Exploring the Screening Power of the Digital Maze Completion Test

by

Dana Mukusheva

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author .... Signature redacted ....
Department of Electrical Engineering and Computer Science
May 26, 2017

Certified by..... Signature redacted ....
Randall Davis
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by.............. Signature redacted ....
Dana Penney
Director Neuropsychology, Lahey Hospital and Medical Center
Thesis Reader

Accepted by .... Signature redacted ....
Christopher J. Terman
Chairman, Masters of Engineering Thesis Committee

Exploring the Screening Power of the Digital Maze Completion Test

by

Dana Mukusheva

Submitted to the Department of Electrical Engineering and Computer Science
on May 26, 2017, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The Digital Maze Completion Test is a novel and unique screening tool for dementia and related cognitive diseases. The test incorporates a combination of a digitizing pen, carefully designed mazes, and sophisticated software. The subject's behavior while solving the maze has potential to reveal the nuances in their cognitive state, which can be used for early diagnosis of impairments such as Alzheimer's disease. In this thesis, we explored the subject's decision making process and planning ability by interpreting and analyzing the relevant data collected by the digitizing pen. We came up with definitions of the associated features that we implemented in the software and extracted from the real-world clinical data. We have evaluated the predictive power of some of the features by applying machine learning classifiers to distinguish the individuals from the various clinical subgroups, such as healthy controls, subjects with Mild Cognitive Impairment, etc. Our key observation is that even a simple subset of the features is quite powerful to perform on par with the traditional screening tools, such as Mini-Mental State Examination. Therefore, we confirmed that the Digital Maze Completion Test is a promising screening tool, the further development and research of which will help to reveal more information about the patients' cognitive conditions.

Thesis Supervisor: Randall Davis
Title: Professor of Electrical Engineering and Computer Science

Thesis Reader: Dana Penney
Title: Director Neuropsychology, Lahey Hospital and Medical Center

Acknowledgments

Foremost, I would like to thank my advisor Professor Randall Davis for his tremendous support, gentle guidance and great patience throughout my work on this thesis. It was an incredible experience learning from him about the concepts of computer science, machine learning, medical applications and life in general. I am genuinely grateful for his motivation and his trust in my abilities to carry out this project.

I would also like to express my gratitude towards my co-supervisor Doctor Dana Penney, from whom I learned immensely about the neuropsychology. Doctor Penney's insightful comments helped me to improve my work and achieve better results.

I would like to thank William Souillard-Mandar for his helpful advice and feedback, as well as his constant willingness to help with my research. I am very grateful to my friends Saadiyah Husnoo and Przemyslaw Pasich for their help with proofreading and editing. Their impeccable writing skills helped me to finish this thesis.

I want to thank my parents and my family for their unconditional love and support. Especially, I want to thank my big brother Daniyar for constantly challenging me and pushing me outside of my comfort zone. I would never be where I am now without his encouragement and faith in me. Last but not the least I would like to thank my beloved fiancé Arman, who was there for me 24/7 during both my high and low moments with his care, love and support.

This work was supported in part by National Science Foundation Award IIS-1404494 and by the Robert E. Wise Research and Education Institution.

Contents

1 Introduction 13

1.1 THink project 13

1.2 Digital Maze Completion Test 15

1.3 Motivation for Digital Maze Completion Test 15

1.4 Contributions 16

2 Related Work 19

2.1 Other maze drawing tests 19

2.2 MMSE and MOCA 20

3 Overview of the Digital Maze Completion Test 23

3.1 Digital Maze Completion Test structure 24

3.2 Software 25

3.3 Data 25

4 Feature Construction 29

4.1 Feature groups 29

4.2 Behavioral features 31

4.3 Additions to software functionality 33

5 Feature Analysis 35

5.1 Machine learning algorithms 35

5.1.1 Experiment setup 35

5.1.2 Baselines 36

5.1.3 Results 38
5.2 Feature importance 40
5.3 Future work 42

6 Conclusion 45

A Feature Names 47

B ROC curves 51

C Test IDs of maze tests used 61

List of Figures

2-1 Example of the Perceptual Maze Test and its solution. 20

3-1 Distribution of the test instances over the primary diagnosis. 28

4-1 Examples of behavioral features: backtracking and hitting the walls. . 32

4-2 Examples of behavioral features: corner cutting and steering towards an alternative choice. 32

4-3 Examples of behavioral features: corridor and deviation from straight path. 33

4-4 Addition to the software interface: speed coloring. 34

4-5 Addition to the software interface: speed graph. 34

5-1 ROC curves for the screening task. 44

B-1 Healthy controls vs. all others 51

B-2 Healthy controls vs. subjects with AD, PD or MCI 52

B-3 Healthy controls vs. subjects with MCI 52

B-4 Healthy controls vs. subjects with amnestic MCI 53

B-5 Healthy controls vs. subjects with AD 53

B-6 Healthy controls vs. subjects with MMSE score ≥ 26 54

B-7 Subjects with MCI vs all others 54

B-8 Subjects with MCI vs. subjects with AD or PD or healthy controls . 55

B-9 Subjects with MCI vs subjects with MMSE score ≥ 26 55

B-10 Subjects with MCI vs. subjects with AD 56

B-11 Subjects with amnestic MCI vs. all others 56

B-12 Subjects with amnestic MCI vs. subjects with AD or PD or healthy controls 57

B-13 Subjects with amnestic MCI vs subjects with MMSE score ≥ 26 . . . 57

B-14 Subjects with amnestic MCI vs. subjects with AD 58

B-15 Subjects with AD vs. all others 58

B-16 Subjects with AD vs. subjects with MCI or PD or healthy controls . 59

B-17 Subjects with AD vs subjects with MMSE score ≥ 26 59

List of Tables

3.1 Distribution of test instances across sites and difficulty levels 26

5.1 Sample set sizes and classes for each classification experiment. 37

5.2 Results for machine learning classifiers: distinguishing healthy individuals from various subgroups. 39

5.3 Results for machine learning classifiers: distinguishing subjects with MCI from various subgroups. 41

5.4 Results for machine learning classifiers: distinguishing subjects with Alzheimer's disease from various subgroups. 42

Chapter 1

Introduction

Every 66 seconds, someone in the United States develops Alzheimer's disease. This number will drop to 33 seconds by 2050, as estimated by Alzheimer's Association in the report from 2017 [1]. Unfortunately, the disease does not yet have any effective treatment that would reverse its symptoms. Nevertheless, an early diagnosis of dementia is critical, as it could give affected people a chance to slow down the progress of the disease, prepare for its outcomes and come up with the caregiving-related plans. Having the early detection of dementia in mind, in this thesis project, we have explored the screening power of a novel and unique tool, the Digital Maze Completion Test. This simple pen and paper drawing test is able to pick up the subtle details of subjects' behavior and infer their cognitive status. Our work is a beginning of the exploration of the Digital Maze Completion Test, which shows enormous potential to effectively measure functioning of various cognitive domains, such as memory, learning and planning. Nonetheless, even with a simple set of extracted measurements, the test is powerful enough to successfully screen the patients and differentiate between certain types of diagnoses, as we present in this thesis.

1.1 THink project

The Digital Maze Completion Test is part of the THink project [7], a collection of neuropsychological tests that currently includes the Digital Clock Drawing Test [18], which is a digitized and operationalized form of the conventional pen and paper Clock Drawing Test [2, 10][1]. In the conventional Clock Drawing Test, a subject is first asked to draw a clock that shows the specified time (command clock) and then to copy another pre-drawn clock (copy clock). The results are then interpreted and manually evaluated by the medical staff based on one of the existing scoring systems (a summary of scoring systems is given in [8]). While the traditional Clock Drawing Test is cost- and time-efficient, as well as relatively accurate in detecting dementia [4], its utility is diminished by the need for subjective judgement as well as ambiguous scoring criteria and mistakes [18].

The Digital Clock Drawing Test solves these problems through the use of a digitizing pen and novel software [7], allowing it to interpret the details of the drawing. In the Digital Clock Drawing Test, the subject completes the drawing tasks with a digitizing pen that records geometrical and temporal data. The software analyzes the pen data and is able to recognize and classify the clock elements, such as individual digits, clock hands, etc. [17]. The software displays the results of the classification in a user-friendly way. Thus, the clinicians avoid the laborious and error-prone sketch interpretation process.

Researchers have also explored the potential of applying machine learning algorithms to the data from the Digital Clock Drawing Test [18]. They developed a large set of features and applied various algorithms to perform a screening task (determining if the subject is healthy or not) and a diagnosing task (determining which one of the specified impairments a person might have). The authors evaluated the predictive power of the new automated Clock Drawing Test and came to the conclusion that it produces relatively accurate, reliable and robust results. For the purposes of our research, we have extended the software from the Digital Clock Drawing Test to work in a new context. Moreover, we explored what useful data can be extracted from the maze drawings and which behavioral phenomena our test is able to detect. Finally, we evaluated the screening power of the test by applying machine learning classifiers to the extracted data.

[1] The THink project is covered in part by the US Patent 8740819.

1.2 Digital Maze Completion Test

The Digital Maze Completion Test is a system that first requires a subject to find their way in a series of mazes by tracing the line from the start to the end locations. Each maze is carefully designed with a special structure that challenges various cognitive domains such as memory, locomotor skills, planning abilities and foresight. The test is conducted as follows: a subject takes the test on paper with a digitizing pen, which looks and feels like a regular ball-point pen, but also records timestamped spatial data. In each test instance, there are three mazes that the subject has to complete: a calibration maze, a choice maze and a no-choice maze. The calibration maze is a simple "corridor" that requires a subject to draw a straight line. The choice and no-choice mazes are more convoluted and have identical solutions, but the choice maze has decision points at which the subject can deviate from the correct path. The data from the pen is then transferred to special software that matches the drawings with the corresponding maze layout. Mazes differ in difficulty level, from easy through intermediate to advanced. The software analyzes and displays various aspects of the drawing process, such as changes in speed, deviating from the correct path (henceforth, *solution path*), or hitting the walls of the maze. We describe our system in detail in Chapter 3.

1.3 Motivation for Digital Maze Completion Test

While there exists a variety of cognitive screening tools for dementia that are commonly accepted in the clinical use, our system provides a few advantages over them. First, the data interpreting software makes the Digital Maze Completion Test free from subjective judgement and human errors, as well as from the laborious scoring process, all of which are inevitable in the manual evaluation process. Moreover, the software component at its current state is able to detect a certain set of behaviors and has a potential to expand this set even further, which opens up more research opportunities. Given that our test is sensitive to very subtle details of the subject's behavior while solving the maze, such as slight deviations from the norm (or from what healthy individuals do), we believe our test may enable detecting dementia in its early stages. Particularly, it can recognize the subjects with Mild Cognitive Impairment, a precursor to dementia [11]. We have studied and reported the diagnostic power of the Digital Maze Completion Test in Chapter 5.

1.4 Contributions

The Digital Maze Completion Test is a new and unique tool that does not have any similar reported counterparts. The focus of this thesis project is to explore the screening power of the test and evaluate its ability to predict the subject's cognitive status. We did this by exploring the features that could be extracted from the maze drawings, using them in state-of-the-art machine learning classifiers, and by improving the software by adding new functionality relevant to the feature analysis.

The main contributions of this thesis project are:

- **Feature construction.** We have defined several features that capture the subject's behavior while drawing. We defined what is considered to be a mistake, such as making or nearly making a wrong turn at the decision point, touching or going through the walls of the maze, clipping the corners while in a rush or not being able to draw a straight line. We came up with precise definitions for such mistakes and implemented them in code.

- **Diagnosis prediction.** We chose a simple subset of features that we used in state-of-the-art machine learning classifiers. We experimented with the tasks for the classifiers, such as separating healthy individuals from the impaired people, separating individuals with amnestic MCI or Alzheimer's disease from each other, from healthy individuals and from the rest of the people in our dataset. We compared our results (precision, recall and f-scores) with predictions of the MMSE scores and found out that a classifier that uses even a small number of features (listed in Appendix A) is able to outperform the MMSE scores.

We also looked at the feature rankings obtained from the trained classifier and noticed that all three parts of the test played a role in the classification process. Particularly, we saw that even the calibration maze, despite its considerable straight-forwardness, does have some screening power.

- **Improvement of the software interface.** We have improved the software component of the Digital Maze Completion Test. It is able to compute and display the behavioral features, as well as show pen speed analysis as a colored drawn path or as a line graph with adjustable sliding window size. Pen speed analysis plays an important role in understanding the subject's behavior while solving the maze because correlating the instances of slowing down and speeding up with the particular locations in the maze can signify one or more behavioral phenomena.

Chapter 2

Related Work

There is a variety of screening methods for dementia and related cognitive diseases, which come in different forms, such as questionnaires or sequences of drawing, spelling or counting tasks. Each of them focuses on different cognitive domains [5]. In this section, we discuss the studies and the research related to our project. We describe other maze drawing tests as well as the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA).

2.1 Other maze drawing tests

To our knowledge, our test is unique in the sense that we use the digitizing pen and the software to interpret and analyze the maze drawing. However, the concept of using mazes as a cognitive screening tool is not new. One such test, the Porteus Maze Test [16], measures the subject's intelligence level, planning capacity and foresight. The subject is asked to complete a series of mazes, which takes from 15 min to 60 min. The test is scored based on the number and character of the errors the subject has made. However, there are no reported studies or experiments where the Porteus Maze Test was specifically used to detect dementia and its early signs.

A second maze-based cognitive screening test is called the Perceptual Maze Test [6]. The structure of this test is different from a conventional maze: the subject is asked to connect the dots sitting on the intersections of a lattice given the certain

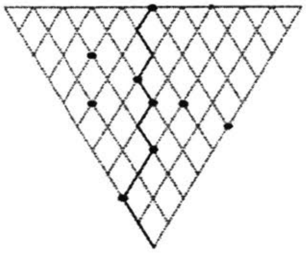


Figure 2-1: Example of the Perceptual Maze Test and its solution.

constraints. The subject moves from top of the maze to bottom, from left to right. He or she cannot move upwards or to the left. The subject cannot cross his or her own path. The goal is to connect as many dots as possible while moving in the allowed directions. Figure 2-1 shows an example of the Perceptual Maze Test. The difficulty of the test is a function of its structure and the pattern of the dot locations. Psychologists use the Perceptual Maze Test to measure the level of intelligence, as for example, in [3], where the test was used to identify subjects with brain damage. However, we are not aware of any recent studies using the Perceptual Maze Test to identify subjects with dementia.

2.2 MMSE and MOCA

Other cognitive screening tests come in the form of a questionnaire. Like the Digital Maze Completion Test, they are time-efficient and simple to administer and score. For such tests, a subject is asked to answer a series of questions or perform certain tasks, such as to name objects, count from one number to another, and so on. Two such tests, the Mini-Mental State Examination (MMSE) [9] and the Montreal Cognitive Assessment (MoCA) [12], are used at the sites where the Digital Maze Completion Test was piloted; therefore, we use the results of these tests as a baseline with which we compare the predictions of our machine learning classifiers.

The MMSE consists of 30 questions, the first 21 assess aspects of cognitive health, such as orientation, registration, attention and calculation, and then 9 questions asses recall and language. The test successfully detects the presence of cognitive difficulties using the thresholds that have been established over years of experience with the test. In our work, we used a score of 28 and higher to identify a test taker as a healthy non-demented subjects and a score between 26 (inclusive) and 28 (exclusive) to identify a subject with Mild Cognitive Impairment (MCI). Our threshold for classifying a subject with dementia is higher than the commonly used score of 24 because the threshold should be adjusted roughly by 2 points for highly educated individuals [13]. In our population, the average number of years of education is 16 (±2.5), hence the use of a cutoff that is higher than the traditional cutoff.

MoCA is a similar screening tool with varying tasks that tests the subject's memory, language, orientation, attention, concentration and visuospatial abilities. The main difference between MMSE and MoCA is that the MoCA was designed to pick up subjects with MCI, who tend to score the same as cognitively normal individuals on the MMSE. Subjects with MCI are at higher risk because it is considered a transitional state to Alzheimer's disease, hence the importance of identifying it early. We used a score of 26 and higher to mark a subject as a healthy control.

In our data set, not all subjects have both MoCA and MMSE scores. We used these scores as a baseline to evaluate the relative performance of our classifiers; therefore, in order to have a reliable baseline, we used MMSE scores for those subjects who did have them and converted MoCA scores to MMSE scores for those subjects lacking an MMSE score. We used a conversion table from [19], where the authors developed a conversion from MoCA to a relatively accurate MMSE score.

Chapter 3

Overview of the Digital Maze Completion Test

The Digital Maze Completion Test is a cognitive screening system that includes multiple components. It is targeted to detect cognitive impairments among adults of age 55 and older. In the test a subject is asked to complete a series of mazes on a paper with a digitizing pen. The instructions specify that the subject should not lift the pen while attempting to solve the test. We use an Anoto Inc. DP-201 pen that records and stores timestamped coordinates indicating the position of the pen point every 12 ms. The data collected by the pen is downloaded to a computer and analyzed by the software component. The software enables users, or more precisely, clinicians, to play back the drawing process, identify and display various features (e.g., hitting the maze walls while drawing), and extract a variety of numerical measures that are useful in assessing the subject's cognitive status. We used state-of-the-art machine learning classifiers to predict whether the subject is likely to have a cognitive impairment, for example, Alzheimer's disease. In this section, we describe the drawing component of the test, present the most important software functionality, and give an overview of the data we have worked with.

3.1 Digital Maze Completion Test structure

Each test has three parts: a calibration maze, a choice maze and a no-choice maze. There are also versions of the test at three levels of difficulty.

The building blocks of each maze are square tiles with a side length of 0.8cm. Each tile can be surrounded by up to three walls. Each maze has a start tile and a goal tile, and there is only one correct solution path that leads from the start tile to the goal tile.

- the **calibration maze** is a simple straight lane. The calibration maze requires the lowest cognitive load among the three mazes and is used both to let the subject get accustomed to the test and to measure the subject's basic locomotor skills. We suspect that this maze does not have any diagnostic power since it is used for a simple and straightforward task.

- the **choice maze** contains multiple decision points (or decision tiles) where the subject must choose which way to continue. The decision points vary in difficulty: some of them lead to dead-ends in fewer steps than others, some require choosing between two alternative paths, while others require choosing among three alternative paths. Some decision points have an embedded choice, in the sense that on the incorrect path there are additional decision points before hitting a dead end. The difficulty of the decision points in the choice-maze distinguishes the three difficulty levels of the Test. The "easy" Test has the fewest decision points with no embedded choices, the "intermediate" Test has more decision points with some embedded choices and the "advanced" Test has the most decision points with even more embedded choices. The varying difficulty of the decision points places a different cognitive load on a test-taker; thus, the behavior around and at such points can potentially help us to evaluate the subject's planning abilities and foresight. We believe that this maze will be the important one in identifying subjects with cognitive impairments.

- the **no-choice maze** is very similar to the choice maze and has exactly the same solution path. However, all of its decision points have only one (correct) choice. Both choice and no-choice mazes make equal demands on the drawing side and motor skills. Comparing various metrics from both mazes can give us insight about memory and planning processes at the decision points.

3.2 Software

The software component of the Digital Maze Drawing Test is an extension of the existing software written to interpret the clock sketch data, presented in [7]. The timestamped data in the pen is grouped in strokes, each stroke belongs to one of the three mazes in a single test instance. The original software allowed users to play back the drawing process at varying speeds, get the geometric coordinates of each point and identify different phenomena in the drawing. Users can also see when and where the subjects touched the maze walls, went across them, made an incorrect decision and backtracked, lifted the pen, etc. However, the software did not have all the functionality necessary to study the behavior around decision points and gain insight to the intention behind one or more phenomena. Therefore, the software was modified and improved to facilitate more thorough analysis, as described in Section 4.3.

3.3 Data

The Digital Maze Drawing Test data has been collected under the THink project, which has several participating sites, including Lahey Hospital & Medical Center and Massachusetts General Hospital (MGH). There are 353 test instances collected from Lahey since 2014 that belong to 108 subjects. There are 206 test instances from 69 subjects from tests at MGH during 2016 and 2017. Among the 559 available tests, 236 belong to subjects that are considered to be "cognitively normal", which were collected either from people who were not subjects or who were not diagnosed with any disorders that have cognitive impact.

Site	Easy	Intermediate	Advanced	Total
Lahey	120	115	118	353
MGH	68	69	69	206
Combined	188	184	187	559

Table 3.1: Distribution of test instances across sites and difficulty levels

The subjects that completed our Maze Test had been diagnosed with 30 different impairments. All subjects have a primary diagnosis, and possibly a secondary and tertiary diagnoses, each labeled with a level of confidence from low to moderate to high. The distribution of the test instances over the primary diagnoses is shown in Figure 3-1. We have used the consensus diagnoses supplied by the clinicians as ground truth in our classification task.

In one set of the experiments, we used our entire sample set. For another set, we only kept four major groups of subjects who had been diagnosed with Alzheimer's disease, Mild Cognitive Impairment (MCI), Parkinson's disease, or healthy controls, for a total of 504 test instances. The subjects with these diagnoses are a part of the study and form a majority of the samples in our data set. Moreover, all of them have clear and well-defined diagnosis. For instance, the subjects with Alzheimer's disease do not have a vascular component to it, which rules out other possible explanations for their memory related problems. Some of the subjects with Parkinson's disease have their diagnosis confirmed through DaT scans. The subjects from the aforementioned group can be clearly separated as they exhibit different symptoms.

In order to increase the set size after filtering out the rest of the samples, we have used other test instances whose secondary diagnosis was one of those mentioned above with at least a moderate level of confidence, which gave us 12 more test instances. These additional samples have Alzheimer's disease (5 tests), unspecified MCI (3 tests) and Parkinson's disease (4 tests) as their secondary diagnoses.

We are particularly interested in detecting the signs of the memory impairment disorders, which includes amnestic MCI and Alzheimer's disease. Diagnosing amnestic MCI, an intermediate stage between healthy aging and dementia [15], is not an easy task due to the subtlety of its symptoms. Subjects with amnestic MCI experience memory problems, but still preserve general cognitive health and are able to carry on with daily life activities. This makes them hard to distinguish from elderly healthy controls. The commonly used screening tests, such as MMSE, do not clearly distinguish early demented individuals from cognitively normal individuals, as shown in [20]. We use the performance of the MMSE as baselines and aim to achieve better results than it does. The results of our experiments are presented in Section 5.1.

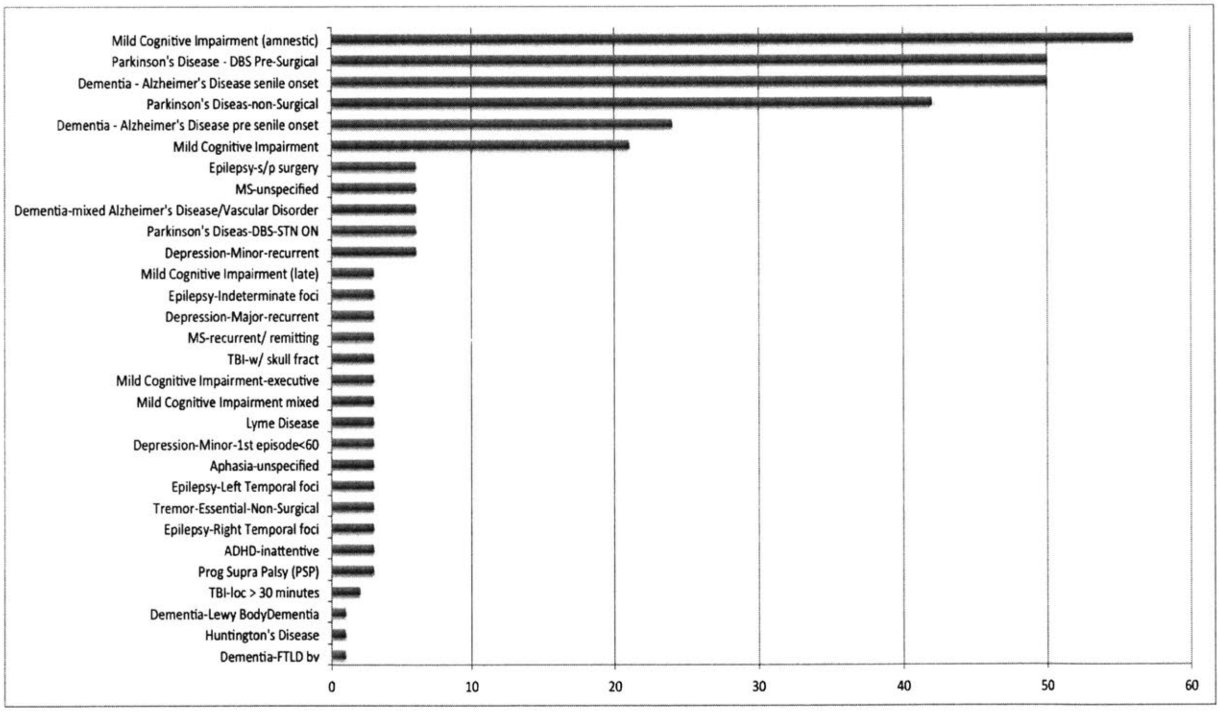


Figure 3-1: The majority of the test instances are from subjects diagnosed with MCI (86 samples), dementia (80 instances) or Parkinson's disease (98 instances). The MCI group includes its various forms, such as amnestic MCI (56 samples), executive MCI (3 samples), mixed MCI (3 samples) and unspecified (24 samples).

Chapter 4

Feature Construction

This chapter gives a general overview of the features that can be extracted from the Digital Maze Completion Test and the changes to the software that we made in order to better understand the subject's behavior during the drawing. Understanding behavior and being able to see the subtle changes in the drawing process is crucial for analyzing a subject's decision making process and can be used to develop new features in the future.

4.1 Feature groups

Currently our software extracts 455 features from each of the mazes (choice maze, no-choice maze and calibration maze), although some features are not defined for the calibration maze (e.g., time in decision tiles). Some of the measures are defined with respect to geometrical location in the maze. For example, we divide the maze into various subsegments and compute the feature values for these subparts of the maze.

We grouped the features into six major categories:

- **Maze completion and number of strokes**. We checked whether the subject has completed the drawing. To get the value for this feature, we checked whether the solution path and the one drawn by the subject path overlap almost entirely. For the calibration maze, we wanted each tile to be visited/marked. For the choice and no-choice mazes, the drawn path might be shorter than the solution path if the subject cut the corners or went through the walls, which results in their skipping some of the tiles from the solution path. In the choice maze, the drawn path might also be longer than the solution path if the subject made the wrong choice at the decision tile and had to backtrack. The maze is considered complete if the subject's path included the goal tile and all but up to n (in our case, $n = 5$) tiles from the solution path. This way we take into account the skipped tiles. The longer paths with extra tiles do not influence this metric.

We also recorded the number of strokes (i.e. pen down to pen up) the subject made. Ideally, the entire path should be one stroke, as the instructions specify that the pen should not be lifted from the paper. However, subjects at times lift the pen, especially when they make a wrong choice at the decision tile.

- **Time measures**: We measured the total time spent on solving each maze and the time spent drawing the path. These values might differ if the subject paused while drawing. We also recorded how much time the subject spent drawing outside of the solution path, at the decision tiles, at the turn tiles and outside of the maze. For the choice and no-choice mazes, we also record the duration of time at the start tiles. We believe that some people pause before attempting to solve the maze because they are planning their path.

- **Path-length measures**: We measured the total length of the drawn path, as well as the length of the path that is outside of solution path. We have also measured the distance of each point from the nearest wall. The idea behind this feature group is that the impaired people might have longer path-lengths due to making incorrect choices at the decision tiles or might not be able to maintain the constant distance from the walls. For example, for people with a tremor, it would be harder to keep the drawn line equidistant from the walls.

- **Speed measures**: For all three mazes and their segments, we extract pen speed as an array of values and compute statistics like mean, median, maximum, minimum and standard deviation. In order to avoid noisy data, we smooth the speed values signal by using a sliding window of size 10 points.

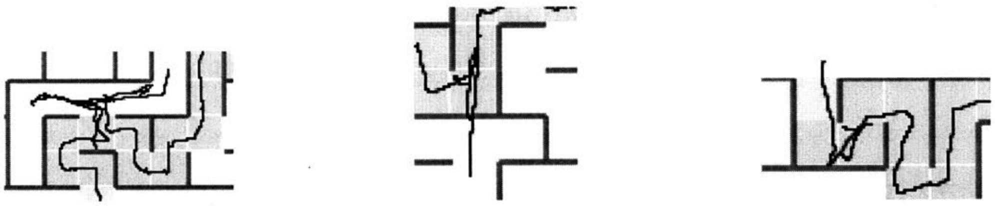
4.2 Behavioral features

Behavioral features attempt to capture different phenomena in subjects' behavior and their mistakes and errors. These features do not have a fixed number of occurrences, as each individual makes a different number of mistakes of various types. Therefore, in the software, we set a cap of how many instances of each behavioral feature we extract. Currently, we set the cap to be 15, as we observed from our sample set that the number of mistakes tends to be less than 10.

One of the most obvious mistakes occurs when a subject makes a wrong choice at a decision tile, which can be a sign of deteriorating planning abilities. The severity of the mistake can be quantified by several features, for example the time the subject took from making a wrong choice until getting back on the solution path. We believe that for healthy people, making a wrong choice is unlikely, but if they do make a wrong turn, they tend to correct themselves and turn to the solution path almost immediately. Figure 4-1a shows an example of backtracking from incorrect decisions.

Another set of behavioral features includes the cases when the subject touched the walls. We consider two scenarios: in the first one (wall collision), the subject went tangential to the wall but did not cross it and in the second one (wall penetration), the subject crossed through the wall and reached the tile beyond it. We suspect that such behavior is a strong sign of poor locomotor skills, and hence will be useful for the diagnosis. In Figure 4-1c, we can see that the subject touched the wall but remained on the tile, which is an example of the wall collision. In Figures 4-1b the subject crossed through the wall, which is an example of the wall penetration.

Sometimes a wall penetration case is not necessarily a cause for concern, but simply a sign that the subject was in a rush. Some people tend to cut corners during the turn in order to get to the goal faster. We categorized such phenomena into a separate feature and called them "corner clips." In Figure 4-2a we give an example

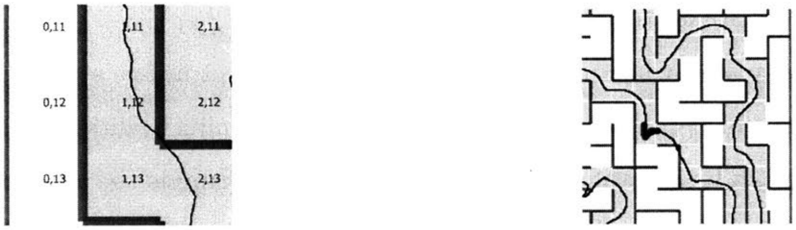


(a) In this maze snippet, the subject deviated from the solution path (colored in light blue) and made an incorrect choice at the decision tile (CIN0639236855).

(b) In this example, the subject went through the wall (CIN0639236855).

(c) The subject only touched the wall but remained on the solution path (CIN0639236855).

Figure 4-1: In this figure we demonstrate the cases of backtracking after deviating from the solution path and the cases of hitting the walls.



(a) Corner clips

(b) Steering towards incorrect decision (CIN1744471031)

Figure 4-2: In (a), the person hit the walls while cutting the corner and staying on the solution path. In (b), the person was drawn towards an incorrect decision but did not deviate from a solution path and made a correct decision.

of a corner being clipped.

While backtracking is an obvious case of making an incorrect decision, it is possible to capture an intent to make an incorrect choice followed by self-correction and resuming on the solution path. We call such a feature "steering", as the subject "steers" towards an incorrect direction when encountering a decision tile, as seen in Figure 4-2b.

Another behavioral feature is related to the structure of the maze: we identify the chunks of the solution path that form a straight path, which we call "corridors". We want to measure the subject's ability to draw a straight line through the corridor. Failing to do so might be a sign of a tremor or other problems with locomotor abilities. We check how much a subject deviates from the line drawn from the first point to

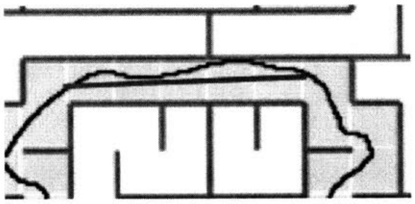


Figure 4-3: In this maze snippet, the subject is on the straight path, which we call a "corridor". The red line shows the optimal path given the start point and the final point of this chunk of the path. However, the subject significantly deviates from the straight line (CIN1476368667).

the last point of the "corridor" (Figure 4-3).

4.3 Additions to software functionality

The existing software inherits its basic stroke rendering functionality from the Digital Clock Drawing Test. It was also adapted to work with mazes and show basic analysis, such as the cases when the subject went tangent to or penetrated the wall, turned away from the correct path, or lifted the pen from the paper. However, it did not capture some other phenomena that we believe are important while analyzing the drawings. For this project, we added new display options such as detection and display of corner clips, steering towards incorrect analysis, and straight chunks of the path. We improved the interface to display a more thorough speed analysis. For instance, we added speed coloring (Figure 4-4), which allows us to immediately see where and by how much the subject's pen speed has changed.

We also added a function that plots a line graph of the speed with varying sliding window size. Looking at the speed values can be crucial for understanding when and where the person slowed down or sped up, which might be related to their decision-making process. In order to smooth the signal, we added an option of choosing the sliding window size, which helps deal with noise. In Figure 4-5, we show an example of the speed graph for a choice maze with decision points highlighted in orange.

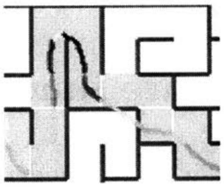
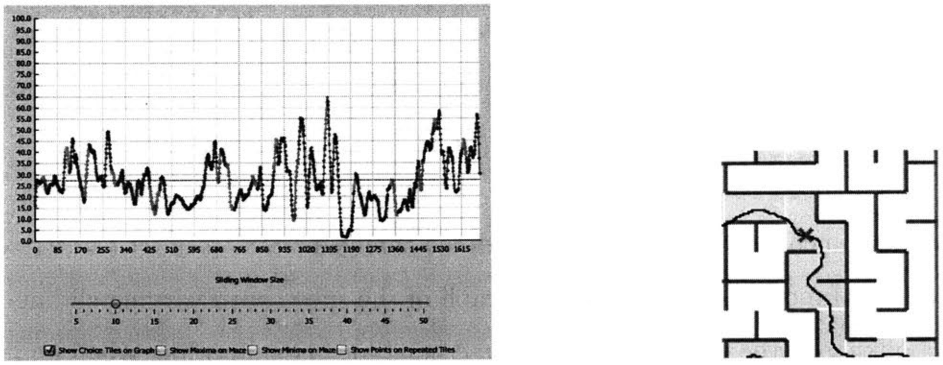


Figure 4-4: We added an option of coloring the drawn path with colors that depend on the speed values. As speed values increase, the color gradually changes from red to green to blue.



☑ Show Choice Tiles on Graph □ Show Maxima on Maze □ Show Minima on Maze □ Show Points on Repeated Tiles

(a) An example of the speed graph with a pink square corresponding to one point on the maze.

(b) An example of cross referencing for the point on the speed graph and maze.

Figure 4-5: The new functionality allows cross-referencing points on the speed graph with points on the maze, as well as looking at the extrema and points that belong to the decision tiles. The straight red line shows an average speed in the entire maze. The sliding window can be adjusted with the slider, which helps control the smoothness of the graph.

Chapter 5

Feature Analysis

In this section we present the results we obtained by applying two machine learning classifiers to our sample set. Our experiments focused on two tasks: screening, which involves distinguishing healthy subjects from the subjects with various cognitive impairments, and identifying subjects with a specific class of memory impairment disorder, such as Alzheimer's disease (AD) and the Mild Cognitive Impairment (MCI). We treated each instance of the Digital Maze Completion Test as a single sample point. We evaluated the accuracy of the predictions made by the classifiers and compared it with the traditional screeners, such as the MMSE and the MoCA. Finally, we looked into which features played the most important role in the classification process.

5.1 Machine learning algorithms

Our choice of machine learning algorithms included two classifiers: regularized logistic regression and support vector machines (SVM). We used the implementation from the python machine learning library `scikit-learn` [14].

5.1.1 Experiment setup

Our experiments used stratified cross-validation with 3 folds. We standardized our feature values and applied the grid-search technique to our training and validation sets to select the best model. We used precision, recall, f-score (the harmonic mean of the precision and recall) and area under the receiver operating characteristic curve (AUC) to compare the predictive power of our classifiers in each task.

For logistic regression, we used l_1 regularization and chose the regularization strength parameter C from the range $\{2^{-10}, 2^{-8}, ..., 2^{4}\}$.

For SVM, we used the values for C (penalty of the error term) from the range $\{2^{-4}, 2^{-2}, ..., 2^{10}\}$, values for γ (kernel coefficient) from the range $\{2^{-6}, 2^{-2}, ..., 2^{10}\}$, and two kernels, linear and radial basis function.

We have conducted 17 experiments with each of these two classifiers, in which we varied which diagnoses were labeled as 1 (we call them *positive samples*). In Figure 5.1, we listed the details of the sample set for each experiment, such as which diagnoses were chosen to be positive samples, how many of such tests there were and how many of the tests we used in total. Out of 559 tests, we excluded 8 due to errors while parsing the pen data or missing MMSE or MoCA scores.

We selected 55 simple features (listed in Appendix A) that represent various metrics of each part of the test. Even though our system is capable of computing many more features, a lot of which are behavioral, we decided to limit our feature space in order to avoid missing values. Missing values occur because some of the behavioral features, such as backtracking details, do not have a fixed number of values, as people make different mistakes. There are two ways to solve this: either remove the samples with missing values or impute missing values. The first option would reduce our already small data set. The second one would distort our sample set, as it would "assign" mistakes to the subjects who did not make any. Therefore, we aggregated behavioral features (e.g. number of corner clipping cases), which are defined for all tests..

5.1.2 Baselines

We used the MMSE scores as the baseline, with cutoffs of 28 and higher for healthy subjects, 26 to 28 for subjects with the MCI and lower than 26 for subjects with dementia. We computed precision, recall and f-score on the same test set.

Exp.	Which samples included	Num. of samples	Positive label	Num. of positive samples
1	All subjects	551	HC	236
2	Subjects with AD, PD, MCI or HCs	504	HC	236
3	Subjects with MCI or HCs	318	HC	236
4	Subjects with amnestic MCI or HCs	291	HC	236
5	Subjects with AD or HCs	321	HC	236
6	HCs and subjects with MMSE score from 26 and higher	474	HC	236
7	All subjects	551	MCI	79
8	Subjects with AD, PD, MCI or HCs	504	MCI	82
9	Subjects with MCI or MMSE score from 26 and higher	482	MCI	82
10	Subjects with MCI or AD	167	MCI	82
11	All subjects	551	amnestic MCI	55
12	Subjects with AD, PD, MCI or HCs	504	amnestic MCI	55
13	Subjects with amnestic MCI or MMSE score from 26 and higher	479	amnestic MCI	55
14	Subjects with amnestic MCI or AD	140	amnestic MCI	55
15	All subjects	551	AD	80
16	Subjects with AD, PD, MCI or HCs	504	AD	85
17	Subjects with AD or MMSE score from 26 and higher	531	AD	85

Table 5.1: For each classification experiment, the first column describes which samples were included in the experiment, the second column contains the sample set sizes, the third column states which diagnoses served as positive samples, and the last is for the number of the positive samples. AD - Alzheimer's disease, PD - Parkinson's disease, MCI - Mild Cognitive Impairment, HC - health control. In experiments 1, 7, 11, and 15 we used only primary diagnoses, while in others we also added the tests that have at least moderate confidence secondary diagnoses, if their primary diagnoses were not originally included in the sample set.

5.1.3 Results

We observed from our experiments that both logistic regression and SVM agree with one another and have similar results in terms of precision and recall. We also learned that in the screening task, our classifiers performed particularly well in distinguishing healthy controls from every other subject in our sample set. Logistic regression had f-score of 0.77 and SVM had f-score of 0.76, both of which are higher than that of the MMSE, which was 0.60, as seen in the first and the second rows of Table 5.2. The success of the screening task confirms that even a small number of features extracted from the Digital Maze Completion Tests is already useful for screening purposes as they provide relatively accurate predictions in comparison with the traditional MMSE. Figures 5-1a and 5-1b show the associated receiver operating characteristic (ROC) curves, in which each curve is a mean over 3 folds.

Our classifiers also perform well in distinguishing the subjects with MCI from various subgroups, such as healthy controls only (Table 5.2, row 3), all other samples (Table 5.3, row 1) and the subgroup of the subjects who are either healthy controls or individuals with Alzheimer's disease or Parkinson's disease (Table 5.3, row 3).

In the task of separating healthy controls from the subjects with the MCI, our classifiers had f-scores of 0.86 (logistic regression) and of 0.82 (SVM), while the MMSE had much lower f-score of 0.72, as seen in the third row of Table 5.2. This task is particularly important, as the MCI is considered to be a transitional stage for dementia, and its amnestic form might later develop into Alzheimer's disease. While the subjects with the amnestic MCI are able to carry on with their daily life activities, they start experiencing memory problems. However, they tend to get high scores on the MMSE, which classifies them as healthy. Particularly, in our sample set, the subjects with MCI have an average MMSE score of 27.7 (±1.6) out of 30 and the subjects with amnestic MCI have an average MMSE score of 28.2 (±1.7) out of 30. Therefore, identifying the early signs of dementia in seemingly healthy individuals is a difficult task. Our results in Table 5.3 show that our classifiers are able to detect MCI subjects relatively well, which is a promising result that can be further

Classes	Exp. num	Classifier	F-score	ROC AUC
HC vs all	1	Logistic Regression	0.77 (0.03)	0.85 (0.05)
		SVM	0.76 (0.05)	0.84 (0.06)
		MMSE scores	0.60 (0.02)	0.68 (0.06)
HC vs AD, PD, MCI	2	Logistic Regression	0.79 (0.03)	0.85 (0.04)
		SVM	0.78 (0.05)	0.84 (0.05)
		MMSE scores	0.64 (0.04)	0.72 (0.06)
HC vs MCI	3	Logistic Regression	0.86 (0.07)	0.90 (0.09)
		SVM	0.82 (0.04)	0.85 (0.06)
		MMSE scores	0.71 (0.06)	0.67 (0.10)
HC vs amnestic MCI	4	Logistic Regression	0.89 (0.05)	0.94 (0.04)
		SVM	0.89 (0.05)	0.91 (0.00)
		MMSE scores	0.75 (0.07)	0.64 (0.06)
HC vs AD	5	Logistic Regression	0.83 (0.03)	0.90 (0.03)
		SVM	0.85 (0.00)	0.91 (0.02)
		MMSE scores	0.84 (0.08)	0.92 (0.05)
HC vs high MMSE scorers	6	Logistic Regression	0.77 (0.04)	0.84 (0.04)
		SVM	0.76 (0.05)	0.83 (0.03)
		MMSE scores	0.53 (0.04)	0.58 (0.09)

Table 5.2: Results for the classification task for machine learning algorithms: separating healthy controls from various sample subgroups: the rest of the sample set, the subgroup with only the subjects with Alzheimer's disease (AD), Parkinson's disease (PD) or Mild Cognitive Impairment (MCI), the subgroup with only the subjects with MCI (amnestic or any type), the subgroup with only the subjects with AD, as well as the subgroup including all the subjects with MMSE score higher than 26. The first column is for the classes, the second refers to the experiment number in Table 5.1, the third is for the classifier names, the fourth is for the mean and standard deviation f-score across 3 folds, and the last is for the mean and standard deviation AUC value.

improved by using more sophisticated features. For a complete set of ROC curves for these experiments, see Appendix B.

Our classifiers have not separated the subjects with Alzheimer's disease from all other subjects as well as the MMSE has, as seen in Table 5.4. While they achieved quite high and reliable f-scores (both logistic regression and SVM had a score of 0.86), the MMSE outperformed them with f-score of 0.92. We suspect that our classifiers did not separate the subjects with Alzheimer's disease from the subjects with other cognitive impairments well enough due to our limited feature space. Some features, while being able to detect the presence of an impairment, do not give precise information about what kind of impairment it could be. For example, both a subject with Alzheimer's disease and a subject with Parkinson's disease will solve the maze more slowly than a healthy control. However, the underlying reasons for slow speed are different: subjects with Alzheimer's disease tend to make more errors that they have to correct, while for the subjects with Parkinson's disease, slow speed is a result of motor skills problems. We believe that expanding the feature set to include more behavioral features will help to overcome this problem, as discussed in Section 5.3.

5.2 Feature importance

In our experiments with the regularized logistic regression and with linear kernel SVM, we extracted the coefficients that were assigned to each feature. Given that we had standardized the feature values, we were able to rank the features according to their coefficient values, and used this to evaluate the relative contribution of each feature to the diagnosis prediction.

Studying the coefficients leads to two observations. First, there does not exist a single subset of features that solely influences the prediction. Each experiment has different features that seem to contribute the most to the prediction. We believe that this finding is consistent with our understanding of the complexity of the problem. There are numerous behaviors in the maze that characterize one or another diagnosis, which does not allow to reduce our feature space. Our second observation is that each

Classes	Exp. num	Classifier	F-score	ROC AUC
MCI vs all	7	Logistic Regression	0.85 (0.00)	0.79 (0.02)
		SVM	0.85 (0.00)	0.68 (0.03)
		MMSE scores	0.76 (0.03)	0.57 (0.10)
amnestic MCI vs all	11	Logistic Regression	0.89 (0.01)	0.80 (0.07)
		SVM	0.90 (0.00)	0.68 (0.08)
		MMSE scores	0.78 (0.01)	0.53 (0.08)
MCI vs AD, PD and HC	8	Logistic Regression	0.85 (0.01)	0.81 (0.03)
		SVM	0.84 (0.01)	0.72 (0.00)
		MMSE scores	0.74 (0.02)	0.54 (0.12)
amnestic MCI vs AD, PD and HC	12	Logistic Regression	0.89 (0.01)	0.82 (0.07)
		SVM	0.89 (0.00)	0.72 (0.03)
		MMSE scores	0.77 (0.02)	0.51 (0.10)
MCI vs AD	10	Logistic Regression	0.77 (0.05)	0.80 (0.02)
		SVM	0.69 (0.05)	0.74 (0.03)
		MMSE scores	0.55 (0.03)	0.87 (0.03)
amnestic MCI vs AD	14	Logistic Regression	0.69 (0.02)	0.77 (0.04)
		SVM	0.67 (0.05)	0.78 (0.05)
		MMSE scores	0.62 (0.04)	0.87 (0.04)
MCI vs high MMSE scorers	9	Logistic Regression	0.83 (0.02)	0.81 (0.06)
		SVM	0.83 (0.01)	0.71 (0.01)
		MMSE scores	0.72 (0.02)	0.65 (0.05)
amnestic MCI vs high MMSE scorers	13	Logistic Regression	0.88 (0.01)	0.79 (0.04)
		SVM	0.87 (0.01)	0.67 (0.07)
		MMSE scores	0.75 (0.03)	0.61 (0.04)

Table 5.3: Results for the classification task for machine learning algorithms: separating the subjects with Mild Cognitive Impairment (MCI, amnestic or any type) from various sample subgroups: the rest of the sample set, the subgroup only with the subjects with Alzheimer's disease (AD), Parkinson's disease (PD) or MCI, the subgroup only with the subjects with AD, as well as the subgroup including all the subjects with MMSE score higher than 26. The first column is for the classes, the second refers to the experiment number in Table 5.1, the third is for the classifier names, the fourth is for the mean and standard deviation f-score across 3 folds, and the last is for the mean and standard deviation AUC value.

Classes	Exp. num	Classifier	F-score	ROC AUC
AD vs all	15	Logistic Regression	0.86 (0.01)	0.76 (0.05)
		SVM	0.86 (0.00)	0.71 (0.05)
		MMSE scores	0.92 (0.02)	0.90 (0.04)
AD vs PD, MCI, HC	16	Logistic Regression	0.82 (0.04)	0.74 (0.03)
		SVM	0.82 (0.03)	0.72 (0.02)
		MMSE scores	0.90 (0.03)	0.90 (0.04)
AD vs high MMSE scorers	17	Logistic Regression	0.84 (0.01)	0.80 (0.04)
		SVM	0.85 (0.01)	0.80 (0.03)
		MMSE scores	0.95 (0.01)	0.92 (0.03)

Table 5.4: Results for the classification task for machine learning algorithms: separating the subjects with Alzheimer's disease (AD) from various sample subgroups: the rest of the sample set, the subgroup only with the subjects with AD, Parkinson's disease (PD) or MCI, and the subgroup including all the subjects with MMSE score higher than 26. The first column is for the classes, the second refers to the experiment number in Table 5.1, the third is for the classifier names, the fourth is for the mean and standard deviation f-score across 3 folds, and the last is for the mean and standard deviation AUC value.

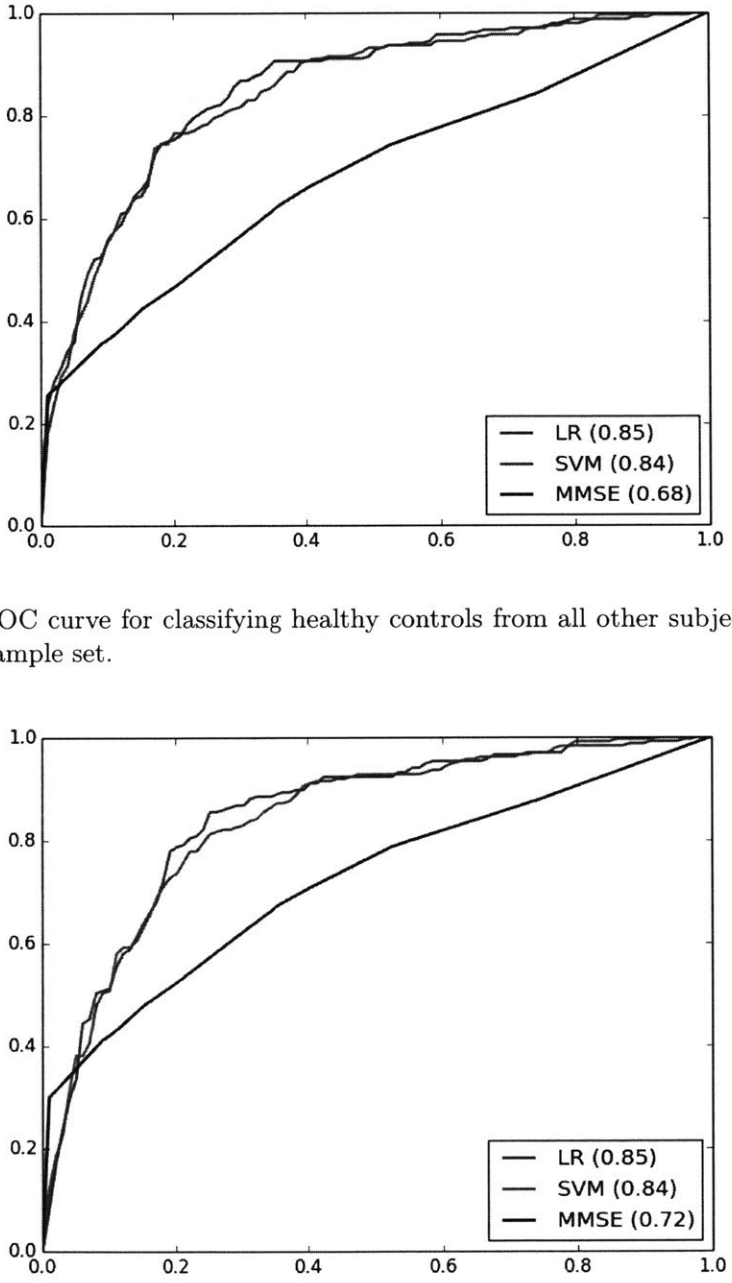
of the three parts of the test and the features associated with them seem to have some diagnostic power. Especially, we noticed that even the calibration maze has diagnostic power, even though it requires the lowest cognitive load and was designed primarily to measure the basic locomotor skills. Given this finding, it is important to further explore the subject's behavior on the calibration maze, which is one of the possible directions for future research.

5.3 Future work

One of the main limitations of this project is a small data set. Given that we have data from fewer than 200 patients, we used each test instance as a single sample and did not take into consideration the fact that each subject completes three difficulty levels of the test. It is likely that combining the data from all three levels for a single individual will result in more accurate predictions. For example, we can measure the learning effect, e.g. the change in parameters such as the average speed or total inking time after the subject took their second and third tests. We expect that healthy individuals will get more accustomed to the test and its requirements, which will result in their speeding up in the later tests. In contrast, the subjects with memory impairments will not exhibit such behavior.

Another possible improvement for our work is expanding the feature set to include more behavioral features, such as the details of the backtracking. Currently, our classifiers use only counts of various mistakes. However, we did not include the features that would quantify and evaluate the severity of these mistakes. For instance, we could compute and record how far the person backtracked if they made a wrong turn on the decision tile. In order to decide which features would be the best candidates to include in our feature space, we could analyze the errors and mistakes that are characteristic for each of the clinical groups and choose the features that efficiently describe those mistakes.

The Digital Maze Completion Test is a new screening tool, which has a remarkable potential to catch subtle behavioral details. Therefore, it opens up plenty of research opportunities to further explore its powers and capabilities.



(a) ROC curve for classifying healthy controls from all other subjects in our sample set.

(b) ROC curve for classifying healthy controls from the subjects with Alzheimer's disease, Parkinson's disease and Mild Cognitive Impairment.

Figure 5-1: ROC curves for the screening task.

Chapter 6

Conclusion

The Digital Maze Completion Test is an innovative screening tool that has potential to screen for dementia on its early stages. The winning combination of the carefully designed mazes, the digitizing pen and the analytical software component allows it to detect the subtle details of the subject's behavior and to infer their cognitive status.

The goal of this research project was to explore the data and measurements that can be extracted from the maze drawings. We extracted some generic measurements, such as information about the subject's speed, time to complete the test, drawing path-length and error count. We have also defined how to identify and operationalize some behavioral phenomena, such as actual mistakes or near-mistakes at the decision points. We implemented these definitions in code, enabling us to analyze behaviors specific to the mazes. We carried out simple experiments with the measurements that we extracted from the real-world clinical data using well-known machine learning classifiers. The experiments demonstrated that our system is capable of performing the screening task (distinguishing healthy individuals from the subjects with cognitive impairments) on par with the traditional and widely accepted screening tools, such as the MMSE.

This thesis is the start of the journey of exploring the potential of the Digital Maze Completion Test and its screening power. We have touched on a small subset of the measurements that could be obtained from its data, while the greater number of them is yet to be discovered and analyzed. The Digital Maze Completion Test offers a rich variety of the research opportunities, each of which can have an impact on timely screening and accurate diagnosis of cognitive impairments.

Appendix A

Feature Names

Here we list all the features that we used in the experiments described in Section 5.1. We first include the glossary to explain the names of the features.

Glossary

Distance from wall - an array consisting of distances from each point to the nearest wall. The values associated with the distances to the nearest walls, such as maximum, standard deviation, etc., are computed from this array.

Ink time - time spent physically drawing the path.

Is completed - a binary value that represents whether the person reached the goal location from the start location, computed by checking whether all but up to n ($n = 5$) tiles of the solution path were included in the tiles of the subject's drawn path.

Solution path - a single correct path from the start location to the goal location.

Speed - an array of speed values for each point in the maze, computed as the average speed in the window of points from x to $x + n$, where n is the size of the sliding window.

Stroke - a continuous line drawn by the pen without lifting it.

Total- - a prefix for features that was taken on the entire maze rather than on one of its segments.

Features for each maze

Calibration maze

- IsCompleted
- TotalNumberOfStrokes
- TotalTimeDuration
- TotalInkTime
- TotalPathLength
- TotalMaxSpeed
- TotalMeanSpeed
- TotalMedianSpeed
- TotalMinSpeed
- TotalStddevSpeed
- TotalMaxDistFromWall
- TotalMeanDistFromWall
- TotalMedianDistFromWall
- TotalMinDistFromWall
- TotalStddevDistFromWall

No-choice maze

- IsCompleted
- TotalNumberOfStrokes
- TotalTimeDuration
- TotalInkTime
- TotalPathLength
- BacktrackingCount
- CornerClipCount
- WallCollisionCount
- WallPenetrationCount
- TotalMaxSpeed
- TotalMeanSpeed
- TotalMedianSpeed
- TotalMinSpeed
- TotalStddevSpeed
- TotalMaxDistFromWall
- TotalMeanDistFromWall
- TotalMedianDistFromWall
- TotalMinDistFromWall
- TotalStddevDistFromWall

Choice maze

- IsCompleted
- TotalNumberOfStrokes
- TotalTimeDuration
- TotalInkTime
- TotalPathLength
- BacktrackingCount
- CornerClipCount
- WallCollisionCount
- WallPenetrationCount
- TimeDeviatingFromSolutionPath
- PathLengthDeviatingFromSolutionPath
- TotalMaxSpeed
- TotalMeanSpeed
- TotalMedianSpeed
- TotalMinSpeed
- TotalStddevSpeed
- TotalMaxDistFromWall
- TotalMeanDistFromWall
- TotalMedianDistFromWall
- TotalMinDistFromWall
- TotalStddevDistFromWall

Appendix B

ROC curves

Here we display ROC curves for the experiments described in Section 5.1. Each plot contains three ROC curves: one for logistic regression, one for SVM, and one for the MMSE. The legend of each plot shows AUC values for each curve. The plots appear in the same order as the experiments summarized in Table 5.1.

AD - Alzheimer's disease, PD - Parkinson's disease, MCI - Mild Cognitive Impairment.

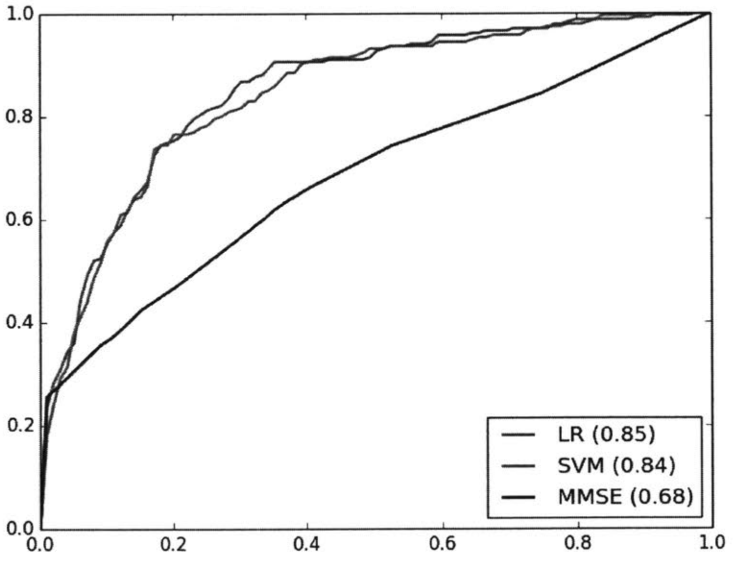


Figure B-1: Healthy controls vs. all others

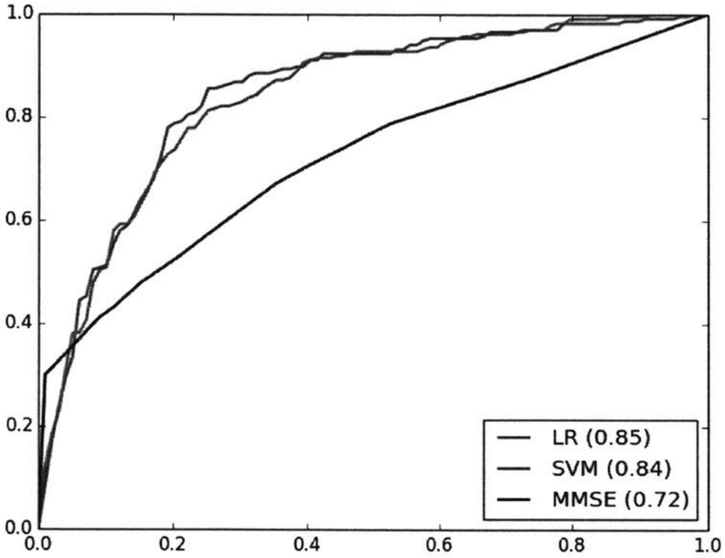


Figure B-2: Healthy controls vs. subjects with AD, PD or MCI

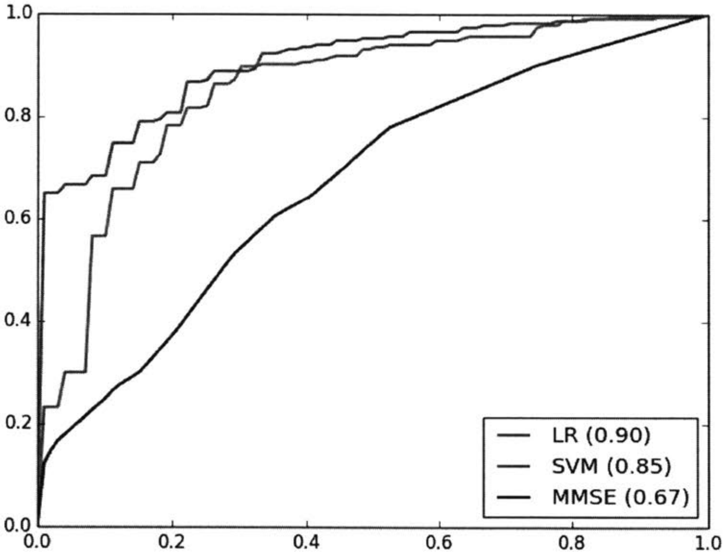


Figure B-3: Healthy controls vs. subjects with MCI

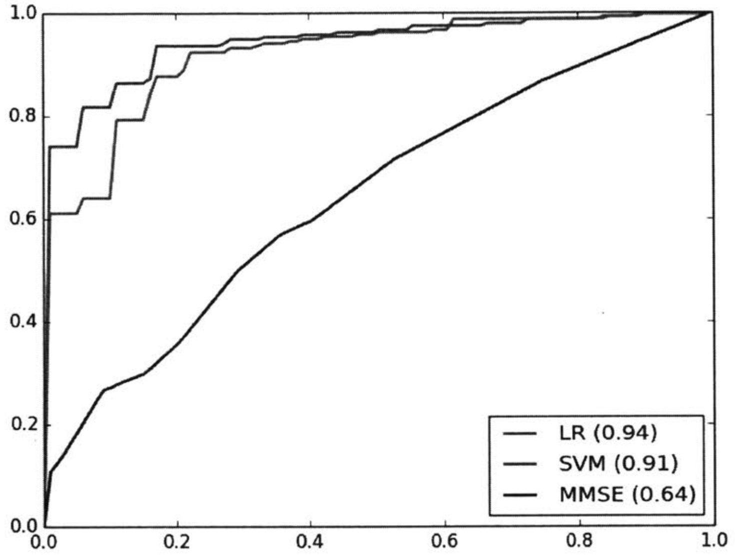


Figure B-4: Healthy controls vs. subjects with amnestic MCI

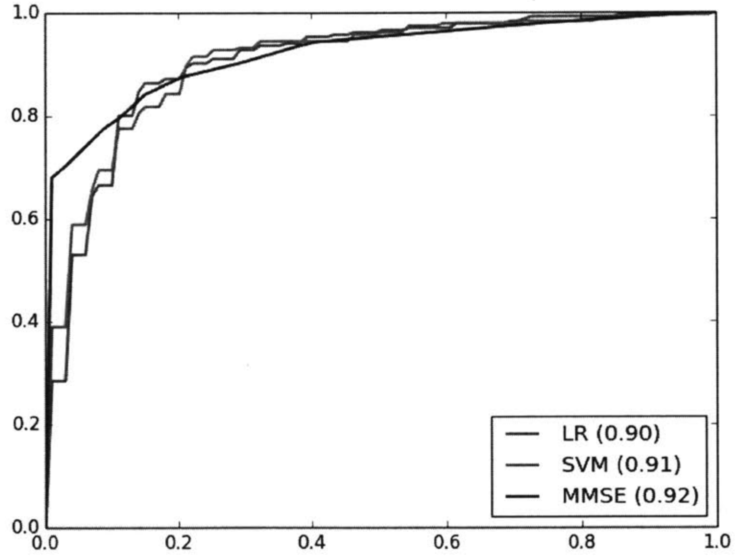


Figure B-5: Healthy controls vs. subjects with AD

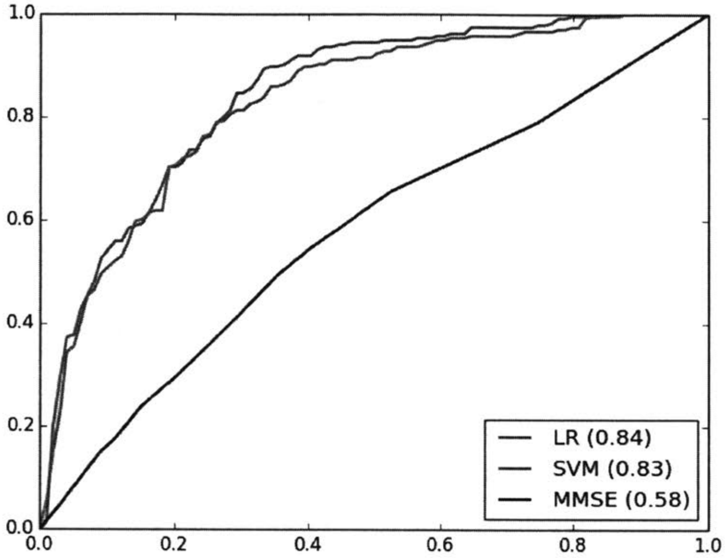


Figure B-6: Healthy controls vs. subjects with MMSE score ≥ 26

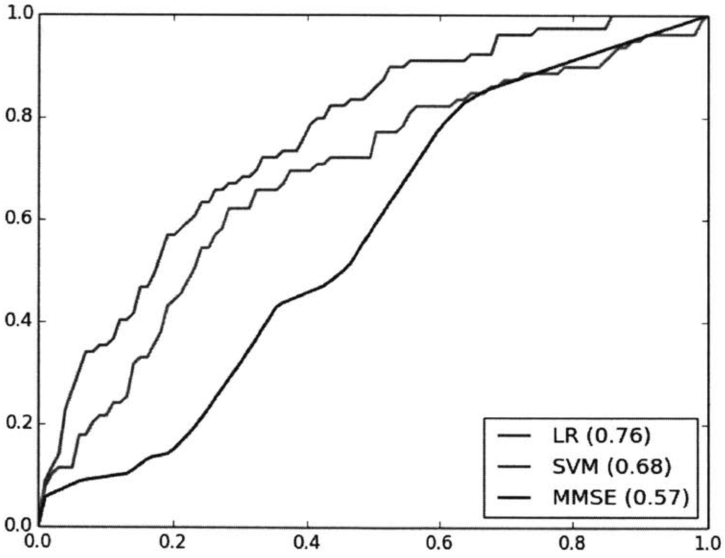


Figure B-7: Subjects with MCI vs all others

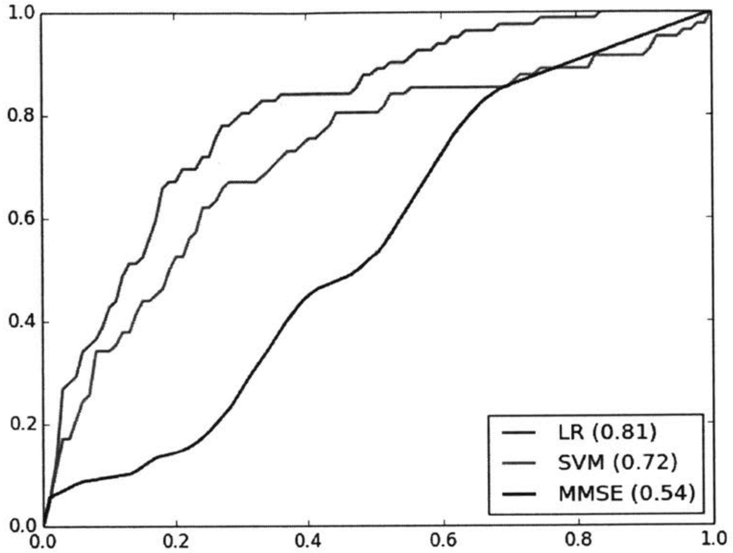


Figure B-8: Subjects with MCI vs. subjects with AD or PD or healthy controls

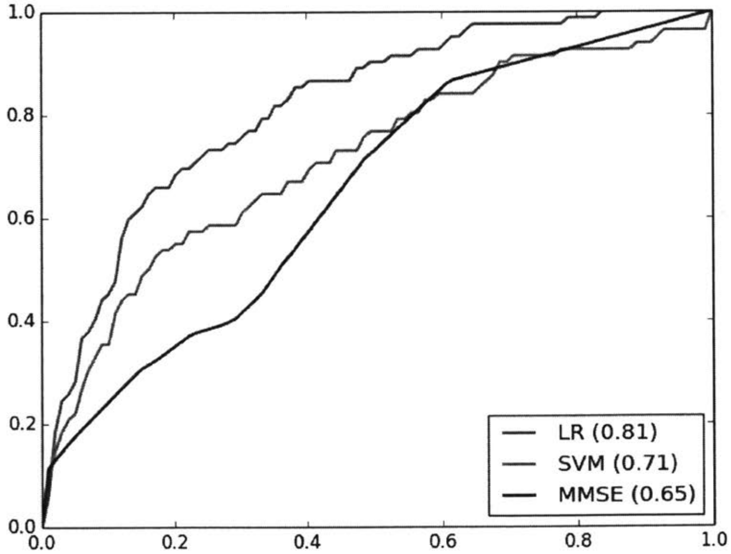


Figure B-9: Subjects with MCI vs subjects with MMSE score ≥ 26

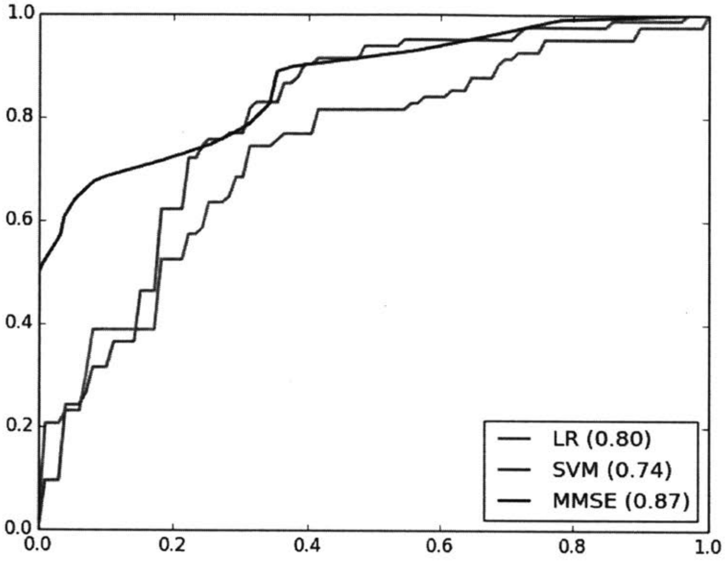


Figure B-10: Subjects with MCI vs. subjects with AD

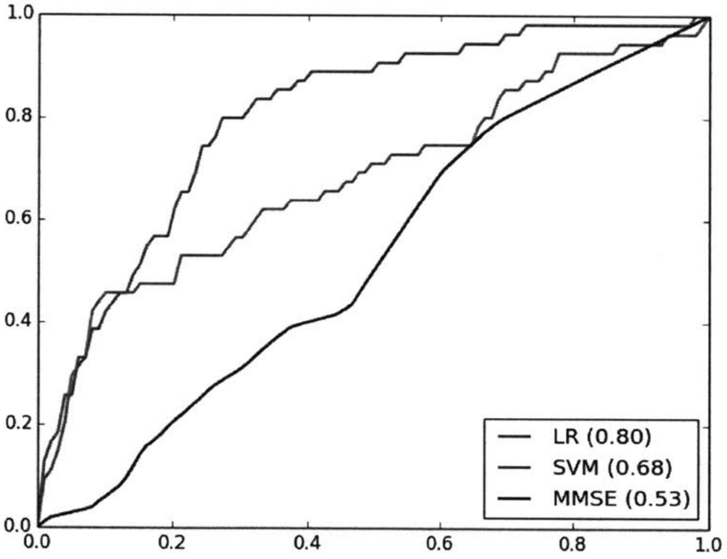


Figure B-11: Subjects with amnestic MCI vs. all others

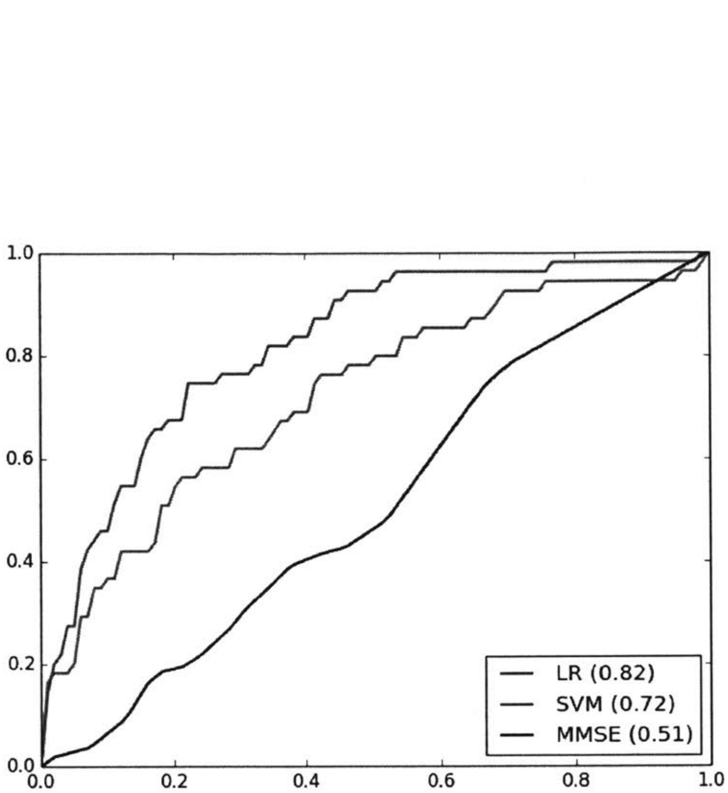


Figure B-12: Subjects with amnestic MCI vs. subjects with AD or PD or healthy controls

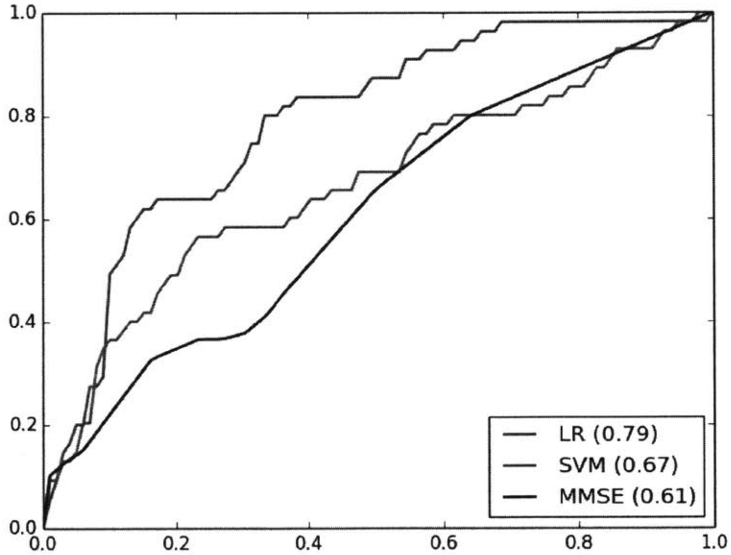


Figure B-13: Subjects with amnestic MCI vs subjects with MMSE score ≥ 26

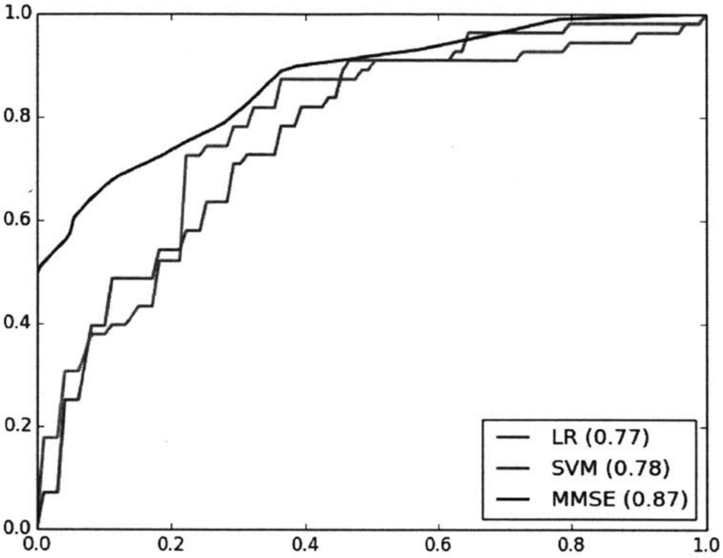


Figure B-14: Subjects with amnestic MCI vs. subjects with AD

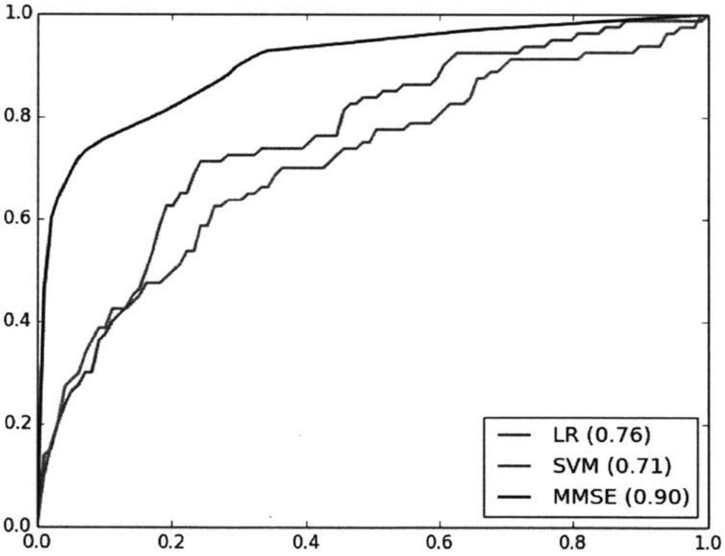


Figure B-15: Subjects with AD vs. all others

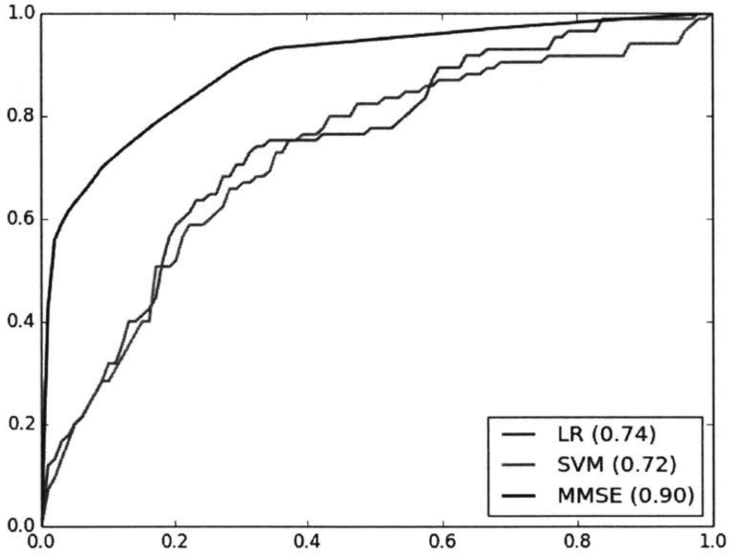


Figure B-16: Subjects with AD vs. subjects with MCI or PD or healthy controls

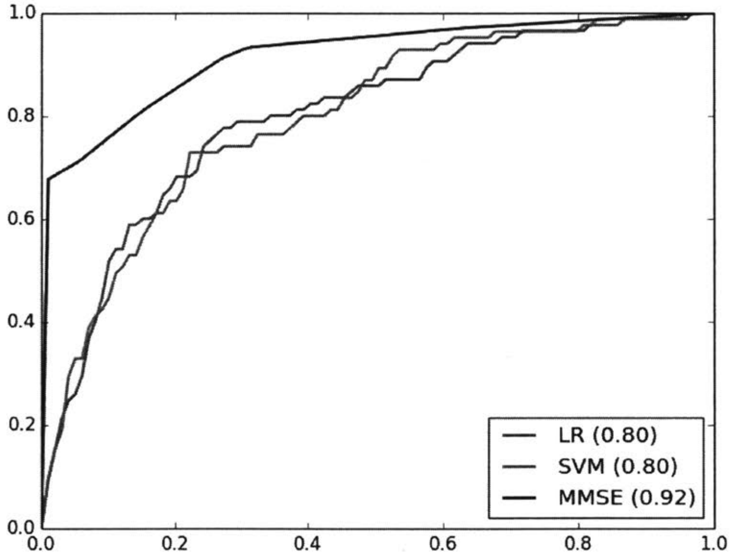


Figure B-17: Subjects with AD vs subjects with MMSE score ≥ 26

Appendix C

Test IDs of maze tests used

CIN0027453305 CIN0030511473 CIN0040753761 CIN0043525354 CIN0045340161
CIN0053804664 CIN0061282119 CIN0064072173 CIN0069468421 CIN0074002309
CIN0100247999 CIN0101456199 CIN0108651308 CIN0111220969 CIN0113871242
CIN0127650113 CIN0143430274 CIN0143830514 CIN0155010036 CIN0156301985
CIN0159327109 CIN0160655681 CIN0198350854 CIN0208838258 CIN0209735954
CIN0211478229 CIN0211896688 CIN0218939228 CIN0223638145 CIN0225620687
CIN0242361456 CIN0257667219 CIN0264441206 CIN0267797456 CIN0273480337
CIN0274243876 CIN0274694469 CIN0279725239 CIN0282356946 CIN0304325114
CIN0306543099 CIN0310484226 CIN0310592942 CIN0320015539 CIN0327719740
CIN0333856160 CIN0336911847 CIN0340986448 CIN0344201181 CIN0347011394
CIN0352763661 CIN0356013375 CIN0377640428 CIN0383782575 CIN0384152275
CIN0394832030 CIN0405482599 CIN0408459509 CIN0432378702 CIN0438669616
CIN0449843758 CIN0451106766 CIN0458921274 CIN0460005819 CIN0466708456
CIN0472650926 CIN0475877544 CIN0483937315 CIN0491514254 CIN0492044693
CIN0493812915 CIN0495490166 CIN0496799447 CIN0502899397 CIN0508284156
CIN0511717296 CIN0532488528 CIN0543293431 CIN0551714413 CIN0557762457
CIN0560415190 CIN0563371687 CIN0567341475 CIN0572899880 CIN0581739426
CIN0583093605 CIN0590235266 CIN0593828261 CIN0597689373 CIN0598585476
CIN0619785477 CIN0620960331 CIN0624948793 CIN0629022312 CIN0630910438
CIN0633284083 CIN0634081704 CIN0639073752 CIN0639236855 CIN0647193771
CIN0654970328 CIN0658997087 CIN0674313789 CIN0681418695 CIN0685186245
CIN0689407144 CIN0692590986 CIN0698877348 CIN0703152628 CIN0703696525
CIN0711343618 CIN0713584310 CIN0717774277 CIN0718397470 CIN0719910018
CIN0725325281 CIN0728657143 CIN0729711183 CIN0732032944 CIN0732479774
CIN0735967298 CIN0740726621 CIN0743714238 CIN0750922822 CIN0753787408
CIN0756114413 CIN0756418191 CIN0768092698 CIN0769209494 CIN0769750644

CIN0770214224 CIN0776088360 CIN0779283762 CIN0800850657 CIN0801997774
CIN0808768264 CIN0810067988 CIN0810078938 CIN0813509055 CIN0815429245
CIN0815784832 CIN0822515449 CIN0824490033 CIN0829752406 CIN084107550
CIN0842578190 CIN0845547847 CIN0855093915 CIN0859051595 CIN0860268641
CIN0876266569 CIN0878002120 CIN0890902759 CIN0892893069 CIN0897017229
CIN0901805831 CIN0903332373 CIN0905528891 CIN0907600305 CIN0909876381
CIN0915153471 CIN0924471590 CIN0943130217 CIN0953721871 CIN0958797916
CIN0961830974 CIN0977045040 CIN0980119626 CIN0983429128 CIN0993734790
CIN1000208092 CIN1005345982 CIN1014610104 CIN1023177495 CIN1024381233
CIN1026042441 CIN1027460162 CIN1030062482 CIN1034254508 CIN1037691433
CIN1041537837 CIN1055737667 CIN1059559453 CIN1063846209 CIN1070836247
CIN1074468994 CIN1074520967 CIN1076454635 CIN1080431405 CIN1086570666
CIN1093402195 CIN1107699723 CIN1109618635 CIN1116788918 CIN1119648474
CIN1125077770 CIN1127845420 CIN1129840295 CIN1136776227 CIN1142776141
CIN1166654052 CIN1166957313 CIN1173803003 CIN1187164912 CIN1188342743
CIN1195745844 CIN1213000923 CIN1216110938 CIN1225051966 CIN1238560499
CIN1252634630 CIN1253773233 CIN1256910021 CIN1268538620 CIN1272301144
CIN1277461397 CIN1279657786 CIN1282315770 CIN1289946017 CIN1291120692
CIN1295850471 CIN1303178866 CIN1304193191 CIN1305059089 CIN1314309706
CIN1314403367 CIN1317406453 CIN1318668363 CIN1323223013 CIN1324038119
CIN1345942278 CIN1364032307 CIN1382473025 CIN1392261787 CIN1398539458
CIN1401006339 CIN1410705315 CIN1413884540 CIN1414407455 CIN1415339366
CIN1423701118 CIN1434995191 CIN1460310555 CIN1464940042 CIN1472507064
CIN1476368667 CIN1484889161 CIN1486926238 CIN1495462411 CIN1502897238
CIN1525891173 CIN1529932164 CIN1548536212 CIN1548911711 CIN1552682580
CIN1552874965 CIN1555239756 CIN1559996627 CIN1577277266 CIN1589152239
CIN1591418879 CIN1592940863 CIN1593793108 CIN1615725566 CIN1616712994
CIN1624500856 CIN1631554297 CIN1645731164 CIN1646204130 CIN1652608832
CIN1655547005 CIN1666635125 CIN1675521659 CIN1676219714 CIN1677445125
CIN1696591134 CIN1698999645 CIN1700910824 CIN1714499402 CIN1718680818
CIN1721327800 CIN1721946539 CIN1726720592 CIN1732533319 CIN1741127383
CIN1741903755 CIN1744471031 CIN1747392422 CIN1751502373 CIN1756891960
CIN1764327790 CIN1767075964 CIN1773476751 CIN1775694095 CIN1783963686
CIN1786560725 CIN1790484151 CIN1798831596 CIN1806322835 CIN1817510006
CIN1825949313 CIN1835895341 CIN1835994286 CIN1836515137 CIN1857460339
CIN1869295836 CIN1871117726 CIN1876952008 CIN1878982131 CIN1880397856
CIN1896230977 CIN1899291958 CIN1904052610 CIN1911835397 CIN1919087753
CIN1922089350 CIN1924169676 CIN1924534824 CIN1927127749 CIN1929882096
CIN1930429340 CIN1935939796 CIN1936889654 CIN1948419166 CIN1950048131
CIN1952903396 CIN1955927816 CIN1959186393 CIN1970645589 CIN1979770752
CIN1982422508 CIN1987197662 CIN1988246011 CIN1995587367 CIN2000253265
CIN2015238121 CIN2019421473 CIN2032223277 CIN2033242976 CIN2040245237
CIN2076553335 CIN2081429799 CIN2083829193 CIN2088280458 CIN2089660071
CIN2094339005 CIN2094879527 CIN2097478402 CIN2115192095 CIN2126143601
CIN2131195185 CIN2136433268 CIN2139402053

SNF0000170401 SNF0015963066 SNF0020119531 SNF0031841418 SNF0039327487
SNF0052547953 SNF0070730860 SNF0079886974 SNF0083589141 SNF0098236749
SNF0138045545 SNF0139363224 SNF0190539304 SNF0217557978 SNF0223639227
SNF0225506695 SNF0242179948 SNF0245696200 SNF0247730213 SNF0290892546
SNF0291390634 SNF0299479777 SNF0311087421 SNF0328080455 SNF0347352582
SNF0352244640 SNF0362485450 SNF0383383302 SNF0394927074 SNF0398101851
SNF0401080315 SNF0421088875 SNF0424328111 SNF0439338577 SNF0441764038
SNF0442847672 SNF0448669642 SNF0450113374 SNF0453185700 SNF0484585945
SNF0499154346 SNF0499808716 SNF0514524207 SNF0544818057 SNF0545810091
SNF0556041375 SNF0565015316 SNF0570598156 SNF0581176675 SNF0585710536
SNF0600831681 SNF0605114913 SNF0608722470 SNF0610903817 SNF0636736400
SNF0652679810 SNF0689950143 SNF0693689096 SNF0696472960 SNF0730955915
SNF0733927587 SNF0740053855 SNF0741426428 SNF0778282644 SNF0794044595
SNF0796259448 SNF0827313661 SNF0871160577 SNF0883560415 SNF0890211282
SNF0895996700 SNF0915769059 SNF0927326479 SNF0939506628 SNF0955418727
SNF0969423930 SNF0980683397 SNF0984719805 SNF1014782548 SNF1038986978
SNF1043621975 SNF1045498879 SNF1051507867 SNF1053642147 SNF1056789110
SNF1064165286 SNF1064464101 SNF1073086710 SNF1080664923 SNF1089408909
SNF1105320493 SNF1115119380 SNF1129011881 SNF1178284599 SNF1260768252
SNF1262140573 SNF1271154745 SNF1288875526 SNF1297130863 SNF1302769793
SNF1318582649 SNF1330362593 SNF1335801196 SNF1349435650 SNF1391527736
SNF1413338361 SNF1421441325 SNF1473809889 SNF1492125105 SNF1501232160
SNF1503506522 SNF1505901276 SNF1525669270 SNF1541069910 SNF1564387853
SNF1570578916 SNF1578891466 SNF1587498011 SNF1637688231 SNF1645473070
SNF1656062327 SNF1656110321 SNF1701307822 SNF1706927091 SNF1715478677
SNF1727730648 SNF1730296050 SNF1732232048 SNF1790382567 SNF1803546174
SNF1843099159 SNF1845413444 SNF1866741631 SNF1892049184 SNF1892274348
SNF1901871546 SNF1910553311 SNF1918771088 SNF1923889381 SNF1943740358
SNF1950239405 SNF1951980442 SNF1977747243 SNF2006581938 SNF2033725053
SNF2034679954 SNF2041272775 SNF2061676800 SNF2078676537 SNF2110510207
SNF2120448437 SNF2146191463 SNF0203694121 SNF0317740612 SNF0362409622
SNF0381228329 SNF0456720560 SNF0457276167 SNF0498995244 SNF0529196096
SNF0533016322 SNF0551919463 SNF0594322048 SNF0605437728 SNF0607508944
SNF0722262410 SNF0723387100 SNF0783636844 SNF0803455253 SNF0808858940
SNF0917493252 SNF0979864217 SNF0993367025 SNF0996240085 SNF1101355647
SNF1105414446 SNF1119320180 SNF1200411159 SNF1260775509 SNF1265109127
SNF1287744746 SNF1289685730 SNF1293264616 SNF1377645176 SNF1384113301
SNF1406624690 SNF1423063868 SNF1440107254 SNF1449430419 SNF1485167067
SNF1485566455 SNF1565435387 SNF1646533492 SNF1657627759 SNF1663478003
SNF1682873659 SNF1705087785 SNF1740992407 SNF1757622425 SNF1782496334
SNF1821483222 SNF1961951668 SNF1993786381 SNF2038529006 SNF2052413634
SNF2071010935

Bibliography

[1] 2017 alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(4):325–373, April 2017.

[2] Berit Agrell and Ove Dehlin. The clock-drawing test. *Age and ageing*, 27(3):399–404, 1998.

[3] A L Benton, A Elithorn, M L Fogel, and M Kerr. A perceptual maze test sensitive to brain damage. *Journal of Neurology, Neurosurgery, and Psychiatry*, 26(6):540–544, 12 1963.

[4] Deborah A Cahn, David P Salmon, Andreas U Monsch, Nelson Butters, WC Wiederholt, Jody Corey-Bloom, and Elizabeth Barrett-Connor. Screening for dementia of the alzheimer type in the community: the utility of the clock drawing test. *Archives of Clinical Neuropsychology*, 11(6):529–539, 1996.

[5] Breda Cullen, Brian O'Neill, Jonathan J Evans, Robert F Coen, and Brian A Lawlor. A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery, and Psychiatry*, 78(8):790–799, 08 2007.

[6] M.G. Davies and D.M. Davies. Some analytical properties of elithorn's perceptual maze. *Journal of Mathematical Psychology*, 2(2):371 – 380, 1965.

[7] Randall Davis, David Libon, Rhoda Au, David Pitman, and Dana L. Penney. Think: Inferring cognitive status from subtle behaviors. In *Proceedings IAAI 2014*, pages 2898–2905, Quèbec City, Quèbec, Canada, July 2014.

[8] Hayley E Richardson and John N Glass. A comparison of scoring protocols on the clock drawing test in relation to ease of use, diagnostic group, and correlations with mini-mental state examination. *Journal of the American Geriatrics Society*, 50(1):169–173, 2002.

[9] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state". a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[10] Morris Freedman. *Clock drawing: A neuropsychological analysis.* Oxford University Press, USA, 1994.

[11] K Kantarci, S D Weigand, S A Przybelski, M M Shiung, J L Whitwell, S Negash, D S Knopman, B F Boeve, P C O'Brien, R C Petersen, and C R Jack. Risk of dementia in mci: Combined effect of cerebrovascular disease, volumetric mri, and (1)h mrs. *Neurology*, 72(17):1519–1525, 04 2009.

[12] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 4 2005.

[13] Sid E O'Bryant, Joy D Humphreys, Glenn E Smith, Robert J Ivnik, Neill R Graff-Radford, Ronald C Petersen, and John A Lucas. Detecting dementia with the mini-mental state examination (mmse) in highly educated individuals. *Archives of neurology*, 65(7):963–967, 07 2008.

[14] 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.

[15] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni. Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine*, 275(3):214–228, 2014.

[16] S.D. Porteus. *The Maze Test and Clinical Psychology.* Pacific Books, 1959.

[17] Yale Song, Randall Davis, Kaichen Ma, and Dana L Penny. Balancing appearance and context in sketch interpretation. *arXiv preprint arXiv:1604.07429*, 2016.

[18] William Souillard-Mandar, Randall Davis, Cynthia Rudin, Rhoda Au, David Libon, Rod Swenson, Catherine C. Price, Melissa Lamar, and Dana L. Penney. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine Learning*, Special issue on Healthcare and Medicine, 2015.

[19] Paula T Trzepacz, Helen Hochstetler, Shufang Wang, Brett Walker, Andrew J Saykin, and for the Alzheimer's Disease Neuroimaging Initiative. Relationship between the montreal cognitive assessment and mini-mental state examination for assessment of mild cognitive impairment in older adults. *BMC Geriatrics*, 15:107, 2015.

[20] Annet W. Wind, François G. Schellevis, Gerrit Van Stavern, Rob J. P. M. Scholten, Cees Jonker, and Jacques Th. M. Van Eijk. Limitations of the mini-mental state examination in diagnosing dementia in general practice. *International Journal of Geriatric Psychiatry*, 12(1):101–108, 1997.