AlzheTect: Bio-Marker Analysis For Early Alzheimer's Classification

Raidel Hernandez
C.S. Masters Graduation Project
Florida State University

AlzheTect Overview

- **Introduction**
- Data Analyzed
- Data Transformations
- Implementations
- Results
- Conclusion

What is Alzheimer's?

- Most Common Cause of *Dementia* (70%)
  - Affects memory, thinking, social ability
- Neurodegenerative progressive disease
  - Worsens over time
- Individuals Affected:
  - 7% of individuals above 65
  - 20% of individuals above 80

AlzheTect's Goal

- Identify important biomarkers indicating Alzheimer's
- Classify individuals based on the stage of the disease
  - Cognitively Normal (CN)
  - Mild Cognitive Impairment (MCI)
  - Alzheimer's Disease (AD)
  - MCI-to-AD
- Predict when an MCI individuals will convert to AD

AlzheTect Overview

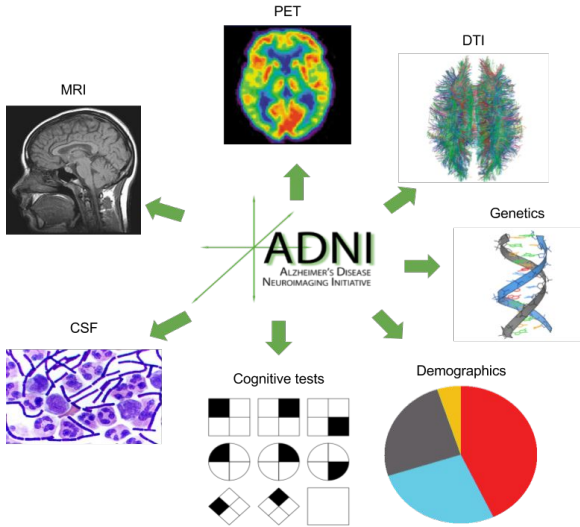
- Introduction
- **Data Analyzed**
- Data Transformations
- Implementations
- Results
- Conclusion

Obtained Data Files

- Data was obtained by the Alzheimer's Disease Neuroimaging Initiative (ADNI) studies:.
  - ADNI-1 (2004)
  - ADNI-GO (2009)
  - ADNI-2 (2011)
- D1 (Training Data Set)
- D2 (Validating Data Set)

Bio-Marker Categories

- Measures of Amyloid-Beta proteins
  - PET, CSF (Cerebrospinal fluid)
- Measures of Damage to nerve cells
  - MRI, FDG-PET scans
- Neuropsychological Tests
  - Cognition, Language, Memory, etc.
- Risk Factors
  - Age, Genetics (APOE4), Gender



Bio-Marker Descriptions

- PET
  - Measures neuron cell metabolism (FDG)
  - Levels of Amyloid-Beta misfolding (plaque)
- MRI
  - Volume of Gray Matter
    - Consists of nerve cells
  - Volume of White Matter
    - Fibers connecting brain cells

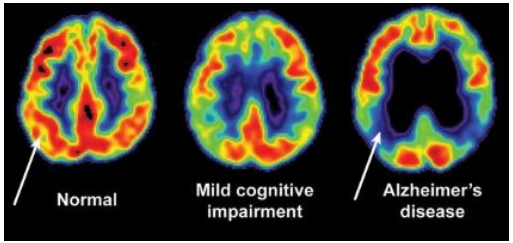


Fig. 1: Cell metabolism in parietal region

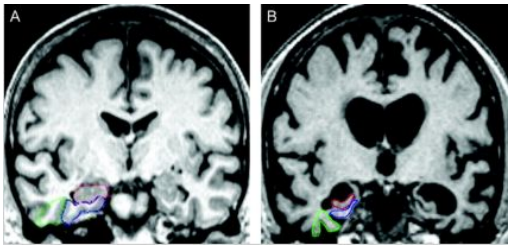


Fig. 2: R-Hippocampus,
B-Entorhinal Cortex, G-Perirhinal Cortex

Bio-Marker Description

- Cognitive Tests
  - Mini Mental State Exam (MMSE)
  - ADAS11, ADAS13
- Risk Factors
  - Age (>65)
    - AD risk doubles every 5 years
  - Genetics (APOE4)
    - AD risk is 10% to 30% greater

Maximum Score	Patient's Score	Questions
5		"What is the year? Season? Date? Day of the week? Month?"
5		"Where are we now: State? County? Town/city? Hospital? Floor?"
3		The examiner names three unrelated objects clearly and slowly, then asks the patient to name all three of them. The patient's response is used for scoring. The examiner repeats them until patient learns all of them, if possible. Number of trials: _____
5		"I would like you to count backward from 100 by sevens." (93, 86, 79, 72, 65, …) Stop after five answers. Alternative: "Spell WORLD backwards." (D-L-R-O-W)
3		"Earlier I told you the names of three things. Can you tell me what those were?"
2		Show the patient two simple objects, such as a wristwatch and a pencil, and ask the patient to name them.
1		"Repeat the phrase: 'No ifs, ands, or buts.'"
3		"Take the paper in your right hand, fold it in half, and put it on the floor." (The examiner gives the patient a piece of blank paper.)
1		"Please read this and do what it says." (Written instruction is "Close your eyes.")
1		"Make up and write a sentence about anything." (This sentence must contain a noun and a verb.)
1		"Please copy this picture." (The examiner gives the patient a blank piece of paper and asks him/her to draw the symbol below. All 10 angles must be present and two must intersect.) 
30		TOTAL

Fig. 1: Example MMSE

AlzheTect Overview

- Introduction
- Data Analyzed
- **Data Transformations**
- Implementations
- Results
- Conclusion

Extracted Classes

- Using DataFrames (*Pandas* Data Analysis Library)
  - Parse ADNI train/test data sets (CSV)
- Entries in Training Dataset:
  - # of measurements: *12,741*
  - # of unique individuals: *1,577*
- Features in Training Dataset:
  - *1,907* (per entry)

Extracted Data

- Diagnosis Classification
  - Provided by ADNI for each individual
  - Classifications Used: CN, MCI, AD, MCI-to-AD
- Months-to-AD (Time) Classification
  - Initial AD diagnosis date - initial MCI diagnosis date (in months)
  - 4 classes were extracted (based on quartiles)
    - 0 : <12 months
    - 1: $12 < X \leq 24$ mos
    - 2: $24 < X \leq 36$ mos
    - 3: >36 mos

Extracted Classes

AD Stage Classifications

Classification	# of individuals
CN	423
MCI	483
AD	332
MCI-to-AD	339

Months-to-AD Conversion Classifications

Classification	# of individuals
<12 mos	76
12 < X ≤ 24 mos	110
24 < X ≤ 36 mos	64
>36 mos	89

Transformations Performed

- Filter DataFrame for a Subset of Features
  - 23 Bio-Markers
  - 59 features (known indicators of AD)
- Binarize Patient Gender
  - 0 - Female
  - 1 - Male

Transformations Performed

- Normalize Data using an Imputer
  - Complete missing/*NaN* values
  - Replace missing values with the mean
- Scaling the Data (Normalization)
  - Mean removal
  - Variance scaling

Feature Selection

- Random Forest Classifier
  - Ensemble of Decision Trees (weak learners)
  - *500* trees
  - *20* max features when considering best split
  - *Gini impurity* measures split quality
  - Cross validation determined the features of importance

Measurement_Type	Field_Name
Cognitive Test	MMSE_bl
Cognitive Test	CDRSB
Cognitive Test	ADAS13
Cognitive Test	ADAS11
Cognitive Test	MMSE
Cognitive Test	RAVLT_immediate
Risk factor	APOE4
PET measures	AV45
Amygdala (L)	LEFT_AMYGDALA_UCBERKELEYAV45_10_17_16
MRI measures	Hippocampus
MRI measures	WholeBrain
Entorhinal Volume (R)	ST83CV_UCSFFSX_11_02_15_UCSFFSX51_08_01_16
Inferior Lateral Ventricle (L)	ST30SV_UCSFFSL_02_01_16_UCSFFSL51ALL_08_01_16 (Volume)
Hippocampus (R)	ST88SV_UCSFFSL_02_01_16_UCSFFSL51ALL_08_01_16 (Volume)
Cuneus Thickness (R)	ST82TA_UCSFFSX_11_02_15_UCSFFSX51_08_01_16 (Average)
Cuneus Thickness (R)	ST82TS_UCSFFSX_11_02_15_UCSFFSX51_08_01_16 (Standard Deviation)
Hippocampus (L)	ST29SV_UCSFFSL_02_01_16_UCSFFSL51ALL_08_01_16 (Volume)
Hippocampus (L)	LEFT_HIPPOCAMPUS_UCBERKELEYAV45_10_17_16
Posterior cingulate thickness (L)	ST109TA_UCSFFSX_11_02_15_UCSFFSX51_08_01_16 (Average)
Medial orbito-frontal thickness (L)	ST39TA_UCSFFSX_11_02_15_UCSFFSX51_08_01_16 (Average)

Diagnosis Classification Data Sets

- Total individuals in data set: 1,577
- Data was shuffled and split as follows:
  - Training - 1,261 (80%)
  - Validating - 316 (20%)

Classification	Training Size	Validating Size
CN	327	96
MCI	395	88
AD	270	62
MCI-to-AD	269	70

Months-to-AD Data Sets

- Total individuals in data set: 339
- Data was split as follows:
  - Training - 271 (80%)
  - Validating - 68 (20%)

Classification	Training Size	Validating Size
<12 mos	59	17
12 < X ≤ 24 mos	92	18
24 < X ≤ 36 mos	50	14
>36 mos	70	19

AlzheTect Overview

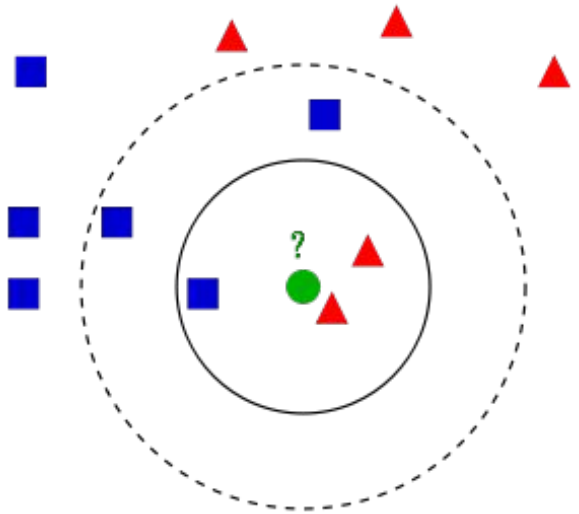
- Introduction
- Data Analyzed
- Data Transformations
- **Implementations**
- Results
- Conclusion

Classifiers Utilized

- K-Nearest Neighbors (KNN)
  - Lazy Learner
- Support Vector Classifier (SVC)
  - Eager Learner
- Deep Neural Network (DNN)
  - Eager Learner

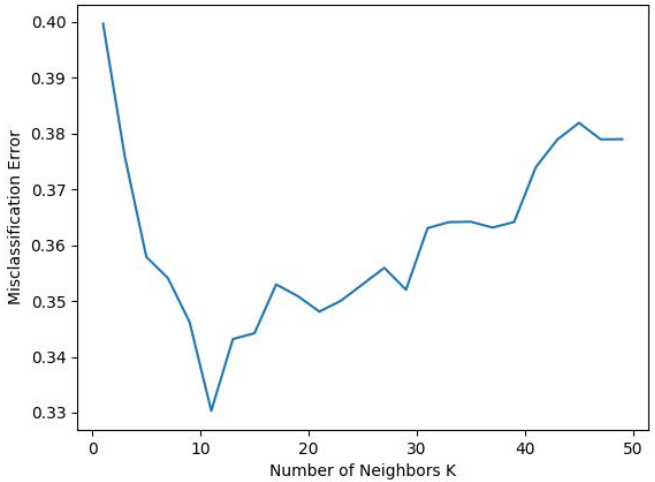
KNN Classifier

- Classifies based on feature similarity
  - How related to K number of neighbors is the data point
- Finding Optimal K
  - 10-fold cross-validation
  - Use K with minimum misclassification error



KNN Parameters (Diagnosis)

- Using K=11
  - Provided the lowest misclassification error
- Distance Metric
  - Euclidean Distance for nearest neighbors
- All neighbors are equally weighed



SVM Classifier

- Defined by a separating line or hyperplane
  - Depending on the number of dimensions (classes)
  - Decision function - distance of X to the separating hyperplane
- Kernel
  - Radial Basis Function (RBF)
  - Maps input vectors into higher dimensional feature space

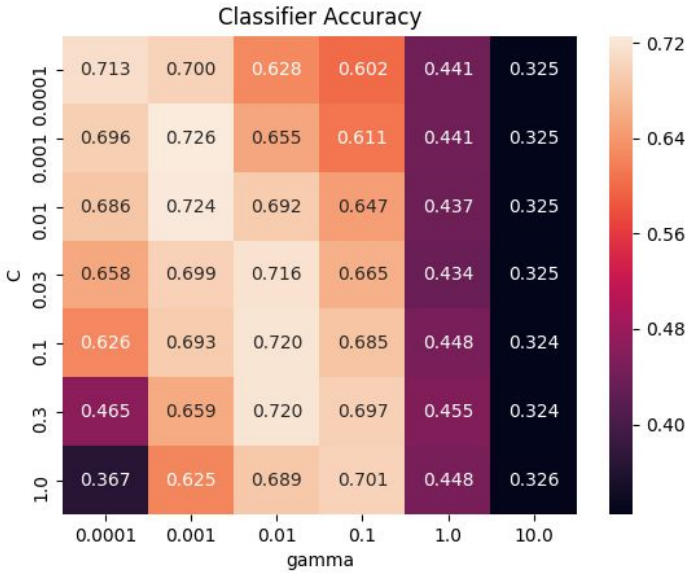
$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$

SVC Boundary Parameters

- Parameter C (Penalty parameter)
  - Larger Value → Small Margin
  - Smaller Value → Large Margin (smooth boundary)
- *Gamma* (Kernel Coefficient)
  - Influence of a Single Training Example
  - High Value
    - Close values affect the decision boundary
  - Low Value
    - Far points affect the decision boundary (smooth boundary)

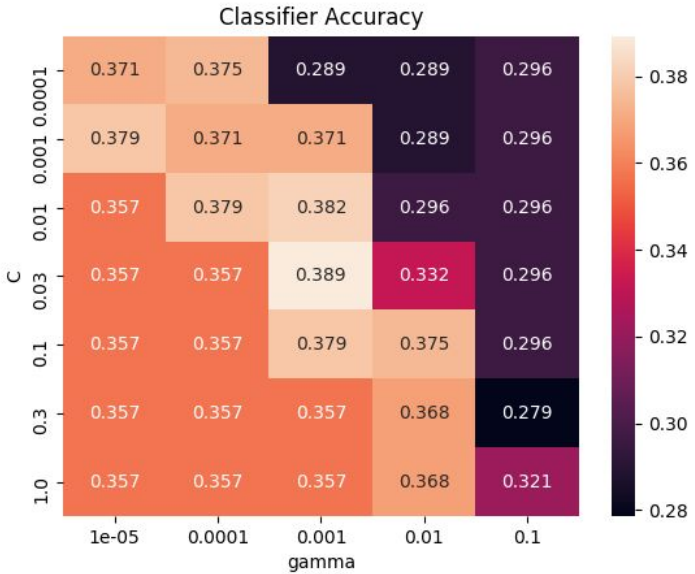
SVC Parameter Tuning (Diagnosis)

- Using a Stratified Shuffle Split
  - Split data into 10 splits
  - 9:1 [Training:Validating]
  - Tested different values for *gamma* and *C*
- Gamma: 0.001
- Parameter C: 0.001
  - C = 1/0.001



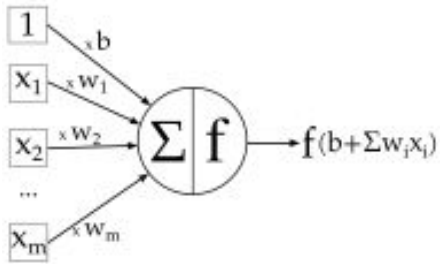
SVC Parameter Tuning (Time)

- Using a Stratified Shuffle Split
- Gamma: 0.001
- Parameter C: 0.03
  - C = 1/0.03



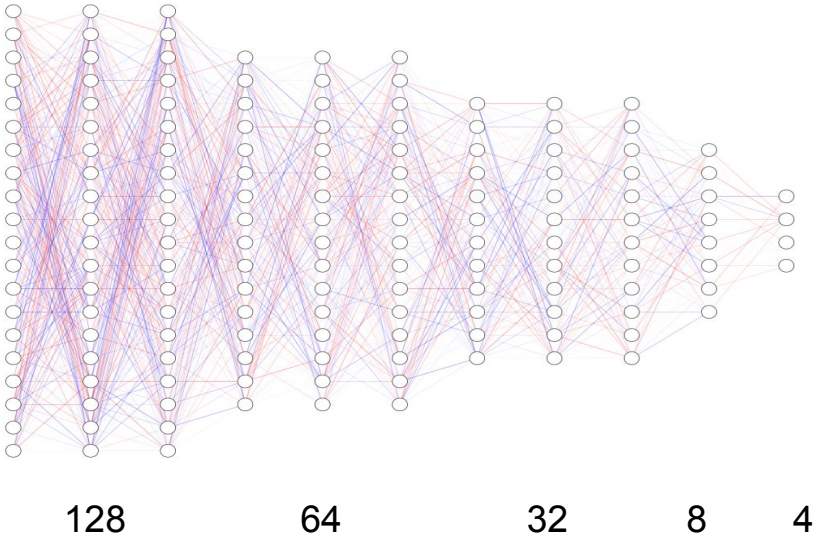
DNN Classifier

- Interconnected Network of Neurons
  - Neurons sum the values of previous neurons + connection weight
- Activation functions used:
  - *relu* : input layer and hidden layers
  - *softmax* : output layer, provides probability prediction



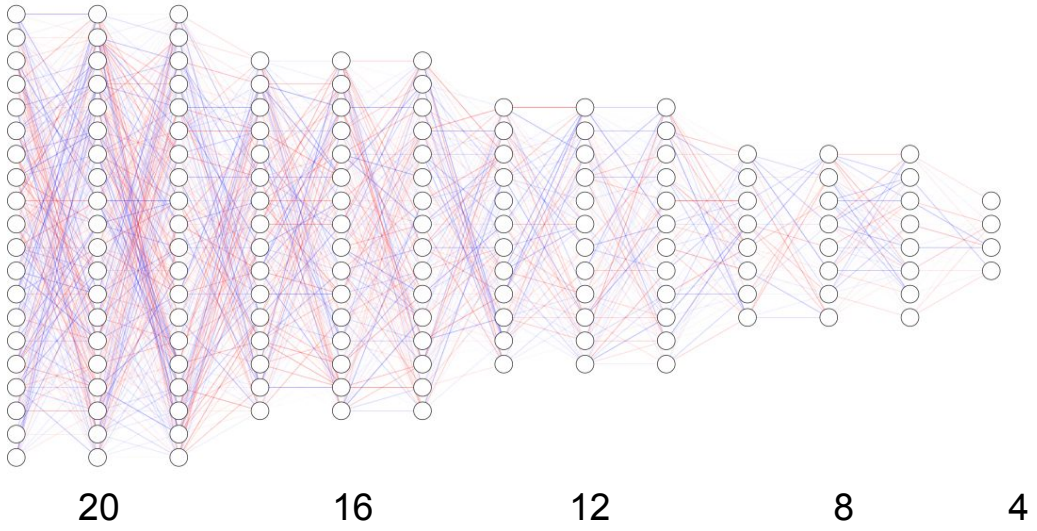
DNN Classifier (Diagnosis)

- Parameter Optimization
  - Adam Optimizer



DNN Classifier (Time)

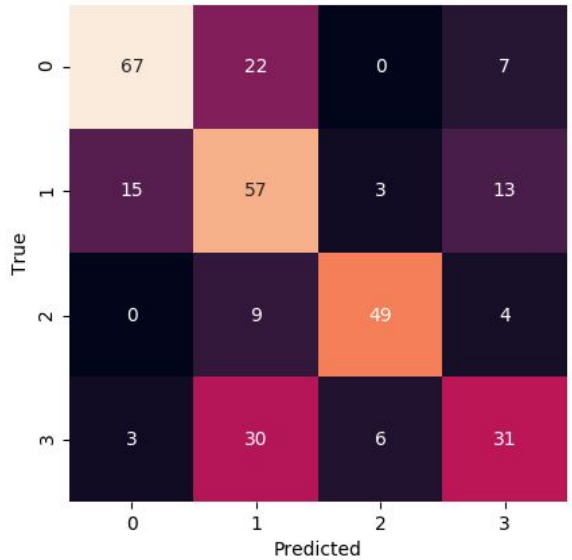
- Parameter Optimization
  - Adam Optimizer



AlzheTect Overview

- Introduction
- Data Analyzed
- Data Transformations
- Implementations
- **Results**
- Conclusion

KNN (Diagnosis)

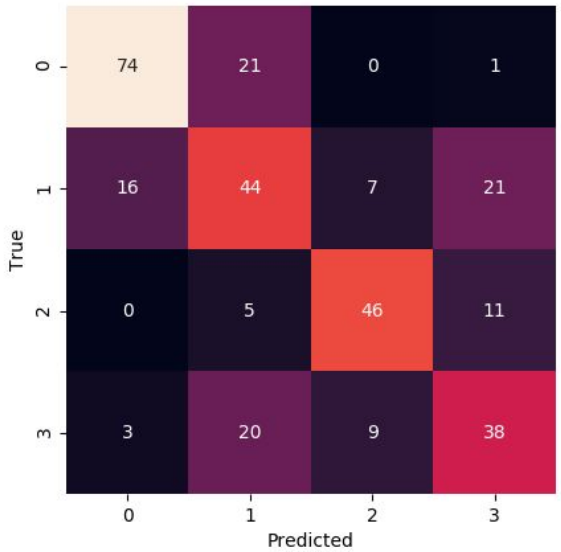


Stage	Correct	Total	% Correct
CN	67	96	70%
MCI	57	88	65%
AD	49	62	79%
MCI-to-AD	31	70	44%

Accuracy:

65%

DNN (Diagnosis)

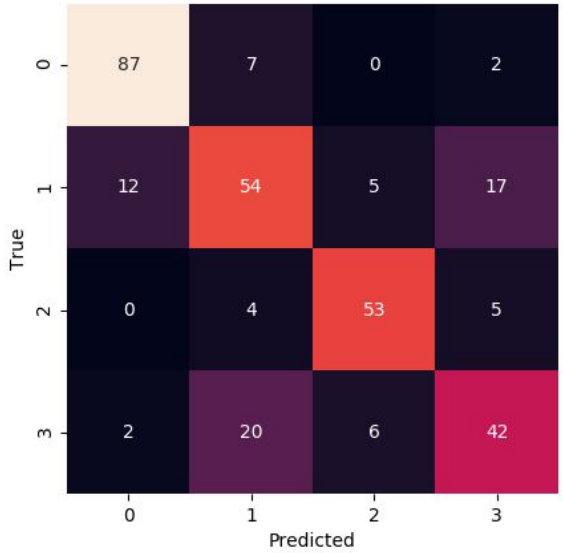


Stage	Correct	Total	% Correct
CN	74	96	77%
MCI	44	88	50%
AD	46	62	68%
MCI-to-AD	38	70	54%

Accuracy:

68%

SVC (Diagnosis)



Stage	Correct	Total	% Correct
CN	87	96	91%
MCI	54	88	61%
AD	53	62	85%
MCI-to-AD	42	70	60%

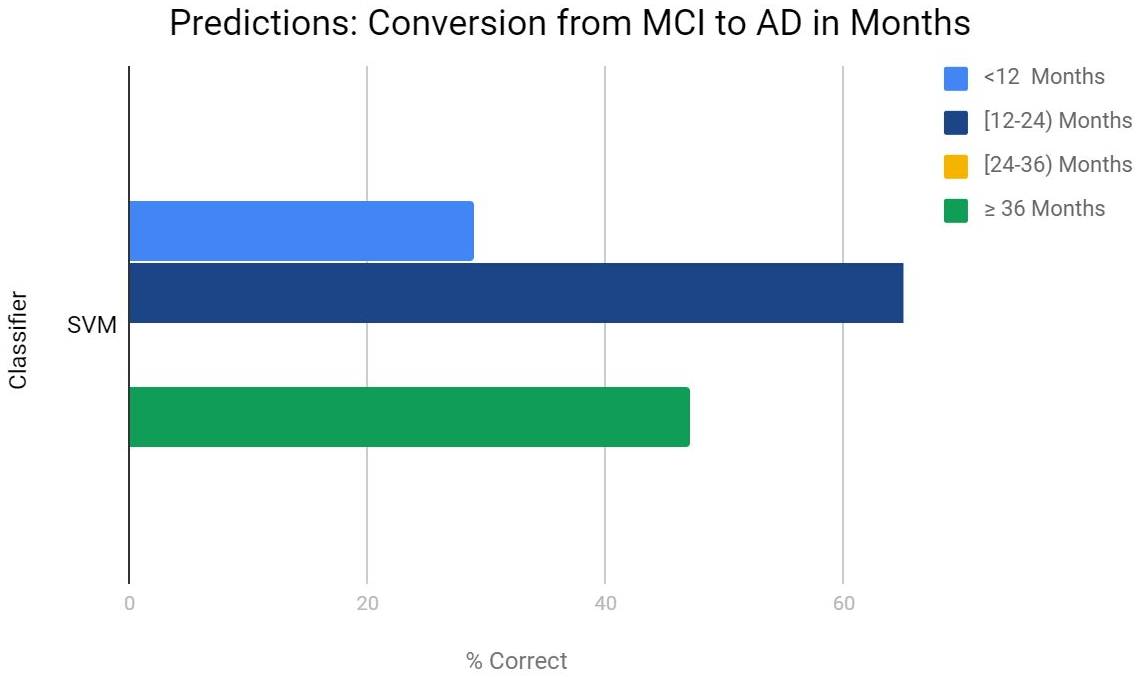
Accuracy:

75%

SVC (Time)

Accuracy:

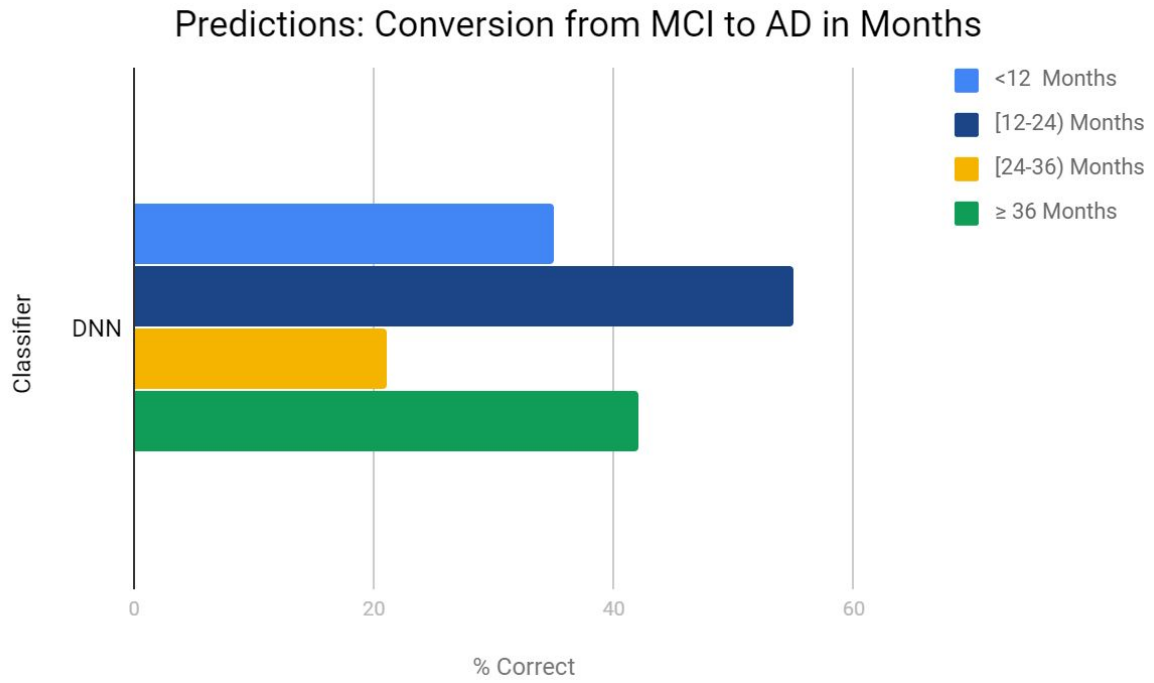
36%



DNN (Time)

Accuracy:

38%



AlzheTect Overview

- Introduction
- Data Analyzed
- Data Transformations
- Implementations
- Results
- **Conclusion**

Conclusion

- Eager learners outperform lazy learners in AD classification
  - SVM outperforms DNN and KNN in AD classification
  - DNN outperforms SVM in MCI-to-AD conversion time classification
- More data sets are required to properly classify MCI-to-AD progression
  - ADNI-3

AlzheTect: Bio-Marker Analysis For Early Alzheimer's Classification

Raidel Hernandez
C.S. Masters Graduation Project
Florida State University