Master’s thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Master of Science UZH in Psychology

Automatic Scoring of the Test des Deux Barrages

Hans Joseph Thalathara
Student ID Nr.: 16-116-220

Examiner: Prof. Dr. Nicolas Langer, Dr. Zofia Baranczuk and Dr. Bruno Hebling Vieira

Department of Psychology – Methods of Plasticity Research

Submission date: 01.06.2023

Abstract

Paper-pencil tests are essential tools in neuropsychology to quantify effects of a disease on cognitive performance. The evaluation of these tests is mostly done manually, which is time-consuming and prone to errors. This thesis proposes an automatic scoring system for the Test des Deux Barrages, a selective attention test regularly used in Switzerland. The goal of the thesis was to investigate the number of errors made in the evaluation and to develop a scoring system that would deliver results similar to the manual evaluation by clinical experts. For this purpose, 23 filled-out test sheets and their evaluations by clinical experts were collected. The clinical evaluations were examined for errors using the ground truth of the patients. The collected test sheets were split into individual symbols and the symbols classified as marked by the patient or not based on their grayscale values. Three models were developed for binary classification. The performance of the models was evaluated on various performance metrics (balanced accuracy, F1 measure, sensitivity and specificity) and compared to the performance of clinical experts using the sign test and Wilcoxon signed-rank test. Of the three developed models, the model with a stricter threshold for non-target symbols produced results consistent with the hypothesis that the performance metrics tested were not statistically different from the clinical experts. The scoring system needs around 13.5 seconds per test sheet which is qualitatively faster than self-reported evaluation times by neuropsychologists for the T2B (10 to 15 minutes). Despite distinct limitations, that need to be addressed in future research and development, this thesis demonstrates the potential of an innovative approach for the clinical field of neuropsychology. The application of computational techniques can streamline and improve psychological assessment processes.

Keywords: Neurological diseases, paper-pencil assessment, Test des Deux Barrages, automatic scoring, grayscale value, binary classification

Table of Contents

1 Introduction

Neurological diseases are among the most common illnesses throughout the lifespan and represent a significant global health concern. According to Maercker et al. (2013), 1.087 billion EUR were allocated to neurological diseases in 2010 in Switzerland, with further increases expected in the future. Moreover, there has been a notable shift in the age distribution over the last centuries due to higher life expectancy and lower birth rates. The older population is projected to increase significantly over the coming decades (Deery et al., 2023). Prevalence rates of neurodegenerative conditions such as Alzheimer’s disease have grown with the increase of older individuals within the population (He et al., 2016; World Health Organization, 2015). These developments highlight and compound the economic and societal burden of neurological diseases, which must be taken into account by decision-makers in health services and finance departments.

Different domains of cognitive functioning – including attention, memory, language, and executive functioning – are negatively affected by neurological diseases. These neurocognitive deficits have a significant impact on an individual’s daily life. Particularly, impairments in attention have been linked to various neurological diseases (Coulthard et al., 2006). For instance, difficulties in sustaining attention reduce the ability to learn and remember new information for individuals with Alzheimer’s disease (Budson, 2022). Shifting attention is impaired in Parkinson’s disease, undermining the ability to multitask and make decisions (Bodis-Wollner, 2003). Similarly, individuals with multiple sclerosis show decreased attention capacity and information processing speed, making daily activities such as driving and working challenging (Amato et al., 2008). With the economic and societal burdens and the impact on individuals in mind, it is important to develop and validate reliable assessment tools to measure the degree of disease severity and monitor changes over time that can inform further therapy planning and rehabilitation.

2 Theoretical Background

2.1 Paper-Pencil Assessment

Currently, a wide variety of paper-pencil tests are used in clinical neuropsychology to assess cognitive functions like attention. Initially, these were adapted from instruments developed to study individual differences in applied settings, such as classifying students or military recruits in the early 20th century. At a time when sophisticated neuroimaging techniques were yet to be developed and made accessible, early results of neuropsychological tests on the basis of specific cognitive and behavioural tests were often impressive in localizing and lateralizing certain lesions of the brain (Miller & Barr, 2017). Despite rapid technological advance in most other fields of medicine, these neuropsychological paper-pencil tests have remained essential to quantifying effects of a disease on cognitive performance. The results gathered from these tests are employed to support neurological diagnosis, to evaluate developments over time and to base decisions on an individual’s working capacity and insurance cover, among other things (Bilder & Reise, 2019). To ensure that decisions based on these tests are not misguided certain psychometric properties are required.

2.2 Psychometric Properties

Psychometric properties refer to the scientific methods used to measure the quality of a psychological assessment instrument. Reliability and validity are considered the main measurement properties of such instruments (Souza et al., 2017). Reliability indicates the consistency of a test’s results over time (test-retest reliability) and across raters (inter-rater reliability), while validity is the degree to which a test measures what it is intended to measure. If a test is valid, it is also reliable. However, a test can be reliable without being valid, if the measures are consistent but do not measure what was intended to be measured (Ginty, 2013). Consequently, a paper-pencil test must display both properties to accurately differentiate between healthy subjects and patients in the presence of a neurocognitive deficit. Crawford et al. (2018) highlight the importance of criterion or discriminative validity of neuropsychological tests. A test is more beneficial in diagnosis when it reliably detects patients with a disease. Construct validity, as in the degree to which a test is able to measure a certain trait, is not regarded as an essential property.

Normative data are used as standard against which test performances are compared and interpreted (Casaletto & Heaton, 2017). Normative data refers to the information about the distribution of scores for a particular population, which allows clinicians to compare an individual's performance to that of the normative group. Neurologically healthy individuals are the normative group for comparison in a neuropsychological setting. Norms are a fundamental tool in clinical neuropsychology and, as such, also need to uphold certain requirements (Capitani, 1997). In theory, normative data must be large and representative enough of the population. It must incorporate additional variables such as age, gender, education, and ethnicity, among others, especially when these variables are expected to have a significant impact on the distribution of test scores (Anagnostou et al., 2013). In practice, however, the normative data is often not large enough, biased, and not representative of ethnic minorities, cultural, or linguistic differences (Loewenstein et al., 1994; Reynolds, 2000). Nonetheless, normative data is used to compare scores for individuals due to a lack of reference data even when risking different clinical conclusions for individuals with different demographics. A country like Switzerland, with different language areas and a large population with migration background, calls for a large and diverse normative sample size.

2.3 Test des Deux Barrages

Ever since its introduction in 2011, the Materialien und Normwerte für die neuropsychologische Diagnostik (MNND) has become the standard psychometric test collection in Switzerland for neuropsychological assessments of patients aging from 18 to 65 years (Balzer et al., 2011a). The Test des Deux Barrages (T2B) is one among many neuropsychological tests in this test collection. The T2B is a selective attention test originally designed by René Zazzo in 1964 (Zazzo, 1964). Similar to Brickenkamps' D2-Test, this test should be independent from cultural techniques such as calculation, reading, and writing. According to a survey sent out to neuropsychologists in Switzerland, around 40 tests are examined per year by a clinician. In the following sections the T2B is described in more detail.

2.3.1 Selective Attention

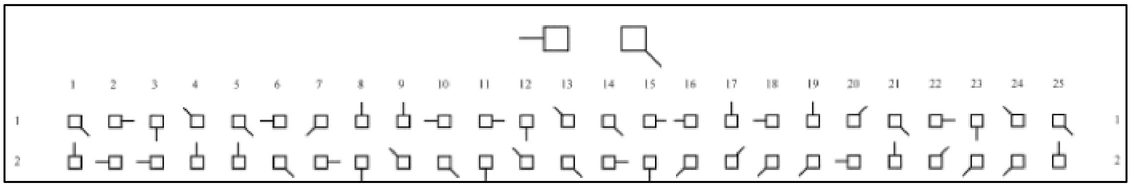
The T2B is used to assess selective attention test. Selective attention is most often associated with focussing attention and refers to the different processing of concurrent sources of information (Goldstein, 2011; Johnston & Dark, 1986). It allows an individual to focus attention on a particular location, object, or message, while simultaneously suppressing irrelevant, competing, or distracting information (Stevens & Bavelier, 2012; Wühr & Wühr, 2022). Different cognitive processes contribute collectively to selective attention. Among the cognitive processes involved in selective attention are filtering mechanisms that reduce cognitive load by prioritizing relevant information (Plebanek & Sloutsky, 2019). Inhibitory processes are required in selective attention as well, to suppress distracting information, preventing interference and enabling individuals to maintain focus on the relevant stimuli (van Moorselaar & Slagter, 2020). Sustained attention is another aspect of selective attention, referring to the ability to maintain focus and vigilance over an extended period, often required in real-life settings (Fisher, 2019). Neurological diseases can cause deterioration in any of these cognitive processes and have a debilitating impact on an individuals' daily life. The T2B assesses different facets of selective attention to quantify possible impairments.

2.3.2 Test Description

A patient is given a DIN A3 sheet vertically on which two target symbols are shown at the top of the sheet (Figure 1). The target symbols are a small square with a line to the left and a small square with a line to the bottom right. From a series of symbols, these two target symbols are to be identified and marked as quickly and accurately as possible, with test duration limited to a maximum of 10 minutes. There are 1000 symbols arranged in 25 columns and 40 rows on the sheet. In addition to the target symbols, other squares with a line to the right, upwards, downwards, upwards to the left, upwards to the right and downwards to the left are listed in random order, but the patient does not have to mark them.

Figure 1

Target Symbols and Top Two Rows



Note. Target symbols depicted on the top of the test sheet and first two rows containing all variations of symbols.

2.3.3 Examination and Protocol

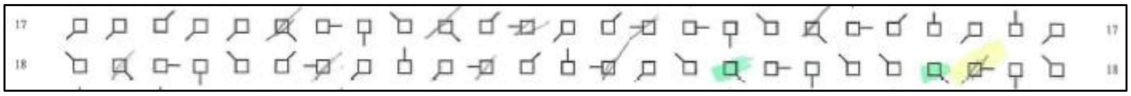
The patient is tasked to mark the two target symbols indicated at the top of the template sheet, from top to bottom and row by row. The patient is asked to work as quickly and as accurately as possible, always working the rows from left to right and both symbols at the same time (not one first, then the other). The first row is considered as a training row, with the test only beginning from the second row onwards. Standard test duration is ten minutes. After each minute, the clinician notes which row and column the patient is currently working through, by reading off the numbered grid around the test sheet. If the patient reaches the end of the sheet before the ten minutes have elapsed, the time during which the test sheet was finished is noted.

2.3.4 Evaluation

Two types of templates for evaluation are provided for this test. Using the first template, the clinician must highlight the target symbols that have been missed by the patient. With the assistance of the second template, the marked non-target symbols must be highlighted next. Afterwards, the exact position (row/column) of the highlighted errors (Figure 2), the position after each minute, and the test duration need to be transferred to an Excel evaluation aid (Figure 3). The number of omissions and the number of mistakes together make up the total errors for each minute. By transferring the errors to the aid, skipped rows are also detected. The evaluation aid compares the relevant test parameters to normative data.

Figure 2

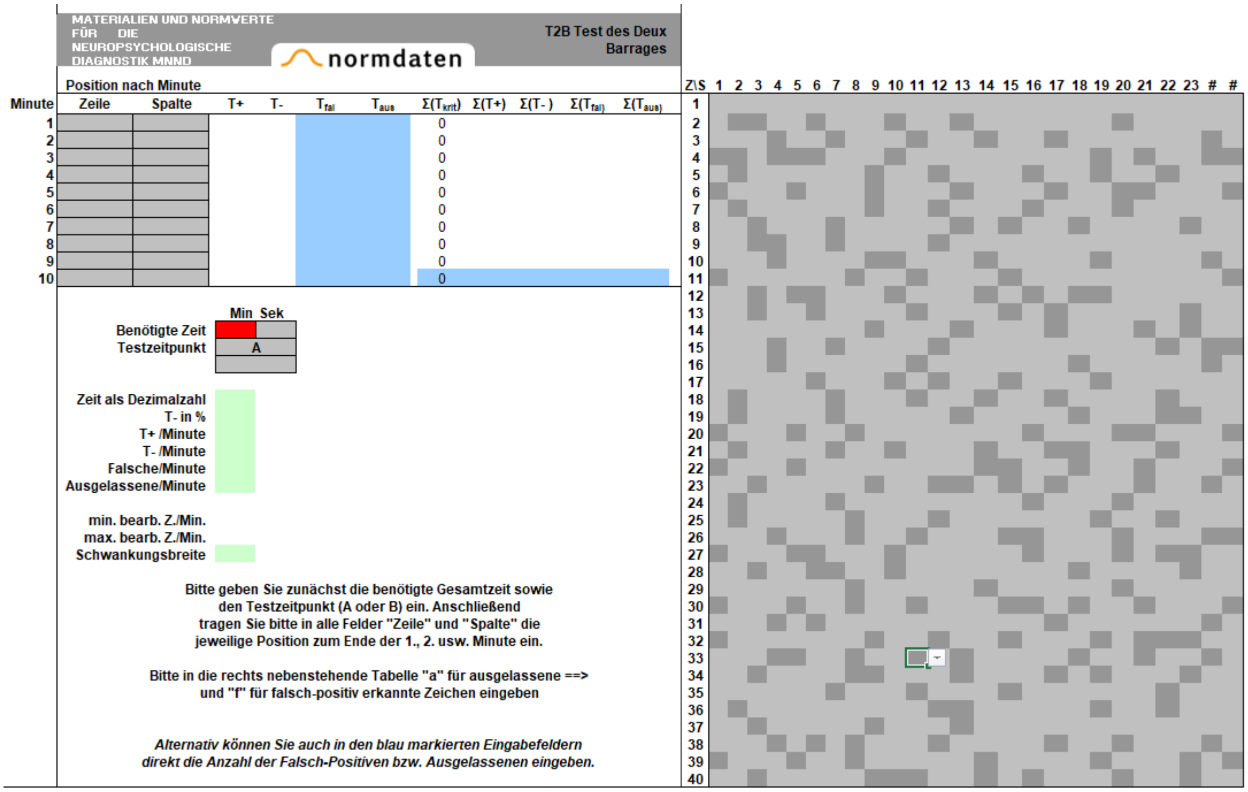
Test Sheet After Evaluation



Note. Highlighted in green are missed target symbols by the patient. Highlighted in yellow is one non-target symbol that has been marked mistakenly by the patient. The exact positions of errors are transferred onto an Excel evaluation aid, which compares the test score to normative data.

Figure 3

Excel Evaluation Aid



MATERIALIEN UND NORMVERTE FÜR DIE NEUROPSYCHOLOGISCHE DIAGNOSTIK MNND

normdaten

T2B Test des Deux Barrages

Position nach Minute

Minute	Zeile	Spalte	T+	T-	T_{fal}	T_{aus}	$\Sigma(T_{krit})$	$\Sigma(T+)$	$\Sigma(T-)$	$\Sigma(T_{fal})$	$\Sigma(T_{aus})$
1							0				
2							0				
3							0				
4							0				
5							0				
6							0				
7							0				
8							0				
9							0				
10							0				

	Min	Sek
Benötigte Zeit		
Testzeitpunkt	A	

Zeit als Dezimalzahl
T- in %
T+ /Minute
T- /Minute
Falsche/Minute
Ausgelassene/Minute

min. bearb. Z./Min.
max. bearb. Z./Min.
Schwankungsbreite

Bitte geben Sie zunächst die benötigte Gesamtzeit sowie den Testzeitpunkt (A oder B) ein. Anschließend tragen Sie bitte in alle Felder "Zeile" und "Spalte" die jeweilige Position zum Ende der 1., 2. usw. Minute ein.

Bitte in die rechts nebenstehende Tabelle "a" für ausgelassene ==> und "f" für falsch-positiv erkannte Zeichen eingeben

Alternativ können Sie auch in den blau markierten Eingabefeldern direkt die Anzahl der Falsch-Positiven bzw. Ausgelassenen eingeben.

Note. Position (row/column) of the highlighted errors need to be transferred by clicking at the respective position on the right and labelling the type of error. The position after each minute and the test duration must be protocolled at designated cells on the left.

2.3.5 Relevant Test Parameters and Interpretation

According to Balzer et al. (2011d) the T2B measures three behavioural components: speed, quality, and temporal course of the performance. The speed of cognitive processing is measured by the number of correctly marked target symbols per minute. Another measure of cognitive processing speed is the test duration. However, since the test is stopped, when the patient takes longer than ten minutes to complete the test sheet, this variable cannot be used to differentiate performance in the lower performance range. The quality or accuracy of processing is recorded with the four variables: Misses per minute, wrongly marked per minute, total mistakes per minute and total mistakes in percentage. Balzer et al. (2011d) emphasize that the last test parameter should be given clear preference when assessing the quality of performance, as it is the only one taking the different probabilities that result from the fact that patients work at different speeds into account (those who work on more symbols in one minute may also miss more target symbols and/or mark more non-target symbols incorrectly in this time). The quality gives insight into a patient's ability to successfully discriminate among competing stimuli and further encompasses the ability to inhibit the processing of irrelevant stimuli. Finally, the range of processing fluctuation provides information about the course of performance over time and about special features of the subject's working behaviour such as initial excitement, constancy, instability, or fatigue. See Table 1 for summary of test parameters.

Table 1

Summary of Test Parameters

Test parameter	**Description**	**Measured function**
Test duration	Total time needed to complete the test sheet in minutes	Cognitive processing speed
Correct per minute	Number of correctly marked target symbols per minute	Cognitive processing speed
Misses per minute	Number of missed target symbols per minute	Quality
Wrongly marked per minute	Number of wrongly marked non-target symbols per minute	Quality
Total mistakes per minute	Sum of misses and wrongly marked symbols per minute	Quality
Total mistakes in percentage	Percentage of mistakes made by the total amount of processed symbols	Quality
Fluctuation	Difference between the highest and lowest amount symbols processed	Temporal course of performance

Note. Adapted from (Balzer et al., 2011d).

2.3.6 Psychometric Properties

Balzer et al. (2011b) provided test-retest reliabilities after four months for all relevant test parameters of the T2B. They ranged from -0.09 for the temporal fluctuation parameter to 0.82 for the number of correctly marked target symbols per minute. Standard error of measurement ranged from 0.42 to 1.04 depending on the test parameter. The authors also presented normative data tables for each test parameter. Each table showing the raw test scores transformed into scores like percentile rank or z-scores, which are commonly used in the clinical field (Balzer et al., 2011c). The normative data was based on a sample of 569 healthy individuals from the German-speaking Swiss population. The sample consists of 292 women and 277 men and covers ages ranging from

16 to 65 years with the average age of the sample being 38.6 years. To correct raw scores from the influence of certain variables control terms were incorporated for age, gender, education, and test version. The authors make no statement about the validity of the test. According to Jäncke (2012), the criterion validity can be considered good. The validity of the similar D2-Test has been studied extensively (Schellig et al., 2009). Intuitively, the reported validity of a similar test would suggest the validity of the T2B measuring selective attention (Schellig et al., 2009). However, to the best of my knowledge, there are no detailed reports of the T2B on its validity or on correlations with other selective attention tests.

2.4 Technological Advances in Neuropsychology

Many researchers have already criticized the delay in technological advances in neuropsychological assessments. Miller and Barr (2017) claim that publishers are culpable for the slow translation of technological developments into the clinical field compared to other medical fields. Instead of considering the implementation of new technological methods to assist and deepen the understanding of patients' cognitive functioning, publishers have so far rather printed 'new' versions with minor changes of the same tests originally derived from the early 20th century. The following sections discuss the opportunities of technological use in neuropsychology.

2.4.1 Complement Traditional Assessment

New technology opens up new possibilities in assessing cognitive deficits in patients with neurological diseases. Libon et al. (2021) have discussed coupling digital technology with traditional paper-pencil tests to identify subtle neuropsychological changes. Pen-pressure or writing speed may be useful as an early indicator before patients meet the diagnostic criteria of mild cognitive impairment or dementia. Tao et al. (2020) have explored the use of eye tracking technology as an addition to traditional cognitive assessments. They concluded that the implementation of eye tracking provides a means of assessing cognitive impairment with higher temporal resolution than cognitive assessment scales. Thus, the use of technology may facilitate more detailed and insightful assessment of neurological diseases.

2.4.2 Share Knowledge Collaboratively

Patients are often tested multiple times over a period of time to estimate disease severity and course, but also to plan and evaluate therapy. Most of the time the results or individual case data are kept exclusively to the testing sites and are only shared for comparison if the same patient is examined again at a different site. Bilder (2011) stated that clinical practice could be drastically improved if the individual case data were collected and shared collaboratively among clinicians and researchers. For example, neuropsychologists and researchers alike could contribute data from examined patients or healthy participants to a database, enabling the comparison of individuals to larger and much more specific reference groups, which current normative data is generally lacking. Computerized tests also enable the assembly of data at item level and the application of modern psychometric approaches, such as item response theory (Bilder & Reise, 2019). Using these approaches, data can be analysed and used to tailor assessment for patients with different demographic characteristics. Data sharing, with informed consent and protected privacy of patients, has huge potential to improve the limitations of normative data and to move away from the one-size-fits-all approach in current neuropsychological assessments.

2.4.3 Place Less Emphasis on Clinicians

Clinicians are highly educated and trained professionals who assess patients with neurological conditions. Their work includes preparation of assessments, aggregation of information about patients, conducting and evaluating assessments, and to write reports that inform doctors, therapists, and insurances among others. While many of these tasks can only be carried out by skilled clinicians, some tasks do not require expertise. The use of technology can serve solutions to alleviate the workload. Computerization has been shown to provide increased ease and standardization of administration, mitigating the effort needed to prepare and instruct tests (Aiello & Depaoli, 2022; Parsey & Schmitter-Edgecombe, 2013). Another example how technology can be used is by simplifying the evaluation of neuropsychological tests. Langer et al. (2022) for example, have demonstrated how machine-learning can be applied to score drawn figures by patients automatically. The trained scoring system from their Rey-Figure Project not only reduced the time to score the drawings but also showcased much lower scoring variance than clinical experts. Besides reducing the workload of clinical experts, these examples proceed to show how the application of technology in the clinical field – and subsequently placing less emphasis on clinicians – can generate more objectivity.

Coming back to the T2B, the evaluation and transfer process may look uncomplicated, but they consume a lot of time for clinicians. Although the templates are provided to aid in the evaluation of the test, the clinician still must check each individual symbol for a mark. According to the survey sent to neuropsychologists in Switzerland, it takes them at least 10 to 15 minutes per test sheet on average. These highly educated clinicians are forced to do this tedious repetitive work on a daily basis, even though the actual evaluation does not require any prior knowledge in the field. While the T2B is a helpful test for the assessment of selective attention over time, the time-consumption and additional workload of the evaluation diminishes its usefulness to some extent. In fact, some clinicians have either replaced the test with a similar one or even discarded it from assessments. In this regard, this thesis proposes an automatic scoring system for the evaluation of the T2B to reduce time-consumption and relieve the workload of clinicians.

3 Research Questions and Hypothesis

Clinicians are assumed to evaluate the test sheets without mistakes of their own. Templates are provided to aid in the evaluation process, but because of the time-consuming and repetitive nature of the evaluation some errors might still occur due to lapses in concentration. The thesis wants to investigate, what the expected or average amount of errors (and types of errors) made by clinicians during the evaluation are.

Secondly, in this thesis, an automatic scoring system will be developed to evaluate the Test des Deux Barrages. The thesis wants to examine, how the automated scoring system performs compared to clinical experts in the evaluation of the T2B. The hypothesis is as follows:

Hypothesis: Performance of the automated scoring system does not differ significantly from clinicians' evaluation.

Thirdly, it is in the interest of the thesis to explore how long on average the developed automatic scoring system takes to evaluate the T2B.

4 Methods

The following sections describe the sample, the procedure and the models developed to classify the symbols on a test sheet and evaluate the Test des Deux Barrages. The code used for the procedure, classification and the results is available at:

https://github.com/Elenoar12/Automatic-Scoring-of-the-Test-des-Deux-Barrages-T2B

4.1 Sample

Test sheets for the thesis were collected from the neuropsychological department at the Bürgerspital in Solothurn. Permission to acquire test sheets was granted by the Solothurner Spitäler AG provided that no information about the patient's identity was visible on the test sheets. Clinicians from the neuropsychological department collected test sheets from every patient they assessed with the MNND. In total a sample size of 23 test sheets was specifically collected for the thesis. Clinicians were asked to scan and save the marked test sheets prior to their evaluation so that the unevaluated test sheet could be used for the development of the scoring system. The scans were done on one scanner at the hospital with no changes made to the default settings and had a resolution of 3308 x 2339 pixels. Clinicians were requested to leave no marks by themselves on the test sheets. Unfortunately, stop signals, marks to signal a mistake made by the patient in the training row, or corrections by the clinicians were still made and could not be prevented on all test sheets. Additionally, for each test sheet, the Excel evaluation aid from the clinicians was gathered. The assessments and evaluations were done by multiple clinicians and one trained intern who worked at the department at the time of collection. To get the ground truth of the patients, each test sheet was evaluated once again by me and an intern at the psychological institute of the University of Zurich. The resulting Excel evaluation aids were thoroughly checked so that the evaluation matched the patients' marks perfectly.

A survey was sent out to neuropsychologists in Switzerland addressing three questions: the length of usage of the T2B (in years), the frequency of usage of T2B (in numbers), and the average evaluation time of the T2B (in minutes). 16 responses were recorded in total. See Table F in Appendix for a summary of responses (survey: https://forms.gle/H1RtnPtm2R2CMwvH9; responses also available at: https://github.com/Elenoar12/Automatic-Scoring-of-the-Test-des-Deux-Barrages-T2B).

4.2 Procedure

In the following sections, the stepwise procedure is described. Test sheets were scanned and saved prior to their evaluation. Each test sheet needed to be processed and prepared before classification.

4.2.1 Image Processing

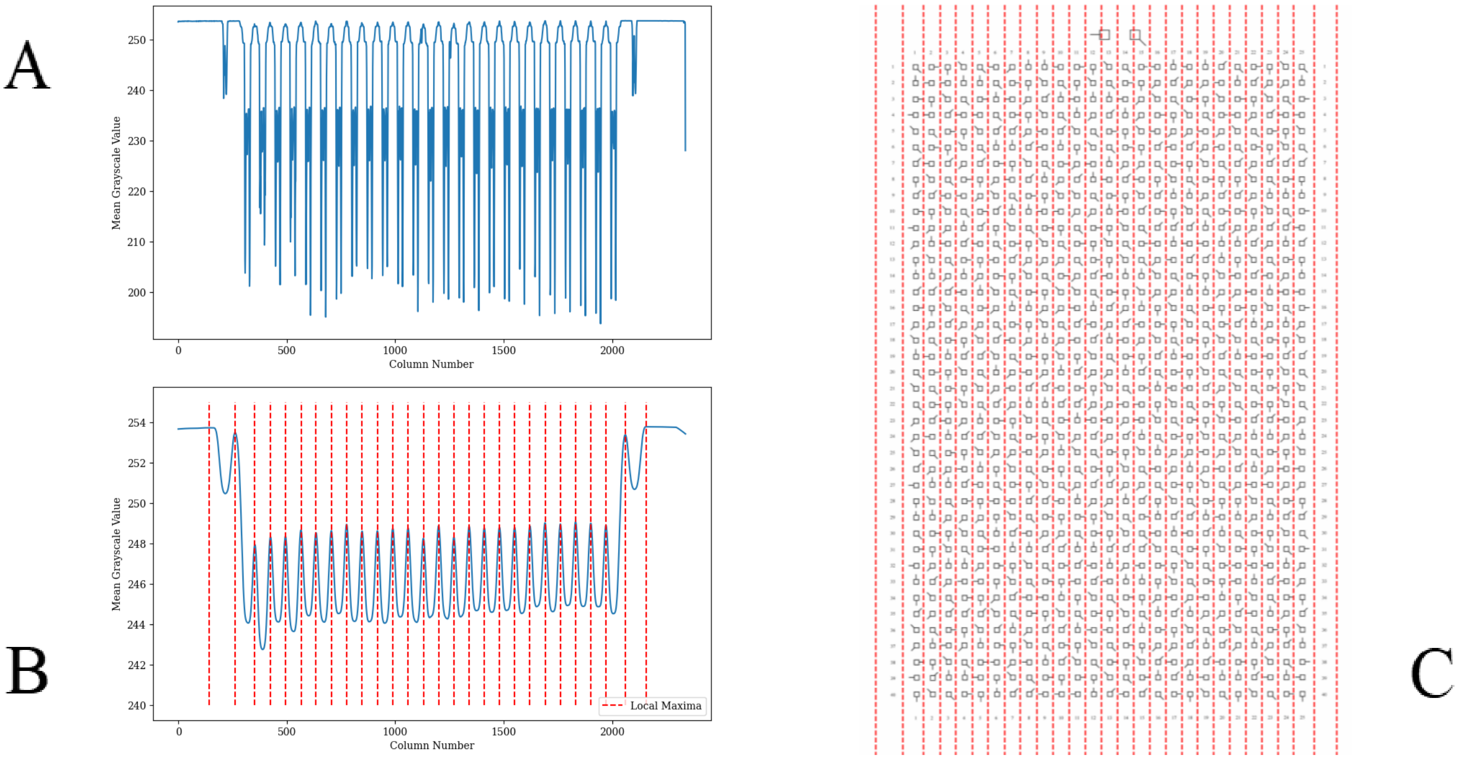
The first stage in the development of the automatic scoring system was to split the collected test sheet into single symbols. A few pre-processing steps were necessary beforehand as the scans obtained by the clinicians were by default saved as pdf files. For image processing and calculations, the scans needed to be converted to an image format (jpeg). The resulting images were then turned into grayscale images, where each pixel of the image has a grayscale value. Grayscale values range from 0 (black) to 255 (white) and can be used to distinguish between the dark ink of the symbols and the brighter areas of the paper around the symbols to split the test sheet into single symbols.

4.2.2 Splitting the Test Sheet

To split the test sheet into single symbols, the areas between the symbols needed to be identified. The symbols on the test sheet are aligned horizontally and vertically and are spaced evenly. Therefore, it can be assumed that the areas between symbols should contain low to no amount of black ink. As brighter pixels have a higher grayscale value than darker pixels, the rows and columns with the highest grayscale values should run right between the symbols. To find the row and columns with the highest grayscale values, each row and column of pixels was examined for their grayscale values and the mean value was calculated and plotted (Figure 4A). The plot was smoothed using a locally weighted smoothing technique (Lowess) to search for the local maxima of the plot. A local maxima or local peak of a function is a point on the graph whose y coordinate is larger than all other points on the graph close by. In the smoothed plot of the mean grayscale values, these maxima would depict the rows and columns of pixels with the highest mean grayscale value (Figure 4B). Lastly, the number of rows and columns was adjusted, so that it matched the number of rows (40) and columns (25) on the T2B, and the distances between the local maxima were adapted evenly. Each symbol could now be split from the test sheet using the row and column information. This splitting method was first tested on an empty test sheet to find the right smoothing parameters (fraction of the data used when smoothing), before being applied to a marked test sheet to check whether the ink of a patient’s marks would skew the splits and the smoothing parameters needed to be changed. After the test sheet has been split into its single symbols, next each symbol was further examined.

Figure 4

Test Sheet Splitting Procedure Shown with Pixel Columns



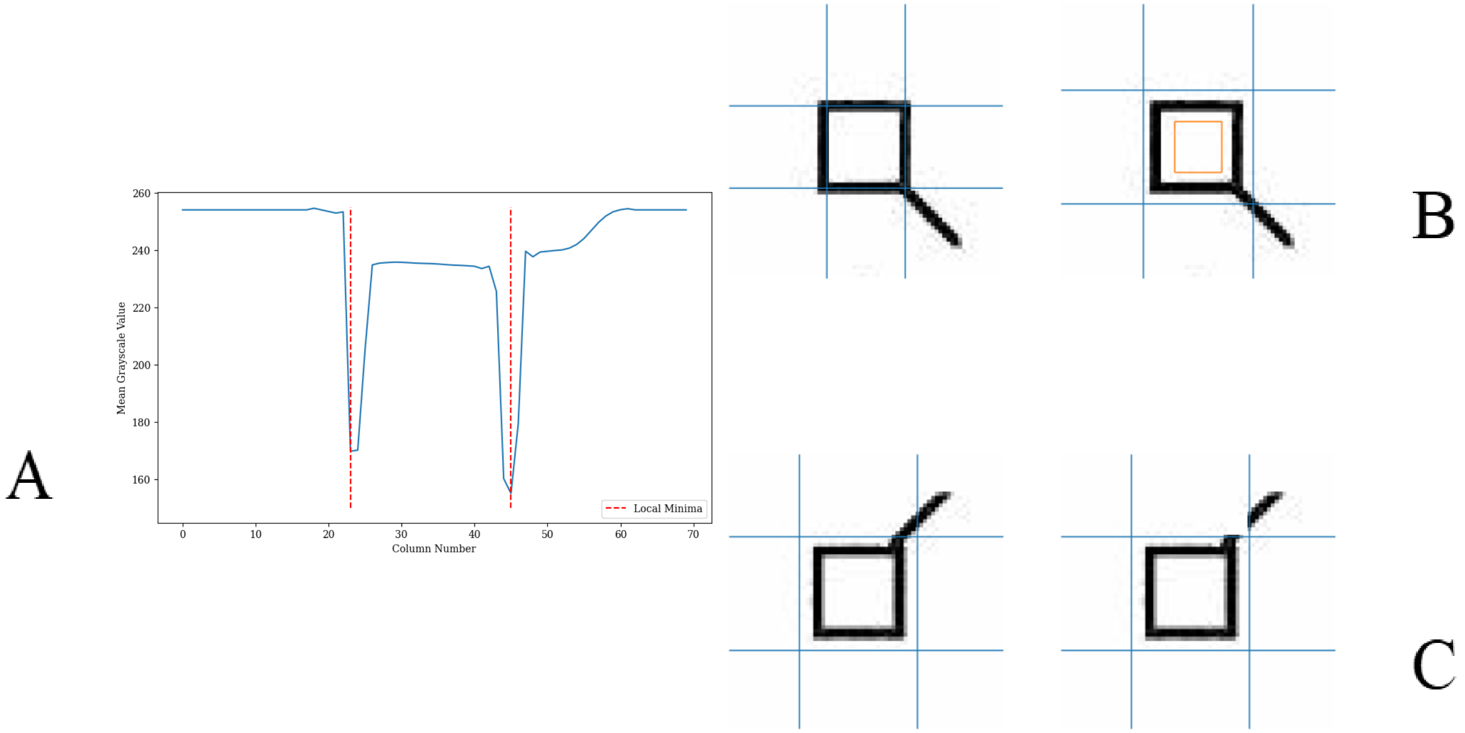
Note. The resolution of the test sheets was always 3308 x 2339 pixels. A: Mean grayscale values of pixel columns plotted. The x-axis shows each column of pixels (2339 columns in total), the y-axis the mean grayscale value for each column. B: Plot after smoothing and local maxima visualized as red dashed lines. Local maxima depict the columns of pixel with the highest mean grayscale values. C: Local maxima or peaks transferred onto the test sheet show that they run between the symbols.

4.2.3 Splitting the Symbols

A mark by the patient is a dark line. To detect such a line the white areas in and around the symbol box were required. Each symbol was split using a similar splitting procedure as in the section before. However, instead of using the local maxima to find the brighter areas, the local minima or local valleys were examined this time to detect the dark borders of the symbol (Figure 5A). After finding the symbol borders using the local minima, the distances between the local minima were either shortened to receive the inner detection box or lengthened to receive 8 areas outside the symbol box, the outer detection boxes (Figure 5B). The goal of adjusting the distances between local minima was to split the symbol into areas that did not contain any dark ink of the symbol borders. These areas would remain white or have a high mean grayscale value if no mark by the patient was present. The symbol variation of a symbol can also be determined using the grayscale value information from the areas outside the symbol box. The area containing the lowest mean grayscale value can be considered as the one area with the variation tail. Using an empty test sheet, the symbol variations for each symbol were determined and saved for the later classification process. After adjusting the distances between local minima, the diagonal symbol variation tails overlapped into neighbouring areas (Figure 5C). To account for overlaps of the diagonal variation tails into more than one area outside the symbol, a small part of the edges of neighbouring areas were set to a grayscale value of 255, depending on the variation. This way the grayscale values of the areas outside the box were corrected from the dark ink of the symbol, which interfered with the calculation of the mean grayscale values and the binary classification by the models described in the next section.

Figure 5

Symbol Splitting Procedure, Detection Boxes and Correction



Note. A: Plot of mean grayscale value of a symbol with local minima or valleys highlighted with red dashed lines. Minima represent the dark ink of the symbol border B: Symbol borders detected using local minima and 9 areas of a symbol after adjusting the distance between the minima (orange = inner detection box, blue = outer detection boxes). C: Overlap of symbol variation tail in multiple areas outside the symbol box before and after correcting the edges of neighbouring areas.

4.3 Models

Three models, a model with inside or outside criteria, a model with inside and outside criteria, and a model with a strict threshold for non-target symbols, were developed for mark recognition with the mean grayscale value as deciding criteria. A threshold was set at a grayscale value of 250, meaning that symbols with a mean grayscale value below the threshold were considered marked. Using the ground truth, all marked and unmarked symbols across all collected test sheets were determined. The threshold was chosen after examining the mean grayscale value distribution of the inner detection boxes of marked and unmarked symbols (Figures 6 & 7). The majority of unmarked symbols showcased a mean grayscale value above 250, which is why the threshold was chosen at a value of 250 and below. Each model uses the criteria and the detection boxes differently. It needs to be mentioned, that the assessment of the performance and shortcomings of the first model, presented in the next section, influenced the other models and their use of the detection boxes. The areas outside the symbol box that contained the symbol variation tail were excluded before running the classification models over the symbols.

4.3.1 Model with Inside OR Outside Criteria

In this model, symbols were first checked for a mark by the patient within the symbol box. If the mean grayscale value was below a set threshold of 250 the symbol was recognized as marked. For any symbol that was not considered marked after evaluating the inner detection box, in a second step the areas outside of the symbol box were checked for their mean grayscale values. If any of the outer detection boxes (apart from the one with the tail of the symbol) had a value below the threshold the symbol was also considered marked.

4.3.2 Model with Inside AND Outside Criteria

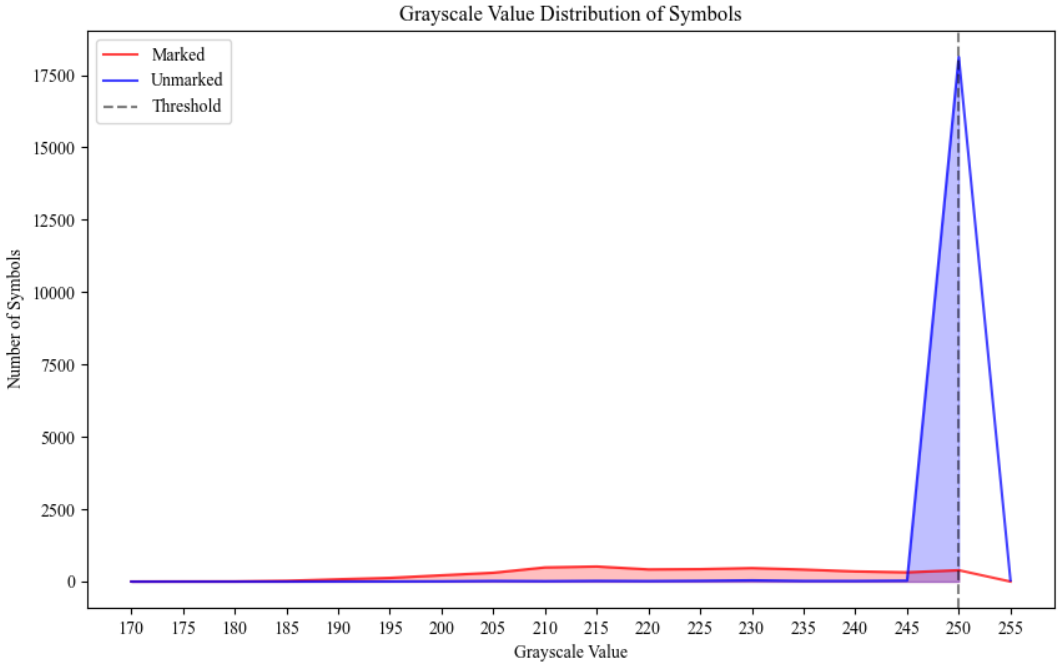
This model evaluated both the inner and outer detection boxes to correct for falsely recognized marks from the first model. For a symbol to be recognized as marked by the patient both types of detection boxes needed to exhibit mean grayscale values below the set threshold of 250. Symbols that did not tick both criteria were not considered marked.

4.3.3 Model with Strict Threshold for Non-Target Symbols

Lastly, this model took the prevalence rate of falsely marked symbols by patients into consideration. The rate of falsely marking a symbol was derived from the ground truth. Overall, 7 out of a possible 17’457 non-target symbols were marked by the patients resulting in a rate of 0.04%. Therefore, the likelihood of marking a non-target symbol can be considered much lower than missing a target symbol. With this in mind, it was decided to set a stricter threshold value of 230 for non-target symbols. As in the model with inside or outside criteria, symbols were first checked for a mark in the inner detection box. Depending on the type of symbol (target or non-target) either the set threshold or the stricter threshold was used for decision-making. Any symbols that were not recognized as marked after evaluating the inner detection box were then checked for marks in the outer detection boxes. Equally, when a non-target symbol was being classified the stricter threshold was used for the outer detection boxes. If any of the outer detection boxes had a value below the respective thresholds the symbol was also considered marked.

Figure 6

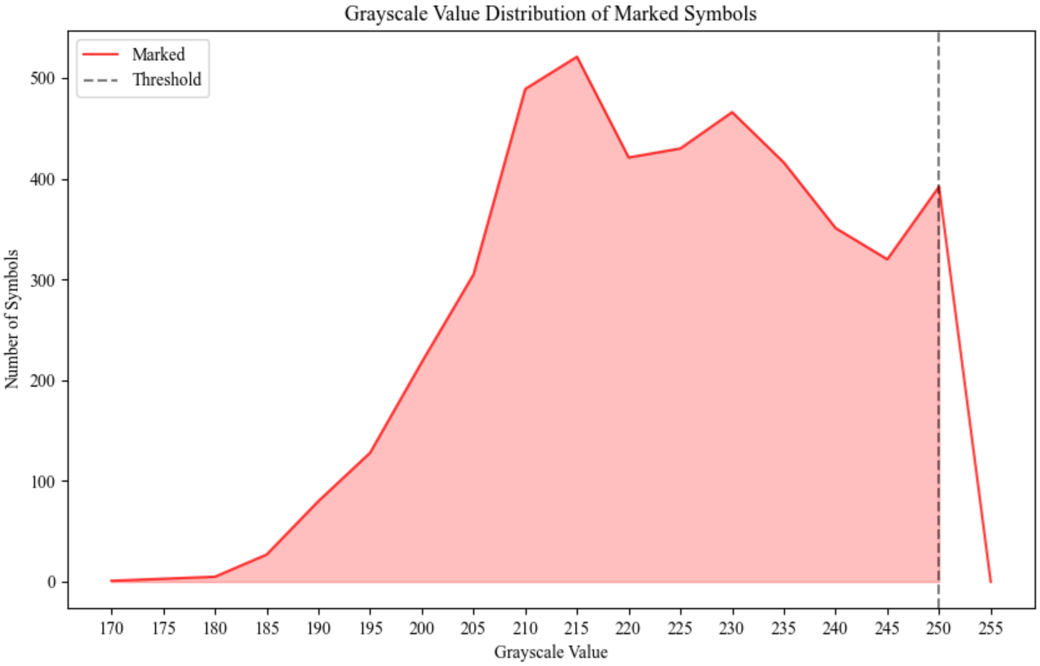
Grayscale Value Distribution of Symbols and Threshold



Note. Grayscale value distribution of all symbols and threshold at 250. Filled out areas under the curves represent the number of symbols that would be detected by the automatic scoring system when the threshold was set at a grayscale value of 250.

Figure 7

Grayscale Value Distribution of Marked Symbols and Threshold



Note. Closer inspection of the grayscale value distribution of marked symbols. Mean grayscale values are distributed more broadly and most importantly mostly below the set threshold of 250. Filled out area under the curve represents the number of marked symbols that would be detected by the automatic scoring system at the set threshold.

4.4 Binary Classification

Signal detection theory provides a model to assess discrimination acuity in decision-making. According to the theory, a signal is only detected when the presence of a signal exceeds a certain threshold (Swets, 2014). Applied to this particular case of classification, the mean grayscale value was used as criteria to decide whether a symbol was marked by the patient or not. Any symbols that had a grayscale value below a set threshold were to be considered marked. The areas inside and outside of the symbol box were used as detection boxes for a patient’s marks. Sensitivity, specificity, and accuracy are key metrics used to assess performance in binary classification (Krupinski, 2017). Sensitivity (SEN) or the true positive rate refers to the model's ability to accurately identify true positive cases (TP = true positive or marked, FN = false negative).

Sensitivity = TP / (TP + FN)

Specificity (SPE) or the true negative rate refers to the model's ability to accurately identify true negative cases (TN = true negative or unmarked, FP = false positive).

Specificity = TN / (TN + FP)

A measure of accuracy can be obtained when combining both these metrics.

Accuracy = (TP + TN) / (TP + FN + TN + FP)

Each model was examined for their accuracy, sensitivity, and specificity when compared to the ground truth of the test sheets. Receiver Operating Characteristic (ROC) analysis is commonly employed to investigate the relationship between the sensitivity and specificity of a binary classifier (P. A. Flach, 2016). Commonly, the true positive rate is plotted against the false positive rate for this. ROC curves at different thresholds were plotted to evaluate the quality of classification when using the mean grayscale value of the inner detection box as criteria.

4.5 Imbalanced Data

The dataset is imbalanced due to the characteristics of the Test des Deux Barrages. Only 241 out of 1000 symbols on a test sheet are target symbols. Extrapolated to 23 test sheets a total of 5543 symbols should have been marked, while 17457 symbols were tasked to be left unmarked. The actual number of marked and unmarked symbols was different because not all patients solved the task at the same speed and within the test duration. When the actual test duration and the last processed symbol (exact position after 10 minutes) of each patient were accounted for, the number of marked symbols was 4576 and the number of unmarked symbols was 15941. For classification, the unprocessed symbols (unmarked symbols after the last processed symbol) were also included in the ground truth, which resulted in 18424 unmarked symbols (Figure 8). If there is a significant imbalance in the distribution of classes, accuracy, and ROC curves may give a misleadingly optimistic impression of an algorithm's performance. In such cases, balanced accuracy (b-ACC) can be used instead. B-ACC is a further development on the standard accuracy metric and calculates the average accuracy for each class rather than combining them. In binary classification, it is equal to the arithmetic mean of sensitivity and specificity (Kelleher et al., 2015).

$$\text{Balanced accuracy} = 0.5 \text{ x } (\text{Sensitivity} + \text{Specificity})$$

In situations where the dataset is heavily imbalanced, Precision-Recall (PR) curves provide a better and more detailed understanding of an algorithm's performance than ROC curves (Davis & Goadrich, 2006). Precision is defined as the number of true positives over the number of true positives and false positives.

$$\text{Precision} = TP / (TP + FP)$$

Recall is defined as the number of true positives over the number of true positives and false negatives. Looking at the equation for recall below it is the same as the equation for sensitivity.

$$\text{Recall} = TP / (TP + FN)$$

The PR curve shows the trade-off between precision and recall or sensitivity when the thresholds are changed. High precision relates to a low false positive rate, whereas high recall relates to a low false negative rate. Closely related to precision and recall is also the F-measure (F1), another metric to assess a model’s accuracy when the classes are imbalanced. According to

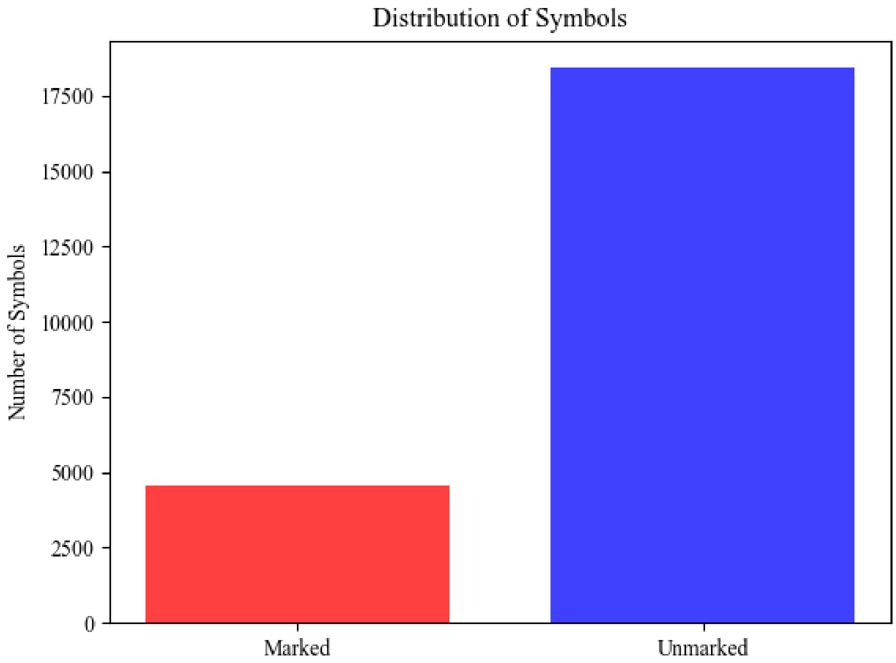
(P. Flach & Kull, 2015), the F1 score can be interpreted as the weighted harmonic mean of the precision and recall and represents both in one metric.

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

In conclusion, PR curves for different grayscale value thresholds were additionally plotted to the ROC curves to evaluate the prediction quality when classes are very imbalanced. For the assessment of a model's performance the standard accuracy metric was replaced by the balanced accuracy and F1 scores to receive a more weighted estimate of the accuracy.

Figure 8

Distribution of Symbols



Note. Distribution of marked and unmarked symbols derived from the ground truth of the patients showing imbalance in dataset/classes.

4.6 Model Comparison to Clinicians

To compare the performances of the different models to the performance of the clinicians the imbalance in data and the fact that all models were performed on the same dataset needs to be taken into consideration. A statistical test that does not require a normal distribution of the data is called a nonparametric test. The sign test and the Wilcoxon signed-rank test are such nonparametric statistical tests that can be used to compare two models, which were evaluated on the same sample (Corder & Foreman, 2014; Scheff, 2016).

4.6.1 Sign Test

According to Corder and Foreman (2014) the null hypothesis, when performing a sign test, conventionally assumes that the distribution of the direction of differences (dichotomous: positive (+) direction, negative (-) direction, therefore the name sign test) is symmetric about zero and the probability of each direction is equal to 0.5, following a binomial distribution. The sign test checks for the number of positive and negative differences between two models and chooses the smaller number of both as *k*. The number of observations *n* used in the analysis consists of the sample size minus the number of zero differences between two models. The *p* value is calculated using *k* and *n* and stands for the probability of finding the observed number of successes *k* or a more extreme number, given that the null hypothesis, that the probability of success is 0.5, is true.

4.6.2 Wilcoxon Signed-Rank Test

When performing the Wilcoxon signed-rank test, the null hypothesis assumes that the continuous distribution of the differences is symmetric about zero corresponding to no difference between the two samples (Rey & Neuhäuser, 2011). Similar to the sign test, the Wilcoxon signed-rank test also checks for the number of positive and negative differences and excludes zero differences from the number of observations. Additionally, this test also ranks the differences according to the size of the difference. Both positive and negative ranks are then summed and as test statistic *W* the smaller of the sums is used.

Each classification model was compared to the clinicians using the sign test and the Wilcoxon signed-rank test to analyse whether the directions and the median of differences of the performance metrics (balanced accuracy, F1 score, sensitivity, and specificity) between the model and the clinicians differed significantly.

4.7 Terminology

In this section, the terminology of errors is described to avoid confusion, when discussing the errors from different sources (patient, clinician, or model). It is important to distinguish between mistakes made by patients, errors made by clinicians, and errors made by the models while assessing the performance. While the patients' performance is compared to a template with perfect score, the performance of the clinicians and the models is compared to ground truth templates derived from the patients' performances.

The perfect score template, with which the patients are evaluated, gives information about the target and non-target symbols. A mistake by the patient is decided by whether the patient has not marked a target symbol or marked a non-target symbol. Therefore, in the following sections, whenever the performance of patients is involved, it will be termed as marked correctly or incorrectly.

The performance of the clinicians and the models is decided by comparing their evaluation to the ground truth. Clinicians not only have to identify mistakes made by the patients on the test sheet but also transfer the exact positions to an Excel evaluation aid to label the mistakes. Since only the unevaluated test sheets and the Excel evaluation aids from the clinicians were made available, it was not possible to discern whether an error occurred in the evaluation process (missing to highlight a mistake by the patient on the test sheet) or in the transfer process (forgetting or misplacing the label of a mistake). As both errors result in incorrectly labelling the mistakes onto the evaluation aid, whenever the performance of the clinicians is presented in the following sections, it will be termed as labelled correctly or incorrectly.

Lastly, the performance of the models is based on the mean grayscale values within the detection boxes. A mark by the patient is recognized as such when the mean grayscale value is below the threshold for recognition. For that reason, whenever the scripts' performance is reviewed in the following sections, it will be termed as recognized correctly or incorrectly. See Table 2 for a summary of terminology for performance.

Table 2

Terminology for Performance

	Compared to	**Error**	**Terminology**
Patient	Perfect score template	Misses target symbol Marks non-target symbol	Marked correctly or marked incorrectly
Clinician	Ground truth	Misses unmarked target symbol or marked non-target symbol Labels marked symbol unmarked	Labelled correctly or labelled incorrectly
Model	Ground truth	Recognizes unmarked symbol as marked Recognizes marked symbol as unmarked	Recognized correctly or recognized incorrectly

5 Results

5.1 Performance of Patients

Test sheets and Excel evaluation aids from 23 patients were collected for this thesis. Each test sheet was evaluated again and checked thoroughly to get the ground truth of the patients. The ground truth of each patient was compared to a template with perfect score to receive the performance of the patients. Patients made a total of 488 errors across all test sheets. 7 non-target symbols were incorrectly marked by the patients at a rate of 0.30 per test sheet. 481 target symbols were missed by the patients at a rate of 20.91 per test sheet. The performance of the patients is included in Table 3 but was not compared to the models or the clinicians using statistical tests.

5.2 Performance of Clinicians

The following sections examine the expected number of errors and types of errors made by clinicians in the evaluation.

5.2.1 Expected Amount of Errors

The clinicians made 35 errors across all test sheets. 23 symbols were labelled marked by the clinicians, even though the patients had not marked them, at a rate of 1 per test sheet. 12 symbols were labelled unmarked, even though the patients had marked them, at a rate of 0.52 per test sheet. The average balanced accuracy across all test sheets is 0.9980, the average F1 score is 0.9959, the average sensitivity is 0.9972 and the average specificity is 0.9988. The performance metrics across all symbols (instead of individual test sheets) give a balanced accuracy of 0.9981, a F1 score of 0.9962, a sensitivity of 0.9974, and a specificity of 0.9988 for the clinicians. Detailed results are summarized in Table 3.

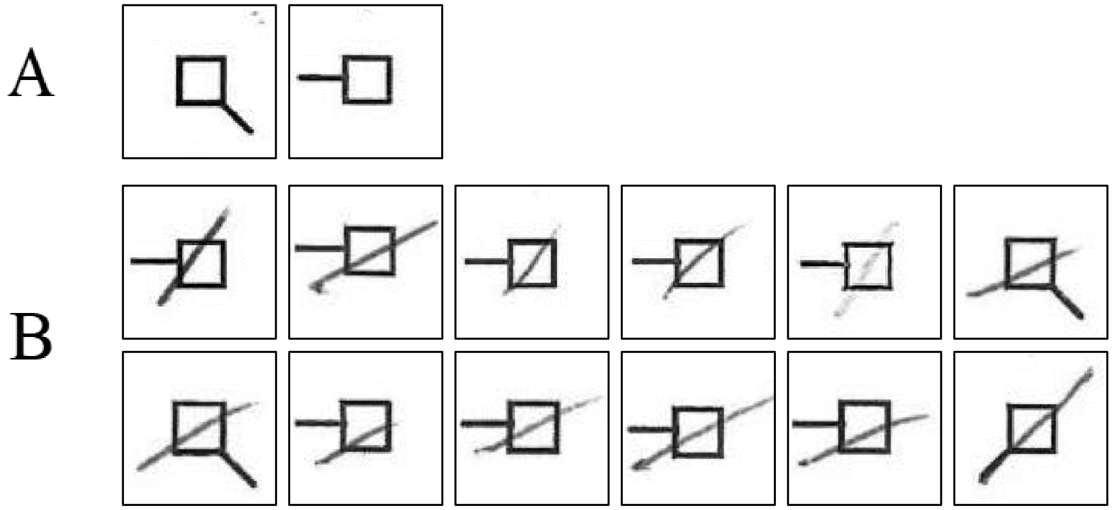
5.2.2 Types of Errors

Clinicians made two types of errors, which either occurred in the evaluation process of the test sheets by hand using the templates or in the transfer process of the exact positions of mistakes by the patients onto the Excel evaluation aid. In total, 23 unmarked target symbols were missed by the clinicians and consequently not labelled as unmarked on the Excel evaluation aid (Figure 9A). These target symbols were given as correctly marked symbols, even though the patients had missed them. This type of error most likely happened while evaluating the test sheet by hand with the templates. They were either not seen or forgotten in the transfer process. It was not possible to discern whether an error occurred in the evaluation process (missing to highlight a mistake made by the patient on the test sheet) or in the transfer process (forgetting or misplacing the label of a mistake) since only the unevaluated test sheets and the Excel evaluation aids from the clinicians were made available.

The second type of error was labelling symbols as unmarked, even though the patients had marked them (Figure 9B). 12 symbols were labelled incorrectly in this way. Depending on whether the symbol was a target or non-target symbol changes the source of the error. For marked target symbols, the error occurred while transferring the exact positions of the mistakes made by the patients. When a symbol is labelled in the wrong position on the Excel sheet an otherwise marked symbol is mistaken as unmarked. Similarly, if the exact position after the test duration is transferred incorrectly, marked symbols are excluded or considered unprocessed and labelled unmarked. For non-target symbols though, the error may have occurred, because the symbol was not seen when evaluating with the templates or forgotten in the transfer process.

Figure 9

Types of Errors Made by Clinicians



Note. A: Missed unmarked target symbols, labelled incorrectly as marked. B: Marked symbols labelled incorrectly as unmarked.

5.3 Model performance

The following sections address the performance of the developed models. The output of each model was compared to the ground truth of the patients and examined for their performance metrics. Performance metrics of each model were contrasted to the performance metrics of the clinicians using the sign test and Wilcoxon signed-rank test.

5.3.1 Model with Inside OR Outside Criteria

This model made 271 errors across all test sheets. 78 symbols were recognized as marked by the method, even though the patients had not marked them, at a rate of 3.39 per test sheet. 193 symbols were recognized unmarked, even though the patients had marked them, at a rate of 8.39 per test sheet. The average balanced accuracy across all test sheets is 0.9778, the average F1 score is 0.9651, the average sensitivity is 0.9597 and the average specificity is 0.9958. The performance metrics across all symbols give a balanced accuracy of 0.9768, a F1 score of 0.9700, a sensitivity of 0.9578, and a specificity of 0.9958 for this method. Detailed results are summarized in Table 3.

The sign test indicated that the median direction of balanced accuracy scores ($k = 4$, $n = 22$, $p = 0.004$), F1 scores ($k = 3$, $n = 22$, $p = 0.001$), and specificity scores ($k = 2$, $n = 19$, $p = 0.001$) was statistically significantly different from the clinicians, indicating better performance of the clinicians. The median direction of sensitivity scores ($k = 4$, $n = 15$, $p = 0.118$) were statistically not significantly different from the clinicians.

The Wilcoxon signed-rank test indicated that the median balanced accuracy scores ($W = 49.5$, $p = 0.007$), F1 scores ($W = 28.5$, $p = 0.001$), and specificity scores ($W = 35$, $p = 0.002$) were statistically significantly different from the clinicians. The median sensitivity scores ($W = 80$, $p = 0.076$) were statistically not significantly different from the clinicians.

5.3.2 Model with Inside AND Outside Criteria

This model made 612 errors across all test sheets. 28 symbols were recognized as marked by the method, even though the patients had not marked them, at a rate of 1.22 per test sheet. 584 symbols were recognized unmarked, even though the patients had marked them, at a rate of 25.39 per test sheet. The average balanced accuracy across all test sheets is 0.9385, the average F1 score is 0.9207, the average sensitivity is 0.8785, and the average specificity is 0.9985. The performance metrics across all symbols give a balanced accuracy of 0.9354, a F1 score of 0.9288, a sensitivity of 0.8724, and a specificity of 0.9985 for the model. Detailed results are summarized in Table 3.

The sign test indicated that the median direction of balanced accuracy scores ($k = 2$, $n = 22$, $p < 0.001$), F1 scores ($k = 2$, $n = 22$, $p < 0.001$), and sensitivity scores ($k = 1$, $n = 19$, $p < 0.001$) were statistically significantly different from the clinicians. The median direction of specificity scores ($k = 9$, $n = 18$, $p = 1.0$) was statistically not significantly different from the clinicians.

The Wilcoxon signed-rank test indicated that the median balanced accuracy scores (*W* = 10.5, *p* < 0.001), F1 scores (*W* = 17.5, *p* < 0.001), and sensitivity scores (*W* = 12, *p* < 0.001) were statistically significantly different from the clinicians. The median specificity scores (*W* = 122.5, *p* = 0.637) were statistically not significantly different from the clinicians.

5.3.3 Model with Strict Threshold for Non-Target Symbols

This model made 102 errors across all test sheets. 22 symbols were recognized as marked by the method, even though the patients had not marked them, at a rate of 0.96 per test sheet. 80 symbols were recognized unmarked, even though the patients had marked them, at a rate of 3.48 per test sheet. The average balanced accuracy across all test sheets is 0.9911, the average F1 score is 0.9875, the average sensitivity is 0.9835, and the average specificity is 0.9988. The performance metrics across all symbols give a balanced accuracy of 0.9907, a F1 score of 0.9888, a sensitivity of 0.9825, and a specificity of 0.9988 for the model. Detailed results are summarized in Table 3.

The sign test indicated that the median direction of balanced accuracy scores (*k* = 7, *n* = 16, *p* = 0.804), F1 scores (*k* = 8, *n* = 16, *p* = 1.0), sensitivity scores (*k* = 3, *n* = 7, *p* = 1.0) and specificity scores (*k* = 8, *n* = 16, *p* = 1.0) were statistically not significantly different from the clinicians.

The Wilcoxon signed-rank test indicated that the median balanced accuracy scores (*W* = 129, *p* = 0.784), F1 scores (*W* = 128, *p* = 0.760), sensitivity scores (*W* = 129, *p* = 0.776) and specificity scores (*W* = 137, *p* = 0.976) were statistically not significantly different from the clinicians.

Table 3

Performance Analysis

		TP	TN	FP	FN	b-ACC	F1	SEN	SPE
Patients*	Average performance								
	Macro performance	198.65 (4569)	693.09 (15941)	0.30 (7)	20.91 (481)				
Clinicians	Average performance					0.9980	0.9959	0.9972	0.9988
	Macro performance	198.43 (4564)	800.04 (18401)	1.0 (23)	0.52 (12)	0.9981	0.9962	0.9974	0.9988
Model with inside OR outside criteria	Average performance					0.9778	0.9651	0.9597	0.9958
	Macro performance	190.57 (4383)	797.65 (18346)	3.39 (78)	8.39 (193)	0.9768	0.9700	0.9578	0.9958
Model with inside AND outside criteria	Average performance					0.9385	0.9207	0.8785	0.9985
	Macro performance	173.57 (3992)	799.83 (18396)	1.22 (28)	25.39 (584)	0.9354	0.9288	0.8724	0.9985
Model with strict threshold for non-target symbols	Average performance					0.9911	0.9875	0.9835	0.9988
	Macro performance	195.48 (4496)	800.09 (18396)	0.96 (22)	3.48 (80)	0.9907	0.9888	0.9825	0.9988

Note. In parentheses total number of respective metrics otherwise presented as rates per test sheet. See Tables A-E in Appendix for detailed performance of models and clinician for each test sheet. *Patients performance was not included in the statistical analysis and is only shown descriptively in this table.

Table 4

Model Comparison with Clinicians

		b-ACC	F1	SEN	SPE
Model with inside OR outside criteria	Sign test	$p = 0.004$	$p = 0.001$	$p = 0.118$	$p = 0.001$
	Wilcoxon signed-rank test	$p = 0.007$	$p = 0.001$	$p = 0.076$	$p = 0.002$
Model with inside AND outside criteria	Sign test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 1.0$
	Wilcoxon signed-rank test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.637$
Model with strict threshold for non-target symbols	Sign test	$p = 0.804$	$p = 1.0$	$p = 1.0$	$p = 1.0$
	Wilcoxon signed-rank test	$p = 0.784$	$p = 0.760$	$p = 0.776$	$p = 0.976$

Note. The *p* values of model comparison with clinicians. Each performance metric was compared to the same performance metric of the clinicians to calculate the *p* values.

5.4 ROC and PR Analysis

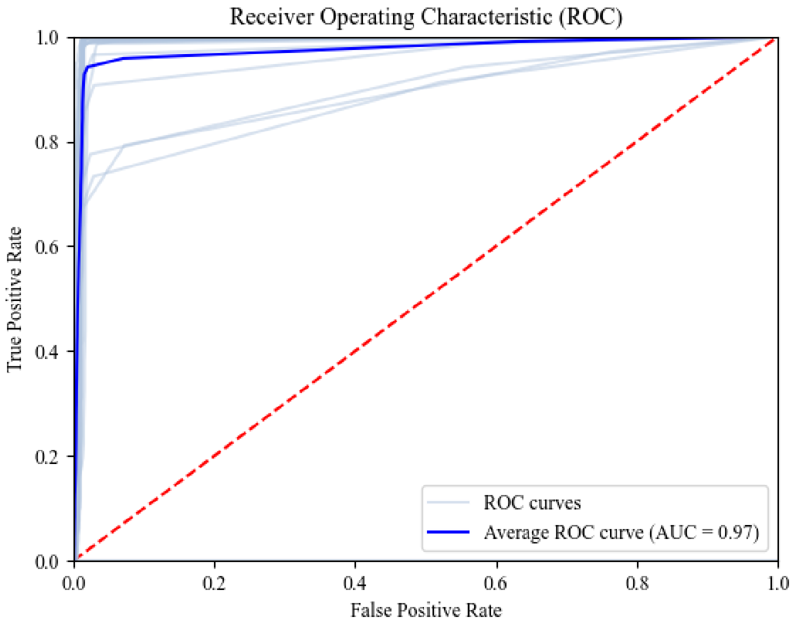
The sections that follow describe the results from the ROC and PR analysis to evaluate the quality of classification when using the mean grayscale value of the inner detection box as criteria.

5.4.1 Receiver Operating Characteristic

To analyse the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) of the classification model at different possible mean grayscale value thresholds ROC curves were plotted for each test sheet and an average ROC curve was calculated with an area under the curve of 0.97 (Figure 10).

Figure 10

Receiver Operating Characteristic (ROC) Curve



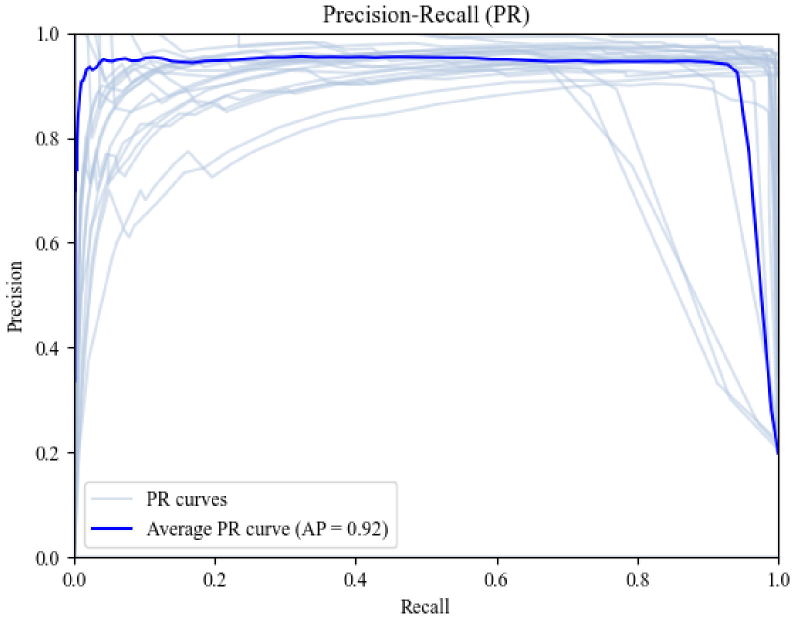
Note. ROC curves for each test sheet (in light blue) and an averaged ROC curve from the individual ROC curves. Red dashed line represents the ROC curve for a random guess.

5.4.2 Precision Recall

Due to restrictions in the analysis of ROC curves in the presence of imbalanced data, an average PR curve is calculated from individual PR curves for each test sheet (Figure 11). The average precision is 0.92.

Figure 11

Precision-Recall (PR) Curve



Note. PR curves for each test sheet (in light blue) and averaged PR curve from the individual PR curves.

5.5 Error Analysis

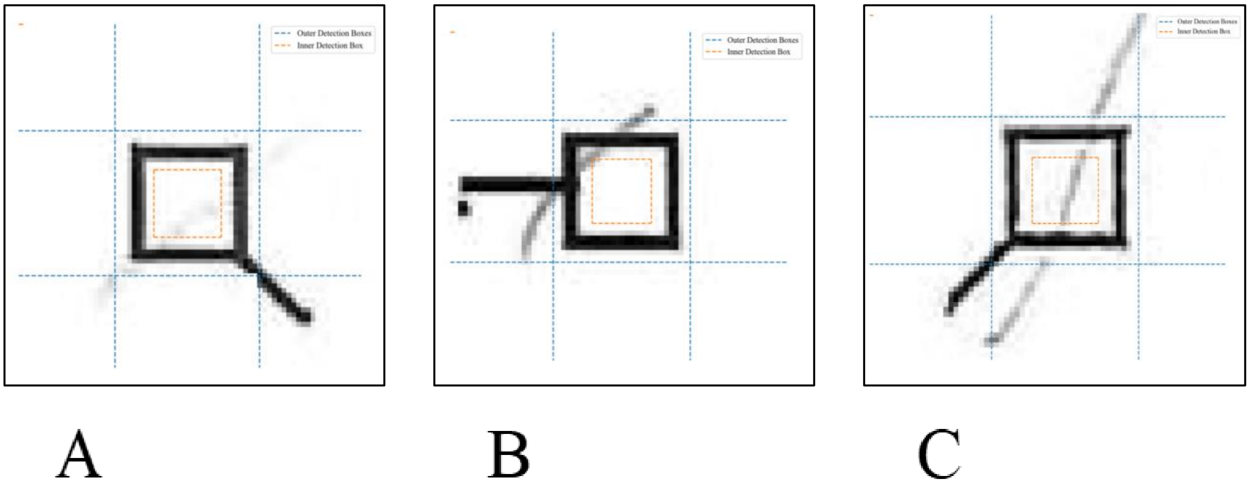
When analysing the errors made by the script some errors occurred across all three models, while others were attributed more to one model or the deciding criteria. Errors made by the model with a strict threshold for non-target symbols are checked in this section to showcase general and specific errors to the model. 102 errors were made by the model. When examined closer these errors can be grouped into types of errors. 80 marked symbols were incorrectly recognized as unmarked, whereas 22 unmarked symbols were incorrectly recognized as marked by the model.

5.5.1 Incorrectly Recognized as Unmarked

Of the 80 falsely unrecognized marked symbols 76 errors can be traced back due to the visibility of a patient's marks (Figure 12A). These symbols were not recognized due to their mean grayscale value being above the set threshold of 250. 3 marked symbols were not recognized, because the patients' marks ran through the corridor or 'blind spot' between the detection boxes (Figure 12B). Both of these types of errors occurred in all models. Finally, one marked symbol was not recognized off the strict threshold for non-target symbols (Figure 12C). This missed symbol was specific to the model with a strict threshold.

Figure 12

Marked Symbols Incorrectly Recognized as Unmarked



Note. Examples of marked symbols that were incorrectly recognized as unmarked A: Error caused by the of low visibility of the patients' mark. B: Error because mark was not sufficiently detected by the detection boxes. C: Error specific to the model with strict threshold for non-target symbols.

Table 5

Error Analysis of Unrecognized Marked Symbols

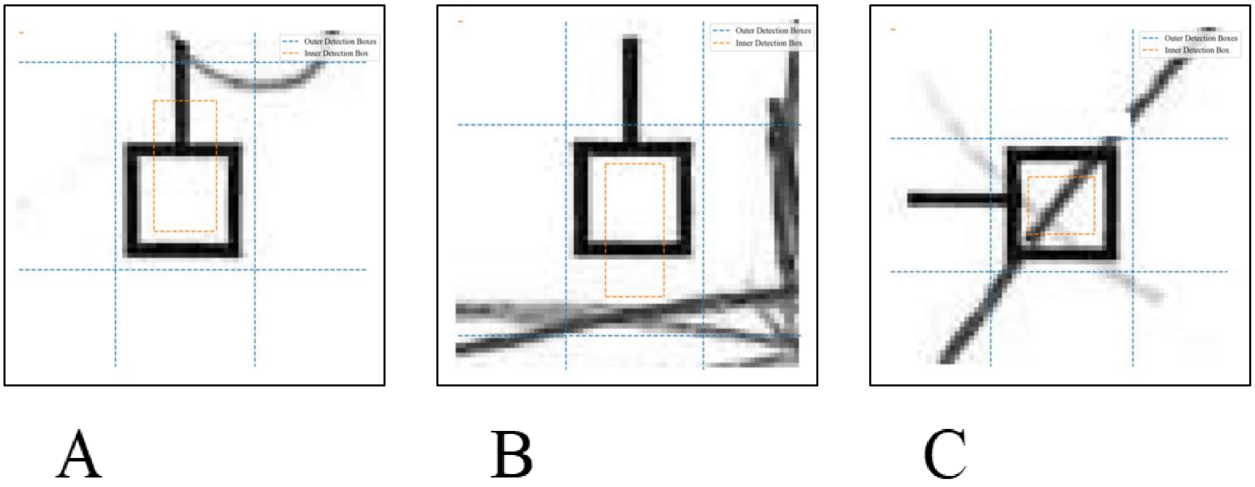
Type of error	**Number of errors**	**Example**
Visibility	76	Figure 12A
'Blind spot'	3	Figure 12B
Model-specific	1	Figure 12C
Total	80	

5.5.2 Incorrectly Recognized as Marked

10 out of 22 unmarked symbols that were incorrectly recognized as marked came as a result of a sign or notice made by the clinicians. A typical example of a clinician's sign was the stop signal made at the last marked or processed symbol by the patient (Figure 13B). These signs by the clinicians were picked up by the detection boxes and led to incorrect recognition (Figure 13). Another 6 were caused by corrections made by the patients. These symbols were incorrectly recognized as marked even though the patients had corrected themselves after marking a non-target symbol (Figure 14A). 2 symbols were recognized as marked because a mark by the patient from a neighbouring marked symbol overlapped into the symbol that was being examined by the model (Figure 14B). Finally, 4 symbols were given as marked because of the scan quality. Scanner irregularities, either through dirt on the scanners' glass or the paper were noticed by the model and recognized incorrectly (Figure 15A). Similarly, creases on the paper left darker imprints on the test sheets which disrupted the recognition process and led to errors (Figure 15B).

Figure 13

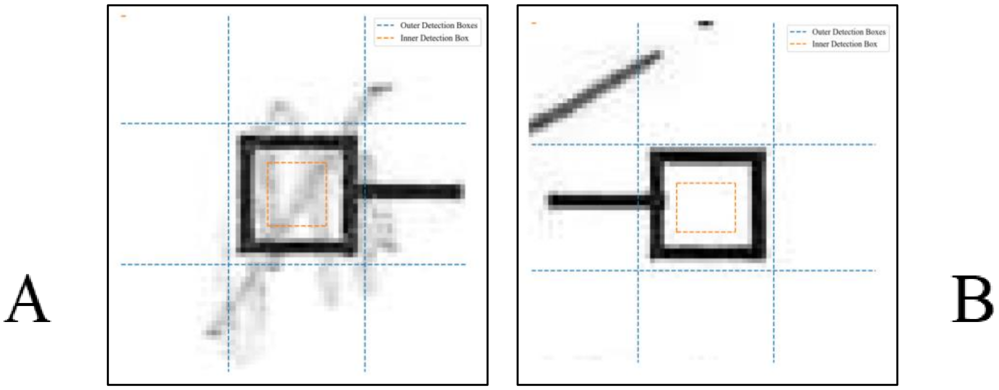
Signs by Clinicians



Note. Signs made by clinicians that lead to errors in the classification of symbols. A: Overlap from a circled symbol that signalled a mistake by the patient in the training row. B: Stop signal made by the clinician at the last processed symbol by the patient. C: Symbol that had been marked by the patient after test duration had run out. Corrected by the clinician.

Figure 14

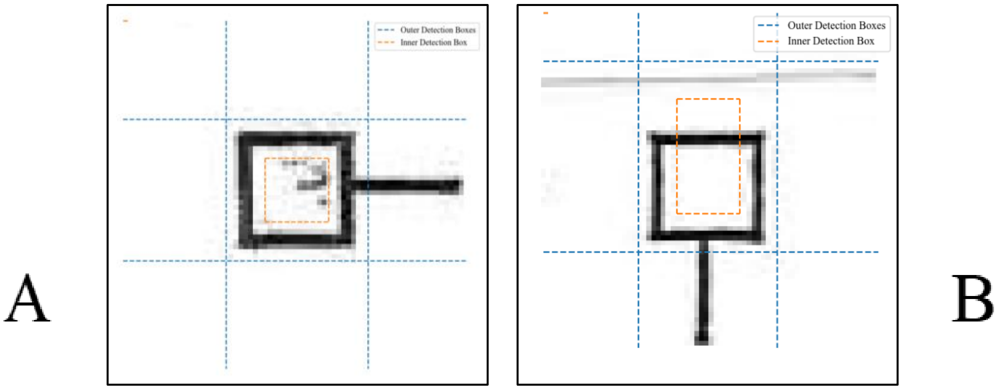
Correction by Patients and Mark Overlap



Note. A: Marked non-target symbol that has been corrected by the patient. B: Overlapping mark by a neighbouring symbol leading to false recognition.

Figure 15

Scanner Irregularity and Scan Quality



Note. Scanner irregularity and scan quality issues. A: Dirt on scanner glass or the paper which was picked up as mark. B: Crease from a folded test sheet, which left a dark imprint on the paper leading to false recognition.

Table 6

Error Analysis of Misrecognized Unmarked Symbols

Type of error	Number of errors	Example
Sign by clinician	10	Figure 13
Correction by patient	6	Figure 14A
Overlapping mark	2	Figure 14B
Scan issues	4	Figure 15
Total	22	

5.6 Execution Time

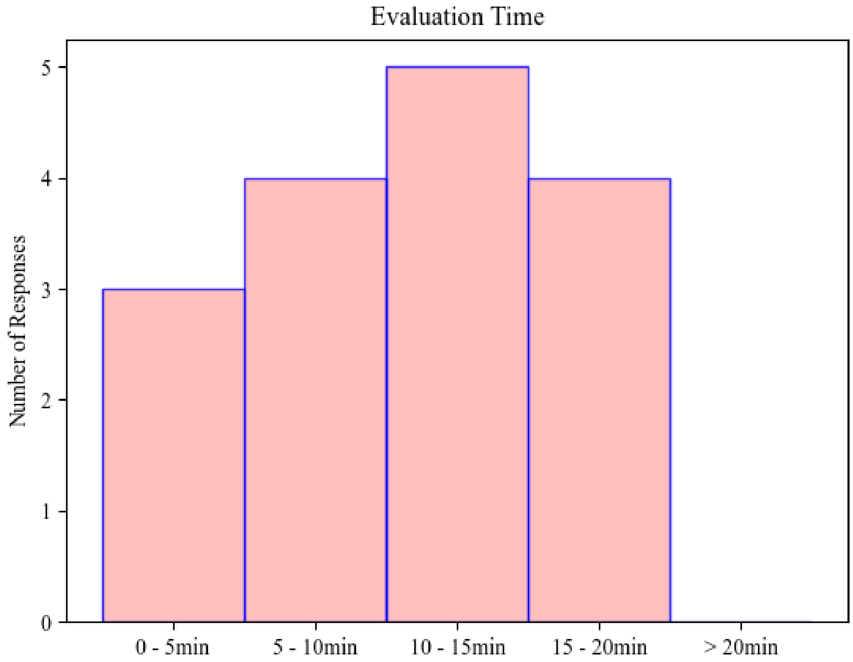
The execution time for the procedure and classification steps was calculated for each test sheet. The average execution time was around 13.5 seconds per test sheet. Execution time does not include time needed for scanning the test sheet, converting the pdf files to images and rotation of the images before the procedure steps and visualization of results, and transfer of errors onto evaluation aid after classification.

5.7 Reported Evaluation Time

In a survey the average evaluation time for the T2B (including the transfer process onto the Excel evaluation aid) was questioned. The survey was sent to neuropsychologists in Switzerland. In total 16 responses were received. Responders could choose from five response categories (1 = 0 – 5 minutes, 2 = 5 – 10 minutes, 3 = 10 – 15 minutes, 4 = 15 – 20 minutes, 5 = over 20 minutes). 10 to 15 minutes for the evaluation was reported the most by the neuropsychologists. No responders reported an evaluation time over 20 minutes. See Figure 16 for number of responses per response category.

Figure 16

Evaluation Time Reported in Survey



Note. Histogram of average evaluation time (evaluation + transfer) reported by neuropsychologists in survey.

6 Discussion

In this thesis, an automatic scoring system was developed to evaluate the Test des Deux Barrages, a selective attention test used in neuropsychological assessments in Switzerland for patients aging from 18 to 65 years. Patients are tasked to mark two target symbols, out of eight possible target symbols, on an A3-sized paper test sheet as fast and accurately as possible. The purpose of the project was to provide an automatic evaluation framework that would return test scores on a comparable level with the results normally gathered manually by clinical experts. For this, three classification models were created and tested on a dataset of 23 test sheets collected from the neuropsychological department at the Bürgerspital in Solothurn. Using grayscale values, test sheets were split into single symbols; each symbol was divided into nine areas, which were used as detection boxes for the recognition of a mark by the patient. The classification models used the mean grayscale value as criteria for mark recognition. The model with a stricter threshold for non-target symbols achieved results that were not inferior to the manual evaluations done by clinicians. The statistical analysis of this model supported the hypothesis that the model performance did not differ significantly from the performance of clinicians. In the following section the previously described results are discussed, and limitations of the thesis are considered.

6.1 Interpretation of results

The Test des Deux Barrages has been a staple in neuropsychological assessments in Switzerland ever since the MNND was introduced in 2011. While it is a helpful test to examine selective attention over time, the current evaluation by hand using templates is time-consuming (10 to 15 minutes per test sheet) and susceptible to errors either when checking for mistakes by the patient on the test sheet or when transferring the exact positions of those mistakes onto the excel evaluation sheet. The clinician’s performance when compared to the ground truth of the patients resulted in near-perfect scores in the performance metrics. This was predictable, given that there is no time limit in the evaluation and the templates were designed to help in the evaluation process. Still, it does not justify the need for skilled and highly educated clinicians to carry out an evaluation that does not require any prior knowledge in the field. Instead, an automatic scoring system could help reduce the workload of clinicians and/or allow them to invest the saved time in more patient-centred care.

The goal of this thesis was to develop a classification model that could produce similar results to clinicians. Performance of both clinicians and models was assessed using the balanced accuracy, F1 measure, sensitivity, and specificity metrics. The output of the developed classification models and the clinicians' evaluation were examined for their performance metrics when compared to the ground truth of the patients. The performance metrics of each model were then contrasted to the performance metrics of the clinicians. The differences in the performance metrics between model and clinicians were tested for significance using the sign test and the Wilcoxon signed-rank test. All three models showed encouraging performances across all test sheets. The performance metrics assessed in this thesis were especially promising for the model with a strict threshold for non-target symbols. This model was developed using the knowledge from the ground truth, that the likelihood of a patient marking a non-target symbol was much lower than a patient missing a target symbol. Both sign test and Wilcoxon signed-rank test supported the hypothesis of a lack of statistical evidence to claim that the model's performance metrics were different from the performance metrics of the clinicians. Conversely, it was more predictable that the performance metrics of the other two models would not emerge as favourable when tested with the sign test and the Wilcoxon signed-rank test. On the one hand, the model with inside or outside criteria was developed first and checked for its performance. The other models were designed to address the shortcomings of this first model. Aside from the sensitivity scores, all other performance metrics performed worse than the clinicians according to the statistical analysis. On the other hand, the model with inside and outside criteria was created in an attempt to reduce the amount of falsely recognized unmarked symbols. The trade-off of a lower sensitivity for a better specificity was predicted, but nevertheless checked. The better specificity of this model was also reflected in the non-significant results of the sign test and the Wilcoxon signed-rank test when compared to clinicians.

Mean grayscale values were used as criteria to determine whether a symbol was marked by the patient or not. All three models used a grayscale value of 250 as threshold for target symbols and delivered a respectable classification output. This was further backed by the ROC analysis at different mean grayscale value thresholds. An area under the curve of 0.97 was achieved using the mean grayscale values of the inner detection boxes as decision variable, which would suggest that the mean grayscale value was indeed highly discriminative in the classification process, meaning it can effectively distinguish between marked and unmarked symbols (Hosmer et al., 2013; Shallcross & Ahner, 2020). Additionally, the PR analysis also indicated that the quality of the classification was given using mean grayscale values even when considering the imbalance in number of marked and unmarked symbols into the equation. An average precision or area under the PR curve of 0.92 demonstrates that the mean grayscale value has high accuracy in classifying marked symbols, while minimizing falsely classified unmarked symbols (P. Flach & Kull, 2015). Lastly, performance of the models would have been even better had the clinicians not made any signs (Figure 13) on the test sheet, which resulted in a higher number of unmarked symbols being misrecognized as marked. Ideally, clinicians do not have to make signs anymore when an automatic scoring system is implemented in the clinical practice.

The automatic scoring system needs 13.5 seconds on average to evaluate a test sheet. Qualitatively, the scoring system is distinctly faster than the self-reported evaluation times (10 to 15 minutes per test sheet). However, the execution time does not account for the time needed before the procedure (scanning, importing, converting and rotating of a test sheet) and after the classification (visualization, and transfer onto Excel evaluation aid). These promising results can therefore only be viewed descriptively and need to be investigated empirically.

To conclude, an automatic scoring system based on mean grayscale values as decision criteria has returned evaluation scores comparable to clinicians. The present findings support that an automatic scoring system can be developed to reduce the workload and time-consumption of clinicians. More generally, the ideas of the thesis can provide a framework that may be applied to similar tests in neuropsychology.

6.2 Limitations

Although the present results support the hypothesis, certain limitations of this thesis need to be addressed. One limitation is that both the sign test and the Wilcoxon signed-rank test were used as nonparametric statistical tests to compare each model's performance to the performance of clinicians. 23 test sheets were collected from October to the end of January to develop and assess the performance of an automatic scoring system. The model comparison to clinicians returned encouraging results in line with the hypothesis, but at a sample size as small as 23, the results need to be judged with caution. When the sample size is small, the power of the Wilcoxon signed-rank test is reduced and may fail to reject the null hypothesis when it is false (Gibbons & Chakraborti, 2020). In the context of the present thesis, this would mean a significant difference between the model and clinicians is not detected. Additionally, it is in the nature of both sign test and Wilcoxon signed-rank test to exclude pairs of observations when their difference is zero, potentially aggravating the sample size issue and loss of power (Pratt, 1959). The performance metrics with zero differences between model and clinicians were not included in the analysis and are effectively ignored, which perturbed the calculation of the p-value. This was one reason why the sign test returned p-values of 1.0 in some cases. Finally, the sign test and Wilcoxon signed-rank test only allow statements about the differences between two models. If the question is whether the models performed similarly or equally good compared to the clinicians, different types of statistical tests are required. The equivalence test would have been an alternative to the nonparametric tests used in this thesis; it could have provided information about the similarity between the models and the clinicians. According to Lakens et al. (2018) the null hypothesis in classic hypothesis tests assumes that an effect (or, in this case, difference) is equal to or around zero. However, in an equivalence test, the null hypothesis is reversed and assumes that the difference between two groups is outside of a predetermined range of practical equivalence, while the alternative hypothesis assumes that the difference falls within that range. Equivalence bounds need to be specified prior to conducting this test. These bounds should represent a range of values (or amount of errors) that would be deemed practically equivalent and within which the differences between models and clinicians can be considered negligible.

Another limitation is the use of specificity as performance metric in the presence of an imbalanced dataset. Specificity was not as meaningful as a metric because the number of unmarked symbols heavily outweigh the number of marked symbols due to the characteristics of the Test des Deux Barrages. In the performance analysis of the scoring system, the main goal was to discern whether the model would reliably detect marks made by patients. Therefore, the specificity was not the focus and high specificity scores were achieved from all three models. It is nonetheless an important measure when the automatic scoring system is applied for evaluation in the clinical field. From the ground truth it was derived that a patient is more likely to miss a target symbol than to mark a non-target symbol. Therefore, it is crucial that the models also detect unmarked symbols correctly.

Moreover, there are limitations of using grayscale values. The results from the ROC and PR analysis indeed support the use of mean grayscale values as decision criteria. Nevertheless, one distinct disadvantage of using grayscale values is that the whole procedure from splitting the test sheet into symbols to mark recognition is heavily reliant on good scan quality. Clinicians were asked to scan the test sheets as cleanly as possible using the scanner available in the hospital. The automatic scoring system was developed using only these test sheets, but never tested in a different setting. In fact, one of the 23 test sheets collected was minimally rotated. In this case, the first steps of the procedure were only possible after correcting the rotation. The uncorrected test sheet would distort the calculation of the row and column peaks used to split the test sheet into its single symbols, resulting in splits being drawn on the test sheet that do not run between the symbols but through the symbols (Figure 17). Similarly, the mark recognition with mean grayscale values was sensitive to small irregularities or creases in the paper (Figure 15), which were falsely picked up as marks by the patient. These errors are proof of the inflexibility of the developed automatic scoring system. In the future, the scoring system does need to be revised for such transformations delivering robust results even when the test sheets are not of the same resolution or quality.

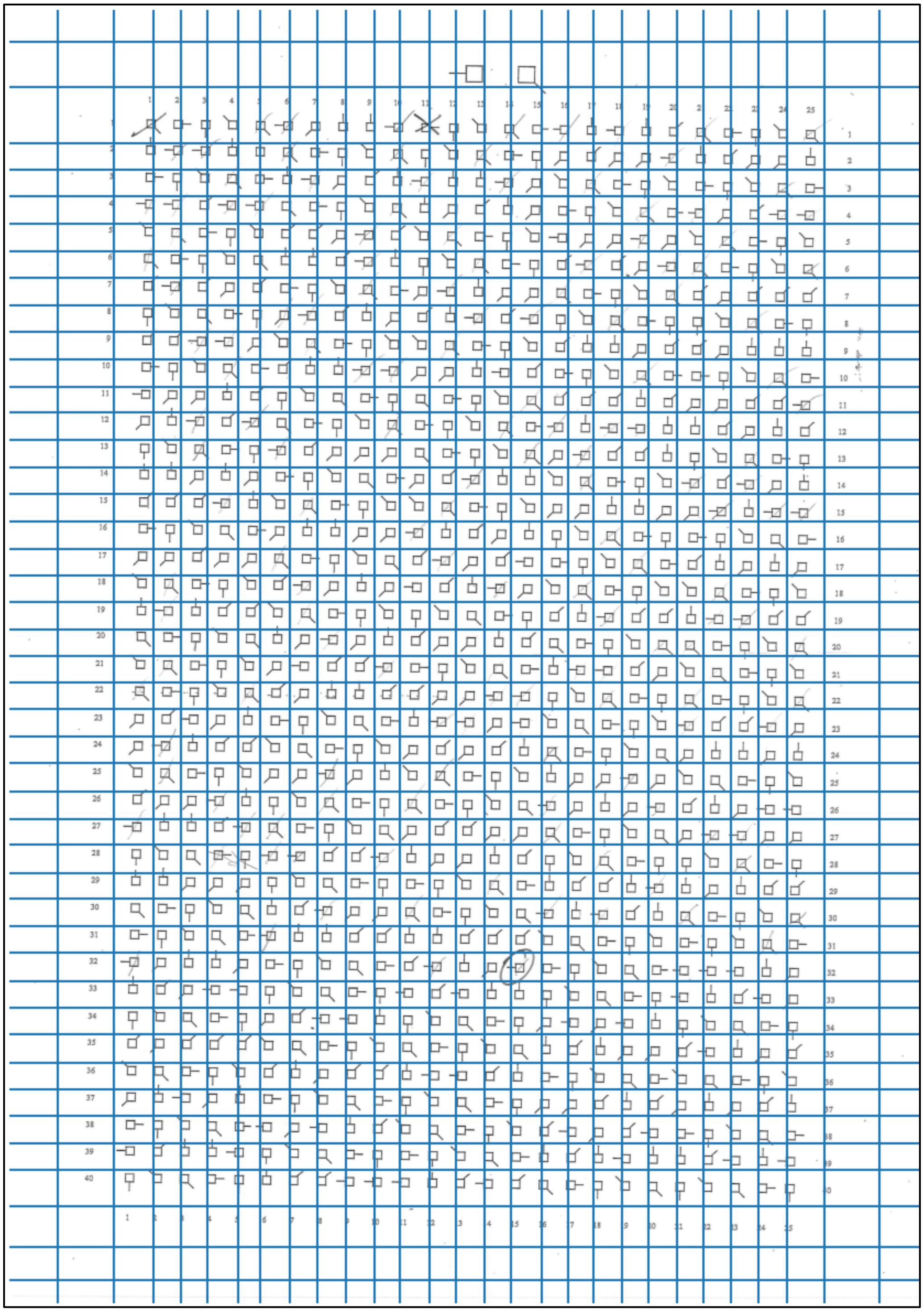
Likewise, a clear limitation of using grayscale values is the inability of the scoring system to detect marks that are not visible enough. Using the mean grayscale value as decision criteria is highly dependent on the mark being dark enough to lower the mean value below the threshold set at 250. Of the 80 missed marked symbols by the model with a strict threshold for non-target symbols, 76 errors occurred due to this limitation. One test sheet in particular contained many misses, because the patient left marks that were too faint for the scoring system to detect (Figure 18). The scoring system needs to be improved for cases, where the visibility of marks is insufficient. For future research, it would also be interesting to know if the faint marks across the test sheet happened by chance or if there was an underlying disease causing the issue. This also emphasizes the relevance of auxiliary information, such as pen pressure or writing speed, that can be captured by new technological measures to aid in the assessment of neurological diseases (Lee et al., 2020).

Furthermore, the models used detection boxes for mark recognition. These boxes were created by finding the dark ink of the symbol borders using local minima and either shortening (inner detection box) or lengthening (outer detection boxes) the distances between valleys. The goal was to avoid any disturbance in the calculation of the mean grayscale value by the dark ink of the symbol border. However, the distances were adjusted by an arbitrarily set distance and did not match the borders of the symbol perfectly. As a result, a corridor or 'blind spot' between the two types of detection boxes was accepted that did not contribute to the calculation of the mean grayscale values. Valuable information was also overlooked by excluding the outer detection boxes containing the symbol variation tails before mark recognition, when marks ran through the respective detection boxes. This led to some marks being missed by the scoring system (Figure 12B). The scoring system does also not account for marked symbols that have been corrected by the patients or for marks that overlap from neighbouring symbols that are included in the mark recognition of another symbol (Figure 14). Finally, model-specific errors were conceded (Figure 12C) for better overall performance but should be observed very closely. The risk of missing a marked non-target symbol should not be underestimated. Few mistakes of this type are usually made by patients. Hence, making many is more heavily weighted or penalized in the assessments.

Lastly, the automatic scoring system took an average of 13.5 seconds per test sheet to execute the procedure and classification steps explained in the methods section. This would suggest an improvement in comparison with the 10 to 15 minutes per test sheet reported in the survey sent to neuropsychologists. However, the execution time does not encompass the time needed before the procedure steps and after the classification. Each test sheet needs to be scanned, rotated, and converted into a grayscale image first before the procedure can be applied. After classification, the exact positions of the mistakes by the patient would still need to be transferred onto the Excel evaluation aid. This represents additional time not taken into account by the execution time alone. Clinical utility was not a primary focus of this thesis. The automatic scoring system indeed provides results in a fraction of the time needed for manual evaluation. Nevertheless, the results are not presented conveniently (see Figure A in Appendix for an example of visualization), nor does it circumvent the transfer process onto the Excel evaluation aid. As such, the automatic scoring system needs to be developed further before any questions about the applicability can be answered.

Figure 17

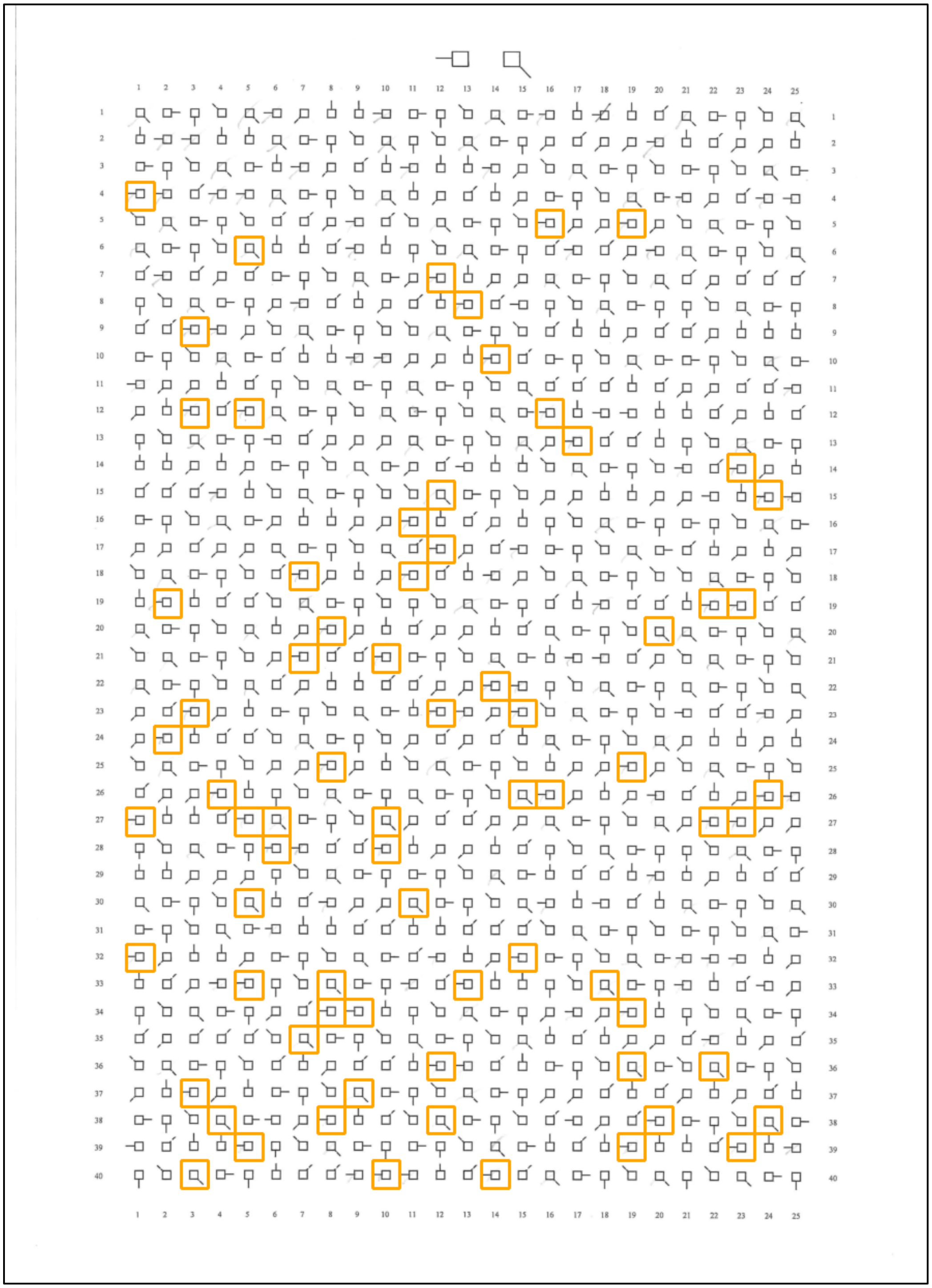
Test Sheet Before Rotation Correction



Note. Row and column valleys depicted on a slightly skewed test sheet before rotation correction.

Figure 18

Missed Marked Symbols on Test Sheet with Low Mark Visibility



Note. Marks by the patient with low visibility resulting in high rate of unrecognized marked symbols by the scoring system (yellow boxes).

6.3 Implications

The automatic scoring system developed in this thesis should be viewed as an example of how the implementation of technology could alleviate the workload of clinicians. Despite its limitations the scoring system demonstrated promising results comparable to the evaluation of clinical experts. The results were generated by using a basic arithmetic measure such as the mean grayscale value, which already proved effective as decision criteria. This section discusses the implications the development of an automatic scoring system carries and highlights its potential contributions to the field of neuropsychology.

Automating the evaluation process eliminates the need for manual scoring, which can be time-consuming and labour-intensive. The reduced time required for scoring can have broader implications for clinical practice. Saved time and resources can be allocated to other critical tasks instead, such as patient care or conducting additional assessments (Langer et al., 2022). This increase in efficiency may also lead to cost savings in terms of personnel and operational expenses amidst increasing pressures to reduce health care spending (Bilder & Reise, 2019; Parsey & Schmitter-Edgecombe, 2013). Moreover, by automating the scoring process, a higher degree of standardization and consistency in the scoring results is ensured. Human scoring can be subjective and prone to inter-rater variability, which can impact the reliability and validity of the test outcomes. The use of an automatic scoring system helps mitigate these concerns by providing consistent and reliable scoring across different administrations enhancing the comparability of results (Bilder & Reise, 2019; Miller & Barr, 2017). Once fully developed and validated, an automatic scoring system can be easily implemented und adopted for widespread use among clinicians and researchers. For the Test des Deux Barrages this would present an opportunity to gain more relevance again and to validate its psychometric properties, which are not disclosed in detail to the best of my knowledge.

In summary, the development of an automatic scoring system demonstrates the potential of an innovative approach for the clinical field of neuropsychology and opens doors for further advancements in this direction. The application of computational techniques can streamline and improve psychological assessment processes.

6.4 Directions for future research

In terms of future research, it would be useful to extend and secure the current findings by examining the performance of the scoring system with a larger and more diverse dataset. Ideally, an equivalence test is performed to compare the model performance with clinicians, with a specified error bound that would represent a practically important difference in the evaluation of the test score when compared to normative data. The scoring system clearly needs to be improved for its flexibility and robustness, to return similar results when scans are not as clean and homogenous as they were in the data set used in this thesis. Visibility issues and variabilities of the test sheet like rotation could be mitigated by image pre-processing steps like increasing the contrast or image co-registration in the future. Instead of relying entirely on grayscale value, different computational techniques could be used for the procedure and classification steps. The test sheet could be split using optical mark recognition or edge-detection techniques. The classification of symbols could be performed by using machine-learning approaches such as the ones used by Langer et al. (2022) in their work. The average execution time also has potential to be much improved. Once a robust and reliable scoring system has been developed, it must also showcase clinical utility. Preferably, the automatic scoring system is implemented as an application, employable on pictures instead of scans, and with the functions of the Excel evaluation aid already embedded to circumvent both evaluation and transfer process. Thereafter, the time- and cost-effectiveness of the automatic scoring system must be examined.

6.5 Conclusion

Based on the results of the current work it can be deduced that an automatic scoring system can be developed for the Test des Deux Barrage. More effort needs to be put into correcting the present limitations, before investigating the clinical utility of the scoring system, which was not a primary focus in this thesis. Paper-pencil tests might currently be the traditional way of assessing patients in neuropsychology. With more digital natives expected among the patient groups in the future the development and administration of digital tests will need much more investment. This represents opportunities to improve upon current clinical practice. Neuropsychology must allow for more interdisciplinary collaboration to not only relieve the workload of clinicians but also enhance the assessment of neurological diseases.

7 References

Aiello, E. N., & Depaoli, E. G. (2022). Norms and standardizations in neuropsychology via equivalent scores: Software solutions and practical guides. *Neurological Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, *43*(2), 961–966. https://doi.org/10.1007/s10072-021-05374-0

Amato, M. P., Zipoli, V., & Portaccio, E. (2008). Cognitive changes in multiple sclerosis. *Expert Review of Neurotherapeutics*, *8*(10), 1585–1596. https://doi.org/10.1586/14737175.8.10.1585

Anagnostou, E., Mankad, D., Diehl, J., Lord, C., Butler, S., McDuffie, A., Shull, L., Ashbaugh, K., Koegel, R. L., Volkmar, F. R., Naples, A., Doggett, R., Hooper, S. R., Casanova, M., Hoffman, E. J., McFadden, K., Anderson, G. M., Gupta, A. R., DiLullo, N. M., . . . Pilato, M. (2013). Normative Data. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 2062–2063). Springer New York. https://doi.org/10.1007/978-1-4419-1698-3_315

Balzer, C., Berger, J. M., Caprez, G., Gonser, A., Gutbrod, K., & Keller, M. (2011a). *Materialien und Normwerte fuer die neuropsychologische Diagnostik (MNND).*

Balzer, C., Berger, J. M., Caprez, G., Gonser, A., Gutbrod, K., & Keller, M. (2011b). Methodik. In *Materialien und Normwerte fuer die neuropsychologische Diagnostik (MNND)* (pp. 31–32).

Balzer, C., Berger, J. M., Caprez, G., Gonser, A., Gutbrod, K., & Keller, M. (2011c). Normwertabellen. In *Materialien und Normwerte fuer die neuropsychologische Diagnostik (MNND)* (pp. 41–47).

Balzer, C., Berger, J. M., Caprez, G., Gonser, A., Gutbrod, K., & Keller, M. (2011d). Testhandbuch. In *Materialien und Normwerte fuer die neuropsychologische Diagnostik (MNND)* (pp. 53–57).

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society : JINS*, *17*(1), 7–13. https://doi.org/10.1017/S1355617710001396

Bilder, R. M., & Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here? *The Clinical Neuropsychologist*, *33*(2), 220–245. https://doi.org/10.1080/13854046.2018.1521993

Bodis-Wollner, I. (2003). Neuropsychological and perceptual defects in Parkinson's disease. *Parkinsonism & Related Disorders*, *9 Suppl 2*, S83-9. https://doi.org/10.1016/S1353-8020(03)00022-1

Budson, A. E. (2022). *Memory Loss, Alzheimer's Disease, and Dementia - E-Book: A Practical Guide for Clinicians* (3rd ed.). Elsevier. https://livivo.idm.oclc.org/login?url=https://ebookcentral.proquest.com/lib/zbmed-ebooks/detail.action?docID=6552155

Capitani, E. (1997). Normative Data and Neuropsychological Assessment. Common Problems in Clinical Practice and Research. *Neuropsychological Rehabilitation*, *7*(4), 295–310. https://doi.org/10.1080/713755543

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society : JINS*, *23*(9-10), 778–790. https://doi.org/10.1017/S1355617717001060

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach* (2. ed.). Wiley.

Coulthard, E., Singh-Curry, V., & Husain, M. (2006). Treatment of attention deficits in neurological disorders. *Current Opinion in Neurology, 19*(6), 613–618. https://doi.org/10.1097/01.wco.0000247605.57567.9a

Crawford, J. R., Parker, D. M., & McKinlay, W. W. (2018). *A Handbook of Neuropsychological Assessment*. Routledge. https://doi.org/10.4324/9780429490316

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on Machine learning - ICML '06* (pp. 233–240). ACM Press. https://doi.org/10.1145/1143844.1143874

Deery, H. A., Di Paolo, R., Moran, C., Egan, G. F., & Jamadar, S. D. (2023). The older adult brain is less modular, more integrated, and less efficient at rest: A systematic review of large-scale resting-state functional brain networks in aging. *Psychophysiology, 60*(1), e14159. https://doi.org/10.1111/psyp.14159

Fisher, A. V. (2019). Selective sustained attention: A developmental foundation for cognition. *Current Opinion in Psychology, 29*, 248–253. https://doi.org/10.1016/j.copsyc.2019.06.002

Flach, P., & Kull, M. (2015). Precision-Recall-Gain Curves: PR Analysis Done Right. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf

Flach, P. A. (2016). ROC Analysis. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1–8). Springer US. https://doi.org/10.1007/978-1-4899-7502-7_739-1

Gibbons, J. D., & Chakraborti, S. (2020). *Nonparametric Statistical Inference*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315110479

Ginty, A. T. (2013). Psychometric Properties. In M. D. Gellman & J. R. Turner (Eds.), *Encyclopedia of Behavioral Medicine* (pp. 1563–1564). Springer New York. https://doi.org/10.1007/978-1-4419-1005-9_480

Goldstein, E. B. (2011). *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience*. Wadsworth Cengage Learning. https://books.google.ch/books?id=Ml4XygEACAAJ

He, W., Goodkind, D., Kowal, P. R., & United States. Bureau of the Census. (2016). *An Aging World : 2015*. *International population reports*. United States Census Bureau. https://books.google.ch/books?id=WvZxuAEACAAJ

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Third edition). *Wiley series in probability and statistics*. Wiley.

Jäncke, L. (2012). Testrezension. *Zeitschrift Für Neuropsychologie*, *23*(1), 55–58. https://doi.org/10.1024/1016-264X/a000062

Johnston, W. A., & Dark, V. J. (1986). Selective Attention. *Annu. Rev. Psychol*, *37*, 43–75.

Kelleher, J. D., MacNamee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. The MIT Press.

Krupinski, E. A. (2017). Receiver Operating Characteristic (ROC) Analysis. *Frontline Learning Research*, *5*(3), 31–42. https://doi.org/10.14786/flr.v5i2.250

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Langer, N., Weber, M., Hebling Vieira, B., Strzelczyk, D., Wolf, L., Pedroni, A., Heitz, J., Müller, S., Schultheiss, C., Tröndle, M., Arango Lasprilla, J. C., Rivera, D., Scarpina, F., Zhao, Q., Leuthold, R., Wehrle, F., Jenni, O. G., Brugger, P., Zaehle, T., . . . Zhang, C. (2022). *The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning*. https://doi.org/10.1101/2022.06.15.496291

Lee, H. S., Youn, J., Cho, J. W., Ahn, J. H., Yoon, J. H., & Na, D. L. (2020). Characteristics of Writing in Parkinson’s Disease: Focused on Pen Pressure, Letter Size, and Writing Speed. *Communication Sciences & Disorders*, *25*(1), 63–74. https://doi.org/10.12963/csd.20691

Libon, D. J., Baliga, G., Swenson, R., & Au, R. (2021). Digital Neuropsychological Assessment: New Technology for Measuring Subtle Neuropsychological Behavior. *Journal of Alzheimer's Disease : JAD, 82*(1), 1–4. https://doi.org/10.3233/JAD-210513

Loewenstein, D. A., Argüelles, T., Argüelles, S., & Linn-Fuentes, P. (1994). Potential cultural bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology*, *16*(4), 623–629. https://doi.org/10.1080/01688639408402673

Maercker, A., Perkonigg, A., Preisig, M., Schaller, K., & Weller, M. (2013). The costs of disorders of the brain in Switzerland: An update from the European Brain Council Study for 2010. *Swiss Medical Weekly*, *143*, w13751. https://doi.org/10.4414/smw.2013.13751

Miller, J. B., & Barr, W. B. (2017). The Technology Crisis in Neuropsychology. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist*, *27*(8), 1328–1361. https://doi.org/10.1080/13854046.2013.834971

Plebanek, D. J., & Sloutsky, V. M. (2019). Selective attention, filtering, and the development of working memory. *Developmental Science*, *22*(1), e12727. https://doi.org/10.1111/desc.12727

Pratt, J. W. (1959). Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *Journal of the American Statistical Association*, *54*(287), 655. https://doi.org/10.2307/2282543

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-Signed-Rank Test. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1658–1659). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_616

Reynolds, C. R. (2000). Methods for Detecting and Evaluating Cultural Bias in Neuropsychological Tests. In A. E. Puente, C. R. Reynolds, E. Fletcher-Janzen, & T. L. Strickland (Eds.), *Critical Issues in Neuropsychology. Handbook of Cross-Cultural Neuropsychology* (pp. 249–285). Springer US. https://doi.org/10.1007/978-1-4615-4219-3_15

Scheff, S. W. (2016). Nonparametric Statistics. In *Fundamental Statistical Principles for the Neurobiologist* (pp. 157–182). Elsevier. https://doi.org/10.1016/B978-0-12-804753-8.00008-7

Schellig, D., Drechsler, R., Heinemann, D., & Sturm, W. (2009). *Aufmerksamkeit, Gedächtnis, exekutive Funktionen*. *Handbuch neuropsychologischer Testverfahren / hrsg. von Dieter Schellig, Renate Drechsler, Dörthe Heinemann und Walter Sturm: Vol. 1*.

Shallcross, N. J., & Ahner, D. K. (2020). Predictive models of world conflict: accounting for regional and conflict-state differences. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *17*(3), 243–267. https://doi.org/10.1177/1548512919847532

Souza, A. C. de, Alexandre, N. M. C., & Guirardello, E. d. B. (2017). Propriedades psicométricas na avaliação de instrumentos: Avaliação da confiabilidade e da validade [Psychometric properties in instruments evaluation of reliability and validity]. *Epidemiologia E Servicos De Saude : Revista Do Sistema Unico De Saude Do Brasil*, *26*(3), 649–659. https://doi.org/10.5123/S1679-49742017000300022

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, *2 Suppl 1*(Suppl 1), S30-48. https://doi.org/10.1016/j.dcn.2011.11.001

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. *Scientific Psychology Series*. Psychology Press. https://permalink.obvsg.at/

Tao, L., Wang, Q., Liu, D., Wang, J., Zhu, Z., & Feng, L. (2020). Eye tracking metrics to screen and assess cognitive impairment in patients with neurological disorders. *Neurological*

Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, *41*(7), 1697–1704. https://doi.org/10.1007/s10072-020-04310-y

van Moorselaar, D., & Slagter, H. A. (2020). Inhibition in selective attention. *Annals of the New York Academy of Sciences*, *1464*(1), 204–221. https://doi.org/10.1111/nyas.14304

World Health Organization. (2015). *World report on ageing and health*. World Health Organization. http://apps.who.int/iris/bitstream/10665/186463/1/9789240694811_eng.pdf?ua=1

Wühr, B., & Wühr, P. (2022). Effects of repeated testing in a pen-and-paper test of selective attention (FAIR-2). *Psychological Research*, *86*(1), 294–311. https://doi.org/10.1007/s00426-021-01481-x

Zazzo, R. (1964). *Le Test des barrages: Mira Stambak. Une épreuve de pointillage*. *Manuel pour l’examen psychologique de l’enfant*. Neuchatel, Delachaux et Niestlé. https://books.google.ch/books?id=dMo5HQAACAAJ

Appendix

Table A

Performance Analysis of Individual Patients

Test sheet	TP	TN	FP	FN	b-ACC	F1	SEN	SPE
T2B_01.12.2022	208	759	0	33	0.9315	0.9265	0.8631	1
T2B_03.11.2022	157	599	1	30	0.9190	0.9101	0.8396	0.9983
T2B_05.12.2022.1	224	759	0	17	0.9647	0.9634	0.9295	1
T2B_05.12.2022.2	175	620	0	23	0.9419	0.9383	0.8838	1
T2B_10.10.2022.1	212	759	0	29	0.9398	0.9360	0.8797	1
T2B_10.10.2022.2	141	517	0	24	0.9273	0.9216	0.8545	1
T2B_10.11.2022.1	203	749	0	32	0.9319	0.9269	0.8638	1
T2B_10.11.2022.2	215	757	2	26	0.9447	0.9389	0.8921	0.9974
T2B_16.11.2022	221	759	0	20	0.9585	0.9567	0.9170	1
T2B_16.12.2022	237	759	0	4	0.9917	0.9916	0.9834	1
T2B_17.10.2022	164	584	0	19	0.9481	0.9452	0.8962	1
T2B_17.11.2022	238	759	0	3	0.9938	0.9937	0.9876	1
T2B_18.01.2023	219	712	0	3	0.9932	0.9932	0.9865	1
T2B_18.10.2022	157	525	0	9	0.9729	0.9721	0.9458	1
T2B_20.10.2022	186	649	0	21	0.9493	0.9466	0.8986	1
T2B_21.10.2022	211	737	0	19	0.9587	0.9569	0.9174	1
T2B_22.11.2022	205	759	0	36	0.9253	0.9193	0.8506	1
T2B_22.12.2022	184	652	3	26	0.9358	0.9270	0.8762	0.9954
T2B_26.10.2022.1	205	649	0	3	0.9928	0.9927	0.9856	1
T2B_26.10.2022.2	234	759	0	7	0.9855	0.9853	0.9710	1
T2B_29.11.2022	196	759	0	45	0.9066	0.8970	0.8133	1
T2B_30.11.2022	230	759	0	11	0.9772	0.9766	0.9544	1
T2B_31.01.2023	147	601	1	41	0.8901	0.8750	0.7819	0.9983
Total	4569	15941	7	481				

Table B

Performance Analysis of Clinicians for Each Individual Test Sheet

Test sheet	TP	TN	FP	FN	b-ACC	F1	SEN	SPE
T2B_01.12.2022	208	791	1	0	0.9994	0.9976	1	0.9987
T2B_03.11.2022	158	841	1	0	0.9994	0.9968	1	0.9988
T2B_05.12.2022.1	224	776	0	0	1	1	1	1
T2B_05.12.2022.2	173	824	1	2	0.9937	0.9914	0.9886	0.9988
T2B_10.10.2022.1	212	784	4	0	0.9975	0.9907	1	0.9949
T2B_10.10.2022.2	141	858	1	0	0.9994	0.9965	1	0.9988
T2B_10.11.2022.1	203	797	0	0	1	1	1	1
T2B_10.11.2022.2	216	783	0	1	0.9977	0.9977	0.9954	1
T2B_16.11.2022	221	779	0	0	1	1	1	1
T2B_16.12.2022	237	763	0	0	1	1	1	1
T2B_17.10.2022	164	836	0	0	1	1	1	1
T2B_17.11.2022	238	761	1	0	0.9993	0.9979	1	0.9987
T2B_18.01.2023	219	781	0	0	1	1	1	1
T2B_18.10.2022	157	842	1	0	0.9994	0.9968	1	0.9988
T2B_20.10.2022	179	813	1	7	0.9806	0.9781	0.9624	0.9988
T2B_21.10.2022	211	789	0	0	1	1	1	1
T2B_22.11.2022	205	794	1	0	0.9994	0.9976	1	0.9987
T2B_22.12.2022	187	813	0	0	1	1	1	1
T2B_26.10.2022.1	205	795	0	0	1	1	1	1
T2B_26.10.2022.2	234	765	1	0	0.9993	0.9979	1	0.9987
T2B_29.11.2022	195	798	6	1	0.9937	0.9824	0.9949	0.9925
T2B_30.11.2022	230	769	1	0	0.9994	0.9978	1	0.9987
T2B_31.01.2023	147	849	3	1	0.9949	0.9866	0.9932	0.9965
Total	4564	18401	23	12				

Table C

Performance Analysis of Model with Inside or Outside Criteria for Each Test Sheet

Test sheet	TP	TN	FP	FN	b-ACC	F1	SEN	SPE
T2B_01.12.2022	208	789	3	0	0.9981	0.9928	1	0.9962
T2B_03.11.2022	157	840	2	1	0.9956	0.9905	0.9937	0.9976
T2B_05.12.2022.1	224	774	2	0	0.9987	0.9956	1	0.9974
T2B_05.12.2022.2	175	822	3	0	0.9982	0.9915	1	0.9964
T2B_10.10.2022.1	87	787	1	125	0.7046	0.5800	0.4104	0.9987
T2B_10.10.2022.2	141	857	2	0	0.9988	0.9930	1	0.9977
T2B_10.11.2022.1	181	790	7	22	0.9414	0.9258	0.8916	0.9912
T2B_10.11.2022.2	217	780	3	0	0.9981	0.9931	1	0.9962
T2B_16.11.2022	220	774	5	1	0.9945	0.9865	0.9955	0.9936
T2B_16.12.2022	233	760	3	4	0.9896	0.9852	0.9831	0.9961
T2B_17.10.2022	164	832	4	0	0.9976	0.9880	1	0.9952
T2B_17.11.2022	238	761	1	0	0.9993	0.9979	1	0.9987
T2B_18.01.2023	217	776	5	2	0.9922	0.9841	0.9909	0.9936
T2B_18.10.2022	155	840	3	2	0.9919	0.9841	0.9873	0.9964
T2B_20.10.2022	185	811	3	1	0.9955	0.9893	0.9946	0.9963
T2B_21.10.2022	211	785	4	0	0.9975	0.9906	1	0.9949
T2B_22.11.2022	174	794	1	31	0.9238	0.9158	0.8488	0.9987
T2B_22.12.2022	187	808	5	0	0.9969	0.9868	1	0.9938
T2B_26.10.2022.1	205	794	1	0	0.9994	0.9976	1	0.9987
T2B_26.10.2022.2	233	765	1	1	0.9972	0.9957	0.9957	0.9987
T2B_29.11.2022	196	802	2	0	0.9988	0.9949	1	0.9975
T2B_30.11.2022	229	769	1	1	0.9972	0.9957	0.9957	0.9987
T2B_31.01.2023	146	836	16	2	0.9839	0.9419	0.9865	0.9812
Total	4383	18346	78	193				

Table D

Performance Analysis of Model with Inside and Outside Criteria for Each Test Sheet

Test sheet	TP	TN	FP	FN	b-ACC	F1	SEN	SPE
T2B_01.12.2022	203	792	0	5	0.9880	0.9878	0.9760	1
T2B_03.11.2022	158	842	0	0	1	1	1	1
T2B_05.12.2022.1	224	775	1	0	0.9994	0.9978	1	0.9987
T2B_05.12.2022.2	168	824	1	7	0.9794	0.9767	0.9600	0.9988
T2B_10.10.2022.1	67	788	0	145	0.6580	0.4803	0.3160	1
T2B_10.10.2022.2	133	858	1	8	0.9710	0.9673	0.9433	0.9988
T2B_10.11.2022.1	131	792	5	72	0.8195	0.7729	0.6453	0.9937
T2B_10.11.2022.2	215	781	2	2	0.9941	0.9908	0.9908	0.9974
T2B_16.11.2022	145	779	0	76	0.8281	0.7923	0.6561	1
T2B_16.12.2022	190	762	1	47	0.9002	0.8879	0.8017	0.9987
T2B_17.10.2022	161	835	1	3	0.9903	0.9877	0.9817	0.9988
T2B_17.11.2022	208	762	0	30	0.9370	0.9327	0.8739	1
T2B_18.01.2023	161	779	2	58	0.8663	0.8429	0.7352	0.9974
T2B_18.10.2022	153	843	0	4	0.9873	0.9871	0.9745	1
T2B_20.10.2022	175	813	1	11	0.9698	0.9669	0.9409	0.9988
T2B_21.10.2022	210	787	2	1	0.9964	0.9929	0.9953	0.9975
T2B_22.11.2022	118	795	0	87	0.7878	0.7307	0.5756	1
T2B_22.12.2022	187	810	3	0	0.9982	0.9920	1	0.9963
T2B_26.10.2022.1	205	795	0	0	1	1	1	1
T2B_26.10.2022.2	232	766	0	2	0.9957	0.9957	0.9915	1
T2B_29.11.2022	196	804	0	0	1	1	1	1
T2B_30.11.2022	220	770	0	10	0.9783	0.9778	0.9565	1
T2B_31.01.2023	132	844	8	16	0.9413	0.9167	0.8919	0.9906
Total	3992	18396	28	584				

Table E

Performance Analysis of Model with Strict Threshold for Non-Target Symbols for Each Test Sheet

Test sheet	TP	TN	FP	FN	b-ACC	F1	SEN	SPE
T2B_01.12.2022	208	791	1	0	0.9994	0.9976	1	0.9987
T2B_03.11.2022	158	842	0	0	1	1	1	1
T2B_05.12.2022.1	224	776	0	0	1	1	1	1
T2B_05.12.2022.2	175	825	0	0	1	1	1	1
T2B_10.10.2022.1	139	788	0	73	0.8278	0.7920	0.6557	1
T2B_10.10.2022.2	141	858	1	0	0.9994	0.9965	1	0.9988
T2B_10.11.2022.1	198	793	4	5	0.9852	0.9778	0.9754	0.9950
T2B_10.11.2022.2	217	781	2	0	0.9987	0.9954	1	0.9974
T2B_16.11.2022	221	779	0	0	1	1	1	1
T2B_16.12.2022	237	763	0	0	1	1	1	1
T2B_17.10.2022	164	835	1	0	0.9994	0.9970	1	0.9988
T2B_17.11.2022	238	762	0	0	1	1	1	1
T2B_18.01.2023	218	780	1	1	0.9971	0.9954	0.9954	0.9987
T2B_18.10.2022	157	842	1	0	0.9994	0.9968	1	0.9988
T2B_20.10.2022	186	812	2	0	0.9988	0.9947	1	0.9975
T2B_21.10.2022	211	788	1	0	0.9994	0.9976	1	0.9987
T2B_22.11.2022	205	795	0	0	1	1	1	1
T2B_22.12.2022	187	811	2	0	0.9988	0.9947	1	0.9975
T2B_26.10.2022.1	205	795	0	0	1	1	1	1
T2B_26.10.2022.2	234	766	0	0	1	1	1	1
T2B_29.11.2022	196	804	0	0	1	1	1	1
T2B_30.11.2022	230	770	0	0	1	1	1	1
T2B_31.01.2023	147	846	6	1	0.9931	0.9767	0.9932	0.9930
Total	4496	18402	22	80				

Survey

A survey was sent to neuropsychologists in Switzerland addressing three questions:

Q1: How long have you been using the Test des Deux Barrages in diagnostics?

Q2: How many Test des Deux Barrages do you conduct per year?

Q3: On average, how many minutes do you need for the complete evaluation of a Test des Deux Barrages (including transfer onto the Excel evaluation aid)?

16 responses were recorded, which are summarised in Table F. Q1 and Q2 were open-response questions, whereas for Q3 five response categories were given (1 = 0 – 5 minutes, 2 = 5 – 10 minutes, 3 = 10 – 15 minutes, 4 = 15 – 20 minutes, 5 = over 20 minutes)

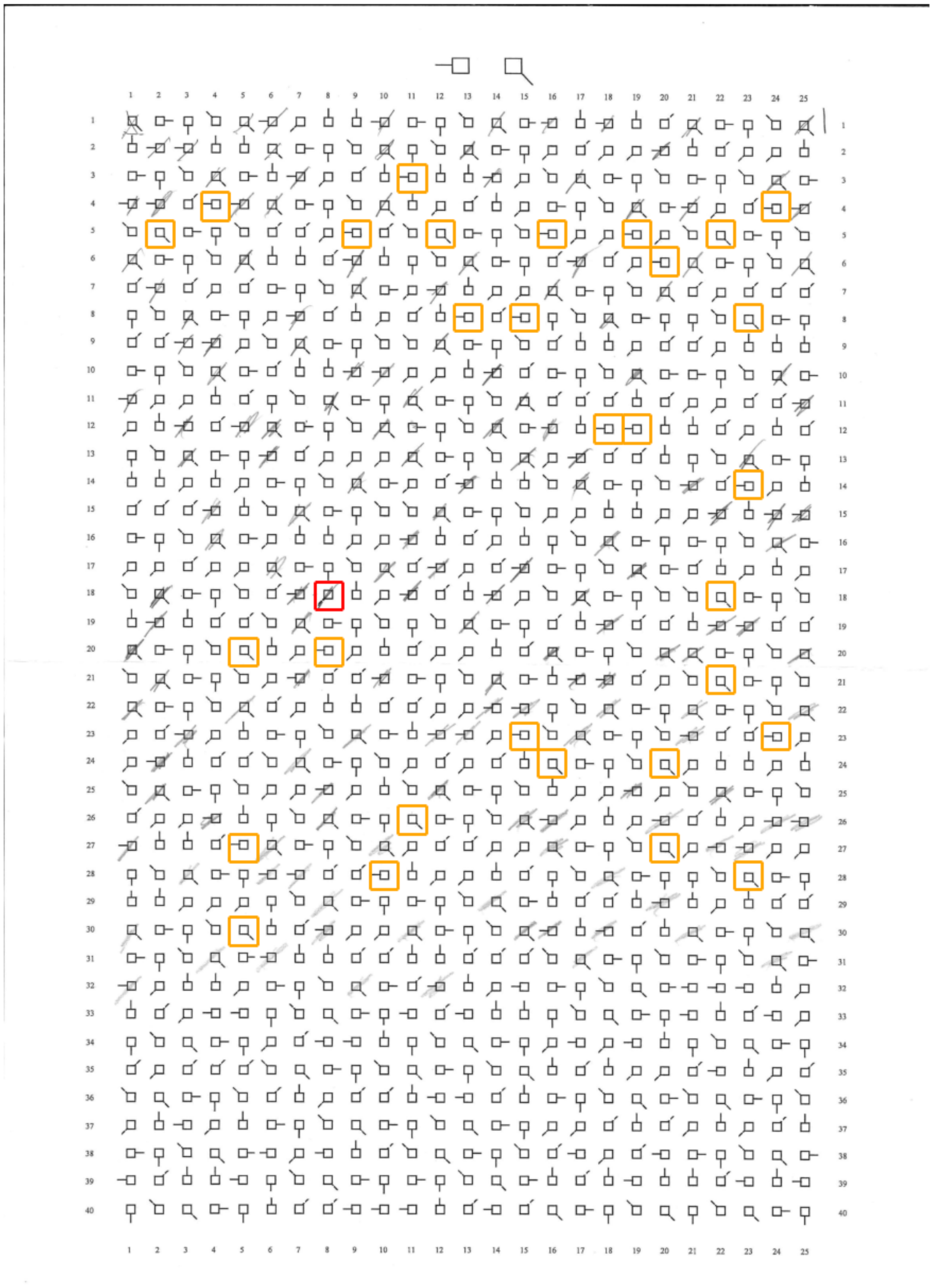
Table F

Survey Responses

Survey Responses			
Responder	**Q1**	**Q2**	**Q3**
1	20 years	80 test sheets	15 - 20min
2	Since the introduction of the MNND in 2011	70 test sheets	10 - 15min
3	Around 15 years	15 test sheets	5 - 10min
4	7 years	40 test sheets	0 - 5min
5	10 years	20 test sheets	0 - 5min
6	3 years	20 test sheets	10 - 15min
7	For 5 years	50 test sheets	5 - 10min
8	Since 2005	25 test sheets	15 - 20min
9	5 years	35 test sheets	10 - 15min
10	4 years	15 test sheets	15 - 20min
11	20 years	20 test sheets	0 - 5min
12	Around 8 years	50 test sheets	5 - 10min
13	Around 10 years (Since the release of the MNND test battery)	30 test sheets	15 - 20min
14	10 years	50 test sheets	5 - 10min
15	For 2 years (Since working as a neuropsychologist)	120 test sheets	10 - 15min
16	For 2 years	40 test sheets	10 - 15min

Figure A

Example of Visualization for Clinicians



Note. Types of errors by patients depicted with different coloured boxes for clinicians. Missed target symbols (yellow boxes) and marked non-target symbols (red boxes).

Declaration of Personal Contribution

Universität ZürichUZH

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Hiermit erkläre ich, dass die Qualifikations-Arbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und dass ich die Grundsätze wissenschaftlicher Redlichkeit eingehalten habe (vgl. dazu: https://api.swiss-academies.ch/site/assets/files/25852/kodex_layout_de_web.pdf).

Zürich, 01.06.2023

....

Ort und Datum Unterschrift