# MACHINE LEARNING CLASSIFICATION OF AUTOPSY PROVEN VASCULAR COGNITIVE IMPAIRMENT AND ALZHEIMER'S DISEASE USING CONNECTED SPEECH SAMPLING

Vassiliki Rentoumi

IIT, SKEL
vrentoumi@iit.demokritos.uk

May 11, 2016

# Language Change in $_{m}$AD and $_{p}$AD

- Language abilities in $_{m}$AD and $_{p}$AD (including connected speech) well documented
- However:
  - diagnostic gold standard difficult to achieve
  - manual language analysis $\rightarrow$ time consuming and subjective
- In this study:
  - Language samples from pathologically confirmed cases
  - Novel computational approach to characterising language

# State of the art in $_{p}$AD vs $_{m}$AD language distinction

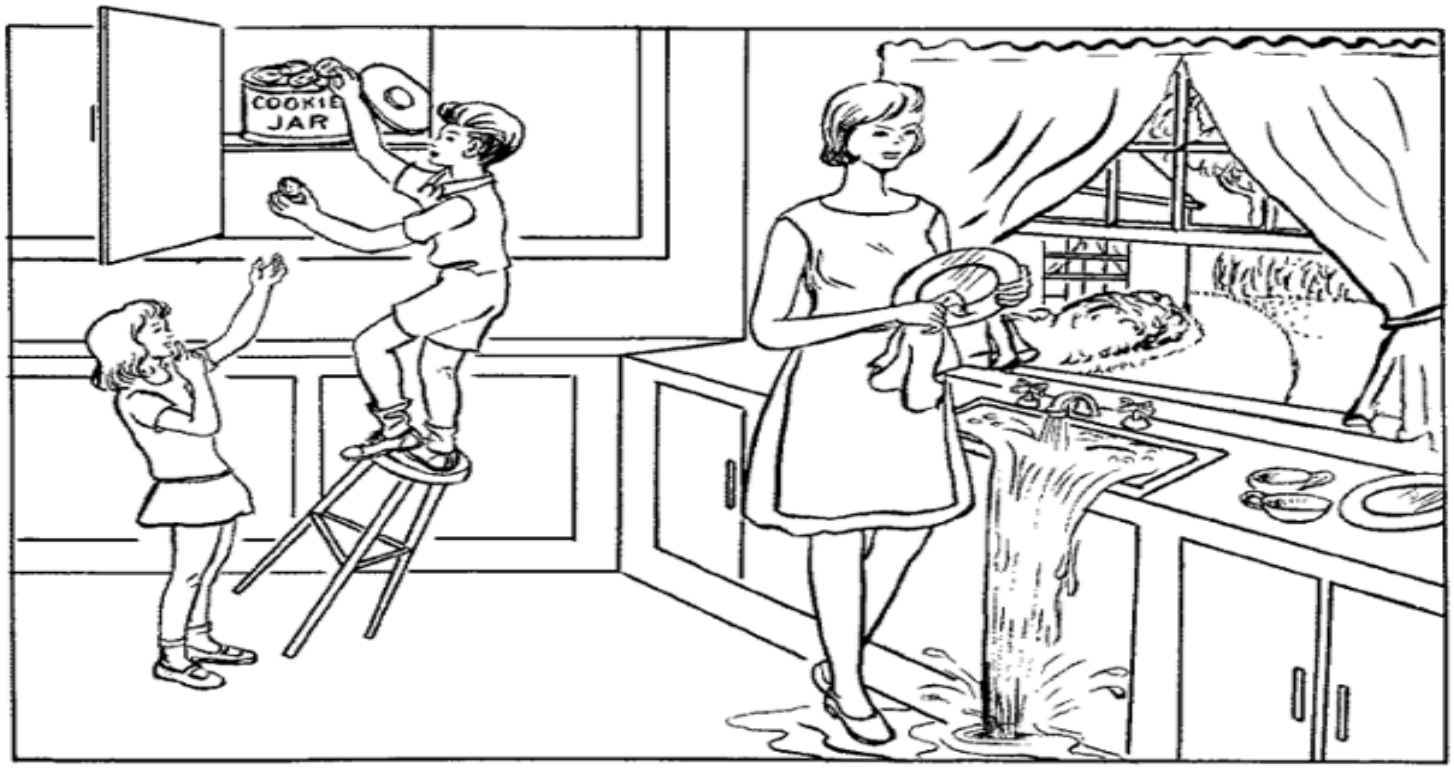
- Language analysis has proved to be a useful tool in understanding the early stages of $_{p}$AD and $_{m}$AD
  - Standardised discourse assessments in $_{p}$AD (Ahmed et al. 2012, 2013) show decline in multiple levels of linguistic analysis.
  - An analysis of the work of author Iris Murdoch demonstrated a decline in linguistic ability in her final work prior to being diagnosed with $_{p}$AD (Garrard et al (2005)).
  - Both $_{m}$AD and $_{p}$AD entail poor semantic processing abilities (Vuorinen et al, 2000).
  - Already proven success of ML based text classification in distinguishing SD from NCs and patients with R>L vs L>R temporal lobe atrophy (Garrard et al. 2013).

# Study Aims

1. To confirm the presence of differences in language produced by $_{m}$AD and $_{p}$AD using quantitative methods of evaluation.
2. To ascertain the most informative sources of language variation between the groups.

# Language Datasets I

- 54 transcribed samples of connected speech obtained from 54 anonymous subjects enrolling OPTIMA (Oxford Project to Investigate Memory and Ageing)
- samples obtained from the Boston cookie theft picture description task (Goodglass & Kaplin, 1982).



# Language Datasets II

- $_{p}$ADs (=18 samples)
- $_{m}$ADs (=18 samples)
- Normal Controls (NC= 18 samples)

All 54 participants followed through until death, and had had diagnosis confirmed post mortem.

# Analytical approach

**Machine Learning**

- Learn purely lexical, Syntactic and Textual features from the Language Data Sets under comparison:

- $_{p}$ADs vs NC
- $_{m}$ADs vs $_{p}$ADs
- $_{m}$ADs vs NC

# Analytical Approach I

***Three Stages:***

- Feature Extraction: Automatic extraction of features to create a set of values representing a number of distinct characteristics of each text.
- Feature Selection: Automatic extraction of significant characteristics that differentiate the transcripts of ${}_{p}$ADsvsNC, ${}_{m}$ADsvs${}_{p}$ADs, and ${}_{m}$ADsvsNC
- Classification: Automatic classification of transcripts to their categories ( ${}_{p}$ADs vs NC, ${}_{m}$ADs vs ${}_{p}$ADs, and ${}_{m}$ADs vs NC)

# Tools implementing methodology

- Feature Extraction: **Word Tokenizer**, **Syntactic Complexity Analyser** (Lu 2010), **Textual Complexity Analyser** (Forsyth, 2012).
- Feature Selection: Correlation based Feature Selection (CFS): uses a correlation-based heuristic to determine the usefulness of features.
- Classification: Use Naive Bayes to assign a transcript to the class that maximizes the posterior probability (the most likely class).

# Features Extracted

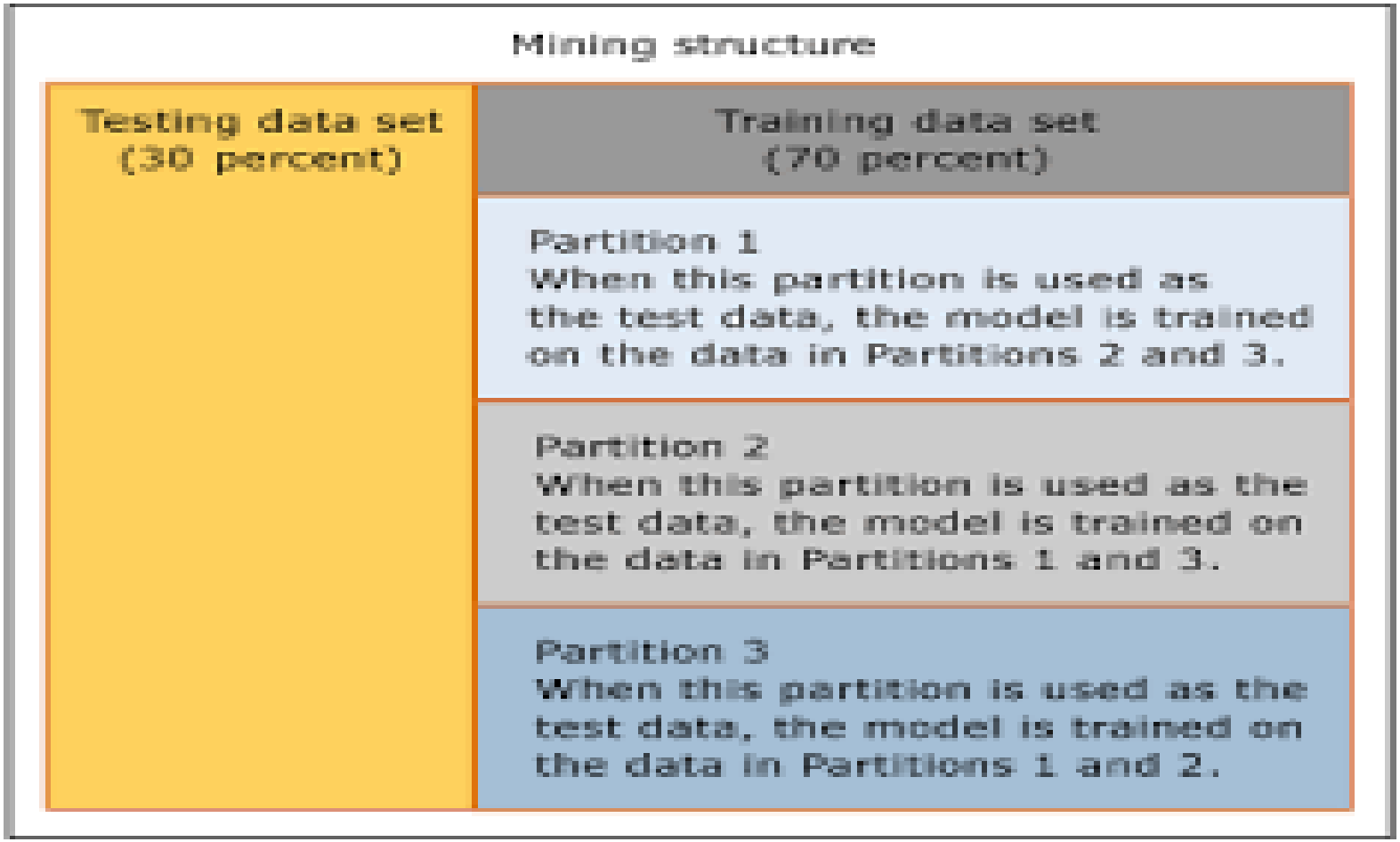
**SYNTACTIC**
MLC (Mean Length of Clause)
MLS (Mean Length of Sentence)
MLT (Mean Length of T-Unit
C/S (Sentence complexity ratio)
C/T (T-unit complexity ratio)
CT/T (Complex T-unit ratio)
DC/C (Dependent clause ratio)
DC/T (Dependent clauses per T-unit)
CP/C (Coordinate phrases per clause)
CP/T (Coordinate phrases per T-unit)
T/S (Sentence coordination ratio)
CN/C (Complex nominals per clause)
CN/T (Complex nominals per T-unit)
VP/T (Verb phrases per T-unit)

**TEXTUAL FEATURES**
Compression Ratio (CR)
Honore R
hapax legomena (once-used tokens)
hapax used word pairs
Simpson index of diversity
Dislegomena divided by vocabulary (i.e.types)
Shannon Entropy

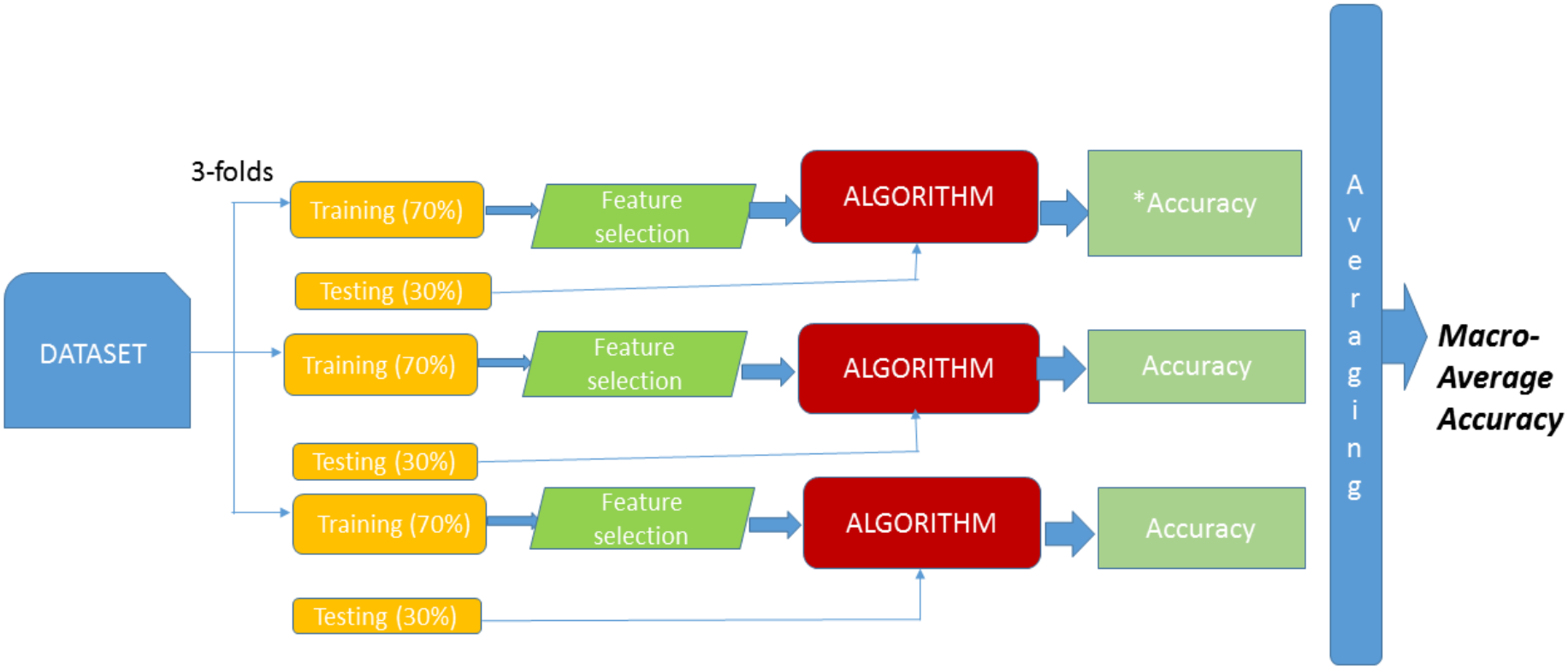
**PURELY LEXICAL**
Presence/Absence of words tokens
Frequency of word tokens

# Evaluation Procedures For Feature Selection & Classification

***Employment of Three-Fold Cross Validation***



# System Description




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

$$Macro - Av.Accuracy = \frac{Acc1 + Acc2 + Acc3}{3}$$

# Feature Selection Results: Selected Discriminative Lexical Features

| mAD vs pAD | mAD vs NC | pAD vs NC |
|---|---|---|
| | | |
| garden pAD>mAD | standing NC>mAD | chair pAD>NC |
| roof pAD>mAD | sister NC>mAD | fall NC>pAD |
| window pAD>mAD | open NC>mAD | giving NC>pAD |
| are mAD>pAD | too mAD>NC | stool NC>pAD |
| when pAD>mAD | while NC>mAD | waiting pAD>NC |
| standing pAD>mAD | | side NC>pAD |
| socks pAD>mAD | | why pAD>NC |
| right mAD>pAD | | much NC>pAD |
| hand pAD>mAD | | plates NC>pAD |
| | | right NC>pAD |

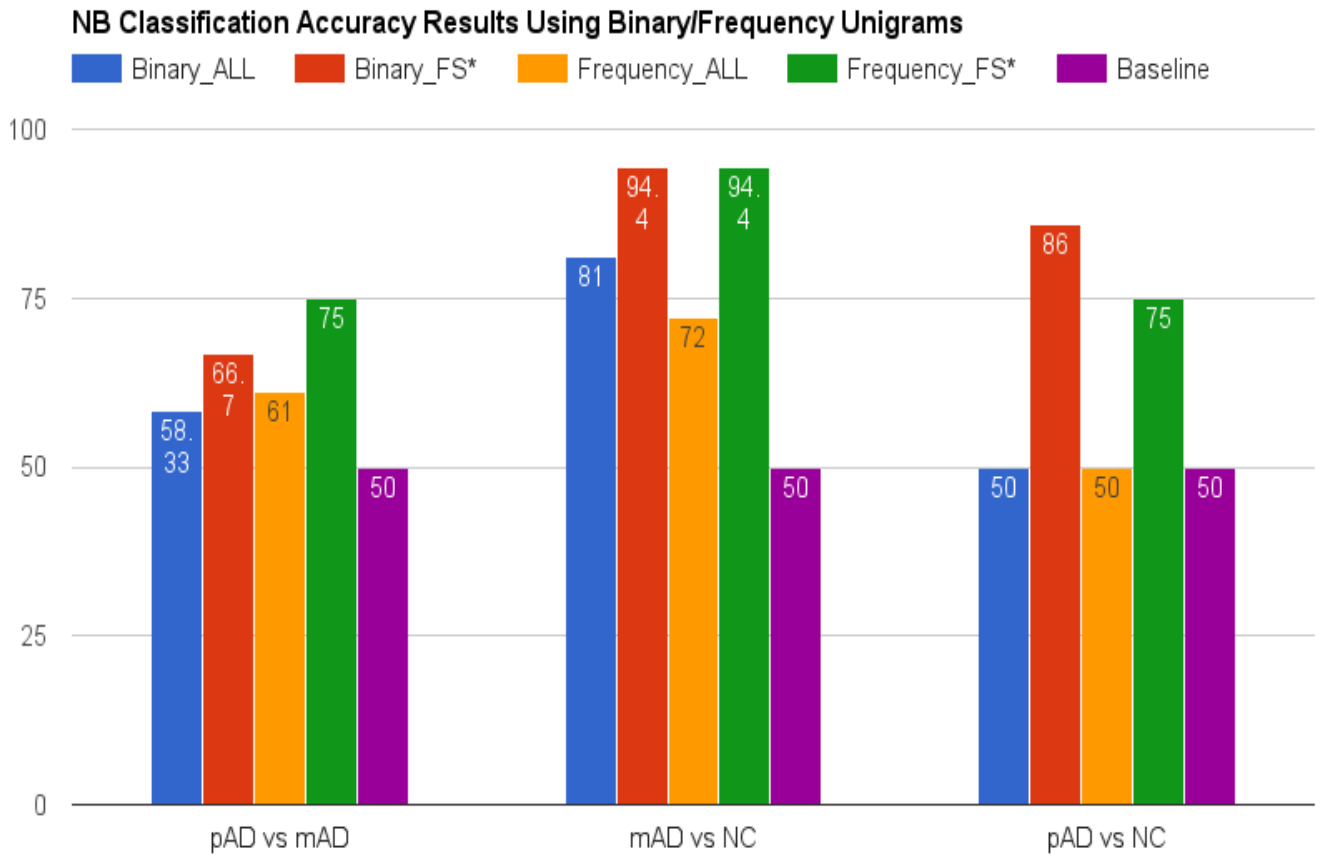
# Feature Selection Results: Selected Discriminative Textual and Syntactic Complexity Features

| Textual & Syntactic Complexity Features | | |
|---|---|---|
| **pAD vs mAD (w/fs)** | **mAD vs NC** | **pAD vs NC** |
| MLS pAD>mAD | MLS NC>mAD | Mean of hapax used word pairs pAD>NC |
| C/S pAD>mAD | CN/C NC>mAD | MLT NC>pAD |
| VP/T pAD>mAD | compression ratio NC>mAD | DC/T NC>pAD |
| MLT pAD>mAD | Honore R NC>mAD | CT/T NC>pAD |
| HonoreR pAD>mAD | Shanon entropy mean NC>mAD | |
| Mean hapaxes pAD>mAD | dislegomena/types mAD>NC | |
| Mean of hapax used word pairs pAD>mAD | | |
| dislegomena/types mAD>pAD | | |

# Syntactic Complexity Example in pADs vs mADs

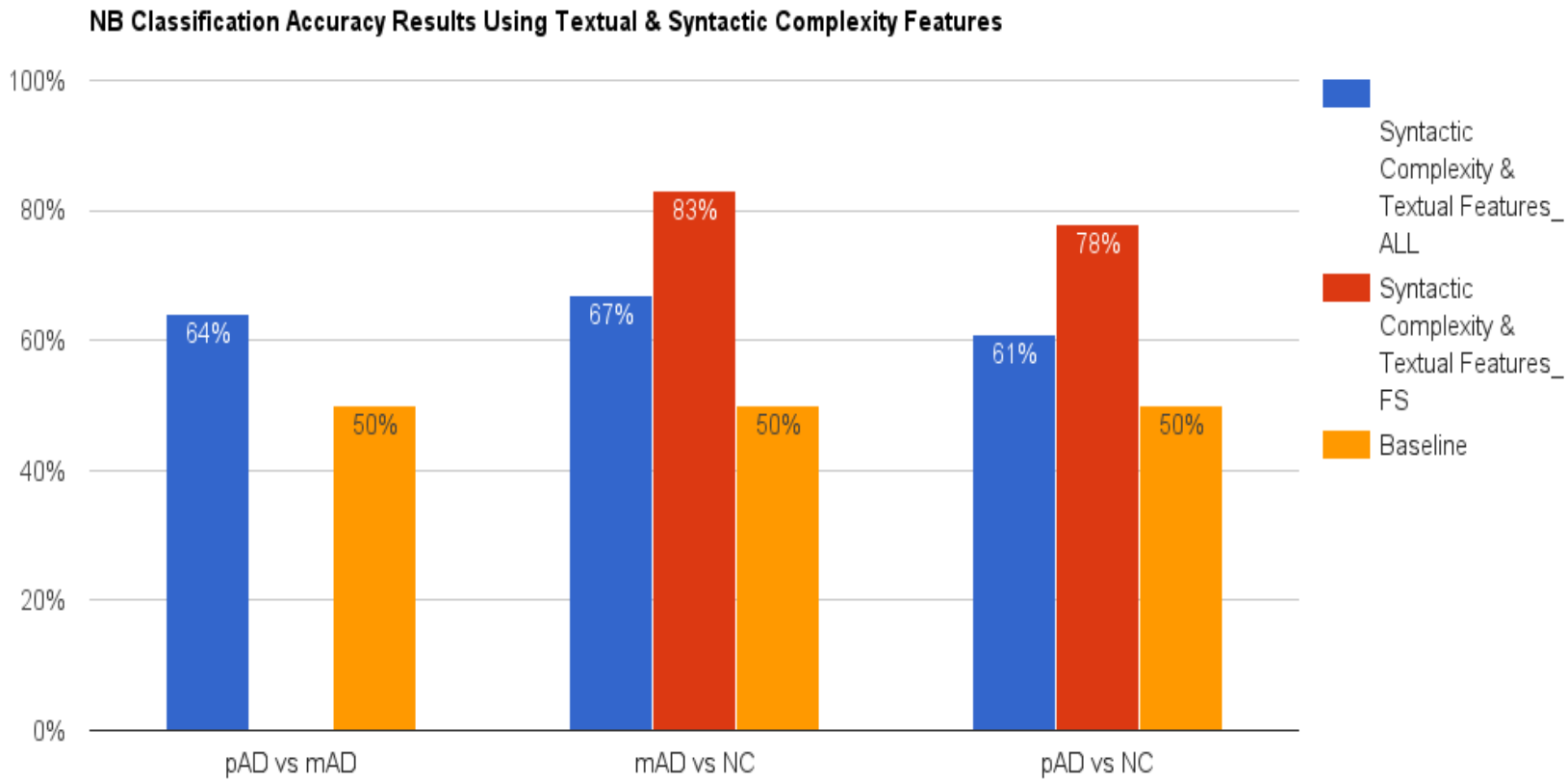
- pAD: ''er it's the boy is um wanting to pinch some biscuits from a top shelf to do so he's standing on a stool which is not a very stable thing to stand on however the girl has got her wits about her because she's got her hand ready to catch the biscuits when the boy has got hold of them providing that he doesn't land on his back and er knock himself out in the meantime the er presumably the mother is getting on with the washing up"
- mAD"looks as if the boy's pinching cakes and the girl's washing up apparently oh there's a flood isn't there yes and the stool is tipping up"

# Classification Accuracy Results I




- pADvsmAD/mADvsNC/pADvsNC: NB > Baseline in most cases (p-values <0.05)
- pADvsmAD/mADvsNC/pADvsNC: NB with Binary_FS or Frequency_FS > NB Binary_ALL, Frequency_ALL (p-values < 0.05)

# Classification Accuracy Results II




- pADvsmAD/mADvsNC/pADvsNC: NB with/without FS $>$ Baseline (p-values $< 0.05$)

# Conclusions

- Pure selected lexical features enhance significantly the distinction of $_{p}$ADsvsNC, $_{m}$ADsvs$_{p}$ADs, and $_{m}$ADsvsNC speech transcripts.
- Both purely lexical and more sophisticated linguistic and textual features offer invaluable insights for the language of each of the two groups $_{p}$AD & $_{m}$AD.
- Practical applications would include the diagnosis or early prognosis of $_{p}$AD & $_{m}$AD diseases.

# List Of Publications

- Journal Articles
  - Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., & Garrard, P. (2014). Features and Machine Learning Classification of Connected Speech Samples from Patients with Autopsy Proven Alzheimer's Disease with and without Additional Vascular Pathology. Journal of Alzheimer's Disease, 42(s3).
  - Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., & Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex, 55, 122-129.

# Acknowledgements


- Dr Samrah Ahmed-Ali, Ladan Raoufian, Yasmin Galbraith
- Leverhume Trust