Early Diagnosis of Alzheimer's Disease using Machine Learning on Cognitive Tests

Timothy Duong, Nicholas Klein, Kyle Naddeo, Thai Nghiem, and Lonnie Souder

Advised by Dr. Robi Polikar

Signal Processing & Pattern Recognition Laboratory, Department of Electrical & Computer Engineering, Rowan University, Glassboro, NJ 08028

duongt3@students.rowan.edu, kleinn0@students.rowan.edu, naddeok5@students.rowan.edu, nghiemt2@students.rowan.edu, souderl9@studens.rowan.edu, polikar@rowan.edu

Signal Processing and Pattern Recognition Laboratory

Introduction & Motivation

- ***Project Motivation*** - *There are 5.2 million AD patients in the US, and the disease can only be diagnosed with certainty via an autopsy, yet early diagnosis can add many years to a patient's life, as well as improving the quality of life in those years.*
- ***Project Goal*** - *Diagnose Alzheimer's and cognitive impairment by analyzing data collected from a cognitive test based on clock drawings.*
- ***Datasets***
  - ***Command Clock****: The patient is audibly instructed to draw a clock representing the time "10 after 11".*
  - ***Copy Clock****: The patient is given a clock with the time 10 after 11 and instructed to copy it.*
  - ***Combined****: All data from both command and copy clock are combined for each patient.*
- ***Diagnostic Categories/Classifications***
  - ***SCI*** *- Subtle Cognitive Impairment*
  - ***MCI1*** *- Mild Cognitive Impairment, relating to amnestic or memory issues*
  - ***MCI2*** *- Mild Cognitive Impairment, mixed diagnosis*
  - ***AD*** *- Alzheimer's Disease*
- ***Features*** - *Each feature in the dataset corresponds to: a construction variable related to the how the clock is drawn, a time variable related to how long it takes to draw each construction variable, or a spatial variable related to where each construction variable is drawn.*

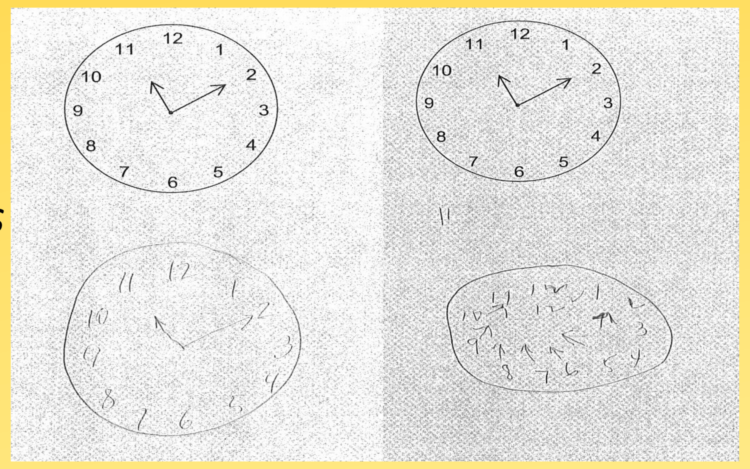


Figure 1: *Raw clock data from the copy dataset.*

Why do we need Feature Selection?

- ***Curse of Dimensionality-*** *As the number of features increases, computational complexity increases exponentially and requires exponentially more data to avoid overfitting.*
- ***Predicting Most Relevant Features-*** *Finding top relevant features helps doctors in understanding what types of clock drawing behaviors are highly correlated with Alzheimer's.*
- ***Reduce Feature Dependency-*** *Many of the clock features are highly correlated with one another which can have a negative effect on training.*

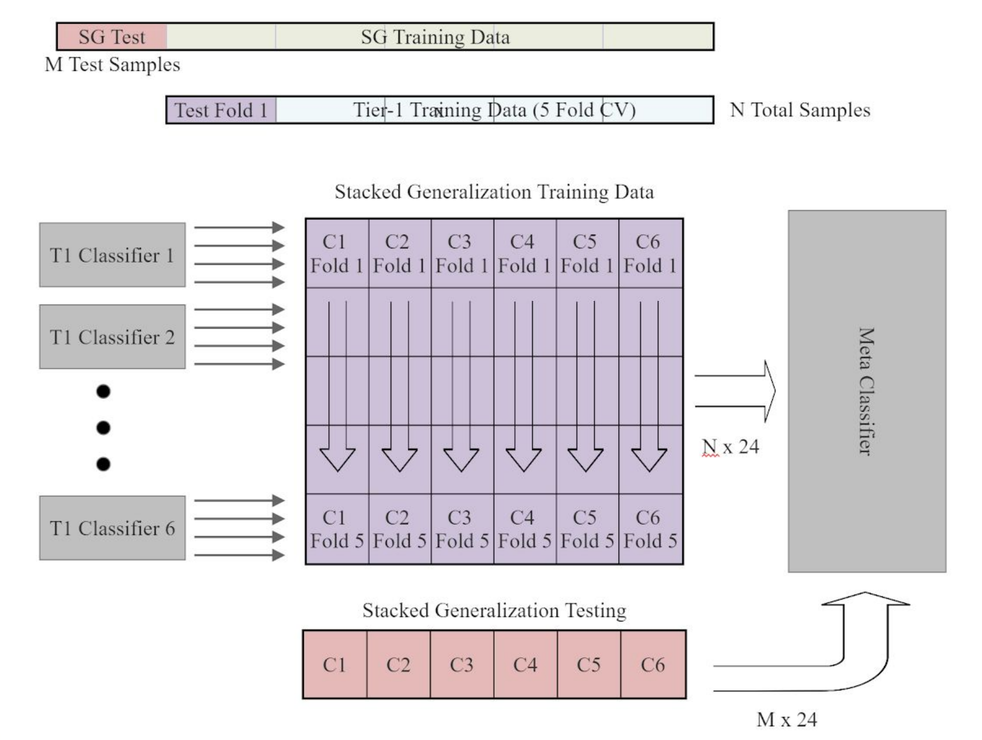


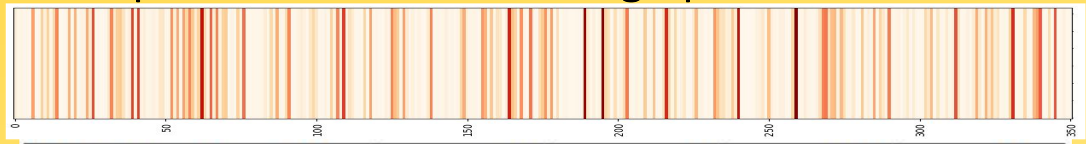
Figure 2: *Stacked Generalization Architecture (4 Class)*

Feature Selection Algorithms

- ***Genetic Algorithm***
  - *A constrained optimization algorithm inspired by Darwin's theory of natural selection*
  - *An individual is represented by its genome; a bit field. The genome represents which features are active in classification.*
  - *The fitness of an individual is the generalization results of a classifier. A simple logistic regression classifier was used for this study.*
  - *The population is a set of individuals from which only the best survive to the next generation. The best individuals have offspring through crossover and exploration is accomplished through mutations.*
- ***Hierarchical Feature Selection***
  - *Using multiple networks in a series to select features, each refining the results of the previous network.*
  - *The features resulting from Information Theory selection are refined using a Wrapper approach.*
  - ***Information Theory Feature Selection***
    - ***Mutual Information :*** $J_{mifs} = I(X_n;Y) - \beta \sum_{k=1}^{n-1} I(X_n;X_k)$
      - *Features chosen based on mutual information between the label and each feature .*
    - ***Minimum Redundancy Maximum Relevancy:*** $$J_{mrmr} = I(X_n;Y) - \frac{1}{n-1}\sum_{k=1}^{n-1} I(X_n;X_k)$$
      - *Checks for redundant features that can be eliminated.*
    - ***Joint Mutual Information:*** $$J_{jmi} = I(X_n;Y) - \frac{1}{n-1}\sum_{k=1}^{n-1}[I(X_n;X_k) - I(X_n;X_k|Y)]$$
      - *Checks feature pairs [Xn & Xk] for single label redundancy.*
    - ***Conditional Mutual Information Maximization:*** $$J_{cmim} = I(X_n;Y) - max_k[I(X_n;X_k) - I(X_n;X_k|Y)]$$
      - *Improves feature scores by pairing low scoring features [Xn] with a second feature[Xk] that maximizes the score.*
  - ***Wrapper Based Feature Selection***
    - *Uses Sequential Feature Selection, recursive greedy search algorithms to select or reject features until optimal set is found.*
- ***Stacked Generalization***
  - *Train a model (Meta Classifier) to learn how to best combine the output of two or more models (T1 Classifiers) trained on the data set.*
  - *Used six T1 classifiers trained on three randomly selected subset of features using RFE, LSVC, MIC.*
    - *Reduce Feature Elimination - Feature ranking with recursive feature elimination and cross-validate the best selection of features.*
    - *Linear Support Vector Classifier - Select features based on the weights of a LSVC.*
    - *Mutual Information Classifier- Estimate mutual information between two random variables and uses entropy estimation from K-nearest neighbor to select features.*

Results

- ***Genetic Algorithm***
  - Genetic Algorithms achieve about 70% accuracy on the 4-class problem using only 50 features or less
    - Improvements have been made to achieve larger search spaces in less time through parallelization



Number of Features	25	50	100	351
Accuracies	57%	69%	63%	57%

Figure 3: *Frequency of Active Genes and Performance of Number of Active Genes*

- ***Hierarchical Feature Selection***
  - For all test cases, the results were comparable to the previous results. However, improvements were found in the SCI vs MCI1 vs MCI2 classification, with an accuracy of 81% (increased 5%).
  - Improvements vs the Information Theory Feature Selection were also found in the 4-class classification.

Feature Selection Method \ Problem Cases	SCI vs MCI1 vs MCI2 (3 classes)			MCI1 vs MCI2 vs AD (3 classes)			SCI vs MCI1 vs MCI3 vs AD (4 classes)		
	Accuracy	*Confidence Interval*	*Feature Used*	*Accuracy*	*Confidence Interval*	*Feature Used*	*Accuracy*	*Confidence Interval*	*Feature Used*
Information Theory **only**	71.64%	± 6.46%	125	75.97%	± 6.19%	50	64.05%	± 4.92%	100
Information Theory **refined** by Wrapper	81.37%	± 5.41%	18	73.52%	± 6.36%	90	68.16%	± 3.01%	73

Figure 4: *Information Theory Vs. Hierarchical result*

- ***Stacked Generalization (SG)***
  - Each T1 Classifiers used a subset of features based on different feature selection algorithms.
  - The Meta Classifier run on 5 folds yielded the results in Fig .5.
  - Using Stacked Generalization yielded 11% increase in performance than individual classifiers for the 4-class problem.

	Performance	Std. Deviation	95% Confidence
SG Score	80.55	6.51	6.53

Figure 5: Results from Stacked Generalization

Conclusion & Future Work

- ***Conclusions***
  - For all classification problems, results in the mid 70% - low 80% are easily attainable with a small network and 100 features selected.
  - Selected features varied between selection algorithms shown in *Fig 7.*

Name	***Feature #***	***Description***
Digit3Angle	66	Angle of 3 on the Command Clock
Digit10Time	130	Time it took to draw 10 on the Command Clock
CopyMH1TotStroke	204	Strokes to draw the Minute Hand on the Command Clock
Digit6NormHT_A	269	Height of Digit 6 on the Copy Clock
Digit10DistCircum_A	313	Distance from Digit 10 to the rim of the clock on the Copy Clock

Figure 6: *Best Features across all tests*

- ***Future Work***
  - Classification accuracies of 80-90% are desired before implementing the clock drawing test in a normal physical.
  - Further analyze how different architectures of SG affect accuracy score such as number of T1 Classifiers, selected features, and T1 architecture.

Figure 7: *Selected Feature Indexes. Top to Bottom: Hierarchical Method Features, Wrapper Method Features, and Genetic Algorithm Features*