University of Zurich[UZH]

Bachelor Thesis

Department of Psychology

Faculty of Arts and Social Sciences at the University of Zurich

Comparison between Support Vector Machine and Convolutional Neural Network for Alzheimer's Disease Classification

Oliver Zingg

16-924-094

Prof. Dr. Nicolas Langer

Chair of Methods of Plasticity Research

01.06.2021

Contents

List of Figures

List of Tables

Abstract

The number of people living with dementia is estimated to triple by the year 2050. This makes Alzheimer's Disease (AD), which accounts for up to 80% of all dementia cases, a crucial area of study. With the trend of biomarkers gaining importance in diagnosing AD, imaging techniques such as magnetic resonance imaging (MRI) have been widely studied. To find disease related patterns, machine learning algorithms have been applied. Furthermore, the interest in automated classification systems assisting AD diagnosis has increased. Support vector machine (SVM) and convolutional neural network (CNN) are the most often used machine learning algorithms for automated classification. Therefore, this review included 10 studies using either SVM or CNN for automatic AD classification with the goal of comparing them. It can be concluded that although both algorithms perform well when classifying AD vs. normal cognition, direct comparison between the two is limited since size of data set, pre-processing steps and evaluation methods vary. Novel research provided a framework to make comparison more transparent. Moreover, all reviewed studies relied on the ADNI database to train their models, which may influence generalizability.

1 Introduction

The Alzheimer Europe Organisation estimates that by the 2050 the amount of people living with dementia will have doubled in Europe, meaning an increase from 9'780'678 to 18'846'286 people (Alzheimer Europe, n.d.). Other estimations suggest that dementia is present in 50 million people worldwide and is predicted to triple by the year 2050 (Alzheimer's Disease International, n.d.). Alzheimer's Disease (AD) is the most prevalent form of dementia and may account for 60-80% of all dementia cases (Leandrou et al., 2018). In elder population, AD has a prevalence of 10% and is considered the most common neurodegenerative disease (Nanni et al., 2020).

Diagnostic criteria for AD have changed multiple times, moving past an exclusively clinical approach of diagnosing AD, towards incorporating biological aspects (Scheltens et al., 2021). The National Institue on Aging and Alzheimer's Association (NIA-AA) putting forward the AT(N) research framework – labeled after the suggested biological markers for AD: β-amyloid, tau and neurodegeneration (Jack et al., 2018) – reflects the trend of separating clinical symptoms from disease pathology and having purely biological criteria for AD (Ahmed et al., 2021; Scheltens et al., 2021). Still, diagnosis is primarily done clinically (Ahmed et al., 2021) and, as Islam and Zhang (2018) describe, requires physical and neurological examinations, a detailed history of the patient as well as tests, like Mini-Mental State Examination (MMSE) (Folstein et al., 1975). A definite diagnosis can only be made post-mortem (Nanni et al., 2020).

In vivo diagnosis accuracy can be affected by the heterogeneity of the clinical symptoms in AD (Ahmed et al., 2021; Frisoni et al., 2011) but increases with the development of the disease (Islam & Zhang, 2018). Still, it is notable that "*10%* to *30%* of individuals clinically diagnosed as AD dementia by experts do not display AD neuropathologic changes at autopsy" (Nelson, 2011, as cited in Jack et al., 2018, p. 538). Although, as Jack et al. (2018) state, the AT(N) framework isn't ready for general clinical usage, it addresses these mentioned challenges by defining AD biologically. This definition includes biomarkers from three categories that form the acronym AT(N): β-amyloid accumulation, pathologic tau and neurodegeneration (Jack et al., 2018). Compared to Positron Emission Tomography (PET), that measures β-amyloid accumulation or tau, Structural Magnetic Resonance Imaging (MRI) is less invasive and can be used to measure neurodegeneration (Wang et al., 2021). Additionally, MRI is always recommended following the clinical evaluation (Scheltens et al., 2021) and is often used as a tool for finding relevant biomarkers (Liu et al., 2020).

Methods such as mass-univariate analyses which are large amount of univariate tests (e.g. t-tests) (Groppe et al., 2011) have been used in combination with neuroimaging studies and led to significant insight into psychiatric and neurological disorders (Vieira et al., 2017). In the case of MRI analysis of AD, hippocampus and other medial temporal lobe structures are of interest (Islam & Zhang, 2018; Scheltens et al., 2021). Vieira et al. (2017) describe Machine Learning (ML) methods as alternative analysis techniques which, in comparison, are multivariate meaning that they take the inter-correlation between voxels into consideration (Vieira et al., 2017). Furthermore, ML methods could allow for individual treatment decisions to be made (Naik et al., 2020), since these methods make statistical inference at an individual level (Vieira et al., 2017). These properties have created a growing interest in ML methods for analysing neuroimaging data (Vieira et al., 2017). Furthermore, ML methods could automate the process of analysing MRI images and consequently safe resources (Islam & Zhang, 2018). Additionally, ML methods have the ability of finding disease-related patterns without prior knowledge about underlying mechanisms (Nanni et al., 2020).

Considering these described aspects, ML methods seem suitable for analysing MRI scans related to AD. Many studies using ML to classify AD have been reviewed in the past (see Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Noor et al., 2020; Tanveer et al., 2020; Wen et al., 2020a). Aiming to provide an updated view on the state of research, this review compares novel studies using ML methods for classifying Alzheimer's Disease.

2 Theoretical Background

2.1 Machine Learning

Shalev-Shwartz and Ben-David (2013) describe ML as automated learning where an algorithm gets training data as input and uses it to optimize certain parameters to find the optimal solution to a given problem. Depending on if the training data is labeled, the learning process can be divided into supervised or unsupervised learning (Shalev-Shwartz & Ben-David, 2013). This review focuses on supervised learning since the input data, MRI-images, are labeled with the patients official diagnosis (see Chapter 2.4).

2.1.1 Support Vector Machine

Support Vector Machine (SVM) is among the most widely used machine learning techniques for classifying AD (Tanveer et al., 2020). It is thought to have good accuracy and sensitivity (terms are described in 2.5) and can handle high-dimensional data well (Toshkhujaev et al., 2020). SVM uses features (independent variables) as input and predicts the class (dependent variable) of an observation (set of independent variable values) (Pereira et al., 2009). For example, Burges (1998) illustrate this with a tree recognition problem where each observation would consist of a feature vector (e.g. pixel values) and a class (e.g. is the observation a tree or not: 1 or -1). SVM then uses these observations to adjust the parameters to solve the binary classification task in an optimal manner. For example, with MRI-images the features used as input can be voxel-based (Tohka et al., 2016), meaning the input comes as vectors with voxel intensity values of the image (Ebrahimighahnavieh et al., 2020). Essentially, as described in Bishop (2006), SVM achieves the classification task through learning the optimal hyperplane that separates the two classes using maximum margin principle. Put simply, a hyperplane can be a line, plane or some other flat subspace depending on the dimensional space it is formed in and margin being the perpendicular distance between the hyperplane and the nearest observation (Nordhausen, 2014). Bishop (2006) states, that to help with generalization (see 2.3) missclassifications of some data points are allowed, which relaxes the hard margin and hence are called soft margin. To find the optimal amount of missclassifications, a tuning parameter that essentially controls the bias-variance trade-off (see 2.3) is chosen via cross-validation (see 2.5) (Nordhausen, 2014). Therefore, SVM's goal becomes to maximize the margin while softly penalizing missclassifications.

SVM can be linear or non-linear depending if the training data is linear separable or not (Burges, 1998). Non-linear SVM follow the idea of mapping an input vector into a feature space with a higher dimension where the input becomes linear separable (Vapnik, 2000). To save computational costs, a kernel function is used to avoid explicitly

transforming, or mapping the input vector into a higher dimensional feature space while achieving the same class separation (Burges, 1998). Non-linear kernel function leads to linear hyperplane in some higher dimension and a non-linear decision boundary in the original dimension of the observations. Therefore, it makes the decision boundary that separates the classes more flexible (Nordhausen, 2014).

For reasons discussed in Mohri (2018) and Vapnik (2000), the core optimization problem of maximizing the margin is equal to solving a convex optimization problem. This is important since it follows that finding any local solution in a convex optimization problem is also a global optimum and reflects a fundamental difference between SVM and Deep Learning (DL) methods (Bishop, 2006).

2.2 Deep Learning

DL is a form of ML (Goodfellow et al., 2016) and is used to describe neural networks with multiple hidden layers (Nordhausen, 2014), therefore, referring to the models "deep" depth (Goodfellow et al., 2016). Neural networks are essentially, as Goodfellow et al. (2016) point out, multiple functions combined in a chain structure that map the input to the output. These functions are termed layers, with hidden layers being the functions between the first (e.g. input layer) and last layer (e.g. output layer). Furthermore, Goodfellow et al. (2016) state that this chain structure or network is described as neural since every hidden layer has many units that are inspired by neurons in the sense that every unit calculates its activation value based on input from units of previous hidden layers. This process somewhat reflects the action potential in neurons. In a review paper Ebrahimighahnavieh et al. (2020) found that Deep Neural Network (DNN), Deep Polynomial Network (DPN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) are four supervised DL methods that have been applied to the AD classification task with CNN being the most often used method. Therefore, this review only focuses on CNN.

2.2.1 Convolutional neural network

CNN (LeCun et al., 1989) is a form of neural network that uses convolution in one or more of its layers (Goodfellow et al., 2016). In the context of neural networks, convolution refers to a mathematical operation performed on the input data using a kernel, or filter, and gives a feature map as output (Goodfellow et al., 2016). These feature maps can be seen as higher level abstraction of the input (Sze et al., 2017) containing relevant information while reducing the size of input (Goodfellow et al., 2016). A kernel, or filter is essentially a matrix with learned weights, that is convolved with the input layer and can be seen as a feature extractor (see 2.4.2) (Wen et al., 2020b). Additionally, multiple

computation (e.g. pooling) are done to reduce dimensionality of the feature map and make the model robust and invariant to small changes in the input (Sze et al., 2017). At the end of each convolutional layer, non-linear activation functions (for example rectified linear unit (ReLU)) are used to introduce non-linearity into the relation between layers (Sze et al., 2017). Normally, fully connected (FC) layers (e.g. every unit from layer n is connected to all units from $n-1$ layer) are then applied after all convolutional layers (Sze et al., 2017). Wen et al. (2020b) point out that these layers use all information from previous layers to perform the given task of the model. For example in a binary classification task the last layer would have two units with the activated unit defining the belonging class of the input. Furthermore, they state that, using a softmax function transforms the belonging to a class into probabilities. Using small sized training data can lead to over fitting (see 2.3) in DL methods (Goodfellow et al., 2016), therefore data augmentation (e.g. generating novel data points through transforming available data of training set (Perez & Wang, 2017)), transfer learning (e.g. training model on generic images (Nanni et al., 2020)) and dropout (e.g. randomly setting the output of units to zero) are three methods of dealing with over fitting (Wen et al., 2020b). Figure 1 shows a simplified CNN with its layers.

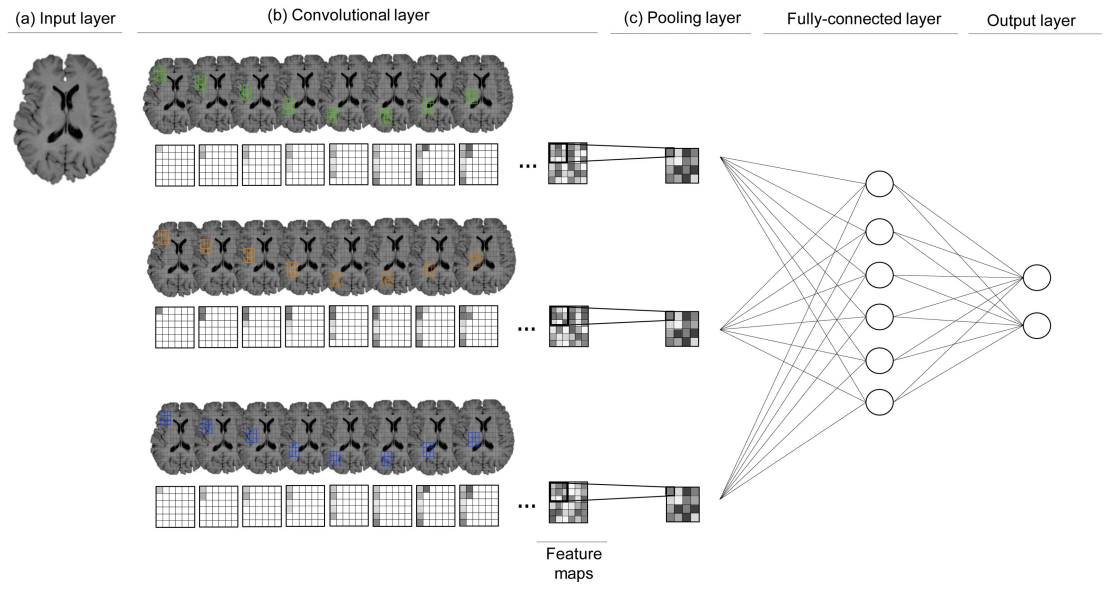

Figure 1: Simplified CNN architecture. Reprinted from "Using Deep Learning to Investigate the Neuroimaging Correlates of Psychiatric and Neurological Disorders" by Vieira et al., *Neuroscience Biobehavioral Reviews*, 74, p.62. Copyright 1969 by Elsevier.

2.3 Bias-Variance Trade-off

Generalization is a fundamental goal of ML techniques and refers to the idea of using a model that has been trained on a finite sample of observation to make accurate predictions about novel data (Mohri, 2018). To assess a models generalizability, test error, that is the prediction error a model makes on novel data, is of importance. The two sources that produce to this error rate are termed bias and variance (Yarkoni & Westfall, 2017). In Nordhausen (2014), variance is described as to what extend the estimated true function learned by a statistical learning method would differ if estimated using different training data. So, for methods that have high variance, small changes in the training set could lead to more variety in the estimated function compared to method with low variance. Bias is described as the error that can happen while approximating complex relations with a simpler model. An example being a linear regression assuming linearity between variable that are likely to be non-linear and thus having some amount of bias in its estimation. Put differently, the bias of a model is a systematic tendency for its predictions to deviate from the true value (Yarkoni & Westfall, 2017). Both variance and bias are connected and increasing one would decrease the other and therefore consideration have to be made on the optimum ratio of bias and variance (Nordhausen, 2014). Over fitting for example occurs with low bias and high variance meaning that the model learns irrelevant patterns from the training set that are caused by random chance (Nordhausen, 2014). Under fitting, on the other hand reflects model with high bias (Goodfellow et al., 2016). Typically, methods that are more flexible have less bias but higher variance (Nordhausen, 2014). In ML choosing a ratio that minimizes the expected prediction error seems the clear goal if data sets are large, assessment of a model's performance can be made objectively and if the used model allows for some control over bias-variance trade-off (Yarkoni & Westfall, 2017).

2.4 Training Data

Figure 2 shows the different steps involved in supervised machine learning. It follows a description of pre-processing step 2.4.1 and feature selection 2.4.2. CNN combines feature selection, feature vectors and classifier, whereas SVM is just a classifier.

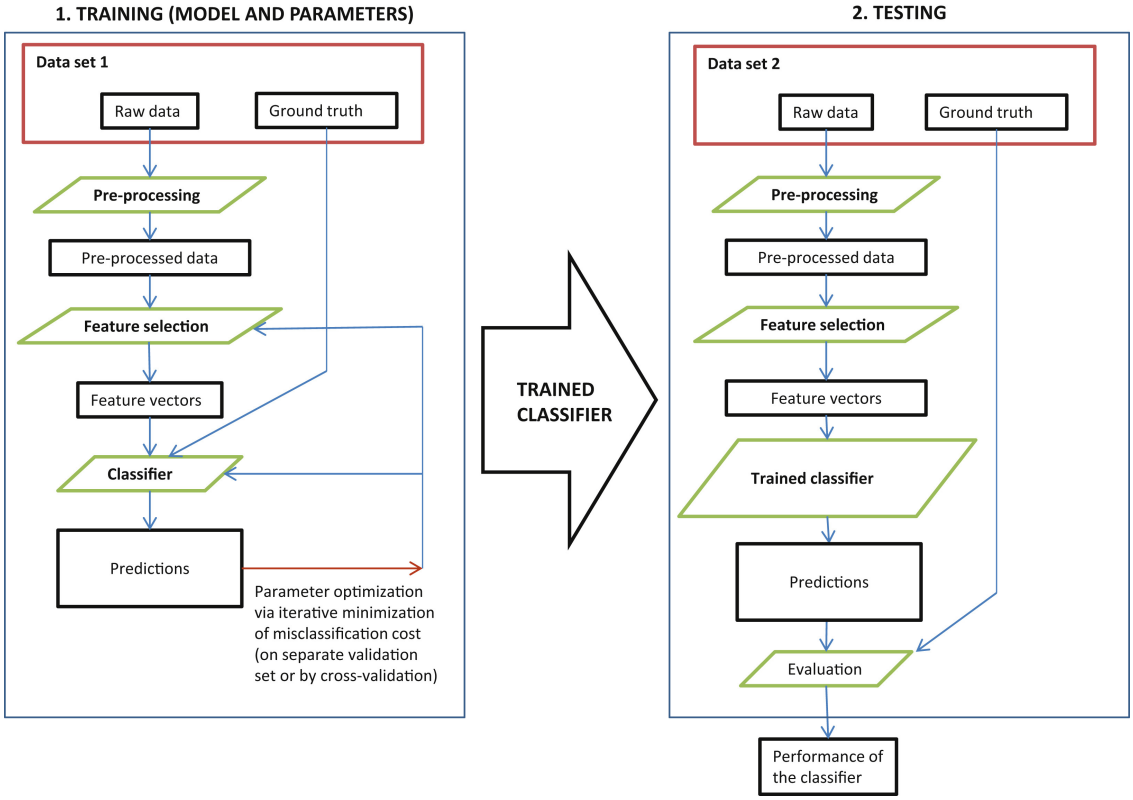

Figure 2: Classification workflow for supervised learning. Reprinted from "Machine Learning of Neuroimaging for Assisted Diagnosis of Cognitive Impairment and Dementia" by Pellegrini et al.,*Alzheimers Dement*, 10, p.520. Copyright 2018 by the authors.

2.4.1 Pre-processing

A core problem with data is that it can be noisy (e.g. containing random and irrelevant information) (Skansi, 2018). Pre-processing can help with reducing noise in data and is a important step in increasing the quality of input data (Noor et al., 2020) with the hope of making it easier for the statistical learning method to find relevant patterns (Bishop, 2006). A review by Ebrahimighahnavieh et al. (2020) states that Intensity Normalization (IN), Image Registration (IR) and Skull Stripping (SST) are among the most often used pre-processing techniques when using MRI-images for AD classification. IN is a normalization technique that helps with comparison between MRI scans and does so by reducing the voxel or pixel intensity variation which can be caused if different scanners are used (as cited in Noor et al., 2020). IR is a second technique that helps with aligning the image scan to the anatomic brain structure so that a voxel position contains the same anatomical structure while comparing MRI-images of different people (Ebrahimighahnavieh et al., 2020). SST is done to remove the skull from MRI images (Ebrahimighahnavieh et al., 2020).

2.4.2 Input Data Management

In a review Ebrahimighahnavieh et al. (2020) found that voxel-based, slice-based, patch-based and Region of Interest (ROI) based approaches are used to manage the input data. The authors describe voxel-based as a method that uses the intensity values of every voxel in a given MRI scan as input, therefore using information of the full brain image. This approach, also sometimes referred to as subject-level approach has the advantage of including spatial information of the whole image but has the drawback of needing to optimize many parameters (Wen et al., 2020a). Reducing feature dimensionality can help in this regards. For example, tissue segmentation can be done therefore only including the values of voxels belonging to a tissue component (e.g. grey or white matter) (Ebrahimighahnavieh et al., 2020). Another approach is the slice-based that extracts a 2D image (i.e. a slice) from of the 3D MRI image. A further discussed approach is called patch-based and reflects the idea of using three-dimensional cubes (i.e. a patches) to form a subset of voxels of the MRI image and are then used for feature extraction. Ebrahimighahnavieh et al. (2020) point out that a challenge is to choose patches that contain all disease relevant information. ROI-based approach, on the other hand, uses prior knowledge about brain structures that are affected by the disease and a brain atlas to group voxels into relevant anatomical regions which are then used as input (Ebrahimighahnavieh et al., 2020).

A further technique that some studies (e.g. (Akramifard et al., 2020; Syaifullah et al., 2021)) make use of is Voxel-based morphometry (VBM) and can be seen as a voxel-based approach. Mechelli et al. (2005) refer to VBM as a mass uni-variate approach that performs voxel-based comparison between images to find significant differences. Normalisation, segmentation of white and gray matter and other pre-processing steps are done to help standardizing the images. Then, statistical analysis (e.g. t-tests and F-tests) find voxel-based differences. This approach can be used to pre-process the data or to find ROI (see Akramifard et al., 2020; Syaifullah et al., 2021).

2.4.3 Database

Alzheimer's Disease Neuroimaging Initiative (ADNI) is a study in which *200* elderly Normal Cognition (NC), *400* with Mild Cognitive Impairment (MCI) and *200* with AD were followed three years while collecting multiple data such as , clinical/psychometric assessments, urine serum, PET. The described aims of the study were to make the information available for the scientific community, standardizing imaging in longitudinal studies, finding optimal methods of analysing images and validating biomarkers (Jack Jr et al., 2008). Weiner et al. (2013) state that since the start of ADNI in North America in 2004, different studies worldwide have been done using the ADNI protocols. For example, Australien Imaging, Biomarkers and Lifestyle (AIBL) was a study done in Australia

with a cohort of *177* NC, *57* MCI and *53* AD subjects. Further, ADNI studies have been done to extend the original cohort (Weiner et al., 2013) and is now in its third instalment ADNI-3 (Weiner et al., 2017). Other online databases that contain neuroimaging modalities are: Open Access Series of Imaging Studies (OASIS) (Kurdi et al., 2017) and Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) (Malone et al., 2013).

2.5 Model Assessment and Model Selection

Model assessment refers to the evaluation of the model's performance (Nordhausen, 2014). Accuracy, specificity and sensitivity are metrics that are used to analyse the performance of a ML classification model (Noor et al., 2020). Accuracy is the most commonly used measurement to test the performance and consists of the fraction of correctly predicted classes in a test set (Pereira et al., 2009). Sensitivity and specificity as further measurements, are described by Noor et al. (2020), as the percentage of true positives that have been predicted as such (sensitivity) and true negative cases that have been predicted correctly (specificity). The term *model selection*, on the other hand, is used for the process of selection a model's right flexibility (Nordhausen, 2014). As mentioned in 2.3, choosing the right flexibility (e.g bias-variance trade-off) is important for ML techniques. Cross-Validation (CV), as describe in Nordhausen (2014), is a method that is used to help with model assessment and model selection. It does so by splitting the training data into a training set and a testing set whereas the model is fitted with the training set and afterwards its performance is evaluated on the testing set. With classification models the accuracy metric is used for evaluation. Furthermore Leave-One-Out Cross-Validation (LOOCV) is described as a variation of CV where a single observation of the training data is left out for validation and the rest are used to train the model. This process is repeated as many times as there are observations in the training data (e.g. n) while differing the observation that is used for validation. Essentially, LOOCV produces n different models that are fitted and validated with different training and testing sets. Averaging the n accuracy rates that LOOCV produce leads to a better estimation of the true error rate the model would have while confronted with novel data. Nordhausen (2014) name k-fold Cross-Validation (k-fold CV) as second variation of CV. K-fold CV randomly divides the training data into k groups so that one group can be used for validation while k-*1* groups train the model. This process is done k times while changing the validation group. Finally, averaging the accuracy rates which results in an estimator for the true error rate. As Nordhausen (2014) point out if k would be set to n, than k-fold CV would be the same as LOOCV.

3 Method

3.1 Archives and Search Queries

PubMed, *Science direct*, *PsycInfo* were used as search archives. An overall scope of the topic was made with search terms "machine learning" <OR> "deep learning" <AND> "mri" <OR> "neuroimaging" <AND> "alzheimer" <AND> "review" <OR> "systematic". After this process the focus of this review was shifted toward the two model SVM and CNN. Therefore, two different groups of search terms were applied to the search engine. The first group of search terms was used to find relevant studies that used SVM model. Search terms consisted of "svm" <AND> "mri" <AND> "alzheimer". The second group of search terms was used to find papers focusing on CNN and consisted of "cnn" <AND> "mri" <AND> "alzheimer". The search was applied to title and abstract section of the paper. Furthermore, the search was limited to studies in English and from year 2020 and 2021.

3.2 Study Inclusion Criteria

Applying the first group of search terms yielded 14 results on *PubMed*. Studies using fMRI or focusing heavily on MCI, or not focus on AD classification were excluded. Results were compared to the same search queries in *Science direct* and *PsycInfo*, finally, arriving at five studies using SVM. Using the second group of search terms resulted in 30 papers on *PubMed*. After cross referencing with other search archives, 17 papers passed inclusion criteria. Since not all 17 studies could be reviewed, further inclusion considerations were made. These considerations were: the study should only use MRI as modality and have a sample size higher than 400 . The sample size is based on reflections made by (Ebrahimighahnavieh et al., 2020) that used sample size to rate the quality of their included studies. Finally, five studies were chosen for further review. Therefore, five studies using SVM and five studies using CNN amounting to a total of 10 studies were included in this review.

4 Results

4.1 Studies using SVM

In a study by Syaifullah et al. (2021) the data set was split into a training set (723 ADNI), validation set (723 ADNI) and four independent testing sets (519 AIBL, 592 Japan , 69 MIRIAD, 128 OASIS). MCI subjects from ADNI were split into stable and progressive MCI. Using the longitudinal information from the ADNI study, the subjects that didn't progress to an AD diagnosis during four or more years were labeled stable MCI. MCI subjects that met criteria given by ADNI (Petersen et al., 2010) for AD diagnosis within the duration of the ADNI study were labeled progressive MCI. To train the model progressive MCI was considered as belonging to the AD continuum and stable MCI not belonging to the continuum. Therefore, the final training set was 321 AD continuum (179 AD, 142 pMCI) and 335 non-AD continuum (271 NC, 64 sMCI). Fine tuning was done on the rest of the set (322 AD continuum, 336 non AD continuum).

VBM was used to normalize the data in a pre-processing step. To help with dimensionality multiple ROI instead of voxel-based were used as input vectors. A nonlinear kernel, radial basis function (RBF), was used to find the optimal hyperplane. LOOCV was applied to estimate the true error rate and reduce over fitting. Classifying observation was done by calculating the distance from the hyperplane which lead to a likelihood score of having AD. A sigmoid function squashed the score into the range (0,1) and the threshold for the classification was set to 0,5 Tested independently on four different databases yielded accuracy ranging from *88% - 94,2%*, sensitivity ranging from *85,1% - 97,8%* and specificity *88,4 % - 90,8%*. For further comparison, ratings of two neuroradiologists as well as a second SVM trained with MMSE as additional input were used. Assisted by an automated technique to help with MRI image interpretations (i.e. Voxel-based Specific Regional Analysis for Alzheimer's Disease (VSRAD)), the highest metrics achieved by the neuroradiologists were: accuracy of *73%*, sensitivity of *72%* and specificity of *79%*. On the other hand, the SVM trained with additional MMSE scores outperformed both the neuroradiologists and the SVM trained with MRI images only and achieved accuracy ranging from *92,5% - 100%*, sensitivity ranging from *93,1% - 100%* and specificity *92,4% - 100%*. The SVM with MMSE scores performed especially well on the MIRIAD database with accuracy of *100%*, sensitivity of *100%* and specificity of *100%*. Performance measures achieved by MMSE scores on its own weren't analysed in their study.

Akramifard et al. (2020) used a first of its kind method to increasing the accuracy of a trained SVM classifier. They did so by repeating the most important feature sets in the input vector therefore emphasizing them. The authors call this method emphasized learning. To train the model the data of 705 people (156 AD, 338 MCI and 211 NC) were taken from ADNI. Testing and validating the model was done on the same data set.

Pre-processing steps included realignment, smoothing, skull-stripping and spatial normalization. For feature extraction white and gray matter were segmented and VBM was used to extract ROIs. The final feature vector included 132 features derived from images and 12 other features that consisted of MMSE score, personal information, Cerebrospinal fluid (CSF) biomarkers and PET voxel values. To evaluate and test the model k-fold CV ($k = 10$) was performed. Principal Component Analysis (PCA) with 25 principle components were used to extract the most valuable features and reduce dimensionality. These features were then repeated as input vector. The optimal amount of repetitions was achieved as soon as k-fold CV values became stable with additional repetitions not increasing or decreasing accuracy. For the classification task SVM was used with a lineal kernel. The decision to use a lineal kernel was made after comparing it to a non-linear RBF that yielded no significant difference in terms of accuracy but was easier and faster to train. By reducing and repeating the data, the initial accuracy of *95.54%* was improved and achieved a maximum accuracy of *98.81%* after having the reduced data repeated five times as input. The maximum sensitivity was at *98.52%* after repeating the reduced data seven times. Specificity was the highest (*99.21%*) while using the reduced data without any repetitions. Furthermore the authors analysed the accuracy that was achieved using each group of data alone (personal information = *60%*, MMSE score = *91,9 %*, MRI data = *86,8%*, CSF biomarker = *59,4%*, PET data = *62,5%*).

With the intention of increasing generalizability of ML models, Li et al. (2020) used an age correction approach to help with classification. Age correction was performed using Human Connectome Project Lifespan-Aging (HCP-A) data set, which consist of 272 healthy adults (146 women, 126 men with age *62.7 +/- 16.8*). For testing and training ADNI ($n = 400$) with 136 AD subjects and 268 healthy control were used. Further testing of generalization was done with an independent dataset of 66 participants (*AD = 41, HC = 25)*. Spatial and intensity normalization, skull stripping and segmentation of grey and white matter were applied as pre-processing steps. In total, 314 features were extracted from MRI images (296 from brain surface, 16 ROIs, Estimated Total Intracranial Volume (eTIV), White Matter Signal Abnormalities (WMSA)). For each feature, a linear regression with its cortical thickness (or volume for , eTIV, WMSA) and age was performed on the HCP-A data set. Yielding *314* different linear regression models, *303* models were kept since features used in those model were significantly ($p < 0.05$) correlated with age. Inputting age of individuals from the training data set ADNI into these linear regression model yielded the difference in predicted volume/thickness and actual volume/thickness of the features. These calculated residuals represent the features without the linear age effect and where used as input vector for the SVM. Optimization of models parameter and estimating the models performance was done with cross-validation. To separate the data a nonlinear Gaussian kernel was used. In ADNI training set, controlling for age in-

creased accuracy from *96.29%* to *97.03%* and sensitivity from *91.91%* to *94.12%* while specificity *98.51%* had no changes. Using the third independent data set for validation revealed lower accuracy, sensitivity and specificity of *84.85%*, *85.36%* and *84%*. The authors then trained a second SVM model using only individuals that were amyloid-β positive (for AD group) and amyloid-β negative (for healthy control group). This second model that was trained with 300 observation (AD = 119, HC = 181) displayed accuracy of *84.85%*, sensitivity of *95.12%* and specificity of *68%* while tested on the independent testing set. Furthermore, analysing the features used for training the authors concluded that volumes of hippocampus and amygdala had the highest classification power.

A study done by Khatri and Kwon (2020) trained two SVMs with 187 individuals from ADNI. After pre-processing the data and extracting cortical thickness, surface area and gray matter volume as MRI features. Then, other modalities such as the Apolipoprotein E (APOE) genotyp, CSF biomarkers and MMSE score were combined into the feature vector. Using *k*-fold CV (*k=10*) with filter and wrapper algorithms the author selected 60 most important features. These were used as input for the classifier. The authors compared 3 different classifiers: Extreme Learning Machine (ELM), a feed forward neural network with one hidden layer, SVM with linear kernal and a SVM with RBF kernal. The classification task of AD vs normal cognition was evaluated using 10-fold cross-validation. With an accuracy of *93.50%*, sensitivity of *95.5%* and specificity of *90.58%*, SVM-linear performed better than SVM-RBF (*91.33%*, *93.33%*, *87.57%*). The ELM on the other hand, outperformed both SVMs with accuracy of *97.31%*, sensitivity of *98.04%* and specificity of *96.28%*. Furthermore, the authors analysed performance measurements while using the groups of feature separately (e.g. cortical thickness, surface area, volume, CSF, APOE + MMSE) which provided much lower metrics.

Emphasizing the importance of feature selection, Richhariya et al. (2020) propose a novel method termed Universum Support Vector Machine based Recursive Feature Elimination (UVSM-RFE). This method adds the ability of using prior knowledge about the data's distribution for the feature selection process. Additionally, it attributes weights to every feature based on its importance within the classification task and eliminating features with low importance. The process of assigning weights and selecting features happens in recursive manner hence the name. Furthermore, the authors claim that since the distribution information is considered while eliminating features this can help with generalization and counteracts the greedy nature of SVM that mainly tries to maximize the margin (see 2.1.1). GM, WM and CSF were extracted from 150 (*AD = 50, MCI = 50, HC = 50*) MRI-images taken from ADNI. To create distribution information, new data points were generated by averaging random data points from the training set. The optimal parameters for feature elimination were chosen with *k*-fold CV. A linear kernel was used to separate the classes. Assessing the models performance was done with an indepen-

dently testing set consisting of 813 MRI images from ADNI (*AD = 187, MCI = 398, HC = 228*). The proposed method UVSM-RFE scored an accuracy of *89.2%*, sensitivity of *84.87%* and specificity of *93.13%* while reducing tissue features by *85%*.

4.2 Studies using CNN

A plea for standardizing the different steps involved in classification with CNN was made by Wen et al. (2020a), providing a open source framework to tackle concerns of biased evaluation, over fitting, reproducability and pre-processing problems. The framework focuses on reproducible evaluation of AD classification using DL methods. It's an extension of their previously proposed framework that focuses on conventional ML for AD classification using different modalities (e.g. PET, MRI) (Samper-González et al., 2018). Using their framework, the authors trained and compared multiple CNN model as well as a linear SVM. The assessed models were trained and validated with ADNI (*n = 1255*) while varying the extend of the pre-processing techniques. Thereby, the author found that intensity rescaling was a crucial pre-processing step and without it the 3D CNN dropped accuracy (from *80%* to *50%*) measured on the validation set. Other pre-processing steps (e.g skull-stripping, non-linear registration) had a small impact on accuracy. Therefore, the authors performed minimal pre-processing (e.g. bias field correction, inentisty rescaling and linear registration) in all other CNN models. Three separate test were done to asses performance using OASIS (*n = 154*), AIBL (*n = 598*) and a randomly selected subset of ADNI (*n = 200*) that was split before training the model. The highest accuracy (*89%*), measured as average from a *k*-fold CV (*k=5*) from the ADNI test set, was achieved by the 3D ROI-based CNN. Linear SVM had the highest accuracy (*88%*) on the AIBL test set, followed by the 3D subject-level CNN with an accuracy of *86%*. Finally, 3D ROI-based CNN yielded the highest accuracy of *73%* tested on the OASIS test set. Sensitivity and specificity measurements weren't reported in their paper but are downloadable (Wen et al., 2019).

Yee et al. (2021) trained a 3D subject-level CNN which was tested for generalizability on four independent databases (subset of ADNI, AIBL, OASIS and MIRIAD) amounting to *7.902* images. Considering baseline and longitudinal information, images of subjects in the databases (except for MIRIAD) were categorized into seven groups with the aim of representing the continuum of AD and differentiating clinical manifestation from patho-physiological brain changes (see table 1 for subgroup descriptions). Three of those groups can be labelled as not related to AD continuum since the subjects of those image didn't receive a AD diagnosis during any follow up screenings. Images from the other four groups can be labelled as AD related because the subjects progressed to AD or already joined the cohort with the diagnosis.

To train the model, they only included baseline and longitudinal images of stable NC (*423*) and stable Dementia of the Alzheimer's type (DAT) (*330*) groups from ADNI. Furthermore, data augmentation was applied to help with over fitting. Validating the model was done with 5-fold cross-validation on the remaining ADNI subjects that weren't used for training. Pre-processing steps, including centering and reslicing the MRI images as well as intensity standardizing were applied to images from all databases with the goal of standardizing the images. The neural network itself consisted of 9 convulutional blocks, every block having a convulutional layer with instance normalization layer and an activation layer (ReLU). Pooling layer and a softmax layer was used as classification layers. The classification threshold was set to *0.5* meaning scores above *0.5* were classified as AD. Metric obtained from cross-validating the training set were: accuracy of *88.1%*, sensitivity of *88.3%* and specificity of *88.1%*. The binary classification task of images (sDAT and sNC) for AIBL, OASIS and MIRIAD yielded accuracy's of *90.7%*, *91.9%* and *95.7%*. In table 2 performance metrics using all subgroup are provided. Additionally, the authors used visualization techniques to compare whether the patterns used by the CNN models were similar to brain structures related to AD. This technique revealed that thalamus, hippocampus and ventricles are key areas the network used for the classification.

Hippocampus has often been studied in relation to AD since it belongs to the first brain regions damaged by the pathological changes that come with AD (Liu et al., 2020). Emphasizing the methodological problems while studying this ROI (e.g. accurate segmentation), Liu et al. (2020) built a DL framework with two CNN models. The first 3D CNN model automatically segments the hippocampus and solves the classification task. A second 3D DenseNet model, a type of CNN, uses the 3D patch generated by the first model to perform its own classification. At the end of both models, a FC layer as well as a softmax layer is applied and performs a final classification combining all relevant learned features. The data set consisted of 449 randomly selected baseline MRI images of ADNI subjects (*96* AD, *233* MCI and *119* NC). SST, IN and IR were used as pre-processing steps. 3D patches surrounding the hippocampus were used as input. Since softmax function deliver probability scores the threshold for the classification was set to *0.5*. To address the over fitting problem, dropout layer were applied. Testing and training was done with k-fold CV ($k = 5$). For performance measures *10%* of the training data was withhold as testing set that wasn't used for training the model. The first model achieved accuracy of *80.1%*, sensitivity of *79.9%* and specificity of *80.3%*. The 3D DenseNet model outperformed the first model with an accuracy of *86.6%*, sensitivity of *79.4%* and specificity of *92.4%*. The proposed framework which includes both CNN models achieved the best performance: *88.9%*, *86.6%* and *90.8%*.

Another study focusing on the hippocampus was done by Wang et al. (2021). The authors extracted the ROI of *933* subjects (*326* AD and *607* CN) and excluded nearly

5% because of problematic segmentation. IN was done as pre-processing step and data augmentation lead to a *6* times larger training set (*5222 subjects*) that was used to train a DenseCNN model which is a lightweight DL model with a total of *243'090* parameters. Dropout layer were included to handle over fitting. Final layers were a FC and softmax layer. Using a k-fold CV ($k = 5$) average performance was: accuracy of *89.1%*, sensitivity of *98.5%* and specificity of *85.2%*.

Nanni et al. (2020) compared three different models: SVM, 3D CNN trained from scratch and an ensemble of five transfer learned 2D CNN. *773* subjects (AD = *137*, NC = *162* and MCI = *474*) were obtained from the ADNI database. Multiple pre-processing steps were applied including SST, IR and segmentation. Whole brain MRI volume was used for all models. The authors provide a method for using 3D MRI volume as input for 2D CNN model while preserving spatial structural information between subjects that normally gets compromised when the input is slice based. The 3D CNN model that was trained from scratch performed badly, which was why the authors didn't consider it for further analysis. The other two model achieved accuracy of *93.2%* (SVM) and *90.2%* (2D-CNN) classifying AD vs. NC.

5 Discussion

Multiple studies have been published in the past reviewing the state of DL and ML for classifying AD (see Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Noor et al., 2020; Tanveer et al., 2020; Wen et al., 2020a). Most of these reviews focus on DL methods, which may be explained by the increasing trend to use DL model for detection of AD (Ebrahimighahnavieh et al., 2020), as well as its ability to achieve high accuracy on high-dimensional and complex data (LeCun et al., 2015; Tanveer et al., 2020). Furthermore, as cited in LeCun et al. (2015), DL models outperformed other models in multiple areas such as image and speech recognition. Especially CNN models are shown to have good performance in the medical field (Ebrahimighahnavieh et al., 2020). SVM on the other hand, are still often used for AD classification because of their robustness, interpretability (Tanveer et al., 2020), are easier to validate and can be trained on a smaller data set (Syaifullah et al., 2021). Considering the *no free lunch theorem*, which states that there is no best model that is superior to other models independently of data set used (Goodfellow et al., 2016; Nordhausen, 2014), this review focuses on both approaches: SVM and CNN. Only studies published in the year 2020 and 2021 were considered for this review since most earlier studies have already been reviewed (see Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Noor et al., 2020; Tanveer et al., 2020; Wen et al., 2020a). A total of 10 studies, with five using SVM and five using CNN, were included in this paper. Tables 4 and 5 provide an overview of the studies included.

5.1 Studies using SVM

An extensive review by Tanveer et al. (2020) found that out of 60 studies using SVM the most frequently applied kernel were: linear (*43%*) and RBF (*32%*). LOOCV (*43%*) and 10 fold cross-validation (*30%*) were the most often used validation methods. They reported accuracy for the binary classification task varying between 72.5 and 100% with most study achieving 90% or higher.

In this review, three studies used a linear kernel while only one study used a RBF kernel. Two of the studies that applied linear kernel also analysed their models using RBF as kernel and found worse (Khatri & Kwon, 2020) or equal (Akramifard et al., 2020) performance. 10-fold cross validation was the most often used (3 out of 5 studies) validation method in this review and LOOCV was only used by one study (Syaifullah et al., 2021). All reviewed studies used the ADNI data set for training and only two studies tested their model on different databases; Syaifullah et al. (2021) used four different databases (AIBL, Japan ADNI, MIRIAD, OASIS) and Liu et al. (2020) used an own data set consisting of subjects scanned at Ruijin Hospital in China. The authors of that study claim that this data set provides additional cross-racial validation since their model was trained

using ADNI where the subjects are from North America (Weiner et al., 2013). Although Richhariya et al. (2020) used an independent data set for testing their model it was still the same database ADNI. Since all MRI images from ADNI are made under the same protocol (Jack Jr et al., 2008), one can argue to what degree testing the model on the same database would reflect clinical practice.

Three studies used additional modalities as feature input for their model. Akramifard et al. (2020) used personal information (for example age and gender), MMSE, CSF and PET data. A further analysis made by the author showed that their model using MMSE score alone could classify AD and NC with an accuracy of *91.9%* while MRI data on its own achieved an accuracy of *86.6%*. A possible explanation for MMSE scores achieving high accuracy could stem from the inclusion criteria for AD subjects in the ADNI study. One criterion being a MMSE scores between 20 and 26 (inclusive), where as the MMSE inclusion criterion of NC subjects in the ADNI study is a score between 24 and 30 (inclusive) (Petersen et al., 2010). Therefore, MMSE scores can distinguish AD from NC with high accuracy. Furthermore, a possible explanation for why the accuracy isn't *100%* could stem from the fact that in the ADNI study, inclusion criterion for MMSE scores of AD and NC overlap. It is likely that this was the case in the study by Akramifard et al. (2020), since they report an average MMSE score of *23.32* in their AD subjects and *27.05* in NC subjects.

The fact that all studies reviewed used ROI as input reflects SVMs limitation of performing badly on raw data and the need for pre-processing and feature extraction (Vieira et al., 2017). The overall performance metrics of the reviewed studies are very high with lowest accuracy of *84.85%* and highest *100%*. The studies that used independent test set all showed over fitting since their performance decreased. One exception was Syaifullah et al. (2021). While testing on the independent data base MIRIAD, the authors found an increase in accuracy (*93.30%* to *94.20%*) and sensitivity (*93.30%* to *97.80%*) but specificity dropped from *93.40%* to *87.00%*. Finally, the absence of an independent test set in Akramifard et al. (2020) may have led to data leakage as well as making it nontransparent to what degree their model is over fitted.

5.2 Studies using CNN

In the past years, multiple reviews on AD classification using DL have been published (see Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Noor et al., 2020; Tanveer et al., 2020; Wen et al., 2020a). They all concluded that CNN are the most widely used model for this classification task with T1-weighted MRI being the most used modality (Tanveer et al., 2020). Although CNN models can perform well on raw data (LeCun et al., 2015) and is seen as an advantage over SVM models (Jo et al., 2019), pre-processing steps and input

management are widely used with CNN models (Ebrahimighahnavieh et al., 2020). In this review all studies included some sort of pre-processing. Wen et al. (2020a) varied the amount of pre-processing and analysed its effect on performance showing that Intensity Normalization is a crucial step in CNN. Ebrahimighahnavieh et al. (2020) finding that IN is the most often used pre-processing technique reflects this importancy. IN is needed since MRI scanners can cause variations in voxel intensity (Noor et al., 2020).

It has been shown that 2D-CNN is widely used DL model and was found to be implemented more often than 3D-CNN in a review by Ebrahimighahnavieh et al. (2020). The only study using 2D-CNN was Nanni et al. (2020). The authors claimed that transfer learning is easier on 2D-CNN since generic images used for training are mostly 2D. Nevertheless, by providing a method, whole brain scans could be used for their 2D-CNN which yielded an accuracy of *90.20%*. All other studies used 3D-CNN models with ROI (e.g. Liu et al., 2020; Wang et al., 2021; Wen et al., 2020a) or voxel as input (e.g. Nanni et al., 2020; Wen et al., 2020a; Yee et al., 2021). Overall, the CNN studies reviewed used larger training data sets compared to SVM studies. Since training set size is shown to influence performance of the model (Wang, 2019, as cited in Ebrahimighahnavieh et al., 2020) having larger training set can be beneficial and this is thought to be especially true for DL methods (Vieira et al., 2017). Researching the importance of training data size, Wen et al. (2020a), varied the size of training data (e.g using only baseline vs. longitudinal images) but no systematic changes in accuracy was found. The sufficient number is still a question of research (Nanni et al., 2020), a vague estimation is 1000 per class but depends on model complexity (Moradi, 2015, as cited in Nanni et al., 2020). None of the reviewed studies had that much data. Wen et al. (2020a) had the largest training set consisting of 1255 subjects from ADNI database. Techniques such as data augmentation and transfer learning can help with simulating larger training set, as well as reducing over fitting. Wang et al. (2021) applied data augmentation which increased their original training set ($n = 933$) six times ($n = 5222$). A second study to use data augmentation was Yee et al. (2021). Transfer learning as other second technique was used by Wen et al. (2020a) and Nanni et al. (2020). Four studies used 5-fold CV for validating their model. The accuracy of the reviewed studies varied between *73.00%* and *95.70%* with most of them being lower than *90.00%*. This result is not in line with previous reviews where most studies performed over *90.00%* on the classification task (Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Tanveer et al., 2020). Even though model comparison is problematic (Bron et al., 2015), a possible explanation is provided by Wen et al. (2020a). In their literature review they conclude that half of their included studies may include data leakage. This form of bias that inflates accuracy is caused by using test data in the training part. The authors provide four main reason where data leakage can happen: "Wrong data split" (e.g. not splitting at subject level when longitudinal is provided), "Late split" (e.g.

feature selection, data augmentation after split), "Biased transfer learning" (e.g. overlap of source and target area) and "Absence of an independent test set" (Wen et al., 2020a, p. 4). In this review Wang et al. (2021) and Nanni et al. (2020) didn't use an independent test set which may open up the possibility of data leakage. The goal standard seems to be splitting the data set into three independent data sets: training, validating and testing set (Nordhausen, 2014; Wen et al., 2020a). Wen et al. (2020a) was the only study splitting the data set into three independent parts where as Liu et al. (2020) and Yee et al. (2021) only used independent test set but validation was done on trianing set using cross-validation. Furthermore, Yee et al. (2021) state that cross-validation was done at subject level to prevent data leakage since the authors were using longitudinal data. Like the studies using SVM, all CNNs were trained using ADNI database with independent testing on different databases only done by two studies (e.g. Wen et al., 2020a; Yee et al., 2021).

5.3 Generalizability

As Mohri (2018) point out generalizability is fundamental to ML and even though cross-validation can be helpful estimating generalizability especially when data is sparse (Nordhausen, 2014), testing on an independent database seems more appropriate in mimicking clinical practice where MRI protocols can vary without the model able to adapt (Bron et al., 2015). Studies like Li et al. (2020), that tested on independent data base, dropped SVMs accuracy by 12 percentage points (*97.03%* to *84.85%*), which indicates the problem of relying only on accuracy measures obtained from training data. Their testing set consisted of elderly Chinese subjects, whereas the model was trained with North American subjects. Even if the used gaussian kernel has been shown to be vulnerable to over fitting (Peng & Nagata, 2020), this decrease in performance could reflect a fundamental issue all studies share: the reliance on one database. Relying on one database can facilitate cross-comparisons between study but also overestimates accuracy (Pellegrini et al., 2018). As discussed in 2.4.3, ADNI study launched an interest in standardizing MRI data sets around the world. This is one strategy of improving generalizability. Other strategies focusing on transparent comparison between trained models are challenges such as CDDementia (Bron et al., 2015) and frameworks that make training and evaluation transparent such as Wen et al. (2020a) provided.

5.4 Training Data

Both methods included in this review are supervised methods. This means that they rely on ground truth values. In the task of classifying AD, these ground truth values are the clinical diagnosis. As seen in the introduction, this diagnosis has changed in the past and still faces limitations. Jack et al. (2018) describe the current AD diagnosis as a diagnosis

of dementia syndrome that can have multiple underlying causes with AD only being one. With a definite AD diagnosis only being possible post mortem and *10%* to *30%* clinically diagnosed with AD not displaying pathological changes post mortem, this ground truth problem is something that needs further research. Furthermore, it enhances the need to interpret the classifier's reason behind its decision. This is especially true for DL models like CNN since they are often described as "black boxes" (Syaifullah et al., 2021). Yee et al. (2021) and Liu et al. (2020) were two studies using CNN that examined the factors contribution towards the decision made by their models. Furthermore, training data used for diagnosis are often elderly subjects where age related neurodegeneration could be a confounding variable. Li et al. (2020) examined this possible problem by creating linear regression models for every selected feature so that the training subject could be compared with healthy individuals of same age. By controlling for age the authors could increase accuracy and sensitivity.

5.5 Clinical usage

The CADDementia challenge provides participants with a data set ($n = 30$) for training. The models of the participants are allowed to be trained with additional information (for example ADNI data) if desired. Since this challenge is focused on clinical usage, a small data set should mimic real world cases of little data available. Furthermore, the task for the classifier is not a binary one but a ternary one meaning the trained model should classify the subjects into AD, MCI and NC. Klöppel (2012) as cited in Bron et al. (2015) state that the potential for clinical usage of ML models are 3 fold. Firstly, they can enhance diagnosis where there is a lack of expert knowledge that is needed to interpret MRI scans and other modalities. Secondly, the speed of diagnosis can be increased. Thirdly, subjects with similar patterns can be recruited for clinical trials.

For clinical usage it should be discussed if sensitivity or specificity of a model should be high. According to Borst (2020), specificity reflects the percentage of people without the disease that are correctly excluded, or classified as not having the disease. Generally speaking, tests or models with high specificity models are good at ruling in a disease while high sensitivity are good at ruling out a disease (Borst, 2020). Focusing on high specificity could minimize unnecessary burden that false positive classification lead to (e.g. treatment, medication, costs, psychological well being).

The question still remains if simple measurements such as age, gender and MMSE scores should be included into classification models. Multiple studies in this review have shown that including different modalities, especially MMSE scores, can lead to better performance (see Akramifard et al., 2020; Khatri & Kwon, 2020). Still, a possible problem in this regard is circularity since accuracy depends on ground truths used to train

the model. These ground truths, on the other hand, are clinical diagnosis that use neuropsychological tests such as MMSE. Nevertheless, since MRI is typically recommended following clinical evaluation (Scheltens et al., 2021), having automated analysis that has shown to be able to outperform experts (e.g. Syaifullah et al., 2021) can be a clinically useful addition.

6 Conclusion

This review showed that both the SVM and CNN achieve good performance in the binary classification task (AD vs. NC). A known problem with training data (e.g. relying on one data base) is a limitation that was found in this review. Furthermore, over fitting seems to be a problem that was more present in the studies using SVM. To conclude what model is better is a difficult question that can't be answered with this review and further research needs to be done in this regard (see table 3 for their strengths and weaknesses). A big constraint holding back comparison is the fact that evaluation needs to be more transparent. Frameworks (e.g. Wen et al., 2020a) and challenges (e.g. Bron et al., 2015) need to considered in future research. A limitation of this review is the small number of included studies and especially studies using CNN are unrepresentative. On the other hand, this reflects the fact that the focus of research in the last two years research was on DL methods such as CNN. Lastly, the diagnosis of AD has its limitations and since both methods reviewed in this paper are supervised learning methods, this may need to be addressed in future research.

References

Ahmed, T. F., Ahmed, A., & Imtiaz, F. (2021). History in perspective: How Alzheimer's Disease came to be where it is? *Brain Research*, *1758*, 147342. https://doi.org/10.1016/j.brainres.2021.147342

Akramifard, H., Balafar, M., Razavi, S., & Ramli, A. R. (2020). Emphasis Learning, Features Repetition in Width Instead of Length to Improve Classification Performance: Case Study—Alzheimer's Disease Diagnosis. *Sensors*, *20*(3), 941. https://doi.org/10.3390/s20030941

Alzheimer Europe. (n.d.). *Dementia in europe yearbook 2019: Estimating the prevalence of dementia in europe*. Retrieved April 13, 2021, from https://www.alzheimer-europe.org/Publications/Dementia-in-Europe-Yearbooks

Alzheimer's Disease International. (n.d.). *World alzheimer report 2018: The state of the art of dementia research: New frontiers*. Retrieved April 13, 2021, from https://www.alzint.org/u/WorldAlzheimerReport2018.pdf

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Borst, M. J. (2020). 2 - evidence-based practice: The basic tools. In C. M. Wietlisbach (Ed.), *Cooper's fundamentals of hand therapy (third edition)* (Third Edition, pp. 15–20). Mosby. https://doi.org/https://doi.org/10.1016/B978-0-323-52479-7.00002-8

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., … Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, *111*, 562–579. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.01.048

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Ebrahimighahnavieh, M. A., Luo, S., & Chiong, R. (2020). Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, *187*, 105242. https://doi.org/10.1016/j.cmpb.2019.105242

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975).
mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/https://doi.org/10.1016/0022-3956(75)90026-6

Frisoni, G. B., Winblad, B., & O'Brien, J. T. (2011). Revised nia-aa criteria for the diagnosis of alzheimer's disease: A step forward but not yet ready for widespread clinical use. *International Psychogeriatrics*, *23*(8), 1191–1196. https://doi.org/10.1017/S1041610211001220

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*, 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Islam, J., & Zhang, Y. (2018). Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics*, *5*(2), 2. https://doi.org/10.1186/s40708-018-0080-3

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., ... Silverberg, N. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4), 685–691.

Jo, T., Nho, K., & Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, *11*. https://doi.org/10.3389/fnagi.2019.00220

Khatri, U., & Kwon, G.-R. (2020). An Efficient Combination among sMRI, CSF, Cognitive Score, and APOE 4 Biomarkers for Classification of AD and MCI Using Extreme Learning Machine. *Computational Intelligence and Neuroscience*, *2020*, e8015156. https://doi.org/10.1155/2020/8015156

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (oasis). *Behavior research methods*, *49*(2), 457–470.

Leandrou, S., Petroudi, S., Kyriacou, P. A., Reyes-Aldasoro, C. C., & Pattichis, C. S. (2018). Quantitative mri brain studies in mild cognitive impairment and alzheimer's disease: A methodological review. *IEEE Reviews in Biomedical Engineering*, *11*, 97–111. https://doi.org/10.1109/RBME.2018.2796598

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Li, B., Zhang, M., Riphagen, J., Morrison Yochim, K., Li, B., Liu, J., & Salat, D. H. (2020). Prediction of clinical and biomarker conformed Alzheimer's disease and mild cognitive impairment from multi-feature brain structural MRI using age-correction from a large independent lifespan sample. *NeuroImage: Clinical*, *28*, 102387. https://doi.org/10.1016/j.nicl.2020.102387

Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., & Xu, M. (2020). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *NeuroImage*, *208*, 116459. https://doi.org/10.1016/j.neuroimage.2019.116459

Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., & Schott, J. M. (2013). Miriad—public release of a multiple time point alzheimer's mr imaging dataset. *NeuroImage*, *70*, 33–36.

Mechelli, A., Price, C., Friston, K., & Ashburner, J. (2005). Voxel-Based Morphometry of the Human Brain: Methods and Applications. *Current Medical Imaging Reviews - CURR MED IMAGING REV*, *1*. https://doi.org/10.2174/1573405054038726

Mohri, M. (2018). *Foundations of machine learning* (Second edition). The MIT Press.

Naik, B., Mehta, A., & Shah, M. (2020). Denouements of machine learning and multi-modal diagnostic classification of Alzheimer's disease. *Visual Computing for Industry, Biomedicine, and Art*, *3*. https://doi.org/10.1186/s42492-020-00062-w

Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., & Castiglioni, I. (2020). Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.576194

Noor, M. B. T., Zenia, N. Z., Kaiser, M. S., Mamun, S. A., & Mahmud, M. (2020). Application of deep learning in detecting neurological disorders from magnetic resonance images: A survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia. *Brain Informatics*, *7*(1), 11. https://doi.org/10.1186/s40708-020-00112-2

Nordhausen, K. (2014). An introduction to statistical learning—with applications in r by gareth james, daniela witten, trevor hastie & robert tibshirani. *International Statistical Review*, *82*(1), 156–157. https://doi.org/https://doi.org/10.1111/insr.12051_19

Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., Danso, S., Muñoz-Maniega, S., Job, D., Pernet, C., Mair, G.,

MacGillivray, T. J., Trucco, E., & Wardlaw, J. M. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *10*, 519–535. https://doi.org/10.1016/j.dadm.2018.07.004

Peng, Y., & Nagata, M. H. (2020). An empirical overview of nonlinearity and overfitting in machine learning using covid-19 data. *Chaos, Solitons Fractals*, *139*, 110055. https://doi.org/https://doi.org/10.1016/j.chaos.2020.110055

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*, S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2010). Alzheimer's disease neuroimaging initiative (adni). *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Richhariya, B., Tanveer, M., & Rashid, A. H. (2020). Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE). *Biomedical Signal Processing and Control*, *59*, 101903. https://doi.org/10.1016/j.bspc.2020.101903

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., & Colliot, O. (2018). Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to mri and pet data. *NeuroImage*, *183*, 504–521. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.08.042

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*. https://doi.org/https://doi.org/10.1016/S0140-6736(20)32205-4

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning. from theory to algorithms. *Understanding Machine Learning: From Theory to Algorithms*. https://doi.org/10.1017/CBO9781107298019

Skansi, S. (2018). *Introduction to deep learning : From logical calculus to artificial intelligence*. Springer.

Syaifullah, A. H., Shiino, A., Kitahara, H., Ito, R., Ishida, M., & Tanigaki, K. (2021). Machine Learning for Diagnosis of AD and Prediction of MCI Progression From Brain MRI Using Brain Anatomical Analysis Using Diffeomorphic Deformation. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.576029

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, *105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., & Lin, C. T. (2020). Machine Learning Techniques for the Diagnosis of Alzheimer's Disease: A Review. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *16*, 1–35. https://doi.org/10.1145/3344998

Tohka, J., Moradi, E., Huttunen, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia. *Neuroinformatics*, *14*(3), 279–296. https://doi.org/10.1007/s12021-015-9292-3

Toshkhujaev, S., Lee, K. H., Choi, K. Y., Lee, J. J., Kwon, G.-R., Gupta, Y., & Lama, R. K. (2020). Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Cortical and Subcortical Features from MRI T1 Brain Images Utilizing Four Different Types of Datasets. *Journal of Healthcare Engineering*, *2020*. https://doi.org/10.1155/2020/3743171

Vapnik, V. N. (2000). *The nature of statistical learning theory* (Second ed.). Springer.

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002

Wang, Q., Li, Y., Zheng, C., & Xu, R. (2021). DenseCNN: A Densely Connected CNN Model for Alzheimer's Disease Classification Based on Hippocampus MRI Data. *AMIA Annual Symposium Proceedings*, *2020*, 1277–1286. Retrieved May 6, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075423/

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, *9*(5), e111–e194.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack Jr, C. R., Jagust, W., Morris, J. C., et al. (2017). The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia*, *13*(5), 561–571.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper, J., Routier, A., Bottani, S., Dormont, D., Burgos, N., & Colliot, O. (2020b). Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation Supplementary Material, 24.

Wen, J., Thibeau–Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., & Colliot, O. (2019). *Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation [Models]*. Zenodo. https://doi.org/10.5281/zenodo.3491003

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., & Colliot, O. (2020a). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. https://doi.org/10.1016/j.media.2020.101694

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning [PMID: 28841086]. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yee, E., Ma, D., Popuri, K., Wang, L., Beg, M. F., Initiative, a. f. t. A. D. N., & flagship study of Ageing, a. t. A. I. B. a. L. (2021). Construction of MRI-Based Alzheimer's Disease Score Based on Efficient 3D Convolutional Neural Network: Comprehensive Validation on 7,902 Images from a Multi-Center Dataset. *Journal of Alzheimer's Disease*, *79*(1), 47–58. https://doi.org/10.3233/JAD-200830

A List of Abbreviations

Symbols | **A** | **C** | **D** | **E** | **F** | **H** | **I** | **L** | **M** | **N** | **O** | **P** | **R** | **S** | **U** | **V** | **W**

Symbols

***k*-fold CV** *k*-fold Cross-Validation. 9, 12–16

A

AD Alzheimer's Disease. 1–4, 7–11, 13–18, 20–22, 32

ADNI Alzheimer's Disease Neuroimaging Initiative. 8, 9, 11–18, 32, 34, 35

AIBL Australien Imaging, Biomarkers and Lifestyle. 8, 11, 14, 15, 17, 32, 34, 35

APOE Apolipoprotein E. 13

C

CNN Convolutional Neural Network. 4–6, 10, 14–19, 21, 22, 35

CSF Cerebrospinal fluid. 12, 13, 18

CV Cross-Validation. 9

D

DAT Dementia of the Alzheimer's type. 15, 32

DL Deep Learning. 4, 5, 14, 15, 17, 18, 21, 22

DNN Deep Neural Network. 4

DPN Deep Polynomial Network. 4

E

ELM Extreme Learning Machine. 13

eTIV Estimated Total Intracranial Volume. 12, 34

F

FC fully connected. 5, 15, 16

H

HCP-A Human Connectome Project Lifespan-Aging. 12

I

IN Intensity Normalization. 7, 15, 16, 19

IR Image Registration. 7, 15, 16

L

LOOCV Leave-One-Out Cross-Validation. 9, 11, 17, 34

M

MCI Mild Cognitive Impairment. 8, 9, 11, 15, 16, 21, 32

MIRIAD Minimal Interval Resonance Imaging in Alzheimer's Disease. 9, 11, 14, 15, 17, 18, 32, 34, 35

ML Machine Learning. 2–4, 6, 9, 12, 14, 17, 20, 21

MMSE Mini-Mental State Examination. 1, 11–13, 18

MRI Structural Magnetic Resonance Imaging. 2, 3, 7, 8, 11–16, 18–21

N

NC Normal Cognition. 8, 9, 11, 15, 16, 18, 21, 22, 32

NIA-AA National Institue on Aging and Alzheimer's Association. 1

O

OASIS Open Access Series of Imaging Studies. 9, 11, 14, 15, 17, 32, 34, 35

P

PCA Principal Component Analysis. 12

PET Positron Emission Tomography. 2, 8, 12, 14, 18

R

RBF radial basis function. 11–13, 17, 34

ReLU rectified linear unit. 5

RNN Recurrent Neural Network. 4

ROI Region of Interest. 8, 11, 12, 14, 15, 18, 34, 35

S

SST Skull Stripping. 7, 15, 16

SVM Support Vector Machine. 3, 4, 6, 10–14, 16–18, 20, 22, 34

U

UVSM-RFE Universum Support Vector Machine based Recursive Feature Elimination. 13, 14, 34

V

VBM Voxel-based morphometry. 8, 11, 12

VSRAD Voxel-based Specific Regional Analysis for Alzheimer's Disease. 11

W

WMSA White Matter Signal Abnormalities. 12, 34

B Supplementary Material

Table 1: Database Stratification in (Yee et al., 2021).

MRI image subgroup	Description
stable NC (sNC)	Images of subjects with NC at all follow up screenings.
unstable NC (uNC)	Images of NC subjects before converting to MCI.
progressive NC (pNC)	Images of NC subjects before converting to AD.
stable MCI (sMCI)	Images of subjects with MCI at all follow up screenings.
progressive MCI (pMCI)	Images of MCI subjects before converting to AD.
early DAT (eDAT)	Images of MCI subjects after converting to AD.
stable DAT (sDAT)	Images of subjects with diagnosis AD at baseline.

DAT = Dementia of the Alzheimer's type, MCI = Mild Cognitive Impairment, NC = Normal Cognition, AD = Alzheimer's Disease.

Table 2: 3D Subject-level CNN Performance Evaluation in (Yee et al., 2021).

Performance	ADNI*	AIBL	OASIS	MIRIAD**
Overall accuracy	72.3%	86.6%	89.9%	94.7%
Overall specificity	68.7%	85.9%	89.9%	94.7%
Overall sensitivity	75.9%	87.6%	71.4%	96.8%

* Validation set consisting of five subgroups: uNC, pNC, sMCI, pMCI, eDAT.
** Testing set consisting of two subgroups: sNC, sDAT.

Table 3: Strengths and Weaknesses of CNN and SVM.

	Strengths	**Weaknesses**
SVM	+ global optimum is guaranteed	- heavily dependent on pre-processing steps
	+ robust	- choosing optimal kernel
	+ good performance on high-dimensional data	
	+ interpretability	
	+ training on smaller data set	
CNN	+ combines multiple steps in classification process	- prone to over fitting
	+ good for high-dimensional and complex data	- global optimum not guaranteed
	+ has shown great performance in other areas	- huge amount of training data needed for optimum results
	+ can handle raw data well	- "black boxes"

C Overview of Studies

Table 4: Overview of Studies using SVM.

Study	**N** training	**Database**	**Model**	**Feature**	**Validation**	**Acc** (%)	**Sens** (%)	**Spec** (%)
Syaifullah et al., 2021	723 (321 AD, 335 NC)	ADNI	SVM (RBF)	ROI	LOOCV	93.3	93.3	93.4
		AIBL*				89.20	94.40	90.10
		JADNI*				88.00	85.10	90.80
		MIRIAD*				94.20	97.80	87.00
		OASIS*				90.60	96.70	88.80
Akramifard et al., 2020	705 (156 AD, 338 MCI, 211 NC)	ADNI	SVM (Linear kernel)	ROI**	10-fold CV	98.81	98.52	99.21
Li et al., 2020	400 (136 AD, 268 NC)	ADNI	SVM (Gaussian kernel)	ROI, eTIV, WMSA	CV	97.03	94.12	98.51
		D3* ($n = 66$)				84.85	85.36	84.00
Khatri and Kwon, 2020	187 (136 AD, 268 NC)	ADNI	SVM (Linear kernel)	ROI**	10-fold CV	93.50	95.33	90.58
Richhariya et al., 2020	150 (50 AD, 50 MCI, 50 NC)	ADNI	UVSM-RFE (Linear kernel)	ROI**	10-fold CV	100	-	-
		ADNI* ($n = 813$)				89.20	84.87	93.13

Acc = Accurracy, Sens = Sensitivity, Spec = Specificity, * = independent dataset, ** = additionally, other modalities were used

Table 5: Overview of Studies using CNN.

Study	**N** training	**Database**	**Model**	**Feature**	**Validation**	**Acc** (%)	**Sens** (%)	**Spec** (%)
Wen et al., 2020a	1255	ADNI	3D-CNN	ROI	5-fold CV	88.00	-	-
		ADNI*(*n = 200*)	3D-CNN	ROI	5-fold CV	89.00	-	-
		OASIS*(*n = 154*)	3D-CNN	ROI	5-fold CV	73.00	-	-
		AIBL*(*n = 598*)	3D-CNN	Voxel-based	5-fold CV	86.00	-	-
Yee et al., 2021	753 (330 AD, 423 NC)	ADNI	3D-CNN	Voxel-based	5-fold CV	88.10	88.30	88.10
		AIBL*				90.70	-	-
		OASIS*				91.90	-	-
		MIRIAD*				95.70	-	-
Liu et al., 2020	449 (96 AD, 233 MCI 119 NC)	ADNI*(*n = 45*)	Two 3D CNNs	ROI	5-fold CV	88.90	86.60	90.80
Wang et al., 2021	933 (326 AD, 607 NC)	ADNI	DenseCNN	ROI	5-fold CV	89.10	98.50	82.20
Nanni et al., 2020	773 (137 AD, 474 MCI, 162 NC)	ADNI	2D-CNN	Voxel-based	CV	90.20	-	-

Acc = Accuracy, Sens = Sensitivity, Spec = Specificity, * = independent dataset

Psychologisches Institut
Bachelor- und Masterstudiengang

Universität Zürich
Psychologisches Institut
Bachelor- und Masterstudiengang
Binzmühlestrasse 14, Box 1
CH-8050 Zürich
www.psychologie.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die Bachelorarbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.uzh.ch/de/studies/teaching/plagiate.html).

Zürich, 1.06.2021 Olivia Zinga

Ort und Datum Unterschrift