and Lima Netto, S. (2021). *Variational Autoencoder*, pages 111–149. Springer International Publishing, Cham.

Puente-Castro, A., Fernandez-Blanco, E., Pazos, A., and Munteanu, C. R. (2020). Automatic assessment of alzheimer's disease diagnosis based on deep learning techniques. *Computers in Biology and Medicine*, 120:103764.

Qiu, S., Miller, M. I., Joshi, P. S., Lee, J. C., Xue, C., Ni, Y., Wang, Y., Anda-Duran, I. D., Hwang, P. H., Cramer, J. A., Dwyer, B. C., Hao, H., Kaku, M. C., Kedar, S., Lee, P. H., Mian, A. Z., Murman, D. L., O'Shea, S., Paul, A. B., Saint-Hilaire, M.-H., Sartor, E. A., Saxena, A. R., Shih, L. C., Small, J. E., Smith, M. J., Swaminathan, A., Takahashi, C. E., Taraschenko, O., You, H., Yuan, J., Zhou, Y., Zhu, S., Alosco, M. L., Mez, J., Stein, T. D., Poston, K. L., Au, R., and Kolachalama, V. B. (2022). Multimodal deep learning for alzheimer's disease dementia assessment. *Nature Communications*, 13(1):3404.

Robert, C. and Casella, G. (2000). Monte carlo statistical method. *Technometrics*, 42.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Schaffert, J., LoBue, C., Hynan, L. S., Hart, J., Rossetti, H., Carlew, A. R., Lacritz, L., White, C. L., and Cullum, C. M. (2022). Predictors of life expectancy in autopsy-confirmed alzheimer's disease. *Journal of Alzheimer's Disease: JAD*, 86(1):271–281.

Settles, B. (2011). From theories to queries: Active learning in practice. In Guyon, I., Cawley, G., Dror, G., Lemaire, V., and Statnikov, A., editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy. PMLR.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Simard, P., Steinkraus, D., and Platt, J. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963.

Singh, P., Singh, N., Singh, K. K., and Singh, A. (2021). Chapter 5 - diagnosing of disease using machine learning. In Singh, K. K., Elhoseny, M., Singh, A., and Elngar, A. A., editors, *Machine Learning and the Internet of Medical Things in Healthcare*, pages 89–111. Academic Press.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2).

Todd, S., Barr, S., Roberts, M., and Passmore, A. P. (2013). Survival in dementia and predictors of mortality: a review. *International Journal of Geriatric Psychiatry*, 28(11):1109–1124.

Venugopalan, J., Tong, L., Hassanzadeh, H. R., and Wang, M. D. (2021). Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports*, 11(1):3254.

Walhovd, K., Fjell, A., Brewer, J., McEvoy, L., Fennema-Notestine, C., Hagler, D., Jennings, R., Karow, D., and Dale, A. (2010). Combining mr imaging, positron-emission tomography, and csf biomarkers in the diagnosis and prognosis of alzheimer disease. *American Journal of Neuroradiology*, 31(2):347–354.

Westman, E., Muehlboeck, J.-S., and Simmons, A. (2012). Combining mri and csf measures for classification of alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage*, 62(1):229–238.

Wilson, A. and Izmailov, P. (2019). Deep ensembles as approximate bayesian inference. *Deep Ensembles as Approximate Bayesian Inference*.

World Health Organization (2023). Dementia. Accessed: 2024-06-08.

Yang, S. and Fevens, T. (2021). Uncertainty quantification and estimation in medical image classification. In Farkaš, I., Masulli, P., Otte, S., and Wermter, S., editors, *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 671–683, Cham. Springer International Publishing.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3):856–867.

Zhang, S.-F., Zhai, J.-H., Xie, B.-J., Zhan, Y., and Wang, X. (2019). Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6.