Université de Montréal

Alzheimer prediction from connected speech extracts: Assessment of generalisation to new data.

par

Genevieve Chafouleas

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Discipline

September 16, 2021

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Alzheimer prediction from connected speech extracts: Assessment of generalisation to new data.

présenté par

Genevieve Chafouleas

a été évalué par un jury composé des personnes suivantes :

Marc Feeley
(président-rapporteur)

Philippe Langlais
(directeur de recherche)

Simona Brambati
(codirecteur)

Jian-Yun Nie
(membre du jury)

Résumé

Plusieurs avancées utilisant le discours obtenu de la tâche de description d'image ont été réalisées dans la détection de la maladie d'Alzheimer (AD). L'utilisation de caractéristiques linguistiques et acoustiques sélectionnées manuellement ainsi que l'utilisation de méthodologies d'apprentissage profond ont montré des résultats très prometteurs dans la classification des patients avec AD. Dans ce mémoire, nous comparons les deux méthodologies sur la scène Cookie Theft du Boston Aphasia Examination en entrainant des modèles avec des caractéristiques sélectionnées à partir des extraits textuels et audio ainsi que sur un modèle d'apprentissage profond BERT. Nos modèles sont entrainés sur l'ensemble de données ADReSS challenge plus récent et évaluées sur l'ensemble de données CCNA et *vice versa* pour mesurer la généralisation des modèles sur des exemples jamais vus dans des ensembles de données différents. Une évaluation détaillée de l'interprétabilité des modèles est effectuée pour déterminer si les modèles ont bien appris les représentations reliées à la maladie. Nous observons que les modèles ne performent pas bien lorsqu'ils sont évalués sur différents ensembles de données provenant du même domaine. Les représentations apprises des modèles entrainés sur les deux ensembles de données sont très différentes, ce qui pourrait expliquer le bas niveau de performance durant l'étape d'évaluation. Même si nous démontrons l'importance des caractéristiques linguistiques sur la classification des AD vs contrôle, nous observons que le meilleur modèle est BERT avec un niveau d'exactitude de 62.6% sur les données ADReSS challenge et 66.7% sur les données CCNA.

Mots-clés: *Traitement automatique des langues, Apprentissage machine, Maladie d'Alzheimer, Apprentissage par transfert*

Abstract

Many advances have been made in the early diagnosis of Alzheimer's Disease (AD) using connected speech elicited from a picture description task. The use of hand built linguistic and acoustic features as well as Deep Learning approaches have shown promising results in the classification of AD patients. In this research, we compare both approaches on the Cookie Theft scene from the Boston Aphasia Examination with models trained with features derived from the text and audio extracts as well as a Deep Learning approach using BERT. We train our models on the newer ADReSS challenge dataset and evaluate on the CCNA dataset and *vice versa* in order to asses the generalisation of the trained model on unseen examples from a different dataset. A thorough evaluation of the interpretability of the models is performed to see how well each of the models learn the representations related to the disease. It is observed that the models do not perform well when evaluated on a different dataset from the same domain. The selected and learned representations from the models trained on either dataset are very different and may explain the low performance in the evaluation step. While we demonstrate the importance of linguistic features in the classification of AD vs non-AD, we find the best overall model is BERT which achieves a test accuracy of 62.6% on the ADRess challenge dataset and 66.7% on the CCNA dataset.

Keywords: *Natural language processing, Machine learning , Alzheimer's disease, Transfer learning*

Contents

Résumé 5

Abstract 7

List of tables 11

List of figures 13

List of Acronyms 15

Dedication 17

Acknowledgements 19

Chapter 1. Introduction 21

1.1. Related Work 23

1.2. Data 25

1.2.1. ADReSS Challenge 25

1.2.2. CCNA 27

Chapter 2. Feature Extraction 29

2.1. Data preparation 29

2.2. Features 32

2.2.1. Linguistic features 32

2.2.2. Acoustic Features 37

Chapter 3. Methodology 39

3.1. Model selection 39

3.1.1. Feature based models 39

3.1.2. Bidirectional Encoder Representations from Transformers (BERT) 40

3.2. Feature Selection 41

3.3. Experiments and Evaluation metrics 42
Chapter 4. Results and Discussion 45
4.1. Publication replication 45
4.2. Results 47
4.3. Feature Importance and Evaluation 50
Chapter 5. Conclusion 57
References 59
Appendix A. Hyperparameter search results 63

List of tables

1.1 ADReSS challenge dataset main characteristics. 27

1.2 CCNA dataset main characteristics. 28

2.1 List of linguistic features extracted from text with their short description. 36

2.2 List of acoustic features from audio files with their short descriptions. 37

4.1 10-fold cross validation results across random seeds on ADReSS validation set using the same hyperparameter as in [**3**] with actual paper results. 46

4.2 10-fold cross validation results across random seeds on ADReSS test set using the same hyper parameter as in [**3**] with actual paper results. 46

4.3 10-fold cross validation best models results across random seeds on ADReSS validation set using Select k Best feature selection method. 47

4.4 10-fold cross validation best ADReSS models results across random seeds on ADReSS test set using Select k Best feature selection method. 48

4.5 10-fold cross validation best ADReSS models results across random seeds on CCNA test set using Select k Best feature selection method. 48

4.6 10-fold cross validation best models results across random seeds on CCNA dataset using Select k Best feature selection method. 49

4.7 10-fold cross validation best CCNA models results across random seeds evaluated on ADReSS challenge test set using Select k Best feature selection method. 49

4.8 Features selected across models and random seeds in Table 4.3 on ADReSS challenge and in Table 4.5 CCNA dataset. Features in bold are common to both settings. 50

4.9 Count of times ICU is mentioned at least once in transcript on both CCNA and ADReSS challenge dataset. 53

A.1 SVM: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection. 63

A.2 Logistic Regression: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection. 63
A.3 RBF-SVM: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection. 64
A.4 MLP: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection. 64

List of figures

1.1 Cookie Theft scene. 26

3.1 Overview of experimental pipeline for feature based models. 42

3.2 Overview of experimental pipeline for BERT model. 43

4.1 Mean valid accuracy when removing and keeping only specific feature groups on ADReSS challenge dataset. 51

4.2 Mean valid accuracy when removing and keeping only specified feature group on CCNA dataset. 52

4.3 Attention head of ADReSS models for specific layer for each random seed of the sentence "I see uh two kids up at the cookie jar, one on a stool the other standing on the floor.cupboard door is opened .mother's washing the dishes.".**Left:** random seed 42 **Middle:** random seed 52. **Right:** random seed 62. 54

4.4 Attention head of CCNA model for specific layer for each random seed of the sentence "I see uh two kids up at the cookie jar, one on a stool the other standing on the floor.cupboard door is opened .mother's washing the dishes.".**Left:** random seed 42 **Middle:** random seed 52. **Right:** random seed 62. 55

List of Acronyms

AD	Alzheimer's Disease
HC	Healthy Controls
CHC	Cognitive Healthy Controls
MLP	Multi Layer Perceptron
SVM	Support Vector Machine with Linear function
SVM-RBF	Support Vector Machine with Radial Base Function
BERT	Bidirectional Encoding Representations from Transformers
ICU	Information Content Unit
ID	Idea Density
CS	Connected Speech
MATTR	Moving-Average Type-Token Ratio

ML	Machine Learning
POS	Part-Of-Speech
TTR	Type-Token Ratio
NLP	Natural Language Processing
VAD	Voice Activity Detection
RFE	Recursive Feature Elimination
MFCC	Mel-frequency cepstral coefficients

Dedication

This thesis is dedicated to my parents Dr. James Chafouleas and Dr. Lisette Lagacé who inspired me to pursue academic research and to everyone who believed in my future.

Acknowledgements

I have received multiple sources of support during my Master's degree and would firstly like to thank UNIQUE for granting me the *Programme de bourse d'excellence UNIQUE*. Secondly, I would like to thank the *Département d'informatique et de recherche opérationelle* (DIRO) for granting me both the *Bourse d'excellence du DIRO* and *Bourse d'étude dans le domaine de l'intelligence artificielle de la Faculté arts et sciences*. These scholarships have provided me with the opportunity to fully focus my attention and time on my Master's thesis and studies.

I would also like to thank co-director Dr. Simona Brambati for the opportunity to work in an exciting and very motivating scientific environment and the guidance and support she gave me but also for the extensive knowledge on the disease she taught me which helped greatly in the understanding of the data and the difficulty of the problem at hand. I would like to thank Antoine Slegers for all the time and advice he gave me on my project and great conversations on the subject. Finally, I would also like to thank my director Dr. Philippe Langlais for the continued guidance and support he gave me to achieve this thesis with success. His expertise in the domain NLP and AI helped me advance my skills and understanding of the application of all the techniques. Thank you everyone again, you have all helped me grow and develop my critical mind.

Chapter 1

Introduction

Alzheimer disease (AD) is the most common form of Dementia. It is a neurodegenerative disease that destroys brain cells causing irreversible language and memory loss over time [**20**]. It mostly affects people in their 60's and 70's but in a small percent it can affect much younger people in their 30's and 40's. According to the 2020 annual report by the Canadian Public Health Agency, more than half a million Canadians are affected with Dementia with two thirds being women [**21**]. The annual cost associated with treatment for Dementia is expected to be $16.6 billion by 2031 [**21**].

Currently, the most common diagnostic tool is based on the results to a battery of neuropsychological tests assessing cognition. One of the most common neuropsychological diagnostic test to asses speech production used is the picture description task. This task requires a patient to describe everything they see in a specific image to extract connected speech (CS). The analysis of CS has shown that specific linguistic biomarkers can be apparent between Healthy Controls (HC) and AD patients as shown by Slegers et al [**28**]. The use of modern technology such as Machine Learning and Natural Language Processing (NLP) has been used to extract and help identify those specific biomarkers characterising the disease. The addition of Machine Learning in the diagnosis can not only help identify AD patients faster and earlier but can also help identify new linguistic patterns characterising AD. These Machine Learning approaches use hand built feature sets created from prior knowledge of how the disease affects changes in the syntax and content of speech. It has been shown that these features perform quite well when fed into Machine Learning models, but some of these features are very specific to the type of image used during the test. This causes the model to be very domain specific which has led to the use of Deep Learning approaches which helps alleviate this by learning the representations directly from the text derived from the CS.

Currently, the state of the art for language models uses very large Deep Learning models with millions of parameters on the most common NLP tasks [**26**]. For these models to generalize well, a very large number of examples are needed to avoid overfitting. Since, the

datasets used in AD prediction are extremely small, with only a few hundred examples, we are faced with an issue that makes the application of such models slightly more challenging. This is a common problem in many areas of medical research, as it is quite difficult and time consuming to collect data. This is where a technique called Transfer Learning could be beneficial, which consists in training a model on a very large dataset typically not necessarily related to the domain of the targeted tasks, then fine tuning it using the target dataset, often much smaller than the training dataset [**32**]. BERT is an example of such a model and has been shown to generalise quite well on the most common NLP tasks [**8**]. This could help resolve the ability to use Deep Learning to classify the small datasets without causing the model to potentially overfit.

The majority of research in the past years has been performed on the Dementia Bank corpus for the binary classification of AD vs HC [**29**]. One inconvenience to publishing with this dataset is that it does not contain a test set for standardisation of the publication of results. This has raised the question of the reproductibility and comparability of the published results. Also, a lack of standardization of the use of the data across publications brings to question the comparability of the results. In 2020, the ADReSS challenge dataset resolved the issue of lack of standardization by creating a standardized version of the Dementia Bank corpus which includes a clear training and test set [**17**].

Given the sparsity of the data, the training set of the ADReSS challenge dataset is extremely small which raises the question of whether we can expect the model to generalize well on CS extracted from a different dataset from the same domain. This is an important question as this would demonstrate the true power of the models proposed in a real world application.

Balagopalan et al [**3**] trains classical machine learning models with the most common feature set in AD prediction compared to a model trained with BERT on the ADReSS challenge dataset. This created an excellent baseline to use to compare results on this new dataset. We thus propose to replicate the "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" [**3**] paper in order to validate how well the new ADReSS challenge dataset helps with standardization in this field.

In this thesis we also propose to resolve this question by training on the ADReSS challenge dataset and evaluating them on both the test set from the ADReSS challenge and the Canadian Consortium on Neurodegeneration in Aging (CCNA) datasets, which includes data from AD and Subjective Cognitive Impairment, meaning that they have noticed potential cognitive changes but the tests show that cognition is normal (i.e: healthy) [**2**], and vice versa as cross dataset evaluation step. The ADReSS challenge dataset refers to their HC as non-AD and we will refer to both groups as Cognitive Healthy Controls (CHC). Both

the ADReSS challenge and CCNA datasets have been extracted by two completely different research groups.

The main objectives of this thesis are firstly to reproduce the paper from Balagopalan et al [**3**] in order to evaluate if the ADReSS dataset proposed helps with the standardization of results in the field. The second objective is to evaluate the power of generalization of the ADReSS dataset given its small size by performing a cross data evaluation step, where we train on the ADReSS dataset and not only use the test set from the ADReSS challenge dataset for evaluation but also the CCNA dataset. The opposite will also be performed where we train on the CCNA dataset and evaluate on the ADReSS dataset. We believe this is a crucial step to determining if the models generalize well to unseen examples given such a small dataset for training and whether the trained models are transferable to new dataset of the same domain.

The use of the most common acoustic and linguistic features in the literature will be applied to Machine Learning models as well as the use of BERT to make a thorough comparison of both Machine Learning vs Deep Learning techniques.

Chapter 2 will dive deeper into the features extracted for the application of the Machine Learning models. Chapters 3 and 4 will focus on the experimental setup and the results extracted. Finally, Chapter 5 will conclude and elaborate on future directions.

1.1. Related Work

There have been many publications on the classification of HC and AD patients using their CS extracts. Many have concentrated their efforts on utilising features derived from the different biomarkers identified from the CS of affected patients, while others have chosen to utilise Deep Learning techniques to classify affected patients directly from the text derived from the CS.

Orimaye et al [**22**] used 242 AD patients and 242 HC from the Dementia Bank corpus and achieved a 74% F1-score with SVM. They extracted a set of linguistic features with the addition of features extracted from the CHAT annotations that are different symbols representing for example: pauses, repetitions, mispronunciations, etc. This makes the methodology less extendable in a real world situation as the CHAT annotations are manually derived from the audio files, therefore increasing the potential bias in the model as well as the dependency on using the specified CHAT format for extracting text from audio files.

Fraser et al [**9**] utilised 370 linguistic and acoustic features extracted from 240 AD patients and 233 HC from the Dementia Bank corpus trained with a logistic regression model. Using 10-fold cross validation they achieved a mean accuracy across all folds of 81.92% using only the top 35 selected features in each fold. They demonstrated that the models perform

relatively the same on the validation set until selecting 50 or more features out of the 370 where it drops drastically in accuracy.

Ammar et al [**4**] used 242 AD patients and 242 HC from the Dementia Bank corpus to extract linguistic features from automatically transcribed text from audio extracts. They achieved an accuracy of 79% with an SVM.

Hong et al have [**11**] trained a bidirectional LSTM with multiple attention layers on 242 HC and 257 AD from the Dementia Bank corpus. Using 10-fold cross validation they achieved a mean accuracy of 83.35%.

Firtsch et al [**10**] trained an LSTM on 255 AD patients and 244 HC from the Dementia Bank corpus. The predictions of patients was determined based on the perplexity of the picture descriptions on both a HC language model and AD language model. Using the perplexity values it allowed them to classify with an accuracy of 85.6%.

Karlekar et al [**15**] trained a CNN-LSTM on 208 AD patients and 243 HC from the Dementia Bank corpus. The transcripts of each patient was separated by utterances to generate more data samples of 11458 AD and 2904 HC. They achieved an accuracy of 84.9% without adding the POS tag data. By adding the POS-tag data to the model they achieved an accuracy of 91.1%.

Pan et al [**23**] trained a hierarchical Bi-LSTM with attention on 222 HC and 255 AD patients from the Dementia bank corpus. Using 10-fold cross validation to split training, test and development sets, they achieved 84.43% F1-score on the manual transcriptions.

Chen et al [**7**] trained a hybrid model using a CNN and a GRU with an attention mechanism on 256 AD patients and 242 HC from the Dementia Bank corpus. They also allow the training of the embedding layer which improves significantly the accuracy of the model. Using 10-fold cross validation they achieved a mean accuracy across all folds of 97.42%.

As one can see from the above related works, there is a significant discrepancy in the number of HC vs AD patients used on the same Dementia Bank dataset. Also, there is not a clear test set and the results are sometimes reported on the validation set which could cause, without purposefully doing so, the optimization of the model on the test set leading to overly optimistic results. Furthermore, the lack of a proper test set makes the results very hard to compare across methodologies as simply the random seed and number of examples has a significant effect on model training and therefore results.

Very recently in 2020, Haider et al [**17**] saw this discrepancy and elaborated a new dataset solving these problems called the ADReSS challenge dataset. This dataset removes the duplicate candidates, standardizes the age and has a clear test set with specific rules for publication. This is an excellent step in standardizing the further publications but comes with a drastic reduction in the number of samples to train on to only 108 examples and a test set of 48 examples.

Balagopalan et al [**3**] set out to use this new ADReSS challenge dataset and compare both classical Machine Learning models trained on linguistic and acoustic features to a Deep Learning approach trained on the raw text extracts. Using 10-fold cross validation to train their models, they achieved 81.3% accuracy using a RBF-SVM and 83.3% using BERT on the test set.

Furthermore, as we can see from the current related work, the test results are published on data from the same dataset. This raises the question given such a small sample size how well does the model generalize to new unseen examples extracted from a different group on the same domain.

Therefore, we propose to use the experimental set up from Balagopalan et al [**3**] and expand it by analysing the importance of the derived features and the learned attentions patterns from BERT. We will then also evaluate the best models trained on the ADReSS challenge dataset on a different test set from CCNA in order to evaluate how well the trained models generalize to unseen examples. We believe that this is a paramount step in determining the true potential of the described models in the literature and how well these could work in a real world application.

1.2. Data

In this research, we use both the ADReSS challenge dataset and the CCNA dataset [1]. Both datasets are descriptions of the Cookie Theft scene from the Boston Aphasia Examination [**25**] as shown in Figure 1.1. Both datasets contain the manual text extractions from audio and the raw audio files. They both differ in the mean age and gender distribution. They also differ in the technique used to manually transcribe the audio files into the text extracts. The ADReSS challenge Dataset utilises the CHAT protocol which adds extra annotations such as exact times of speech from both interviewer and patient and POS tags, etc [**18**] while the CCNA dataset only extracts the text verbatim from the audio files with minimal additional annotations without any information on the time of speech from both the interviewer and patient.

1.2.1. ADReSS Challenge

The ADReSS challenge dataset is a standardized version of the classical Dementia Bank dataset from Talk Bank [**17**]. Each of the audio files are manually transcribed using the CHAT protocol [**18**]. The dataset contains only AD and control extracts with an equal amount of women to men with standardized age gaps. The audio files are normalized for

[1]Many experiments and analyses were performed in the first part of this research using the Dementa Bank dataset [**29**] to replicate the current literature to analyse the inconsistency in the dataset. This thesis will not be describing these experiments as the ADReSS challenge dataset solves the inconsistency issues in the aforementioned papers.

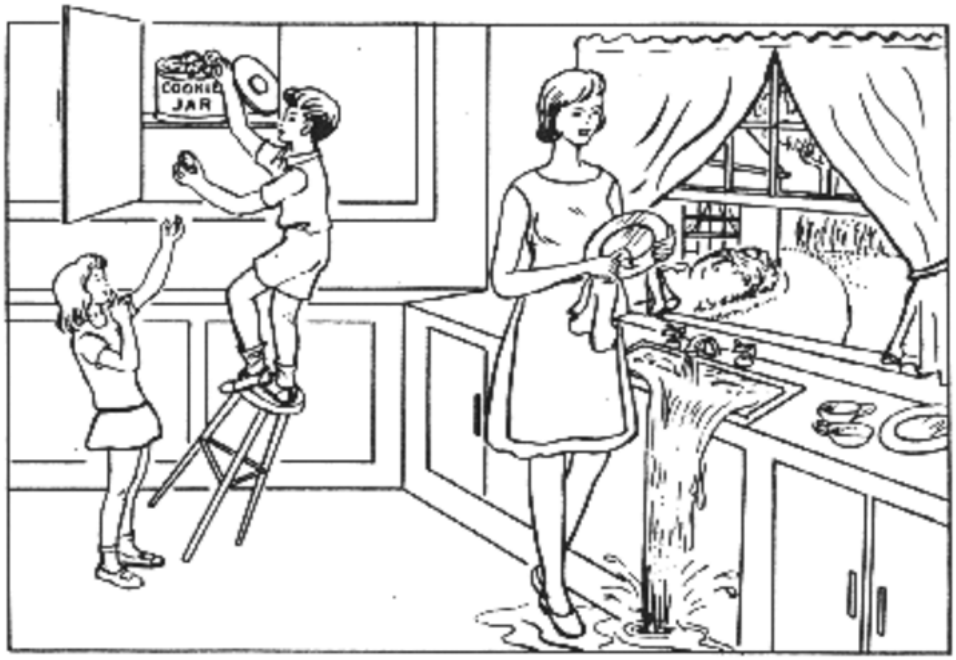

Fig. 1.1. Cookie Theft scene.

noise and Voice Activity Detection (VAD) is used to extract the audio chunks of each patient audio file. As shown in Table 1.1 the distribution of age and gender between both HC and AD patients is the same. Here is an example of the input data:

CHC : *"well there's a mother standing there &uh &uh washing the dishes an(d) the sink is overspilling [: overflowing] [* s:r] an(d) &uh the window's open . and outside the window there's a <walk with a> [//] &c curved walk with a garden . and you can see another &uh &uh building there . looks like a garage or something with curtains and the grass in the garden . and there are dishes [//] &uh &uh two cups and a saucer on the sink . and &uh she's getting her feet wet from the overflow of the water from the sink . she seems to be oblivious to the fact that the &s sink is overflowing . she's also oblivious to the fact that her kids are stealin(g) cookies out o(f) the cookie jar . and the kid on the stool is gonna fall off the stool . he's standing up there in the cupboard takin(g) cookies out o(f) the jar, handin(g) them to his [//] &uh a girl about the same age . the kids are somewhere around seven or eight years old or nine . an(d) the mother is gonna get shocked <when the> [/] when the [//] he tumbles and the cookie jar comes down . an(d) I think that's about all . [+ exc] "*

AD : *"okay . [+ exc] &uh we see a [/] &uh a &b little boy climbed up on a stool reaching for the cookie jar . and &uh the stool <is about to or> [//] is falling .&uh he is trying to get a cookie for himself and also one for his sister . &uh his sister is telling him to be very*

quiet . she's goin(g) shh@o . can't write that down . [+ exc] telling him to be quiet . [+ gram] and &uh let's see . [+ exc] in the meantime &uh the mother is <washing the dishes or yes> [//] washing dishes and the sink has &uh overflowed and &uh is pouring water on the floor . (..) I don't think I see anything else . [+ exc] okay . [+ exc] "

	AD (n =78)	**CHC (n=78)**
Age	66.6 (6.8)	66.3 (6.6)
Gender (male/female)	35/43	35/43
Vocabulary size (number different tokens)	1381	1584
Mean utterance length in number of tokens	143.6	139.8

Table 1.1. ADReSS challenge dataset main characteristics.

1.2.2. CCNA

CCNA datasets text extracts are derived from the audio files verbatim with annotations on whether the patient laughs, coughs, etc [**2**]. The audio files are not normalized for noise and volume. From Table 1.2 we can see that the number of both CHC and AD patients is almost equal. The ratio of male to female is very different between both groups with a lot more females to males in the CHC group. We can also see from Table 1.2 that the age gap between AD and controls is still 5 years. Here is an example of the input data:

CHC : *"Okay. Um, the boy is taking cookies from the cookie jar, and giving it to a girl, I am assuming his sister. The stool is about to tip over. It's a three legged stool. The mother is washing the dishes, and the sink is overflowing. There are some dishes on the counter. She is, [pause] it looks like she has got a dish in her hand perhaps, wa, washing it.Is there anything else I can see? [Pause]. The cupboards, I said the cupboards are closed at the bottom. Um. It looks like it is [pause], this looks like a, a very happy family. [Laughing]. Typical children, trying to steal cookies from the cookie jar. Um. [Pause].Yes. Yeah."*

AD : *"Well, there's a little boy standing on a stool that's about to fall over, trying to get a cookie out of a cookie jar. He's taken the lid off the cookie jar, and I think he's probably getting it for his sister, who is standing below him. While his mother is drying dishes, but all of a sudden, the sink is overflowing.Uh, it looks like the mother is going to be in horrible mess. [Laugh]. Um, she doesn't look happy. The little girl looks as if she's saying, oh my goodness. Um, but she's waiting to get the cookie in her hand. The window is open. There's a path going outside. Um, there's some trees outside.Okay."*

	AD (n = 32)	CHC (n=33)
Age	75.2 (7.5)	70.9 (7.0)
Gender (male/female)	17/15	5/28
Vocabulary size (number different tokens)	1186	922
Mean utterance length in number of tokens	134.9	112.3

Table 1.2. CCNA dataset main characteristics.

Chapter 2

Feature Extraction

The size of the dataset has a great impact on the methodology used to train and evaluate models. A very interesting advantage of Deep Learning models is that they can learn representations directly from the raw data, but to learn good representations a very large dataset is required to generalize well to unseen examples and avoid overfitting. Classical Machine Learning models can work quite well on hand built features extracted from a very small dataset. Another advantage to using models trained on a hand built feature set compared to the inferred feature space from the Deep Learning models is the ease of interpretability. The interpretability of the learned features can be critical when building a model for a diagnostic purpose. In this section we present a detailed explanation of all the features extracted which will be used by our classical Machine Learning models in our experiments.

2.1. Data preparation

An extremely crucial part in NLP is the data preparation step of the text. This step is very important prior to building features or to use directly in a model. The preprocessing is required on the raw data in order to remove any noise or anomalies that can be found in text such as special characters, etc. In this section we present the preprocessing done to the data prior to extracting the feature set.

The text from the ADReSS challenge dataset is extracted from `.cha` files created using the CHAT protocol [**18**] which is a transcription protocol that describes very specific rules on how to transcribe audio files to the corresponding text in their specific `.cha` files format. This permits to have a standardized file format of the translated audio files to text. The CHAT protocol has a set of very specific rules for annotations for each of the text. For example, here is a small text extract from a CHC with CHAT annotations from the ADReSS challenge dataset:

"well there's a mother standing there &uh &uh washing the dishes an(d) the sink is overspilling [: overflowing] [s:r] an(d) &uh the window's open . and outside the window there's a <walk with a> [//] &c curved walk with a garden . and you can see another &uh &uh building there . looks like a garage or something with curtains and the grass in the garden . and there are dishes [//] &uh &uh two cups and a saucer on the sink . and &uh she's getting her feet wet from the overflow of the water from the sink . she seems to be oblivious to the fact that the &s sink is overflowing . she's also oblivious to the fact that her kids are stealin(g) cookies out o(f) the cookie jar . and the kid on the stool is gonna fall off the stool . he's standing up there in the cupboard takin(g) cookies out o(f) the jar, handin(g) them to his [//] &uh a girl about the same age . the kids are somewhere around seven or eight years old or nine . an(d) the mother is gonna get shocked <when the> [/] when the [//] he tumbles and the cookie jar comes down . an(d) I think that's about all . [+ exc] "*

From the example above we can see that the addition of annotations from the transcriber are added. For example, an(d) would be an annotation that the speaker said an but it should actually be and. Also, the annotation &uh means that it was the sound uh that was pronounced by the patient. An other example of specific annotations is *[: overflowing] [* s:r]* which is a correction to the previous word. The correction is in brackets *[: overflowing]* and *[* s:r]* is the error code so if the error is semantic, phonetic, etc. Many more annotations are available in the CHAT protocol such as *[/]* for repetition, etc and can be found in their manual [**18**].

We normalized the raw data by removing all annotations added by the transcriber to remove as much as possible any potential bias from the transcriber such as [/], [//], [+ exc], etc. For example, here is the normalized text extract:

"well there's a mother standing there uh uh washing the dishes an the sink is overspilling. an uh the window's open. and outside the window there's a walk with a c curved walk with a garden. and you can see another uh uh building there. looks like a garage or something with curtains and the grass in the garden. and there are dishes uh uh two cups and a saucer on the sink. and uh she's getting her feet wet from the overflow of the water from the sink. she seems to be oblivious to the fact that the s sink is overflowing. she's also oblivious to the fact that her kids are stealin cookies out o the cookie jar. and the kid on the stool is gonna fall off the stool. he's standing up there in the cupboard takin cookies out o the jar, handin them to his uh a girl about the same age. the kids are somewhere around seven or eight years old or nine. an the mother is gonna get shocked when the when the he tumbles and the cookie jar comes down. an I think that's about all. "

Furthermore, the annotated `.cha` files have the timestamps of when each speaker in the audio file making the extraction of phonation time and pause time very easy to determine. We used the `pylangacq` library from python to extract the manual transcription from the `.cha` derived from the audio [**16**].

The text from the CCNA dataset is extracted from the audio files with the addition of annotations such as [Pause], [Laughing], etc., meaning that a transcriber will listen to the audio file and transcribe word for word what is being said by the patient. Unlike the ADReSS challenge dataset the CCNA dataset does not follow a specific protocol such as the CHAT protocol for the naming and type of annotations used. For example here is a small text extract from a CHC with annotations from the CCNA dataset:

"Okay. Um, the boy is taking cookies from the cookie jar, and giving it to a girl, I am assuming his sister. The stool is about to tip over. It's a three legged stool. The mother is washing the dishes, and the sink is overflowing. There are some dishes on the counter. She is, [pause] it looks like she has got a dish in her hand perhaps, wa, washing it.Is there anything else I can see? [Pause]. The cupboards, I said the cupboards are closed at the bottom. Um. It looks like it is [pause], this looks like a, a very happy family. [Laughing]. Typical children, trying to steal cookies from the cookie jar. Um. [Pause].Yes. Yeah."

All annotations are removed from the text such as [Pause], [Laughing], etc. for equal comparison between both datasets. For example, here is the normalized text extract:

"Okay. Um, the boy is taking cookies from the cookie jar, and giving it to a girl, I am assuming his sister. The stool is about to tip over. It's a three legged stool. The mother is washing the dishes, and the sink is overflowing. There are some dishes on the counter. She is, it looks like she has got a dish in her hand perhaps, wa, washing it. Is there anything else I can see?. The cupboards, I said the cupboards are closed at the bottom. Um. It looks like it is , this looks like a, a very happy family. Typical children, trying to steal cookies from the cookie jar. Um. Yes. Yeah."

For the audio files, no preprocessing was done for the ADReSS challenge dataset as the audio files have already been normalized for noise and volume as mentioned in [**17**]. The patient audio was extracted using the timestamps from CHAT annotations from the `.cha` files.

For the CCNA dataset we also normalized the audio files for noise and volume. We normalize for volume using the `ffmpeg-normalize` python package and use the `EBU R128` normalization algorithm. For the extraction of the patient audio, the CCNA dataset text files do not contain the timestamps of when the patient is speaking as is present in the ADReSS

challenge datset. Therefore, the patient audio was extracted by cutting the interviewer audio out manually.

As we can see the format of both datasets is quite different given the different types of annotations provided. The choice of removal of all annotations of both datasets is a critical decision as keeping any of them may skew the results in favor of one dataset and may add bias to the dataset prior to any feature extractions. Also keeping these annotations would interfere with the tokenization of the text. Therefore, we can see that the removal of all annotations is required to level the playing field and make sure the text is as pure as possible prior to the feature extraction process. It was decided to normalize the text so that the text would closely resemble an automatically translated audio. If it was automatically translated no additional annotations from the user would be added as they are not present in the audio recording. It would only be the text without annotations of corrections and pauses, etc. Furthermore, no specific details on the text preprocessing is elaborated by Balagopalan et al [**3**] which is why we decided this direction for the text preprocessing.

Furthermore, it is important to note that this work does not perform Automatic speech recognition from the audio files in order to extract the raw text. The linguistic features are extracted from the manually transcribed text derived from the audio files given in the dataset. This choice was made as it is the procedure currently done by most of the literature and the one utilised by Balagopalan et al [**3**]. To maintain consistency we opted for this approach to have a better comparison of results.

2.2. Features

2.2.1. Linguistic features

Linguistic features are some of the most used types of features in the literature for the diagnosis of AD. Throughout the research, different anomalies present in the CS of patients have been identified and elaborated into meaningful Machine Learning features. For this research we chose our set of linguistic features based mostly on the features extracted by both Balagopalan et al [**3**] and Fraser et al [**9**] which relate to the anomalies reported by Slegers et al [**28**]. We also added a feature developed by Ilya Ivensky in his Masters thesis [**13**] that is not found in both papers that also suggests differentiation between AD and CHC called vector norm. Furthermore, the majority of the linguistic features used were also tested on a different dataset and connex task of positive amyloid AD vs negative amyloid AD by our team, Slegers, Chafouleas et al [**27**] which also demonstrated good differentiation between both groups. Slegers et al did a thorough evaluation of the most common speech anomalies found in the literature that affect patients with AD by comparing 44 different articles which include the ones used by Fraser et al [**9**]. According to this review, AD

patients use more pronouns, word categories (i.e: nouns, verbs, etc) with higher frequency which gave rise to extracting features groups such as POST tag counts, Noun chunks, Noun ratios, verb ratios, proportions, etc as described in Table 2.1 [**28**]. Also, according to Slegers et al [**28**] it was shown that AD patients have trouble finding specific words which is described by the feature groups Miscellaneous count which count words related to hesitations, word finding, dietetics, uncertainty as described in Table 2.1. It was also pointed out that AD patients provide fewer information units about the picture which derived the feature group ICU Count and also convey less information than CHC in their sentences which derived the feature group Semantic Idea Density as described in Table 2.1 [**28**]. These semantic differences led to the development of some meaningful features used in the literature for the classification of AD patients. For example if we look at the following extracts from both CHC and AD we can appreciate the intuition behind the features elaborated in Table 2.1:

CHC : *"well there's a mother standing there uh uh washing the dishes an the sink is overspilling. an uh the window's open. and outside the window there's a walk with a c curved walk with a garden. and you can see another uh uh building there. looks like a garage or something with curtains and the grass in the garden. and there are dishes uh uh two cups and a saucer on the sink. and uh she's getting her feet wet from the overflow of the water from the sink. she seems to be oblivious to the fact that the s sink is overflowing. she's also oblivious to the fact that her kids are stealin cookies out o the cookie jar. and the kid on the stool is gonna fall off the stool. he's standing up there in the cupboard takin cookies out o the jar, handin them to his uh a girl about the same age. the kids are somewhere around seven or eight years old or nine. an the mother is gonna get shocked when the when the he tumbles and the cookie jar comes down. an I think that's about all. "*

CHC : *"okay . uh we see a uh a b little boy climbed up on a stool reaching for the cookie jar . and uh the stool is falling .uh he is trying to get a cookie for himself and also one for his sister . uh his sister is telling him to be very quiet . she's goin . can't write that down . telling him to be quiet . and uh let's see .in the meantime uh the mother is washing dishes and the sink has uh overflowed and uh is pouring water on the floor . (..) I don't think I see anything else . okay . "*

We can notice in the both of these extracts from the ADReSS challenge dataset that the AD patient utters more *uh* than the CHC extract which corresponds to the hesitation feature described in Table 2.1 which counts the number of tokens corresponding to hesitations (i.e :

uh, um, etc) which was also pointed out by Slegers et al [**28**]. Also, it seems that the extracts have a difference in the utterance length between both AD and CHC which corresponds to the Mean utterance length feature described in Table 2.1. It is also apparent that the sentences of the AD patient is much simpler and contains less information than the one uttered by the CHC, this can also give an intuition on why the Semantic Idea density feature described in 2.1 is used to discriminate AD patients in the literature. Furthermore, we notice that in both extracts they mention ideas related to the image being described. As mentioned by Slegers and al [**28**], AD patients mention less key information about the image than CHC. In this example the CHC is a lot more detailed than the AD patient in describing the image and conveying more information related to the image which refers to the commonly used feature ICU count described in Table 2.1. As mentioned by Slegers and al [**28**], AD patients use more pronouns, word categories (i.e: nouns, verbs, etc) with higher frequency which is slightly apparent in this example where AD uses a slightly higher proportion of verbs than the CHC group. This characteristic is extracted from the common feature groups for all POS tag count and POS Tag ratio features described in Table 2.1 which extract keys features related to the count and ratio of specific POS tags including the verb tag. These are only a few of the characteristics and intuitions of the extractions of this set of features on the classification of AD vs CHC in the literature.

The features were hand-built based on the description of the features from the literature. In order to build all the features many open source libraries were used. The `python` implementation of most of the linguistic features were reused from the ones already implemented by Ilya [**13**] and Slegers et al [**27**] as we had access to the source code and wanted to reuse for standardization. The reminder of some of the linguistic features such as lexical norm-based features were added in implementation. The `Spacy` library was used as the tokenization tool and POS tagging mechanism [**12**] for building the majority of the features. For the idea density features, the use `glove 50-d`[1] [**14**] and `300 google news`[2] [**1**] as the word embedding dictionaries.

[1] `https://nlp.stanford.edu/projects/glove/`
[2] urlhttps://code.google.com/archive/p/word2vec/

Feature group	#Features	Description
POS Tag	43	Count of different POS tag (i.e: pronoun, noun, verb, adverb, adjective), out of vocabulary words, L2norm, pronoun ratio, mean left and right children)
Tree structure	5	Tree structure of texts using `Spacy` to determine the trees. (dep diversity, total dep, avg dep distance, average parse tree height, average tree width)
Semantic Idea Density	4	Calculate the average cosine distance of all pairwise combination of word embeddings within a sliding window [**27**]. This is calculated with a window size of 10. Done for both base tokenization and only lemmas using 50-dimensional and 300-dimensional embeddings.
Miscellaneous Count	6	Count of hesitations (i.e: uhm), personal pronouns, word finding (i.e: 'know', 'remember', 'unable'), dietetics (i:e 'this', 'that', 'here'), uncertainty (I.e: 'think', 'look', 'like', 'kind', 'seem', 'maybe', 'can', 'something')
ICU count	25	Boolean if a specific word related to the image is mentioned during the description. We also have the total number if ICU's mentioned and the word to ICU proportion.
Mean length utterance (MLU)	1	Number of morphemes over total number of utterances per transcription. (using `Spacy` each token is a morpheme containing the prefix and suffix)
Noun chunks	3	Noun chunks are flat phrases having noun as their head. noun chunks (mean number of noun chunks per transcripts), NP -> PRON (ratio of all noun chunks where root is PRON), noun chunk length (mean length of noun chunks)
Subordinate count	1	The sum of all subordinate counts (i.e: 'csubj', 'xcomp', 'ccomp', 'advcl', 'acl')

Feature type	#Features	Description
TTR	5	Type token ratio (TTR). This is the ratio number of unique tokens (types) over number total tokens. We also add types and tokens as own features. We also do this for lemmas. [**6**]
Honore score	1	Score to show vocabulary richness. [**6**]
Brunet Score	1	Score to show vocabulary richness. [**6**]
MATTR10 & 3	4	Average TTR in a moving window size (i.e: 10 or 3) after filtering punctuation's.
BiMATTR50	2	Average number of unique bigrams in a moving window size of 50.
Proportions	34	Ratio of the following token type ('nouns', 'pronouns', 'verbs', 'adjectives', 'adverbs', 'leftChildren', 'rightChildren', 'types', 'subordinates', 'nounChunks', 'nounChunkLength', 'inflectedVerbs', 'hesitation', 'personalPronoun', 'wordFinding', 'deictics', 'uncertainty') with respect to the token count. Also ratio with respect to the ROOT token type count.
Vector norms	5	Average of vector norms over transcript of all words, verbs, adjectives, nouns and adverbs [**13**].
Dependencies	25	Ratio of count tokens that are tagged as dependencies by tokenizer over the count of 'ROOT' token.
Noun ratios	2	Ratio of nouns over verbs, nouns and adverbs. Ratio of nouns over nouns, verbs, adverbs and adjectives.
Verb ratios	4	Ratio of verbs over verbs, nouns and adverbs. Ratio of verbs over nouns, verbs, adverbs and adjectives.
Lexical norm-based	12	Average norms imageability, age of acquisition, familiarity and frequency across all words, noun and verbs [**3**]
Sentiment rating	9	Average sentiment rating on valence, arousal and dominance across all words, noun and verbs.

Table 2.1. List of linguistic features extracted from text with their short description.

2.2.2. Acoustic Features

Extracting acoustic features directly from the audio files has been attempted in the literature. Both Balagopalan et al [**3**] and Fraser et al [**9**] have extracted the mean, variance, skewness and kurtosis of the Mel-frequency cepstral coefficients (MFCCs) and also the pause duration. Balagopalan et al [**3**] also extracted the mean, variance, skewness and kurtosis of the zero-crossing rate and fundamental frequency. For our experiment, the set of acoustic features were hand-built based on Balagopalan et al [**3**] as shown in Table 2.2. For example, it was pointed out by Slegers et al [**28**] that AD patients have a higher word finding difficulty which can be described by pauses and also that AD patient had affected phonation rates which is described the feature group pause and phonation described in Table 2.2. The acoustic features were extracted with the help of the `librosa` [**19**] library from python. Unlike both Balagopalan et al [**3**] and Fraser et al [**9**] where the pause features are extracted using the timestamps from the CHAT annotations of the `.cha` files, we extracted the pause features using Voice Activity Detection (VAD) using the `pyannote` python library [**5**] which takes the audio directly and gives the timestamps from when there is voice activity, i.e when someone is speaking. We can therefore use this to evaluate the phonetion time and pause time of the audio file. This will allow a more universal approach to pause extraction that does not require human timestamp annotation. For the replication of "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" the pause features were extracted using the timestamps from the CHAT protocol as done by Balagopalan et al [**3**] for comparison.

Feature type	#Features	Description
Pause & phonation	7	number pauses, total phonetion time, pause phenetion time, total pause time, mean pause time, number short pauses (< 1s), number long pauses (> 1s)
MFCC	168	Mean, standard deviation, skewness and kurtosis of 42 MFCC coefficients.
Zero-Crossing rate	4	Mean, standard deviation, skewness and kurtosis zero crossing rates.
Fundamental Frequency	4	Mean, standard deviation, skewness and kurtosis of fundamental frequency.

Table 2.2. List of acoustic features from audio files with their short descriptions.

To our knowledge there does not exist a tool to extract both the linguistic and acoustic features and are assumed to be hand-built in the literature. As far as we are aware the

way the features are built is very much left to the interpretation of the reader which makes reproduction very difficult as the choice of implementation may have a very big impact on the results. The standardization of the most common set of features reused throughout the literature would benefit and help the advancement in the literature. We therefore created an open source project containing the code to extract features from both text and audio transcriptions in order to help forward the standardization across publications. [3]

[3] https://github.com/gchafouleas/Connected-Speech-Feature-Extraction/tree/readme

Chapter 3

Methodology

In this section we present the details of the classification models, feature selection process and evaluation metrics.

3.1. Model selection

The main goal of our experiment is to be able to differentiate AD patients from CHC using the text descriptions from the Cookie Theft scene. The complexity and potential subtle differences between CHC and AD patients makes this less of a clear separable classification problem as the subtleties in the disease may become more prevalent as the disease progresses. This brings the question of what models are best suited for the task. Both classical Machine Learning models and Deep Learning models have been used to perform the classification of AD patients from their text extract in the literature. In this section we describe the different models used in our experiments.

3.1.1. Feature based models

The choice of the model used on a feature set can have a great impact on the results. Each model has different learning characteristics that may be better suited for different situations. The choice of the models were based on models that have had success in the literature for the classification of AD patients from CS.

Logistic Regression and SVM are the most common types of models used with linguistic and acoustic features in the literature. We therefore chose to also utilise a Logistic Regression and SVM with Linear Kernel (SVM) in order to have a baseline for comparison. Both the Logistic Regression and SVM assume the data is linearly separable but differ in their objective function. Logistic Regression tries to find the best line separating all the data and is usually used as a good baseline. The SVM's objective function maximizes the margin between the support vectors therefore maximizing the distance of the data points from the decision boundary. The larger the margin the better the model is at separating the data points.

One main advantage with using an SVM is that you can optimize the hyperparameter C which determines how hard you want the margin to be. Meaning how tolerant you are to errors which is beneficial for any classification problem but especially important in medical classification, as you would want a large C value to tolerate the least amount of errors and points very close to the decision boundary. Another advantage to using an SVM is that you can change the type of kernel used. The kernel is an inexpensive way to transform the data into a higher dimension in order to solve a non-linear problem. Therefore, we also opted for the radial base kernel SVM (RBF-SVM) as used by Balagopalan et al [**3**] which showed good results when used with linguistic and acoustic features.

Finally, we opted for a Multi Layer Perceptron (MLP) used on the linguistic and acoustic features. This is a good option when the data may not be linearly separable. It is also a very flexible model as the number of layers and hidden units can be optimized as well as the use of different activation functions.

3.1.2. Bidirectional Encoder Representations from Transformers (BERT)

Language based models using only the raw text as input have been shown to have great performance on a wide range of NLP tasks [**26**]. The advantage of these models is that they can learn the representations of the language themselves without prior knowledge of the domain. One caveat is that these models are very big and have an extremely large number of parameters to train and need very large amounts of examples to generalize well. The more training data there is, the better it will be at matching the distribution of the test data. Unfortunately, it is not always possible to retrieve a large enough set of labelled data. This is especially true in the medical field as it is quite difficult and time consuming to retrieve labelled data.

The use of transfer learning has been shown to solve this issue by training a very large model on a very large dataset in a domain that can be transferable to another domain. For the transfer of information to occur both domains need to have a connection [**32**]. This permits smaller datasets to benefit from the learned representations from one domain and transfer them to the classification task of a different domain therefore reducing the amount of time needed to converge to proper solution. The pre-trained model is used to fine-tune its weights to the new domain task.

We chose BERT as our transfer learning model as it was shown to perform very well on classification of textual inputs [**8**]. It uses multi-layer bidirectional recurrent transformer encoders. BERT utilises bidirectional self-attention for its transformer, meaning that every token can attend to both the right and left side tokens. BERT is trained on two different tasks. The first is predicting a percentage of the masked words of a sentence. The second is

a binary prediction of the next sentence, i.e predicting whether the next sentence is either sentence A or B. The pre-trained model can be used directly with a classification layer as the output for a classification downstream task. A new model can be fine-tuned on the classification task by updating the weights of the model for that particular classification task.

For this research, we used `BERT-base-uncased` from the `hugging face` [**31**] python library for the implementation to fine-tune a binary classification model, as it is more light weight in terms of parameters to train. This model contains 12-layer, 768-hidden, 12-heads, 110M parameters and is trained on lower-cased English text [**31**].

3.2. Feature Selection

Feature selection in a Machine Learning pipeline is a crucial component as it helps determine which features are the most important in separating the data. This is especially important when the number of features is larger than the number of examples. Many different methods can be used to select the best features for a model. The main two categories are selected based on statistical significance and selected based on an external estimator which has feature ranking possibilities (ex: SVM, random forest, etc). We opted for four different techniques which are described below.

Select K Best: The features are sorted according to their ANOVA f-value score between labels and features from `sklearn` library [**24**]. The top k features with the highest scores are kept.

Multicollinearity: From Slegers et al [**27**] with methodology as follows "1) perform Welch's t-tests across groups; 2) correct the associated P values for multiple comparisons by the Benjamini-Hochberg method for False Discovery Rate (FDR) at .05; 3) subset only significant features at p $<.05$ after correction for FDR; 4) starting with the largest Welch's coefficient, enter the features in the model in a stepwise manner at the condition that the added feature is not correlated $>.75$ to a previously entered feature with a higher t statistic " where the value 0.75 is a hyperparameter to be tuned.

Recursive feature elimination (RFE): Using an external estimator (i.e SVM) the estimator is trained on the whole feature set and the features with the smallest coefficients are removed. This is done recursively on the subset of features until we have the number of features desired using `sklearn` library [**24**].

Select from Model: Using an external estimator (i.e: SVM) the estimator is

trained on the whole feature set and the top k features with the biggest coefficients from the estimator are selected using `sklearn` library [**24**]

The feature selection process is only used with models that use the linguistic and acoustic feature set as input. The input for the BERT model is the preprocessed text and therefore we do not perform feature selection during the cross validation stage.

3.3. Experiments and Evaluation metrics

The experimental pipeline is based on "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" by Balagopalan et al [**3**]. We utilise 10-fold cross validation with grid search to evaluate the best hyper parameters for each model. The feature selection described in Section 3.2 is done at each fold of the cross validation process. The validation accuracy of the model is used to evaluate the best hyper parameters during cross validation. The 10-fold cross validation grid search was performed for three different random seeds (i.e 42, 52, 62) each seed selecting the best hyper parameters and best features across all folds. Figures 3.1 and 3.2 describe the experimental pipelines for both BERT and feature based models.

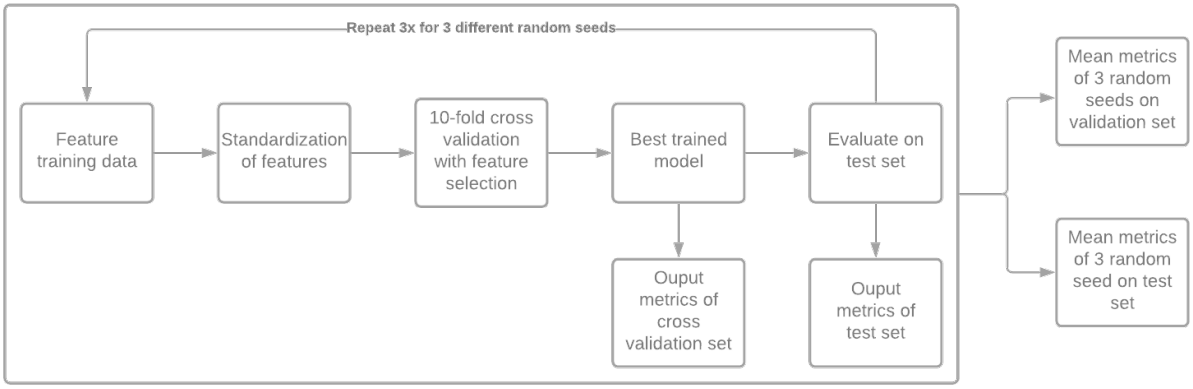

Fig. 3.1. Overview of experimental pipeline for feature based models.

As shown in Figure 3.2, the input is the preprocessed text data and 10-fold cross validation is performed to find the best hyperparameters of the BERT model. Once the best hyperparameters are selected we do a final training with all of the training data with the selected hyperparameters to use as the model for evaluation. This is done three times for three different random seeds.

As shown in Figure 3.1, feature selection is added at each fold of the cross validation process for all models that are described in section 3.1.1 and have as input the feature set. Also, prior to the feature selection process, the features are standardized using the standard

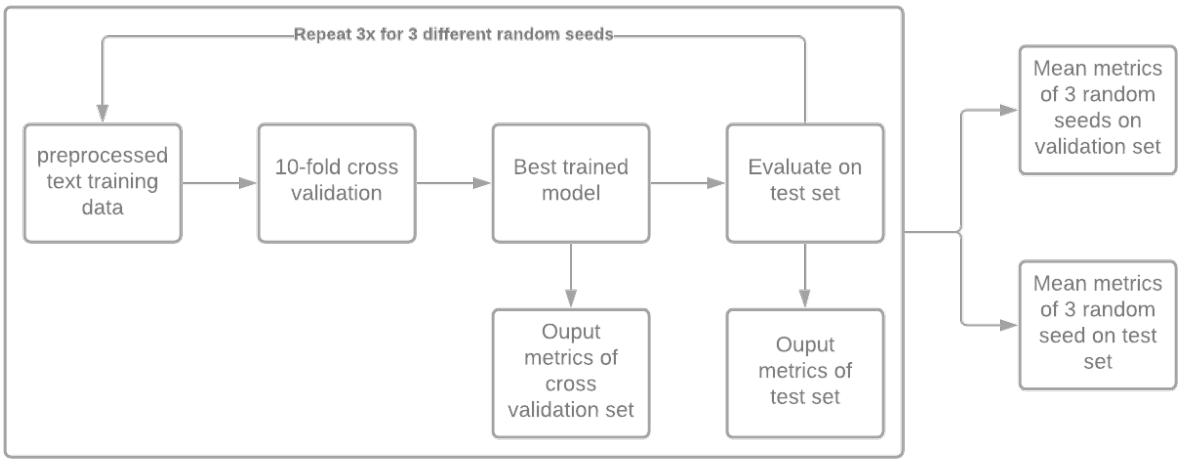

Fig. 3.2. Overview of experimental pipeline for BERT model.

scalar method from `sklearn` for our pipeline. This ensures that all features have the same variance and therefore, there is no feature that stands out more because the ranges are significantly different. In "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" by Balagopalan et al [**3**], it is not specified if any normalization of the features is applied and therefore, no normalization of the features is done prior to replicating the paper.

For the evaluation metrics we use accuracy, precision, recall, F1 and sensitivity across all folds. The average across all random seeds. Accuracy alone can sometimes give an overly optimistic overview of the performance as it does not discriminate the type of missclassifications made. Recall, precision, sensitivity and F1 can become more insightful on the true performance of the model. Below are the metrics formulas:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.3.1}$$

$$recall(positive\,class)/sensitivity = \frac{TP}{TP + FN} \tag{3.3.2}$$

$$precision = \frac{TP}{TP + FP} \tag{3.3.3}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.3.4}$$

where

TP stands for True Positive, TN stands for True Negative,

FP stands for False Positive, FN stands for False Negative

Recall takes into account not only the true positive but the false negatives for the metric which is an important metric for a medical model as one would want to minimize the number of false negatives, i.e patients affected with the disease diagnosed as healthy and *vice versa*. While as shown in equation 3.3.3, precision measures how accurate your model is at predicting a specific class. In an ideal world we would like a high precision and recall but there is a trade off between both metrics. By increasing recall the precision goes down and *vice-versa*. This is where the F1 score can become useful as it is the weighted average of precision and recall [**24**]. This score is very useful when one wants to find an optimal value of both recall and precision. The unweighted average of recall and precision of both negative and positive classes is reported in our results. Given that we would like to also have a metric to track how correct we are at predicting only the AD class, sensitivity will be used in our results. Sensitivity measures the recall associated with only the positive class in a binary classification setting.

The mean metrics for accuracy, precision, recall, sensitivity and F1 across all random seeds are presented as shown in Figure 3.2. The best models are selected based on the best mean validation metrics across all random seeds. The test results are evaluated on the best selected models for each random seed as shown in Figure 3.2. In the results section we demonstrate both the mean validation metrics of the 10-fold cross validation and the mean test metrics across all random seeds in order to properly compare the cross data evaluation. Without the mean validation metrics it is difficult to see if the model is generalizing well to the unseen examples on the different dataset and therefore helps to visualize the power of generalization of the models to unseen examples.

Chapter 4

Results and Discussion

4.1. Publication replication

In order to determine how well the ADReSS challenge dataset can help the standardization of results across the literature we wanted to replicate "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" by Balagopalan et al [**3**]. They use both linguistic and acoustic features trained with classical Machine Learning models and compare them to an implementation of BERT trained only on the transcriptions. We compare our results to theirs on a RBF-SVM, MLP and BERT model using the same hyperparameters described in section B of the publication. Therefore, the same hyperparameters were used across all random seeds. Note that the exact random seeds and particular details of BERT are not clearly defined, therefore these were arbitrarily assigned to achieve best results. As shown in Tables 4.1 and 4.2 the paper only describes the recall and precision for the positive class compared to our unweighted average. Therefore for comparison of the results, the sensitivity metric in the tables is equivalent to their recall. This choice was made as we wanted the macro evaluation of both classes to get a better overview of the overall model performance and the addition of the Sensitivity to track the AD specific metric performance.

As we can see from Tables 4.1 and 4.2 we achieve very similar results for the RBF-SVM as Balagopalan et al [**3**] with a validation accuracy of 80.6% and test accuracy of 81.25% compared to the paper of 79.6% for valid accuracy and 81.3% for test accuracy. The results differ slightly for BERT as the valid and test accuracy are inverse to the paper with 83.5% for valid accuracy and 81.25% for test accuracy. The main difference is the valid accuracy of the MLP with 73.5% compared to 76.2% and test accuracy of 80.6% compared to 77.1%. The difference in results of both the MLP and BERT models may be subject to the choice of random seeds used, as the choice of the random seed can have a great impact on the start of the parameter space when training a model. Furthermore, our best model is also BERT but only slightly better than the SVM-RBF and MLP as observed by Balagopalan et all [**3**].

Our results						
Model	**# features**	**Accuracy**	**Precision**	**Recall**	**Sensitivity**	**F1**
SVM-RBF	10	80.6 (1.1)	82.8	80.6	79.6	80.3
MLP	10	73.5 (5.5)	76.1	73.5	67.3	72.2
BERT	–	**83.5 (1.9)**	**83.7**	**83.6**	**83.3**	**83.6**

Paper results [3]						
Model	**# features**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1**
SVM-RBF	10	79.6	81.0	78.0	82.0	79.0
MLP	10	76.2	77.0	75.0	77.0	76.0
BERT	–	**81.8**	**84.0**	**79.0**	**85.0**	**81.0**

Table 4.1. 10-fold cross validation results across random seeds on ADReSS validation set using the same hyperparameter as in [**3**] with actual paper results.

Our results						
Model	**# features**	**Accuracy**	**Precision**	**Recall**	**Sensitivity**	**F1**
SVM-RBF	10	81.3 (0.0)	81.7	81.3	75.0	81.2
MLP	10	80.6 (2.6)	81.8	80.6	70.8	80.4
BERT	–	**81.3 (1.7)**	**85.1**	**81.3**	**66.7**	**80.7**

Paper results [3]						
Model	**# features**	**Accuracy**	**Precision** $(^{\text{CHC}}_{\text{AD}})$	**Recall** $(^{\text{CHC}}_{\text{AD}})$	**Specificity**	**F1** $(^{\text{CHC}}_{\text{AD}})$
SVM-RBF	10	81.3	83.0 80.0	79.0 83.0	-	81.0 82.0
MLP	10	77.1	78.0 76.0	75.0 79.0	-	77.0 78.0
BERT	–	**83.3**	**86.0** 81.0	79.0 **88.0**	-	83.0 **84.0**

Table 4.2. 10-fold cross validation results across random seeds on ADReSS test set using the same hyper parameter as in [**3**] with actual paper results.

Another possible difference between our results and the publication, is how the initial text is preprocessed as this might have a great impact on the tokenization and therefore affect many of the features and lead to discrepancies in the results. For example, the term utterance is very different if using the annotations from the CHAT files or the ones extracted using

the `Spacy` library therefore affecting for example the MLU feature. Also, depending on the tokenization tool used this may have an effect on the POS tag, proportion and tree features as they may identify certain tokens as different types of tokens. It would be important to standardize the most commonly used features as to make sure the comparison is more tangible.

Overall the replication of the paper is achieved and confirms the implementation of the Machine Learning pipeline for further experiments. As we have shown the results are comparable but with some discrepancies even with a properly designed benchmark. We can therefore conclude that this dataset is a great start for comparison between publications, but future standardization of the features as well as more detailed documentation on the exact text preprocessing done prior to feature extraction may help this discrepancy.

4.2. Results

In this section we present the results of performing grid search on our pipeline. For each of the models a 10-fold grid search cross validation is performed to find the best hyper parameters across all folds. The full list of all results for each feature selection technique on the ADReSS challenge dataset is reported in the Appendix A. The addition of both a Logistic Regression and SVM is used. The best models are reported in the tables with the best feature selection method being Select K best also observed by Balagopalan et al [**3**]. After, multiple hyperparameter searches for BERT, the best model achieved is the same as the one reported in Section 4.1.

Model	**Accuracy (std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SVM	84.9 (0.4)	87.0	84.9	84.6	84.5
LR	87.0 (0.8)	90.0	87.0	84.6	86.9
RBF-SVM	88.3 (0.4)	87.0	88.3	88.9	88.1
MLP	**89.8 (0.8)**	**91.5**	**89.8**	**88.9**	**89.7**
BERT	83.5 (1.9)	83.7	83.6	83.33	83.6

Table 4.3. 10-fold cross validation best models results across random seeds on ADReSS validation set using Select k Best feature selection method.

From Table 4.3 we observe that the best accuracy and overall performance on the validation set is obtained by the MLP model with a valid accuracy of 89.8%. From Table 4.4 it is apparent that the MLP model does not generalize as well as expected on the test set with an accuracy of 81.9% which is significantly lower than the valid accuracy. We notice the same behavior with the RBF-SVM and the Logistic Regression. From Table 4.4, we observe that the best model on the test set is the Logistic Regression with an accuracy of 84%.

Model	Accuracy (std)	Precision	Recall	Sensitivity	F1
SVM	81.9 (2.6)	84.1	81.9	69.4	81.6
LR	**84.0 (2.0)**	**85.6**	**84.0**	**73.6**	**83.9**
RBF-SVM	80.6 (1.0)	82.0	80.6	70.8	80.3
MLP	81.9 (1.0)	84.0	81.9	70.8	81.7
BERT	81.3 (1.7)	85.1	81.3	66.7	80.7

Table 4.4. 10-fold cross validation best ADReSS models results across random seeds on ADReSS test set using Select k Best feature selection method.

Model	Accuracy (std)	Precision	Recall	Sensitivity	F1
SVM	50.3 (1.5)	51.1	49.7	10.41	40.6
LR	53.9 (1.3)	58.9	53.3	15.6	45.2
RBF-SVM	60.0 (4.5)	67.8	59.5	25.0	53.5
MLP	55.9 (5.1)	59.4	55.35	19.8	48.6
BERT	**62.6 (8.4)**	**76.1**	**62.0**	**26.0**	**55.2**

Table 4.5. 10-fold cross validation best ADReSS models results across random seeds on CCNA test set using Select k Best feature selection method.

This is higher than 81.3% observed by Balagopalan et al [**3**] but only by a small margin. Furthermore, the recall and precision are very similar which is favorable as we would want to have high recall and precision being 84.0% and 85.6% on the test set.

From Tables 4.3 and 4.4 we observe that the models with the most stable performance seem to be the SVM, Logistic Regression and BERT when evaluated on the ADReSS challenge test set as the valid and test set accuracies do not differ greatly. We notice that the sensitivity of the models are quite low compared to the average recall on the ADReSS challenge test set. For example, from Table 4.4 the SVM recall is 81.9% for which the sensitivity is 69.4%. This may signify that the model is better at classifying CHC than AD patients.

However, from Table 4.5 we observe that the models trained on the ADReSS challenge dataset do not generalize well to a new dataset. The highest test results are achieved by BERT with an accuracy of 62.6% but with a very high standard deviation between the random seeds showing that there is a very large difference in the results when changing the random seed. Another observation is that the sensitivity of the models is extremely low, averaging between 10% to 26% which indicates that the model has a very poor ability to discriminate the AD patients at all. Overall, the models perform very badly when evaluated on the CCNA dataset.

A set of models trained on the CCNA dataset using the same feature selection procedure Select K Best and fine tuning hyperparameters was done and evaluated on the ADReSS challenge test set. Tables 4.6 and 4.7 show these results.

Model	**Accuracy (std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SVM	81.0 (1.9)	81.7	80.6	76.6	79.2
LR	82.6 (1.5)	87.4	82.4	77.1	81.5
RBF-SVM	81.5 (1.1)	88.6	81.3	74.7	80.4
MLP	**86.2 (2.2)**	**88.3**	**86.1**	**85.5**	**85.6**
BERT	78.7 (1.5)	79.4	78.9	84.8	78.9

Table 4.6. 10-fold cross validation best models results across random seeds on CCNA dataset using Select k Best feature selection method.

Model	**Accuracy (std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SVM	57.6 (1.0)	58.7	57.6	40.3	56.3
LR	52.1 (2.9)	36.0	52.1	11.11	40.16
RBF-SVM	55.6 (5.2)	49.2	55.6	54.2	47.5
MLP	54.9 (3.5)	48.8	54.9	25.0	47.5
BERT	**66.7 (1.7)**	**71.5**	**66.7**	**88.9**	**64.7**

Table 4.7. 10-fold cross validation best CCNA models results across random seeds evaluated on ADReSS challenge test set using Select k Best feature selection method.

From Table 4.6 we can observe that the best model trained on the CCNA dataset is MLP with the highest overall metrics. This is the same observation when models are trained on the ADReSS challenge dataset. The results on the CCNA validation set are a little lower than the ones observed on the validation set from the ADReSS challenge dataset but only by a few percentages bringing all models with a valid accuracy above 80% which follows the results in the literature. From Table 4.7 we can observe that the models trained on the CCNA dataset also do not generalize well to the ADReSS test set. The best results on the ADReSS test set is observed by the BERT model with a test accuracy of 66.7% with similar precision and recall scores, but again this is a very low score barely above the 50% decision boundary which is equivalent to chance. Also, BERT has a very high sensitivity which indicates that it is good at classifying correctly the AD patients.

Overall the deep learning model BERT seems to perform better when used on a different dataset. The reason for this is potentially because it is initially a transfer learning model and

is made to be used on multiple cross domain applications making it more robust to subtle changes in the dataset.

Overall both datasets give very good results on the validation set but perform very poorly when evaluated on a different dataset from the same domain. This raises the question as to whether the models selected meaningful features and are these selected features too specific to the ADReSS challenge dataset.

4.3. Feature Importance and Evaluation

In this section we evaluate the validity of the feature selection process and whether the features selected are the expected biomarkers affecting AD patients we measure. Which linguistic and acoustic feature group are the most discriminating in the feature selection process.

ADReSS Selected Features		**CCNA Selected Features**	
Feature name	**group**	**Feature name**	**group**
ratio of pronouns	POS tag	ratio of neg	Dependencies
average parse tree height	Tree structure	number of verbs	POS tag
ICU: stool	ICU count	number of inflected verbs	POS tag
ICU: window	ICU count	verb word vector norm	Vector norm
ICU: curtains	ICU count	proportion of pronouns	Proportions
ICU: taking	ICU count	proportion of pronouns/ROOT	Proportions
total ICU count	ICU count	proportion of verbs	Proportions
proportion of noun	Proportions	proportion of verbs/ROOT	Proportions
proportion adverbs	Proportions	noun ratio to noun/verb/adverb	Noun Ratio
proportion of types	Proportions	noun ratio to noun/verb/adjective	Noun Ratio
proportion inflected verbs	**Proportions**	**proportion of inflected verbs**	**Proportions**
proportion of inflected verbs/ROOT	**Proportions**	**proportion of inflected verbs/ROOT**	**Proportion**
ratio of determiner	dependencies	verb ratio to noun	Verb Ratio
ratio of preposition	dependencies	verb ratio to noun/verb	Verb Ratio
ratio of object of preposition	dependencies	verb ratio to noun/verb/adverb	Verb Ratio
		verb ration to noun/verb/adjective	Verb Ratio
		mfcc skew 6	MFCCs
		mfcc skew 39	MFCCs
		mfcc skew 40	MFCCs
		mfcc skew 41	MFCCs

Table 4.8. Features selected across models and random seeds in Table 4.3 on ADReSS challenge and in Table 4.5 CCNA dataset. Features in bold are common to both settings.

Table 4.8 shows the top features selected across all of the best models for both the ADReSS and CCNA models. We can observe that linguistic features are selected by the ADReSS models and mostly linguistic features and some of the MFCCs accoustic features are selected by the CCNA models. These are interesting observations as it would have been expected that the models would select pause features as one of top features as it was shown by Slegers et al [**28**] where AD patients had 100% of the time word-finding difficulties which may

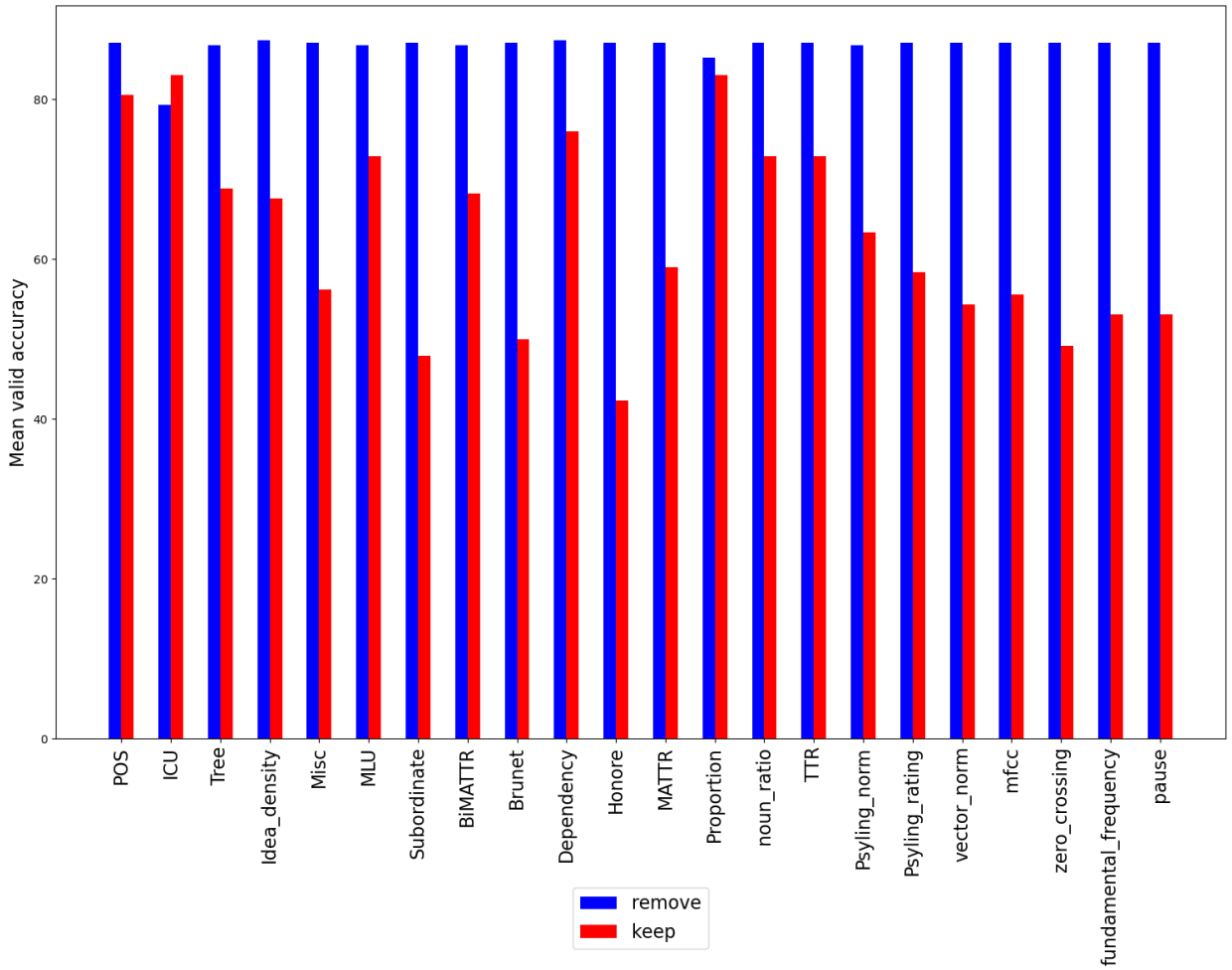

Fig. 4.1. Mean valid accuracy when removing and keeping only specific feature groups on ADReSS challenge dataset.

be indicated by a pause. Balagopalan et al [**3**] also observed the selection of only linguistic features on the ADReSS challenge datset which correlates with our findings. From Table 4.8 we can observe that there are only two common features that are selected between both CCNA and ADReSS challenge datasets which are the inflected verbs proportions features. As we can see, the features selected by both datasets are completely different to build the best models on the validation set which may explain why the performance on the test sets is poor.

From Table 4.8 we can observe that POS tag, ICU count, proportion and dependencies feature groups are selected when using ADReSS challenge dataset. While from Table 4.8 we can observed that POS tag, Proportions Noun ratio and verb ratio are the top selected feature groups when using CCNA dataset. Figures 4.1 and 4.2 show the mean valid accuracy when only training with the features of a specified feature group, for example only training the best model using only POS tag features, etc as well as when training with all features

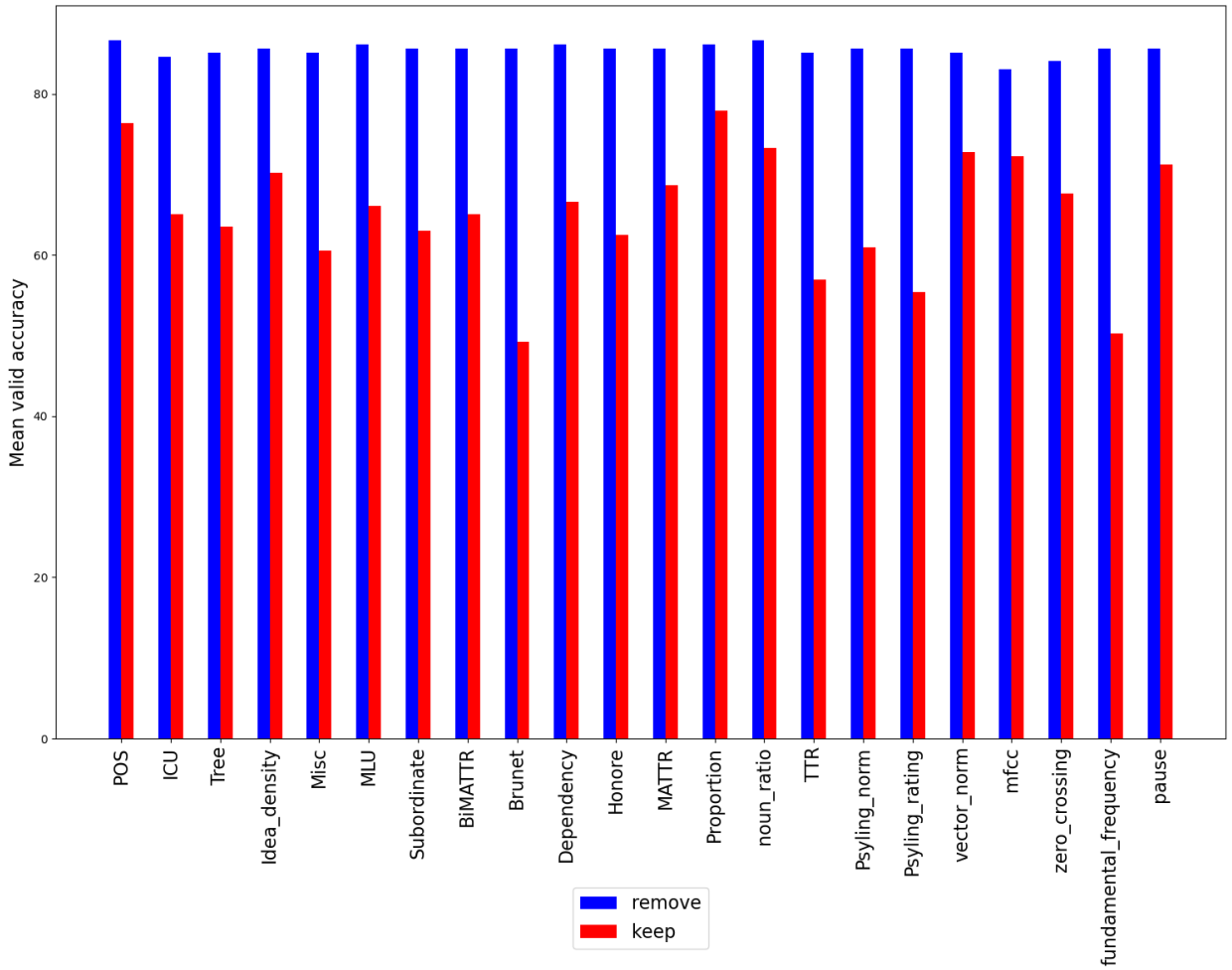

Fig. 4.2. Mean valid accuracy when removing and keeping only specified feature group on CCNA dataset.

except a specific feature group. Figure 4.1 shows that the three feature groups with the most impact when training separately are ICU, POS Tag and Proportion feature groups when using the ADReSS challenge dataset. We can also see that removing ICU count has a slight drop in mean valid accuracy which reinforces the importance of this feature group, but that overall removing only a specific feature group does not have significant effect on the overall mean valid accuracy. This may suggest that the most significant feature group selected when training on the ADReSS challenge dataset is in fact the ICU count feature group. Given that this feature group determines if a specific token is mentioned out of a predefined list of Information tokens, example Cookie, Jar, etc concerning the image, it is possible that this feature group is very variable from one dataset to the next. Given that the list of ICU tokens is human made, it may induce bias to the feature causing it to potentially not work well on all datasets. Table 4.9 shows this discrepancy between both datasets for the top selected ICU values. We can see that there is not a very significant difference between

both AD and CHC on the CCNA dataset also supporting why it may have performed so poorly.

	ADReSS		CCNA	
ICU name	**AD**	**CHC**	**AD**	**CHC**
Stool	74	48	31	24
window	45	16	12	11
curtain	32	1	6	6
taking	49	15	21	12

Table 4.9. Count of times ICU is mentioned at least once in transcript on both CCNA and ADReSS challenge dataset.

While when training on the CCNA dataset, we can observe in Figure 4.2 that keeping only the ICU count feature reduces drastically the performance of the model compared to that observed when trained on the ADReSS dataset. Also we can observe that the features group that seem to discriminate the best between both CHC and AD groups are POS tag, Proportions and noun/verb ratios which is consistent with the top selected features observed in Table 4.8 when trained on the CCNA dataset but very different from the observed ones when trained on the ADReSS challenge dataset. Given that both datasets have selected drastically different features to train their best models may be the reason why the Machine Learning models perform so poorly when evaluated on either datasets.

Table 4.8 also shows the high importance of using linguistic features as they form the majority of the selected features. Moreover given that the CCNA models also selected some of the acoustic features suggests their importance as well. It is apparent, therefore that both categories of features have some importance in the classification of AD vs CHC.

In order to evaluate the learned representations of BERT in the classification of AD vs CHC the interpretation of the attention heads is necessary. The model view from `bertViz` [**30**] is used to visualize all attention heads at each layer. This permits us to identify specific attention heads that may show interesting patterns from the tokens. Figures 4.3 and 4.4 shows a specific attention head at a specific layer for each of the three random seed models. The most interesting patterns that are observed in most of the attention heads when trained on the ADReSS dataset, but very visible in the ones in Figure 4.3, are the attentions to specific words like cookie, dishes, stool, kids, washing, cookie jar, mother, floor. This is very interesting as they have the most attention and represent the ICU tokens used to build the ICU count feature group. This may indicate that the model also learnt that there is attention to be made on the ICU tokens to predict AD patients. Furthermore, we observe in both CCNA and ADReSS models the attention to the word "uh" which is interesting as AD

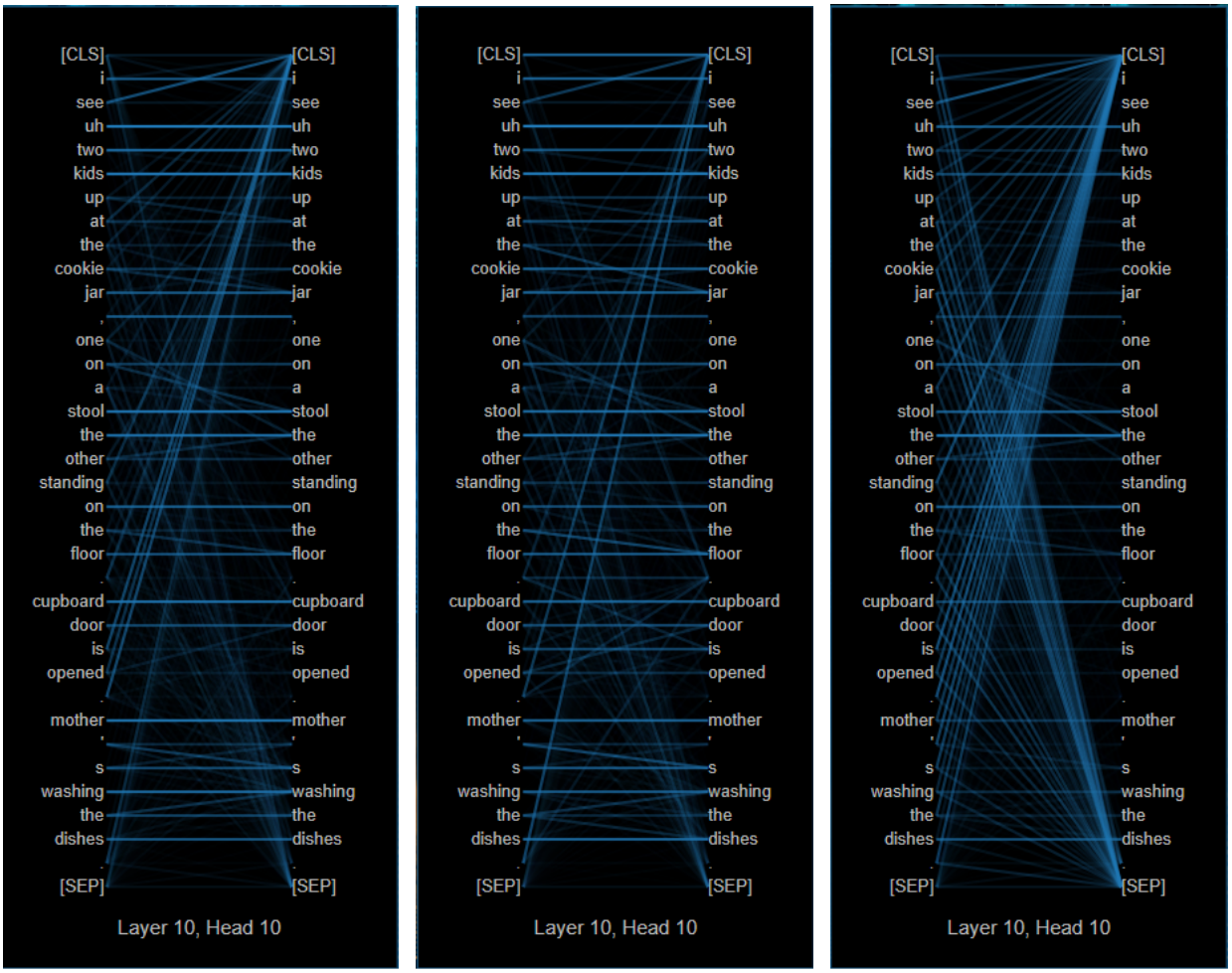

Fig. 4.3. Attention head of ADReSS models for specific layer for each random seed of the sentence "I see uh two kids up at the cookie jar, one on a stool the other standing on the floor.cupboard door is opened .mother's washing the dishes.".**Left:** random seed 42 **Middle:** random seed 52. **Right:** random seed 62.

patients present more word finding difficulties which can be apparent by pauses or the use of an indeterminate term such as "uh" as described by Slegers et al [**28**]. From Figure 4.4 we can observe that the most interesting attention when trained on CCNA data is the attention to again ICU words but mostly for Cookie Jar, mother, dishes, stool which is also noted on the ADReSS models attention. The main difference is there seems to be a slight increase in the attention made to verbs such as washing, opened, standing, etc. This correlates quite well with the top selected features from the machine learning models trained on the CCNA data as they selected verb ratio and verb proportion features. From the attention patterns it is possible that some representations that correlate with the linguistic differences between AD and CHC have been learnt by BERT on the ADReSS challenge dataset and CCNA dataset. It is apparent that there is a difference in the learned representation for

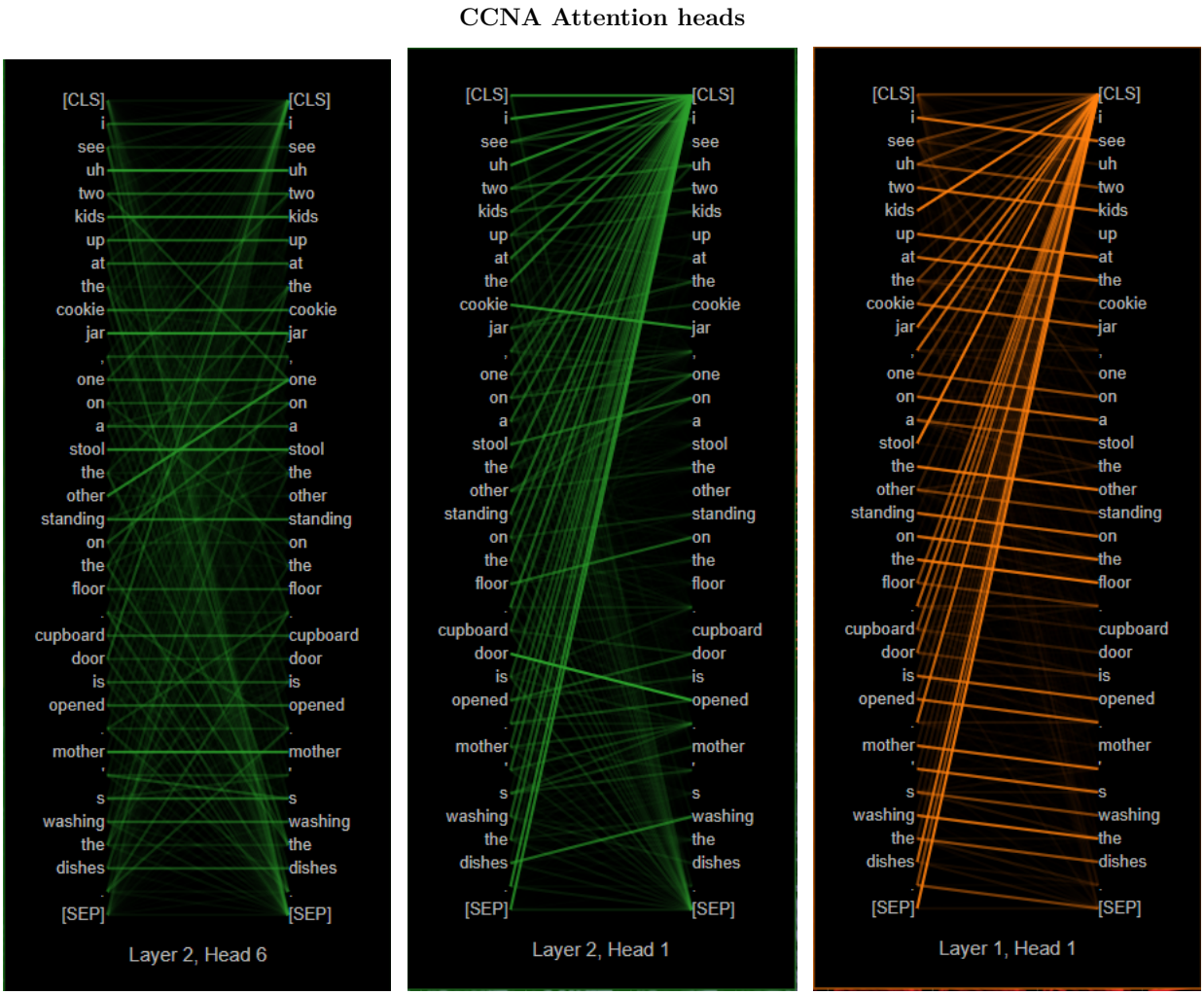

Fig. 4.4. Attention head of CCNA model for specific layer for each random seed of the sentence "I see uh two kids up at the cookie jar, one on a stool the other standing on the floor.cupboard door is opened .mother's washing the dishes.".**Left:** random seed 42 **Middle:** random seed 52. **Right:** random seed 62.

both CCNA and ADReSS models which could explain the poor performance but there are still some overlapping learned representations which could also explain why the BERT model performs better when used on a different dataset compared to the Machine Learning models. Another possible reason why BERT performs better when evaluated on a different dataset may be due to the fact that it is a Transfer Learning model which is built to be applicable to transferable domain data.

Both the selected features and the learnt representations from BERT seem to be more domain specific to the task of connected speech of the Cookie Theft scene in English. This is very domain and language specific and seems to not be transferable to another dataset from the same domain.

Chapter 5

Conclusion

To conclude, the replication of "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection" [**3**] was achieved using the ADReSS challenge dataset. It was shown that potentially a more detailed and standard way of extracting acoustic and linguistic features should be done to have proper comparison between new techniques. A more thorough model selection was performed on the ADReSS dataset and evaluated on a different CCNA dataset and *vice versa*. It was shown that the trained models on the ADReSS challenge and CCNA dataset are not transferable to a different dataset on the same domain.

We further explored the significance of the selected features by the best models and their significance to the disease. It was shown that the selected features are very related to the differentiation between the disease but are drastically different between models trained on ADReSS and CCNA datasets. This may suggest that the two datasets are ,too small to capture the distribution of both CHC and AD groups when evaluated on a new dataset.

In the future, it is crucial that given the use of very small datasets, that the evaluation of the results be performed on a different dataset from the same domain. This is significant in order to determine the generalization of the model. The potential combination of both CCNA and ADReSS training set could be used to train the models. Also, continuing research of predicting AD vs CHC with less attention to the specificity of the picture description task would benefit greatly in building more robust and accurate models for the diagnosis of AD as it was shown that the deep learning approach performed better when evaluated on a different dataset. A potential research avenue could be the investigation of using Adapters which could perform well on very small datasets as there are a significantly smaller number of parameters to tune. Furthermore, exploration of the potential of the raw audio files should be continued as these would help reduce the reliance on the transcription of the audio files to text and the knowledge of the domain.

References

[1] Google code: Archive word2vec, 2015. Last accessed 26 March 2021 `https://code.google.com/archive/p/word2vec/`.

[2] "ccna database", 2020. Last accessed 28 May 2021.

[3] Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. 07 2020.

[4] R. Ben Ammar and Y. Ben Ayed. Speech processing for early alzheimer disease diagnosis: Machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8, 2018.

[5] Hervé Bredin. pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, August 2017.

[6] Romola Bucks, S. Singh, J.M. Cuerden, and Gordon Wilcock. Analysis of spontaneous, conversational speech in dementia of alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14, 01 2000.

[7] J. Chen, J. Zhu, and Jie ping Ye. An attention-based hybrid network for automatic detection of alzheimer's disease from narrative speech. In *INTERSPEECH*, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] K. Fraser, J. Meltzer, and F. Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease 49 (2016) 407–422*, 2016.

[10] J. Fritsch, S. Wankerl, and E. Nöth. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845, 2019.

[11] S. Hong, L. Yao, W. Cheah, W. Chang, L. Fu, and Y. Chang. A novel screening system for alzheimer's disease based on speech transcripts using neural network. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2440–2445, 2019.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[13] Ilya Ivensky. *Prediction of Alzheimer's disease and semantic dementia from scene description: toward better language and topic generalization*. Master's thesis, University of Montreal, 2020.

[14] Christopher D. Manning Jeffrey P., Richard S. Glove: Global vectors for word representation, 2015. Last accessed 05 March 2021 `https://nlp.stanford.edu/projects/glove/`.

[15] Sweta Karlekar, Tong Niu, and Mohit Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models, 2018.

[16] Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. Working with chat transcripts in python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago, 2016.

[17] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020.

[18] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates*. 2000.

[19] B. McFee, M. McVicar, S. Balke, , and et al. "librosa/librosa: 0.6.3", 2020. Last accessed 26 March 2021 `https://zenodo.org/record/3955228#.YF47kq9KiUk`.

[20] Alzheimer Society of Canada. Alzheimer society of canada, 2021. Last accessed 29 March 2021 `https://alzheimer.ca/en/about-dementia/what-alzheimers-disease#Alzheimer's_disease_is_the_most_common_type_of_dementia`.

[21] Public Health Agency of Canada. A dementia stategy for canada, 2020.

[22] Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[23] Yilin Pan, Bahman Mirheidari, Markus Reuber, Annalena Venneri, Daniel Blackburn, and Heidi Christensen. Automatic hierarchical attention neural network for detecting ad. pages 4105–4109, 09 2019.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Carole Roth. *Boston Diagnostic Aphasia Examination*, pages 428–430. Springer New York, New York, NY, 2011.

[26] Sebastian Ruder. Nlp-progress, 2017. Last accessed 26 February 2021 `http://nlpprogress.com/english/language_modeling.html`.

[27] A. Slegers, G. Chafouleas, C. Bedetti, M. Montembeault, A.E. Welch, G. Rabinovic, M.L. Gorno-Tempini, and S.M. Brambati. Connected speech markers and identification of high amyloid burden in primary progressive aphasia. 2020. In review.

[28] A. Slegers, Renée-Pier Filiou, M. Montembeault, and S. Brambati. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's disease : JAD*, 65 2:519–542, 2018.

[29] Becker J. T., Boller F., Lopez O. L., and Saxton J.and McGonigle K. L. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

[30] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[32] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. 2020.

Appendix A

Hyperparameter search results

SVM	**Accuracy(std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SelectKBest	**84.9 (0.4)**	**87.1**	**84.9**	**84.6**	**84.5**
Correlation	85.5 (0.4)	87.9	85.9	84.6	85.3
RFE	83.0 (2.9)	86.8	83.02	80.9	82.6
SelectFromModel	84.9 (0.4)	87.1	84.9	84.6	84.5

Table A.1. SVM: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection.

Logistic Regression	**Accuracy(std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SelectKBest	**87.0 (0.8)**	**90.0**	**87.04**	**84.6**	**86.9**
Correlation	85.8 (0.9)	87.0	85.8	86.4	85.6
RFE	83.0 (1.9)	84.3	82.7	83.02	
SelectFromModel	86.1 (0.8)	87.6	86.1	87.0	85.8

Table A.2. Logistic Regression: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection.

RBF-SVM	**Accuracy(std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SelectKBest	**88.3 (0.4)**	**87.0**	**88.3**	**88.9**	**88.1**
Correlation	87.04 (1.3)	86.3	87.0	90.7	86.8
RFE	84.9 (2.3)	85.8	84.9	85.8	84.6
SelectFromModel	86.4 (1.9)	88.33	86.4	86.4	86.1

Table A.3. RBF-SVM: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection.

MLP	**Accuracy(std)**	**Precision**	**Recall**	**Sensitivity**	**F1**
SelectKBest	**89.8 (0.8)**	**91.5**	**89.8**	**88.9**	**89.7**
Correlation	88.6 (0.9)	88.7	88.6	90.1	88.4
RFE	85.8 (0.9)	85.1	85.8	88.9	85.5
SelectFromModel	87.3 (1.2)	88.9	87.3	88.9	87.0

Table A.4. MLP: 10-fold cross validation results across random seeds on ADReSS validation set with hyper parameter search and different feature selection.