Dissertation: A multitask multimodal ML algorithm for PET feature forecasting and AD classification

Edward Martin: C1629237

5th November 2021

Course: MSc Advanced Computer Science, School of Computer Science and Informatics, Cardiff University
Supervisor: Dr. Xianfang Sun

Contents

1 Abstract **4**

2 Introduction **5**

3 Alzheimer's Disease (AD) **5**
3.1 Amyloid Precursor-Protein (APP) and the βAmyloid (βA) 5
3.2 AD bio-markers 6
3.2.1 Cognitive testing 6
3.2.2 Magnetic Resonance Imaging (MRI) 7
3.2.3 Positron Emission Tomography (PET) 7
3.2.4 cerebral-spinal fluid (CSF) 8
3.2.5 Background information (BGI) 9

4 Artificial Intelligence (AI) & Machine Learning (ML) **9**
4.1 Deep Learning and Neural Networks 9
4.1.1 Convolutional Neural Network (CNN)s 10
4.1.2 Recurrent Neural Network (RNN)s 11

5 The model-scape **11**
5.1 Axis 1: Number of tasks 11
5.2 Axis 2: Number of data modalities 12

6 Literature Review **13**
6.1 Neuroimaging input methods 13
6.2 Problem definitions 14
6.2.1 The data problem 14
6.2.2 Evolution of disease classification methods 14
6.2.3 Multi-task models 16
6.3 Evolution of neuroimaging input methods and technologies 17
6.3.1 Summary of current trends and the state-of-the-art 17
6.4 General observations 17
6.5 Individual study reviews 18
6.5.1 Study 1: Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks - Basaia S. et al, 2019 18
6.5.2 Study 2: Predicting Alzheimer's disease progression using multi-modal deep learning approach - Lee G., et al., 2019 18
6.5.3 Study 3: Anomaly Analysis of Alzheimer's Disease in PET Images Using an Unsupervised Adversarial Deep Learning Model - (Baydargil H. et al, 2021) 19
6.6 Where to begin 20
6.7 Trust in ML 20

7 Proposed model **21**
7.1 The base 21
7.1.1 Advantages 22
7.1.2 Limitations 23

7.1.3 Other . . . 23
7.2 My model . . . 23
7.2.1 Justification . . . 23
7.2.2 Design overview: . . . 24
7.2.3 Experimentation . . . 25

8 Conclusions **27**
8.1 Successes . . . 27
8.2 Limitations . . . 28
8.3 Future work . . . 28
8.4 Final thoughts . . . 28

9 Bibliography **28**

1 Abstract

Alzheimer's Disease (AD) affects millions worldwide to an obscene cost both financially and to individuals. AD is currently incurable and irreversible, as a result early diagnosis plays a key role in allowing for preventative care to either delay the onset of or prevent fully onset AD. Machine Learning (ML) technologies are being increasing used in research relating to computer-aided diagnosis (CAD) and disease progression forecasting, with the aim of eventually creating tools to aid clinicians in saving lives and tackling AD. The goal is this research is to specifically investigate the role that neuroimaging data, and models which use it, plays in this area. The aim is to perform a thorough review of the relevant literature, to critically analyse it and to draw conclusions which will point our own research efforts in the correct direction. The end goal is to develop our own ML system utilising neuroimaging data in some form, and to identify future areas for researchers to explore.

2 Introduction

The aim of this research is to investigate the role of neuroimaging data in ML systems for the diagnosis, detection or progression forecasting of AD. AD is a disease which affected an estimated 46.8 million people globally in 2015, at an estimated cost of $818 billion (The Good Care Group, 2015). As large data repositories become available an increasing amount of research into the application of statistical AI and ML becomes more feasible, and the overall systems developed continue to improve in performance and utility. Whilst the overall goal of this research area is to progress towards systems which are capable of diagnosing patients and forecasting their potential disease progression, this research aims to investigate and understand at a holistic level what paths may be most fruitful to explore and apply cutting-edge technologies to them. The goal here is not to develop a system with unrivaled performance, but to identify regions of potential performance and utility gains and experiment with them.

3 AD

To understand the tools and techniques for the analysis of AD using AI and ML, it is first necessary to understand the biology behind the disease itself and the bio-markers used by clinicians for it's diagnosis. But firstly, a critical point to remember about AD is that the disease is as of yet, irreversible and incurable. That being said, moderately effective treatments exist for early stage AD which will delay and sometimes prevent the full onset of the disease. This treatment is only effective if caught in the Mild Cognitive Impairment (MCI) state. Furthermore, not all patients experiencing a MCI will progress to having AD, indeed this may mean they are resistant to the disease or that the MCI is being caused by a different ailment.

Given this we can conclude that the classification of AD in patients may be useful and provide key information relating to the disease, however recognising the prodromal stages of the disease (MCIs which progresses to AD) may prove more fruitful. This however has been shown to be a more difficult task, and will be explored later in section 6.2.

3.1 APP and the βA

The APP sits within the neural cell wall and has two tail ends, one of which is outside the neuron, and one which is inside. The sucratase family of enzymes is responsible for the cutting of the APP at different points to create soluble proteins which can be broken down and replaced to ensure correct neuronal function. α-sucratase and γ-sucratase 'snip' the APP at different points to create the αAmyloid (αA) protein and γAmyloid (γA) protein. Both of these are soluble and their production is considered normal, sustainable functions.

β-sucratase however 'snips' the APP at a point which creates the βA protein, a non-soluble compound which can be described as 'chemically sticky'. The βA will flout in the synapses between neurons and, over time, accumulate to form 'plaques'. The effect of these plaques is that they begin to disrupt the normal communication of neural signals between effected synapses and degrade neuronal function. Furthermore, the βA plaques cause hyper-polarisation of specific molecules in the cito-skeletal structure of the neuron, causing it to collapse over time. The collapse prevents nutrients flowing throughout the neuron and will essentially starve it, causing the neuron to die. As the disease progresses, more and more neurons die and the brain becomes permanently damaged. Eventually, vital neural

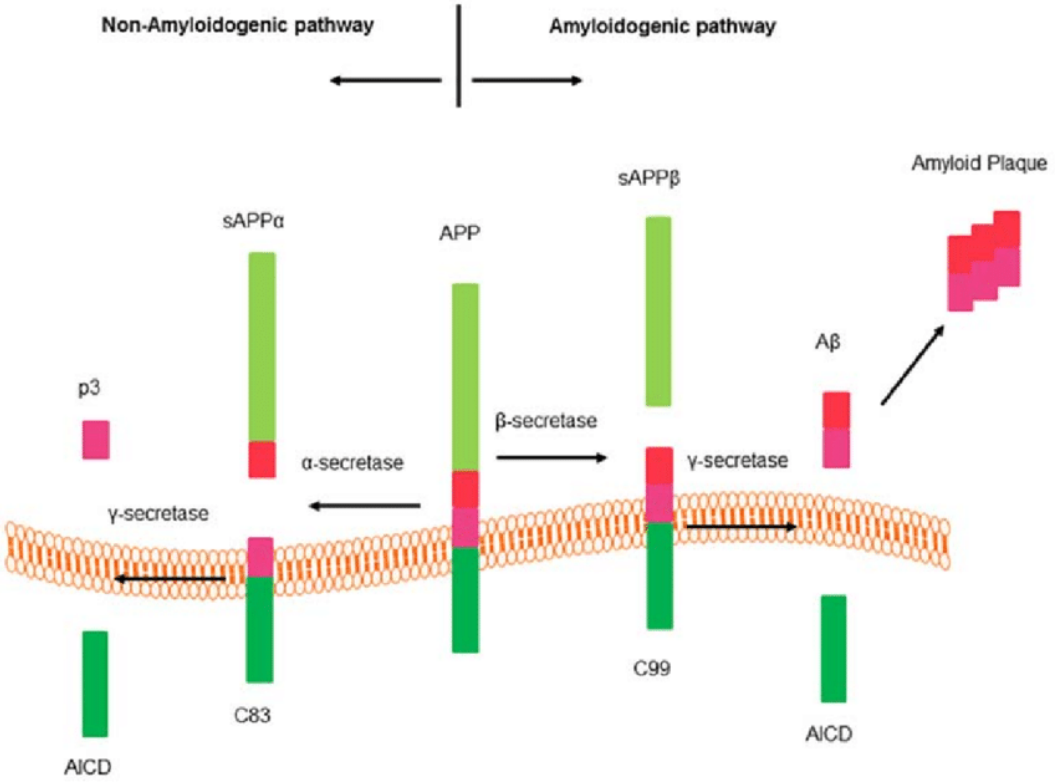

Figure 1: APP cleavage illustration from (ref image)

pathways relating to key survival mechanisms within the brain will fail also, causing the patients death.

Analysis of AD patients brains through neural imaging has revealed that the hippocampus is the first affected brain region in AD patients, and that the disease is centrifugal, meaning it begins in the centre (hippocampus) and spreads outwards.

3.2 AD bio-markers

Given this prior knowledge we can begin to understand the bio-markers used by physicians for AD diagnosis and treatment. We can classify a bio-markers effectiveness along three axes, information resolution, cost of obtaining the bio-marker, time taken to collect it, and the stage of AD at which the bio-marker is useful.

3.2.1 Cognitive testing

The primary complaint of early-stage AD patients is memory impairments. Specifically, difficulty in forming new memories rather than forgetting long-term memories such as your spouses name, a common miss-conception. This squares up perfectly to our knowledge that the hippocampus is affected first in AD patients as the hippocampus plays a key role in the

formation of new memories. As such, a number of cognitive tests have been developed, aimed at quantifying a patients mental state and cognitive abilities. Through observations over time a clinician can utilise these tests to more effectively quantify the patients symptoms and therefore aid in diagnosis.

Cognitive testing is the most inexpensive and easiest of the bio-markers to obtain and is used in virtually all AD diagnoses. However it should be noted that these tests aim at quantifying AD symptoms and as such are susceptible to interference and present relatively low resolution information. Furthermore, each patients baseline 'healthy' cognitive state is unique and therefore a set of cognitive tests at a given time does not necessarily give clear a indication of AD presence. This is overcome through continual testing over a period to time to measure cognitive deterioration, or lack of it. As previously stated, no known cure or effective treatment exists for late stage AD and the only effective treatments must be undertaken as early as possible to prevent further deterioration, or to at least slow its onset. A bio-markers usefulness is therefore in part determined by the amount of time required to obtain an accurate diagnosis, and cognitive testing is seriously lacking in this respect. A distinct advantage however is that as cognitive impairment usually precedes neural degeneration and brain damage, cognitive scores can be hypothetically useful for early stage AD diagnosis.

3.2.2 MRI

MRI is a neural-imaging technique used to obtain high-resolution images of a patients neural structure. It is currently the best widely available and used technology for structural brain imaging. As AD progresses the neural damage can be clearly seen in the structural brain changes. MRI can therefore provide much more direct and high-resolution information relating to the progression of AD. Generally, MRI is used to confirm a clinician's suspicions of a positive AD diagnosis, or simply to gain extra clarity if cognitive testing is inconclusive.

Obtaining high-resolution images of the hippocampus for example can provide key information, however the issue remains that neural damage comes after impairment, and therefore MRI is hypothetically less able to distinguish between healthy and AD patients in the earlier stages of the disease. Through various statistical methods different critical brain regions and functional areas have been identified for increased accuracy in AD diagnosis. Also, the cost of obtaining MRI information is very high and waits can be extremely long.

Figure 2: Example MRI scan from *brainfacts.org*

Whilst MRI is currently the best in-vivo solution for structural neural imaging, it is not able to display the minuscule βA plaques or neuro-fibrilary tangles responsible for the diseases symptoms. If neural imaging to this scale was possible it would be fair to assume that extremely high amounts of information could be extracted from it, but as this is not the case this bio-markers utility is still hypothetically capped.

3.2.3 PET

PET is another type of neural imaging technology, however unlike MRI which focuses on brain structure, it measures proxies for brain activity. Specific, tailor made compounds which are injected into the patients blood stream and these compounds interact with different neural receptors and upon doing so cause an annihilation reaction, producing two

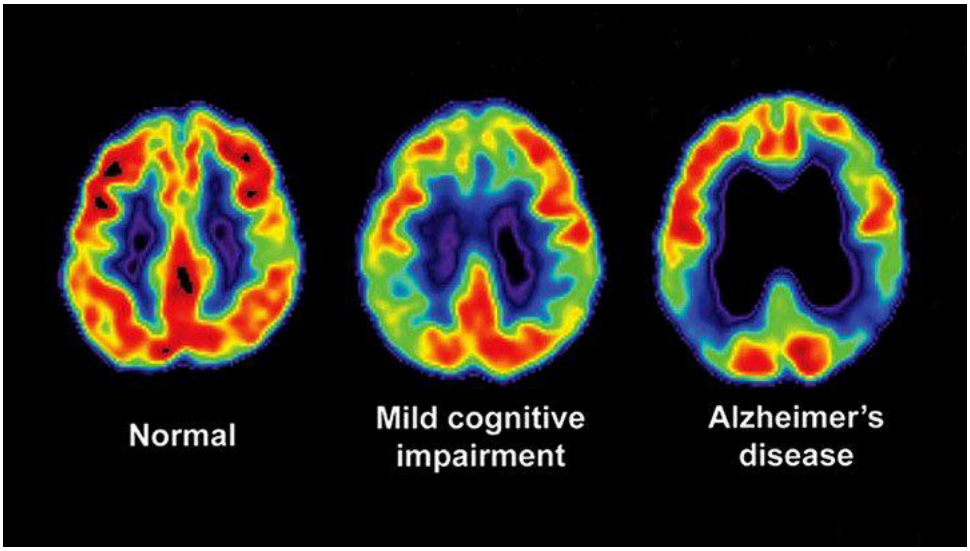

Figure 3: Example PET scan from *www.everydayhealth.com*

positrons which are be measured and used to see which regions of the brain are utilising the injected compound. Each compound has a specific use-case, for example fluorodeoxyglucose (FDG) is used to measure glucose utilisation, a proxy for brain activity. Given that AD it's initial stages impacts synaptic transmission prior to neural collapse and damage, FDG-PET is therefore potentially extremely useful in detecting early AD. Other compounds have also been used to measure βA accumulation. For this reason, PET is considered a useful bio-marker in AD diagnosis and treatment.

PETs glaring weakness however is the method by which the proxies are measured. Known as 'coincidence detection', it's extremely vulnerable to positron scattering and general interference. The final images created by PET are therefore much more noisy and lower-resolution than those of MRI for example. Furthermore, PET is even more expensive than MRI, presenting similar yet more severe financial challenges.

3.2.4 CSF

Another known biomarker is the genetic profile of a patient, extracted from cerebral-spinal fluid, often simply referred to as CSF. It has been shown that specific genetic profiles can make individuals more susceptible to developing AD later in life. The most well known example of this the gene which is responsible for production of the beta-sucratase enzyme. It is for this reason that many people with downs-syndrome (who naturally possess an extra copy of the chromosome containing gene Beta-Secretase 1 (BACE1)) will develop AD by age 40, a truly tragic fact.

Given that an individuals genetic profile remains constant throughout ones life, CSF is a time-independent biomarker for AD. It is far from the most costly biomarker to procure also, however it offers the relatively little information pertaining to AD symptomology for a given patient, however it does provide useful background information on a subjects likelihood of developing the disease.

3.2.5 BGI

What should be considered background information is up for debate. For our purposes we will define it as any known biomarkers which contain supplementary (complementary) rather than primary information about a patient. In other words, biomarkers which do not cross an arbitrary minimum threshold for latent information gain, or do not change over time. For example, a patient experiencing a MCI at age 20 is high unlikely to be experiencing prodromal AD, but at age 80 the inverse is true. In this case, age provides context to the symptoms experienced, but alone is virtually useless. An astute reader may concur that CSF should also be in the BGI category, but CSF is known to be a strong biomarker and thus is treated separately from demographic information which is mostly considered to be BGI.

4 AI & ML

To understand this research a baseline knowledge of the underlying technologies is essential. AI systems are computational systems which appear to be able to make 'intelligent' decisions. The simplest example of this are expert-systems, in which a series of inputs is used to traverse a deterministic tree-like structure where what branch of a tree is chosen at a given node is analogous to a decision. As such the system appears to be reasoning intelligently about the problem, however in reality they offer little utility outside of automating very simple tasks.

Of interest to us is a subset of AI known as ML. ML refers to a set of statistical methods used to produce complex, high-dimensional models which can be used to extract latent information about a given input and subsequently utilise this information for a specific task. The key distinction here is that the 'decision making' parameters are learned through statistical methods rather than being predetermined by the programmer. Removing this burden from programmers shoulders makes it easier to leverage extremely high dimensional data to make decisions.

4.1 Deep Learning and Neural Networks

The perceptron (F. Rosenblatt, 1958) is an algorithm designed to imitate the functionality of neural synapse. Here, a number of inputs (feature vector) are given to a single node which contains a set parameters (weights) which are applied to the inputs. This weighted input is passed through an activation function and a single output is given. Combined with learning algorithms such as back-propagation, a 'loss' can be calculated, that being the difference between the true value and the algorithms output, and the internal parameters (weights) can be adjusted to 'teach' the perceptron the underlying inputs data distribution and how it relates to it's required output. As such, given enough training data the perceptron can effectively perform simple tasks.

When multiple perceptrons are organised into layers as can be seen in figure 4.1 where all inputs are connected to every node (perceptron) in the following layer, when combined with non-linear activation functions at each node a system can be created which harnesses the 'learning' capabilities of the perceptron but expands it's representational abilities to highly complex non-linear data. These networks are referred to as Neural Network (NN)s and are heavily inspired by the structure of the human brain.

Deep Learning (DL) is a further subset of ML systems which utilise 'deep' architectures. Here, 'depth' refers to a sequence of ML techniques feeding into one another which are

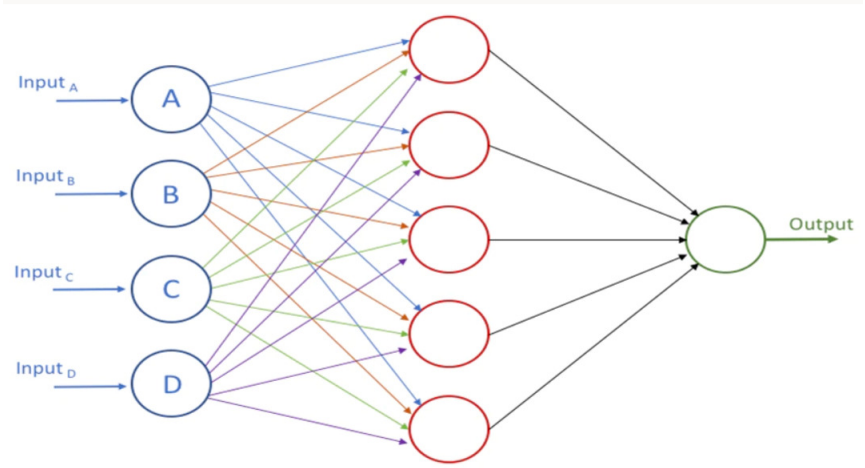

Figure 4: Example NN illustration from *developer.oracle.com*

usually some form of neural network. These NNs form the basis of most modern ML systems. Their exists a diverse and complex set of NN architectures and ML techniques which can be applied to a variety of problems. Before reviewing the relevant literature we shall explore common NN techniques in their most basic forms, and discuss how these fit into functioning models.

4.1.1 CNNs

Computer Vision has historically been an extremely difficult task. Teaching computer to see, or more specifically to take image data and extract useful information from it is no trivial task. Indeed, there is nearly an infinite combination of possible interpretations of any given input. Furthermore, high-resolution image processing by classic NN architectures can lead to millions of learnable parameters for a single node in a network, making the task virtually impossible to solve without infinite training data and computational resources. This is because classic fully-connected NN architectures will learn parameters for every pixel of every colour channel in an image.

One solution to this problem is the CNN, a ML technique used to consider points in the input data in context and produce abstracted representations of them. CNNs use three core constraints to achieve this: local receptive fields, spatial subsampling and shared weights. Here, the " receptive field is a defined segmented area that is occupied by the content of input data that a neuron within a convolutional layer is exposed to during the process of convolution" (*Towards Data Science*, 2020). As each filter acts as a feature detector across all local receptive fields they are capable of detecting features present in any part of the input, and this drastically decreases the number of parameters required by the network. Through a sequence of convolutional layers the network may then extract successively more complex features before in which features position relative to one another is considered rather than their absolute position. Between layers sub-sampling or dimensionality reduction is common, in which the feature maps are reduced in size to simplify the network and reduce computational costs. The combination of these factors allow CNNs to provide some degree

of invariance to translation, rotation, scale, and deformation. As a result, CNNs have overseen a massive leap forwards in computer vision capabilities. CNNs provide a promissing candidate for useful ML techniques to be used our our neuroimaging data.

4.1.2 RNNs

Classic feed-forward NNs have significant difficulty learning from data in which latent information is hidden within a sequence. Each piece of information is considered at a singular position, without knowledge of what proceeds or comes before it. For example, in sentiment analysis in Natural Language Processing (NLP) a words meaning is heavily dependant on the context of the sentence. 'That movie was great', and 'great, that movie is over' both use the positive adjective 'great' yet the sentiment is completely different.

RNNs are a class of NN architectures which aim remedy this lack of contextual knowledge by implementing 'memory' into the system through recurrent connections. As such, each computation performed at a recurrent layer takes into account the result from the previous piece of data, providing a kind of short term memory.This allows the RNN to extract latent information from the sequence of data itself rather than simply it's component parts. To extend the RNNs capabilities different variations have been created to address it's shortcomings. The Long Short-Term Memory (LSTM) network employs purpose built, learnable memory units to allow for longer range contexts to be exploited. The Bidirectional Long Short-Term Memory (Bi-LSTM) network stacks forward and backward pass LSTM layers to allow for bi-directional analysis of the input data, further increasing it's learning capacity.

5 The model-scape

Here I wish to establish a conceptual framework from which we can gain a better holistic view of the literature we want to review. Whilst this framework was developed over the course of the literature review it bares discussing now such that we can better critically analyse the research and plot our own course.

Firstly, we will broadly split the approaches towards applying ML to AD along two axis, the number of tasks it aims to complete, and the number of information modalities it utilises to this end. The figure below is what I will refer to as the model-scape, that is deep-learning research models into AD for diagnosis and progression forecasting placed on a graph of two axes.

5.1 Axis 1: Number of tasks

A single-task model can be defined as a ML model which uses extracted latent information as it's input to optimise a single classifier or predictor, i.e. having a single goal/task. A multitask model on the other hand utilises the extracted latent information in multiple submodels, each of which applies to a different task which can be of very different types. For example, the extracted features may be fed into a classifier network as well as a regression network for forecasting purposes.

As we move along the axis and the number of tasks the model, as posited by El-Sappagh et al (2020), becomes increasingly generalised as the tasks act at implicit regularisers. Theoretically, models which have more objectives will be less prone to overfitting as the latent information relative to all tasks is prioritised and reinforced more than those which only

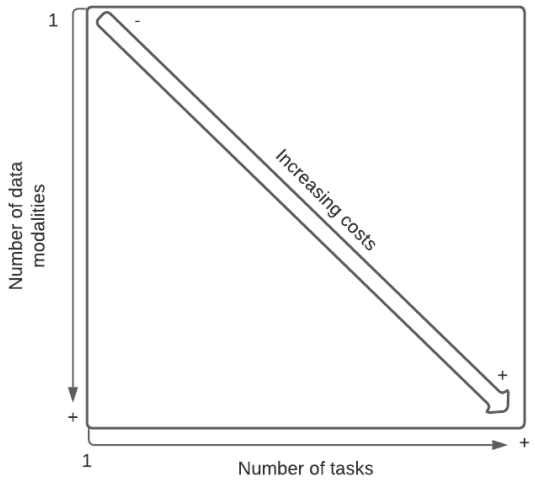

Figure 5: Model-scape illustration

help one task, evidence for which was given by El-Sappagh et al (2020). In reality however, given limited training data this competition can prevent genuine improvements in task performance as tasks compete for learnable information.

5.2 Axis 2: Number of data modalities

A mono-modal model is a one which aims to take a single data modality and use it to perform one or more tasks. Models of this type generally aim to find the optimal method of extracting latent information from that single data modality, such as PET or MRI. The benefits of this are many; the technology used in the model can be more specifically suited to the data modality and thus extract more latent information, generally achieving higher performance than multi-modal models in their specific modalities. Furthermore, the simplicity of this approach allows for a closer inspection of the technologies interaction with the data, giving further insight in to potentially optimal methods of information extraction.

This approach is however very limited by it's single data modality, and it is unlikely that any future viable system for AD prediction for example will utilise only a single modality due to it's lack of complete latent information relating to the task at hand. Furthermore, it may well prove to be that a given data modality only becomes suitably valuable once examined in conjunction with another, providing the motivation to move towards multi-modal models; models which use more than a single data modality. When performed correctly this can have very positive impacts on a models performance, but comes at increased overall costs, including the amount of data, computing power and time needed to train and run such a system, incurring significant financial costs.

6 Literature Review

In this section we will perform a review of the relevant literature and attempt to understand a number of questions. In order to begin building a useful ML system we first need to understand what problem we wish to solve, as well as what options are available to us in terms of data and technologies.

Also, generally speaking only deep-learning approaches will be considered here. Numerous studies including that done by Venugopalan J. et al (2021) have demonstrated deep learning's clear superiority in this area of research, thus to reduce the reviews scope we will limit it to only deep learning.

6.1 Neuroimaging input methods

Survey papers can give us an excellent overview of the current state of research into ML for AD. The ML techniques used will be heavily dependant on the input vector form, and Ebrahimighahnavieh A. et al. (2020 give us a thorough examination of input methods for neuroimaging techniques. Specifically, this review concerns research into AD detection (classification). These techniques are split into four broad categories, these being Region of Interest (ROI) based, Patch-based (PB), Voxel-based (VB) and Slice-based (SB) methods.

ROI based methods use only data from partitioned, hand selected regions of the brain. As we know some areas of the brain effected by AD at different stages of the disease, as well as how these present, we can reduce the size of the input vector to just those already established regions of interest from the clinical literature, thus greatly reducing the number of parameters required in a given NN. Here, the burden of responsibility for feature selection and reduction is almost entirely placed on the programmers shoulders, thus allowing for simpler ML models to be used. Example regions of interest include the hippocampal texture and volume. This lowers costs, costs here referring to factors which scale with input vector size (computational load, amount of data required and so forth). This also has the effect of removing 'noise' from our model inputs, potentially improving model performance. Furthermore, the resultant ML should be more interpretable given it's constrained inputs, potentially improving faith in the system.

This method also has many flaws as vast number of assumptions are made when applying a ROI based approach. Firstly, the regions themselves must be hand-selected and crafted into features through a labour intensive process performed by those with the relevant expert knowledge. This is not only costly but it assumes that the expert can select better regions and craft better features than the model-itself, essentially handicapping it. The expert will undoubtedly miss useful latent information. Rather than utilising in-model mechanisms, this approach simply excludes large quantities of data, most likely placing a ceiling on model performance. Overall, ROI-based methods present the simplest potential solution in terms of model complexity.

VB methods however represent the opposite approach to ROI-based methods, in which the whole image is given as an input. As such the responsibility of identifying key features within the data is entirely off-loaded onto the ML model, alleviating previously mentioned costs. This allows for maximum utilisation of the most modern ML techniques and lift the models performance ceiling to it's maximum (given the input data).

The increased reliance and workload on the ML system however comes at increased computational costs and model complexity. Dimensionality reduction is incredibly important here. Furthermore, input vector pipelines must be increasingly complex to transform the

basic input data to something more usable. Also, this global approach could miss more nuanced local information.

PB methods sit in-between the two previously mentioned methods, in which the brain is segmented into a series of patches of uniform size and spacing. Whist normally all patches are supplied to the ML model, the model will perform patch-selection to reduce the required number of network parameters and select the most informative ones. This method offer a decent half-way-house, allowing for the capture of local data within a patch as well as inter-patch data. Thus this method is more generalised than ROI based, but more specific than PB methods. It's computational demands and other related costs also sit in-between the two.

Finally, SB methods assume that the ML task can be solved using only 2-dimensional brain slices, rather than the 3-dimensional approach of the previously described methods. SB methods benefit from decreased numbers of required network parameters but will undoubtedly miss some amount of information.

With this understanding of the various input methods used within the literature, we can begin to exam it and see how methods have evolved through time and the reasons for this evolution.

6.2 Problem definitions

6.2.1 The data problem

Firstly, a brief note on data. A fundamental problem in AD ML research is a lack of quality data in large enough quantities for training advanced ML algorithms. Indeed, advancements in the collection of high-quality, uniformly collected and processed data have been a crucial bottleneck. Research initiatives such as have produced standardised procedures for biomarker collection across many data modalities, producing cohesive data sets for ML research. Given the time-scale on which AD operates (sometimes tens of years) data collection is unfortunately slow, however enough exists for truly advanced ML algorithms to be used. Whilst this is true for all data modalities, it is especially true for neuroimaging data. Small variations in neuroimaging collection can have massive consequences on how comparable data is. Furthermore, neuroimaging data is expensive to collect and further processing, discussed in section 3.2.2, is time-consuming and requires large amounts of expert knowledge. As such, ML researchers in the AD space have been especially constrained.

6.2.2 Evolution of disease classification methods

<u>Cross-sectional Studies</u>

Computer-Aided Diagnosis (CAD) is the utilisation of technology for aiding in the diagnosis of specific diseases. CAD for AD using ML technologies aims to take one or more data modalities and output a classification in regards to the patients disease status. This problem is also commonly referred to as AD detection, however they are both functionally identical.

Much initial research attempted to classify simply between those with, and with clinically onset AD. Faturrahman, M et al. (2017) present a Deep Belief Network model for classifying between and AD patients. Indeed a number of early papers on ML for AD took this approach. Given our knowledge of AD progression and the necessity of early-detection and treatment however, this output is of little value. Small variations on the AD vs. classification included dividing the AD subjects to those with early-AD and late-AD (Jha D

and Kwon G. 2017). This approach is uncommon at the present for reasons which will be discussed shortly.

Later research such as Li X. et al (2017), Rajapakse R. (2021), and Fan Z. et al (2020) acknowledged this short-coming and further divided the subjects into , AD and MCI groups. The goal here is to detect the onset of AD as early as possible, greatly increasing the utility of their systems. For example, Fan Z. et al (2020) used support-vector machines to classify between , AD and MCI groups. Classification performance across this research demonstrated the increased difficult of distinguishing MCI subjects from others. One reason for this is that given only a percentage of patients experiencing MCIs will go on to development AD. To further complicate matters, it can be extremely unclear when a patient has a perceived MCI whether or not it relates to AD or if another ailment is causing it. Not only this, but patients subjected to preventative AD care post MCI diagnosis have varying responses to it. It is also worth noting that a patient noted to have been experiencing a MCI in a data-set may have progressed to fully onset AD after the fact.

This indicates that the MCI category may be insufficient for the specific requirements of ML classifications. A less popular dividing of the MCI category was in to early-MCI (e-MCI) and late-MCI (l-MCI), such as in Jha D. and Kwon G., (2017) in which a spare auto-encoder was used for classification. A far more common approach in the most recent research however is spliting MCI into stable-MCI (s-MCI) and progressive-MCI (p-MCI), meaning those who's MCI does not lead to AD and those whos MCI does progress to AD. This has proved to be a significantly more challenging task than vs. AD classification, with all research attempting to distinguish between the two displaying the worst performance of any one-vs-one classification. This hypothetically also provides the most useful output of any task mentioned thus far given that a patients first visit to a clinician and subsequent diagnosis and/or treatment occurs when the experience a perceived MCI. This spitting of MCI into stable/progressive categories is now present in the majority of AD classification (Liu S. at al (2014), Sakvatore C. et al (2015))

The progressive/stable MCI dichotomy is however not perfect. Those counted as s-MCI subjects may progress to developing AD, and thus would fall into the p-MCI category. In reality, the s-MCI group may include genuinely stable patients, but also most likely encompasses those whose progression is slower (more stable) than those of the p-MCI group.

Longitudinal Studies

Up until this point only studies utilising cross-sectional data have been discussed, as in the early stages of research into ML for CAD of AD there was too few high-quality longitudinal data sets of sufficient size to warrant large research efforts. Hinrichs C. et al (2011) demonstrated an early attempt at ML models utilising longitudinal MRI and PET data for classification between CN, AD and MCI subjects. Their application of Multi-Kernel Learning here is a perfect demonstration of the limitations within longitudinal research at this time had to work around. Contrasted with Aghili M. et al;s (2018) study some years seven later, and we see the drastic difference greater data availability makes on the studies choice of ML model, in this case RNNs, specifically LSTM networks.

As outlined by Martí-Juan G. et al (2019) when our previously discussed CAD problems are cast as longitudinal ones we see a blanket performance boost across all classification tasks. Not only this, but multiple studies have noted greater improvements in 'harder' problems (such as s-MCI vs. p-MCI). These findings have been consistent across multiple studies.

Longitudinal data utilisation however brings some significant challenges, especially when dealing with missing data. As discussed by Martí-Juan G. et al (2019), temporal alignment

of biomarkers is necessary but not sufficient. The rate of disease progression is not constant across all patients, and many factors including age, genetic profile and lifestyle have a large impact. As such, 'reference variables' are commonly used to build a context within which neuroimaging data can be examined. There are many studies utilising multiple data modalities in combination with neuroimaging data.

Temporal alignment can also refer to the regular intervals between data collection. Many ML models for learning from temporal data implicitly rely upon uniform spacing between patient data collection.

A larger problem however is missing data. Martí-Juan G. et al (2019) broadly outline three methods of handling missing data which are useful to us here. A key distinction however must be made between data that is absent or missing. For example, if a patient has a perceived MCI but it diagnosed as a , then future follow up visits with subsequent testing is not necessary. As such, this follow-up data may not be considered missing in the same sense that if a patient diagnosed with AD misses their regular 6 month checkup. These two types of missing data can be considered as missing data and data missing at random (MAR).

- *Removing incomplete data*: This is the simplest approach in which any data series which is not complete is simply discarded. This however discarded both missing data and data which is MAR, discarding potentially useful information. Indeed, this may in fact introduce biases into the model if patterns exist in the discarded data which is not MAR.

- *Data inference*: A number of methods exist for data imputation in longitudinal datasets, including forward-filling values and using averages from across a classification. This approach is often most useful for dealing with small amounts of data missing at random, but as noted by (ref long survey) scales poorly.

- *Robustness to missing data* - Another method is to use models which are by nature robust to missing data.

It is important to bear these options in mind and examining which are selected when critically analysing studies.

6.2.3 Multi-task models

In cross-sectional studies, CAD is generally the problem to be tackled. With longitudinal studies however another set of tasks has seen great progress in recent years.

Disease progression forecasting involves the use of regression models to predict some patient biomarker(s) within a given time frame. The most common instantiation of this is the prediction of cognitive scores. Fisher, Smith and Walsh, (2018) developed a conditional RBM for the prediction of total ADAS cognitive scores for example. What is relatively new however is the use of neuroimaging data, especially using whole-image input methods, to predict various patient outcomes and variables. These models tend to also be multi-modal, that is using multiple data modalities to aim in prediction or forecasting.

Lee G. et al (2019) present an excellent example of this type of model. They utilised MRI images as well as CSF and EHR (Electronic Health Record) information. What characterises many models of this type is uniformity across multiple data pipelines. This is most likely to promote the learning of features in a similar enough fashion such that they can be merged into a single latent vector in an effective manner and used for multiple tasks. As

such, extremely specialised network types and architectures such as 3d CNNs become less common, and RNNs and their variations take over for learning temporal features.

6.3 Evolution of neuroimaging input methods and technologies

Early studies in this area predominantly used ROI based methods out of necessity. As discussed previously, the high dimensionality of imaging data makes whole-image inputs for traditional ML algorithms and technolgies impractical. Indeed, during this time the was extremely popular and had little capabilities in the way of dealing with whole-image inputs, at least not in a manner which could compete with ROI based methods in performance terms.

As time progressed however, the CNN became increasingly popular to the point where now it dominates the research space. Ebrahimighahnavieh M. et al, (2019) observed that of 2d and 3d CNN were by far the most common ML models to be used for neuroimaging data in AD diagnosis. Indeed, the conclusion of many survey papers (Ebrahimighahnavieh M. et al, 2019; Martí-Juan G. et al, 2019) was that CNNs currently present the best performance in cross-sectional neuroimaging studies, with individual studies consistently showing performance gains over more traditional methods.

6.3.1 Summary of current trends and the state-of-the-art

6.4 General observations

As observed by Martí-Juan G. et al, (2019) this research space currently suffers from a lack of reproducability. The majority of studies either do not provide code, or data, or either in order for other to reproduce and build upon their work. Furthermore, studies commonly do not fully explain their models or provide sufficient detail to recreate their findings. The reason for this is unclear, but it presents a significant issue for researchers such as myself who work alone and with limited time and resources.

Also, the vast majority of studies reviewed used the ADNI data-set for training and testing their models. As such there is a serious risk that biases in the ADNI data-set could be influencing the vast majority of research in this area, acting as a single point of failure. The only remedy to this issue is the creation of other competing high quality data-sets, however not many exist that rival ADNI. Some studies do make an effort to utilise proprietary data for validating models, such as Silvia B. et al (2019), however these data-sets are often not public and thus make reproducing a studies findings more difficult still.

As such my goal should not be to compete with their models in terms of maximum performance, as this would surely be beyond the scope and scale of my research. Instead, an investigation into potential routes of research and experimentation with new ideas may be most fruitful.

Now that we have established a general overview of the field we can begin to dive deeper into individual studies.

6.5 Individual study reviews

6.5.1 Study 1: Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks - Basaia S. et al, 2019

Basaia S. et al, (2019) present a system for the automated classification of AD patients using ML. The only data modality utilised for the training of this model was whole MRI images from patients of one of four classes. These classes were NC, s-MCI, p-MCI and AD, acknowledging the failings of the older 3 class split discussed in section 6.2.2.

Prior to this study only one other paper had attempted to classify p-MCI using MRI images without any prior feature selection, that other paper being by Suk H. et al (2017). In other words, they the first to use the VB method (section 6.3 for neuroimaging inputs in this regard. This means that no pre-processing was necessary wit the acquired MRI images, but instead placing the burden of feature detection and selection on the model. As with virtually all SB or VB input based models, they utilised a CNN for feature learning, in this case a 3d CNN. 3d CNNs abilities in feature detection as discussed in section 4.1.1 make it ideal for this task.

Interestingly, Basaia S. et al, performaed data augmentation, a technique not too common in this space. They performed augmentation through image flipping, scaling, cropping, deformation and rotation.

They utilised an 'all convolutional NN', a type of CNN which removes the need for max-pooling and has been shown to increase performance. In terms of concrete performance, as would be expected the NC vs AD classification was almost trivial, with an accuracy of 98%. p-MCI vs NC accuracy however was lower at 86%, and p-MCI vs s-MCI classification only achieving 75% accuracy. These results are expected as outlined in section 6.2.2

Limitations - The study was limited in a number of ways, but primarily by it's heavily specialised architecture. As with all highly specialised models it is unable to take advantage of certain areas where performance gains may lie. For example, this study did not utilise any time-series data, and would be unable to without a significant overall of the model. Furthermore, it's 3d CNN structure prevents it from utilising other data modalities without, once again, a serious architectural overall.

Advantages - As with all SB or VB imaging input methods, the training data for the model required no preprocessing or labour intensive feature creation and selection. The studies usage of all-convolutional NNs was also interesting, with comparable performance to a previous study, indicating that it is a valid technique to use in this situation. The performance of the model was comparable to many multimodal systems that had come before it, suggesting a large leap in latent information learned considering this study utilised only a single data modality.

6.5.2 Study 2: Predicting Alzheimer's disease progression using multi-modal deep learning approach - Lee G., et al., 2019

Whilst discussed breifly earlier in section 6.2.3 this study is extremely interesting and deserves a deeper inspection.

Albright J. presents a multimodal network for forecasting p-MCI conversion rates using RNNs and time-series data. As would be expected from a multimodal study, the model itself is far less specialised towards a single data modality. Indeed it is able to learn from multiple data modalities with relatively little variation between each data pipeline. The advantage

to this however is that more temporal and inter-data-modality data can be extracted, as is demonstrated by the performance of the network. When compared to cross-sectional data used for the same task a clear performance gain can be observed, bolstering the claim the longitudinal data is superior for CAD for AD.

Whilst not explicitly stated, we can assume that the input method was ROI given that no CNNs are present within the network, the that the MRI information used was phenotypic information. Whilst CNNs appear to remain the best current technique for extracting latent information from whole image inputs, VB or SB inputs, ROI methods in combination with other data modalities remains a competitive strategy.

This is also one of few studies to present an interesting strategy for dealing with irregular data, unlike most which opted for simple methods as outline in section 6.2.2

6.5.3 Study 3: Anomaly Analysis of Alzheimer's Disease in PET Images Using an Unsupervised Adversarial Deep Learning Model - (Baydargil H. et al, 2021)

Baydargil H. et al (2021) present a far less common framing of the ML for AD problem. The authors aim here was to create an adversarial model to train a classifier to perform a type of anomaly detection. By creating a generative model to create synthetic PET images and utilising an adversarial training scheme, a discriminator can be taught to distinguish between real and fake images. The entire model however is only trained on one class for a given one-vs-one classification task. Once the model has been fully trained, the generative element is discarded and the discriminators loss on test data represents how different the input image is than the class it was trained on.

The models generative element consists of an encoding element to learn features from the input images followed by a decoder which takes the latent information data and produces the synthetic images. Unlike our previous model, this model takes entire images as inputs rather than only ROIs, and uses 2d CNNs to extract information from it. The encoder uses a parallel model of almost identical CNN pipelines, with one regular CNN pipeline and one diluted CNN pipeline, the difference between the two being slightly different parameters which allow for better local feature detection in the former, and better global feature detection in the latter. These parallel pipelines series of repeating sets of convolutional layers with batch normalisation, followed by max-pooling layers for feature reduction. After 4 consecutive blocks the features of either pipeline are fused to create the latent vector. The decoding element begins at the latent vector and through a series of 2d convolutional layers and upsampling blocks, eventually produces a synthetic image.

The models discriminative element is similar to an auto-encoder in structure, using a series of repeating CNN and maxpooling layers to learn features from the input, either a fake or real image. After the final CNN block a series of fully-connected layers are used before a final fully-connected layer with a single node and a sigmoid function to output the predicted class.

Advantages - Firstly, the parallel encoder architecture extracts better features than non parallel version, increasing the amount of local and global information learnt from the input images. As such, the reconstructed image can become closer to the real thing. The generative aspect of the model

6.6 Where to begin

Suppose that in the coming years this research topic is solved, and a viable commercial solution has been developed. Given the vast array of data modalities and powerful machine learning techniques it is unlikely that guessing with essentially random combinations of the two will drop this solution in our laps, thus a search of this model-scape seems logical. The question then becomes, from where do we begin our search?

Many papers have been published on AD ML research using the data-set. Consider again the model-scape graph shown earlier. The majority of research in this area falls into the top left corner of that graph, meaning single task and using a single data modality.

I hypothesise that most best locations would be at either extreme of the graph, that being mono-modal, monotask models or multitask, multimodal models. For my situation and research specifically I believe that beginning at the bottom right extreme provides distinct advantages to its top left counterpart. This is due largely to the unpredictability of an additive search approach in which certain combinations of data modalities for example could hypothetically create drastic and hard to explain changes in model performance. Or in the case of a monotask model could be moving towards overfitting rather than an optimal generalisable solution. A iteratively reductive approach however would theoretically begin at the position of greatly increased interaction between data modalities and task regularisation, and more accurately represent the quantity and quality of useful latent information they contain. This should make it simpler to traverse the aforementioned search tree (and that general area of the graph), or at least make it simpler to interpret changes as elements of the model are removed. Therefore, for the purposes of research such as ours, the logical position to begin in is one close to the bottom right extreme, allowing maximum possible inter-data-modality interaction and providing excellent sandbox in which to experiment with different technologies. In short, we can negate the effect of a lack of diverse data-sets for training and testing by building models which implicitly generalise across multiple tasks.

Fortunately, this aligns excellently with my needs as a researcher. Currently I do not posses the large amount of cutting edge knowledge required to perform research into the most advanced technologies applied to single data modality, single task models. Furthermore, interpreting their behaviour for me would be significantly more difficult when adding or removing components. Multitask, multi-modal models however are generally limited in how specialised and therefore advanced they can be. Compromises must be made in order to accommodate the multiple modality formats and the learning of features which can be fused together into a cohesive system. These compromises and necessary simplifications make this area of research much more accessible to me.

6.7 Trust in ML

Trust in ML systems has become a famously divisive topic in recent years. Automated vehicles provide an excellent example of this. Whilst being driven by Tesla's automated vehicles, one is ,on the whole, much safer than being driven by the average human driver, there are many people who swear they will never get in of these vehicles. Whilst not a perfect parallel, doctors similarly are reluctant to hand over responsibility to, and trust in, automated ML systems. The reasons for this are clear, a primary concern being the difficulty of explaining the actions of ML systems, especially in safety-critical matters.

This problem is usually posed as 'explainable ML' but what does this really mean? Generally, this term is used to describe the inability to look inside a complex system such as

a NN model and make sense of its millions of parameters. It is for this reason NNs are often referred to as black boxes which spit out results whilst hiding their internal representations from closer inspection. The framing of this common interpretation of the issue however is rather misleading, as the output of the ML system is itself is a high-level representation of the systems internal state. Thus we could reframe the problem as not an absence of interpretability, but rather an insufficient amount of it.

A simple analogy would be as such: Person A is a car designer and asks Person B if he thinks a car he has designed is "cool" with the aim of understanding what makes a car "cool". If Person B gives a yes/no response, Person A can interpret little other than that the arrogate of the cars features are appealing. If Person B also says he likes it because it is red, looks expensive and has a spoiler then Person A can begin to make further interpretations about Person B's likes/dislikes, in other words, their internal state.

For Person A, understanding exactly why the car being red has an impact on its appeal is not important. He does not need a perfect understanding of the infinite combination of psychological, economical, sociological or biological reasons for Person B's preferences. He only needs a functional interpretation of the persons internal state which can be improved through the collection of auxiliary information.

One potential solution to this re-framed problem is to produce more than a single output, and have models which produce auxiliary information about related tasks already known to be linked to the main task. which can be used to increase model interpretability by representing different latent information. In combination with the idea of increasing ML generalisability through multiple tasks acting as regularisation parameters in our model, we can both increase our models generalisability and provide insight into it's internal workings.

For example, in a clinical setting a CN v MCI v AD prediction for a patient 12 months after a given time may offer useful information, utilising the latent information extracted to also predict a patients score in a MMSE exam, or ADAS score could greatly increase a clinicians trust in the system as this auxiliary data can be used as supporting evidence, offering a higher resolution representation of a patients health in the future and creating a system in which the overall output is greater than the sum of its individual parts.

To extent this analogy to our domain of research and ML in general, if we can supply the 'user' of the algorithm with enough auxiliary information we may be able to increase trust in the system through refining and improving their functional interpretation of the system.

7 Proposed model

The proposed model is derived from a single model, with a modified version aiming to create a model which can utilise latent information extracted from multiple data modalities to generate synthetic pet information.

7.1 The base

The model presented in (El-Sappagh S. et al, 2020) is a multimodal, multi-task ensemble model which utilises 1d CNNs and Bi-LSTMs to learn features from time-series data and combines this with background data. The model has five outputs. A classification of CN vs sMCI vs pMCI vs AD, and four other predicted cognitive scores. The models inputs are 5 different time-series data modalities, cognitive scores, neuropsychological data, neuropathology data and MRI and PET imaging data. A 6th input is also given, this being background baseline information about the patient such as demographic information.

CNN Bi-LSTM pipelines: The purpose of the CNN Bi-LSTM pipelines is to learn local and temporal features within the data. Up to 15 time-steps are given as the input to a given pipeline, each step being at regular six month intervals. As discussed in section 4.1.1, the CNN layers learn local features within each time-step, producing increasingly abstract feature maps. The following Bi-LSTM layers can then learn temporal features from the resultant set of feature maps. After a set of fully-connected layers the information from each individual pipeline, which are learned separately, is fused together by a series of fully-connected layers.

This model differs from the vast majority of studies using neuroimaging data in that it is ROI based. As described in section 6.1 this hold a number of advantages including a large reduction in the number of parameters required and also strips large amounts of useless data from the input. The trade-off here is of course is that useful information is likely also discarded, indeed in the study we see that single-modality performance for MRI data is significantly worse than it's counterparts, as well as far worse than other studies using only MRI data with highly specialised networks. This trade-off appears worthwhile however as this models performance was state-of-the-art in multimodal multitask models upon it being published. Furthermore, using a ROI based input method allows for 1d CNNs to be used and thus keep each CNN Bi-LSTM pipeline identical.

Background pipeline: The background data pipeline is comparatively simple. The background data here is a combination of demographic and other static information about a patient, as well as some statistical information relating to their other time-series data. This background data is fed into a series of fully-connected layers, the output of which is concatenated with the final output of the CNN Bi-LSTM pipelines.

This final concatenation is then fed through more fully-connected layers to fuse together the learned features and produce the latent feature vector. This latent feature vector is the final representation of the underlying data. From here a set of separate pipelines is defined for a number of tasks. All pipelines utilise several fully connected layers, with only the final layer being different. The purpose of the fully-connected layers in each pipeline here is to learn what features from the latent feature vector are relevant to the pipelines specific task. For the classification task, a fully-connected layer with 4 nodes is use in combination with a softmax activation function to obtain a vector with 4 values, each corresponding to the models predicted likelihood that the input falls into a given category. The other regression pipelines end in a fully-connected layer with one node and a sigmoid activation function to produce a normalised prediction for a given cognitive score.

7.1.1 Advantages

Advantages of this approach related to hit spoken about before The advantages of this approach are discussed in the original study itself and are thoroughly backed up by the surrounding literature. Firstly, it utilises time-series data which has been shown to have higher performance than cross sectional models. It is also an ensemble model which uses multiple data modalities, which have also been shown to achieve higher performance than single-modality models as discussed in section 6.2.3. Furthermore, as it is a multitask model it is hypothetically more generalisable and stable than other single-task models. On top of this it also provides auxiliary information (cognitive scores) to would be users of the system, thus improving trust in the system. This extra information gives makes the model more interpretable through providing information which allows for the building of a functional interpretation of the systems internal learned features. Finally, the models classification

between sMCI and pMCI patients is excellent relative to it's other classification.

7.1.2 Limitations

The study has several limitations. Firstly, the authors imputed vast quantities of data to train and test the model. The general advantages and disadvantages of this method were outlined in section 6.3, but it bares mentioning that that over 80% of all MRI data for example was imputed through one method or another. Whilst the authors back-up this method with solid reasoning and reference to established practices for AD data imputation, it is difficult to assess if this is introducing biases into the data due to a lack of cross referencing across different data sets as discussed in section 6.4.

7.1.3 Other

Whilst the model has many advantages from a performance and learning perspective, there are also a number of features which make it a highly suitable choice of model to adapt specifically for my situation. As discussed in section 6.6, this model presents a suitably complex and sophisticated system whilst remaining accessible with my current skill set. For example, each of the five CNN Bi-LSTM pipelines is identical to allow for features to be learned which can be fused together in an effective manner. Whilst more information could be learned from each modality with more specialised pipelines, this would come at the cost of creating a cohesive system and would likely negatively impact performance.

7.2 My model

Whilst this section is meant to provide specific details as to the model, the data and it's performance, it is also to describe the process I undertook to reach the end result. My experiences seem to be indicative of what one can expect from an area of research plagued by studies which are extremely difficult to reproduce. This section also offers a reflection of mistakes I made along the way, and the discoveries I made from said mistakes.

7.2.1 Justification

As mentioned in section 6.7, we can theoretically improve trust in ML systems by providing extra auxiliary information to the end user. El-Sappagh S. et a (2020) present a great example of this whilst also setting itself in the previously discussed sweet-spot of being multitask and multimodal. It presents an excellent base to work from.

The inspiration for this comes from the study discussed in section 6.5.3. Personally I find the most interesting part of this of this study to be the discarded generative element rather than its discriminative counterpart. If we could use time-series patient data to generate synthetic PET images in order to predict what a patients PET scan would look like after a given time period since a visit, we could give clinicians an invaluable tool. This would not only increase trust in the ML system, but potentially hold some performance benefits.

Unfortunately, circumstances have forced me to scale back my expectations and instead devise a model which can predict variables relating to brain activity in FDG-PET images. Whilst the forecasting of these values has been done before, their prediction in a multitask model which also performs classification and uses six separate data modalities has been less researched.

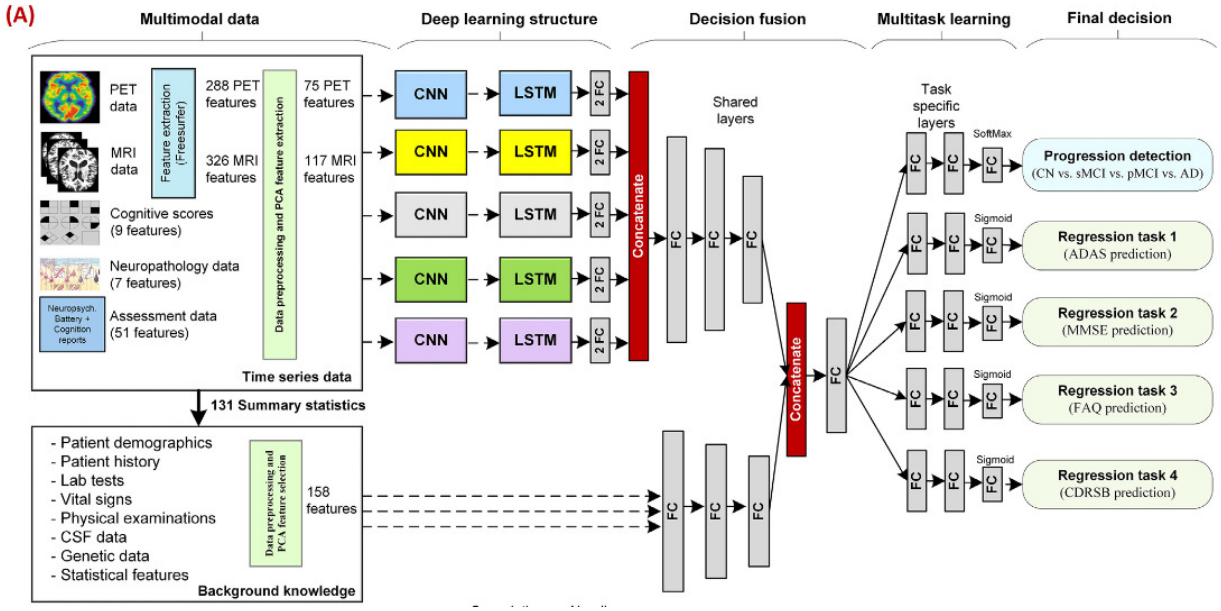

Figure 6: Model-ase illustration from El-Sappagh S. et a (2020)

7.2.2 Design overview:

The machine learning model itself was implemented with the Keras API for Tensorflow. All feature reduction was performed with the sklearn python package.

Base model: Our base model from which modules will be added and investigated is identical to the model presented by (ref ensemble model). The figure below shows taken from El-Sappagh S. et a (2020) show the exact structure of the network. The functionality of the system we have already explored at length. A more detailed description can be seen in figure 7.2.2.

Each CNN Bi-LSTM pipeline consists of one 1d-convolutional layer with 128 filters, each of size 4, and a stride of 1. The activation in the CNN layers as 'relu', and padding was kept at 'same', followed by a 1d max pooling layer with a pool size and stride of 2 and 'same' padding. All other parameters were left as default. This is followed by three stacked Bi-LSTM layers with 128 units each, 'tanh' activation, a dropout rate of 0.1 and l2 regularisation set to 1e-2. Finally, 2 fully-connected layers with 64 nodes each, rectified linear (relu) activation and l2 regularisation of 1e-2. The final layer of each pipeline is then concatenated into a single fully-connected layer with 320 nodes. This is followed by 3 more fully-connected layers with identical parameters to the previously mentioned fully-connected layers, before being concatenated with the background data pipeline.

The background data pipeline consists of 3 fully-connected layers with 64 nodes, l2 regularisation set to 1e-2, and a dropout rate of 0.2. The final concatenation is passed through one more fully connected layers with 128 nodes, l2 regularisation set to 1e-2 and a dropout rate of 0.1. The end result is the latent vector from which each individual task pipeline comes.

The classification pipeline is structured as follows. 2 fully-connected layers with 32 nodes

each, softmax activation, a dropout rate of 0.2 and l2 regularisation set to 1e-3 lead to a final fully-connected layer with (params). The output of this final layer is therefore a 1d of shape (1,4), in which every value represents the models predicted probability that the given input falls in to a given category (AD, sMCI, pMCI or AD). The metric used for this task was categorical crossentropy.

The four regression pipelines are have 2 identical fully-connected layers. The final layer however is a fully connected layer with 1 node and a sigmoid activation function. The final output of each regression pipeline is the normalised expected output for a given cognitive score. The metric used for this task was mean absoute error.

Finally, the optimiser used was 'Adam' with a learning rate of 0.00005.

7.2.3 Experimentation

Issues with data replication

The first test was to build the model which is almost identical to that in the previously mentioned study. The building of the model was simple enough, after a few online tutorials in utilising the Keras functional API I was able to begin. What came as more of a surprise however was the difficulty of replicating the data used in the study.

Figure 7.2.2 gives a very high-level overview of the data used for training the model, with further details provided in a supplementary file to the study. The exact ADNI subject ids was given for the subjects used, however this was the only clear information the study provided.

Firstly, gathering the data and developing pipelines for it proved a nightmare for myself given my lack of data science skills and prior experience with only toy ML data-sets. A significant amount of time was spent simply collating the data I could find into usable structures with the 'pandas' python package. Firstly, the variable listed by the study as existing in specific files were not always in the places they claimed to be. On top of that, some input variables such as 'BMI' for the background data simply did not exist in any files I could locate.

These problems plagued the data collection and processing process. For example, the study gives a number of features used for the MRI data, but does not say which exact variable are used. Furthermore, following the steps given in said study, such as removing features if 30% of it's data is considering MAR yields significantly fewer features than required. Furthermore, multiple versions of each file existed and it was not stated which should be used. In some cases such as with the 'MOCA' variable, there was no data for any patients from the ADNI1 protocol. Finally, when the study claims to use statistical data from each of the time-series modalities in the background information, simple maths will show that a number of these statistics can not have been used.

Overall I learned an important lesson. One cannot rely upon the data pipelines given to us from such studies unless they are suitably explained or code is provided for them. As neither was the case here, I ended up wasting massive quantities of time desperately trying to replicate the un-replicatable. Whilst this is most surely a failure on my part, one cannot help but notice this issue being pointed out also in numerous survey papers, and indicates a genuine problem within this area of research. That being said, I will not be making the same mistake again.

Data used

All data used here is provided in the accompanying zip file, with the final inputs being stored in '.npy' numpy files, and the original pandas dataframes stored in the '.csv' files.

In terms of preprocessing, all data was normalised using their 'zscore' which were calculated with the sklearn 'StandardScaler' function. The MRI and PET data was also reduced using principle component analysis, a standard feature reduction technique. This was acheived through the use of sklearn's PCA function with 100 components.

Only classification testing

The first batch of testing was done with only one task, that being classification between AD, p-MCI, s-MCI and AD patients.

The maximum achieved performance here was 96%, most notably with 95% accuracy in p-MCI classification. This was not the most common result however. The model generally suffered from severe overfitting to the s-MCI and early stopping was necessary in many of these cases.

Commonly, the model would score 0% accuracy on a single class whilst have relatively high accuracy in other classes. For example on two separate tests the following results were obtained

Test 1: Overall accuracy - 76%

- NC - Precision: 1.00; Recall: 0.99; f1-score: 0.99
- s-MCI - Precision: 0.96; Recall: 1.00; f1-score: 0.98
- p-MCI - Precision: 0.00; Recall: 0.00; f1-score: 0.00
- AD - Precision: 0.47; Recall: 0.95; f1-score: 0.63

Test 2: Overall accuracy - 71%

- NC - Precision: 1.00; Recall: 0.98; f1-score: 0.99
- s-MCI - Precision: 0.95; Recall: 1.00; f1-score: 0.98
- p-MCI - Precision: 0.00; Recall: 0.00; f1-score: 0.00
- AD - Precision: 0.47; Recall: 0.95; f1-score: 0.63

The final average accuracy over a series of tests was 75%.

Analysis

Initial experimentation led to alterations of the original base model such as decreasing the learning rate and increasing the number of nodes per fully-connected layer for the classification task. Each of these helped stabilise the model and decrease the frequency of early stopping being necessary, however the issue of one class being essentially ignored persisted. Particularly the MCI classes were common for this sort of behaviour.

The most likely reason for this erratic model behaviour is an insufficient amount of latent information in the training data. My failure to replicate the original studies training data and heavy reliance upon it for feature selection most likely caused large amount of latent information to be excluded from the data. On reflection, I should have used more variables in the training data and employed automated feature selection tools such as PCA to perform this task for me.

Training the model with all 5 original tasks

One again early stopping was a necessity here as the model continued to experience severe overfitting. The frequency of these events however decreased overall and overall performance gains were seen.

Peak performance reached an accuracy of 84%, however the regression tasks performed far worse, with an average mean absolute loss of 0.54. Repeated testing gave an average of 78% accuracy and an average mean absolute loss of 0.69 across all regression tasks.

Analysis

The regression performance is surprisingly weak considering what was achieved in the original study, however as expected we did see improved generalisation in our classification task, with getting 0% accuracy on a given class being uncommon.

Training for new PET regression tasks

Here the regression pipelines of the original study were replaced with 5 identical regression pipelines for predicting the mean values of a set of 5 pet variables.

The best performing model obtained an accuracy of 73% across all classes and an average mean absolute error of 0.596 across all regression tasks. The models still suffered from extreme overfitting and early stopping was necessary in the majority of training runs.

Model stability in these tests appeared comparable to that of the previous multitask tests, but definitely worse, and the classification accuracy was worse. Interestingly, the regression tasks performed better than that of the cognitive score regression tasks.

Analysis

There are many potential reasons for this. This could be due to the fact that neuroimaging data accounted for the largest amount of overall features in the training modalities. More likely however is that as each regression task is performing more similar tasks than the separate cognitive score regression tasks, that the model is learning more specialised features. As a result, the regression tasks have a lower mean absolute error.

A potential issue with this is that the models over generalisability is being reduced compared to the model predicting cognitive score. Indeed, this is most likely the reason that the classification tasks performance also decreased and suffered from overfitting more. Closer inspection of the individual class accuracy's reveals that commonly one or even two entire classes were achieving 0% accuracy.

8 Conclusions

8.1 Successes

In terms of model performance, this research has been a clear demonstration of the ability of multitask models to generalise well when compared to single-task models. Since the aim of this research was not to compete with state-of-the-art models but to explore the possibilities of different ways to utilise neuroimaging data, specifically PET data, we have seen some success. As discussed in section 6.7 we have provided another piece of auxiliary information that may be useful to clinicians in need of CAD for AD.

In terms of a holistic understanding and discussion of the area of research, I believe we have uncovered some interesting insights in to the direction of progress as well as current limitations and negative trends. It is clear from the literature review that multitask, multimodal ML models will continue to be explored and offer distinct advantages in the way of model performance, but also in investigating data interactions. Furthermore, these models appear to be offering a chance to create better generalised models in a landscape with potentially massive biases lurking under the surface thanks to essentially a single point of failure, that being the ADNI data-sets.

8.2 Limitations

The work is limited in a number of ways. Firstly as previously stated my reliance upon the El-Sappagh S. et al (2020) study for data collection and processing led to a data-set which is lacking in latent information. As a result the performance of the model was poor relative to the current state-of-the-art. A potential remedy to this has already been mentioned, that being utilising more variables and feature selection and reduction tool, however by the time this was realised the timing constraints upon this research meant that it was too late.

Secondly, the model itself limited in the same fashion as other multimodal studies. For example, requiring the time-series data pipelines to be extremely similar for feature extraction and effective fusion inevitably reduces the maximum amount of learnable from each data modality. This is of course however necessary a trade-off.

8.3 Future work

The most obvious place for future work is a re-examination of training data and a different approach to data generation involving feature reduction and selection across a broader range of variable. This would undoubtedly improve model stability and performance.

Aside from performance however, the ideas put forward relating to trust in ML systems and how to improve it through allowing the user to build a functional interpretation of the systems internal structure and reasoning through providing auxiliary information seem promising. Future work involving multitask models may find this useful and provide said systems with increased utility. For example, as was the original goal of this research, one could generate synthetic PET images.

8.4 Final thoughts

This research was well aimed, but poorly executed. Whilst the reasoning and motivation behind it and the system which was built was solid, it's ultimate downfall was my lack of experience in dealing with ML data. That being said, it was an invaluable experience in utilising first-principles to reason through a complex area of research involving the most state-of-the-art machine learning technologies, and I am extremely grateful to have undertaken it.

9 Bibliography

Martí-Juan, G., Sanroma-Guell, G. and Piella, G., 2019. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease.

Aghili, M., Tabarestani, S., Adjouadi, M. and Adeli, E., 2018. Predictive Modeling of Longitudinal Data for Alzheimer's Disease Diagnosis Using RNNs. [online] Elsevier. Available at: ¡https://link.springer.com/chapter/10.1007

Albright, J., 2019. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. [online] PMC. Available at: ¡https://www.ncbi.nlm.nih.gov/pmc/articles/ [Accessed 5 November 2021].

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R. and Filippi, M., 2018. Automated classification of Alzheimer's disease and mild cognitive im-

pairment using a single MRI and deep neural networks. [online] pubmed.gov. Available at: ¡https://pubmed.ncbi.nlm.nih.gov/30584016/¿ [Accessed 5 November 2021].

Baydargil, H., Park, J. and Kang, D., 2021. Anomaly Analysis of Alzheimer's Disease in PET Images Using an Unsupervised Adversarial Deep Learning Model. [online] PubMed.gov. Available at: ¡https://www.mdpi.com/2076-3417/11/5/2187¿ [Accessed 5 November 2021].

Ebrahimighahnavieh, M., Luo, S. and Chiong, R., 2020. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. [online] Elsevier. Available at: ¡https://www.sciencedirect.com/science/article/pii/S0169260719310946?via%3Dihub¿ [Accessed 5 November 2021].

El-Sappagh, S., Abuhmed, T., Riazul Islam, S. and Kwak, K., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. [online] Science Direct. Available at: ¡https://www.sciencedirect.com/science/article/pii/S0925231220309383¿ [Accessed 4 November 2021].

Fan, Z., Xu, F., Qi, X., Li, C. and Yao, L., 2020. Classification of Alzheimer's disease based on brain MRI and machine learning. [online] Springer Link. Available at: ¡https://link.springer.com/article/10.1007/s00521-019-04495-0¿ [Accessed 5 November 2021].

Faturrahman, M., Wasito, I., Hanifah, N. and Mufidah, R., 2017. Structural MRI classification for Alzheimer's disease detection using deep belief network. [online] Ieeexplore.ieee.org. Available at: ¡https://ieeexplore.ieee.org/document/8265643¿ [Accessed 2 November 2021].

Fisher, C., Smith, A. and Walsh, J., 2018. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. [online] arXiv. Available at: ¡https://arxiv.org/abs/1807.03876¿ [Accessed 4 November 2021].

Hinrichs, C., Singh, V., Xu, G. and Johnson, S., 2011. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. [online] ScienceDirect. Available at: ¡- https://www.sciencedirect.com/science/article/pii/S1053811910015806¿ [Accessed 5 November 2021].

Ieeexplore.ieee.org. 2017. Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data. [online] Available at: ¡https://ieeexplore.ieee.org/document/8211486¿ [Accessed 5 November 2021].

Ieeexplore.ieee.org. 2017. Structural MRI classification for Alzheimer's disease detection using deep belief network. [online] Available at: ¡https://ieeexplore.ieee.org/document/8265643¿ [Accessed 2 November 2021].

Jha, D. and Kwon, G., 2017. Alzheimers Disease Detection Using Sparse Autoencoder, Scale Conjugate Gradient and Softmax Output Layer with Fine Tuning. [online] Ijmlc.org. Available at: ¡http://www.ijmlc.org/vol7/612-IP018.pdf¿ [Accessed 5 November 2021].

Jha, D. and Kwon, G., 2017. Alzheimer's Disease Detection Using Sparse Autoencoder, Scale Conjugate Gradient and Softmax Output Layer with Fine Tuning. [online] Ijmlc.org. Available at: ¡http://www.ijmlc.org/vol7/612-IP018.pdf¿ [Accessed 2 November 2021].

Ju, R., Hu, C. and Zhou, P., 2019. Early Diagnosis of Alzheimer's Disease Based on Resting-State Brain Networks and Deep Learning. [online] Ieeexplore.ieee.org. Available at: ¡https://ieeexplore.ieee.org/document/8119531¿ [Accessed 5 November 2021].

Ju, R., Hu, C., Zhou, P. and Li, Q., 2019. Early Diagnosis of Alzheimer's Disease Based on Resting-State Brain Networks and Deep Learning. [online] Ieeexplore.ieee.org. Available at: ¡https://ieeexplore.ieee.org/document/8119531¿ [Accessed 2 November 2021].

Lee, G., Nho, K., Kang, B., Sohn, K. and Kim, D., 2019. Predicting Alzheimer's disease progression using multi-modal deep learning approach. [online] nature.com. Available at: ¡https://www.nature.com/articles/s41598-018-37769-z¿ [Accessed 5 November 2021].

Li, X., Li, Y. and Li, X., 2017. Predicting Clinical Outcomes of Alzheimer's Disease from Complex Brain Networks. [online] Springer Link. Available at: ¡https://link.springer.com/chapter/10.1007

Liu, S., Liu, S. and Cai, W., 2014. Early diagnosis of Alzheimer's disease with deep learning. [online] Ieeexplore.ieee.org. Available at: ¡https://ieeexplore.ieee.org/document/6868045¿ [Accessed 5 November 2021].

Martí-Juan, G., Sanroma-Guell, G. and Piella, G., 2019. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease.

Rajapakse, R., 2021. Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data. [online] Ieeexplore.ieee.org. Available at: ¡https://ieeexplore.ieee.org/document/82 [Accessed 5 November 2021].

Rosenblatt, F., 1958. THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN. [online] Citeseerx.ist.psu.edu. Available at: ¡https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3398rep=rep1type=pdf¿ [Accessed 4 November 2021].

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M., Quattrone, A. and Castiglioni, I., 2021. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. [online] Frontiers. Available at: ¡https://www.frontiersin.org/articles/10.3389/fnins.201 [Accessed 5 November 2021].

Suk, H., Lee, S. and Shen, D., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. [online] Nation Centre of Biotechnical Information. Available at: ¡https://pubmed.ncbi.nlm.nih.gov/28167394/¿ [Accessed 5 November 2021].

Venugopalan, J., Tong, L., Hassanzadeh, H. and Wang, M., 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. [online] Nature.com. Available

at: ¡https://www.nature.com/articles/s41598-020-74399-w¿ [Accessed 5 November 2021].

Towards Data Science. 2020. Understand Local Receptive Fields In Convolutional Neural Networks. [online] Available at: ¡https://towardsdatascience.com/understand-local-receptive-fields-in-convolutional-neural-networks-f26d700be16c¿ [Accessed 5 November 2021].

Ebrahimighahnavieh, M., Luo, S. and Chiong, R., 2019. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. [online] Pubmed. Available at: ¡https://pubmed.ncbi.nlm.nih.gov/31837630/¿ [Accessed 5 November 2021].