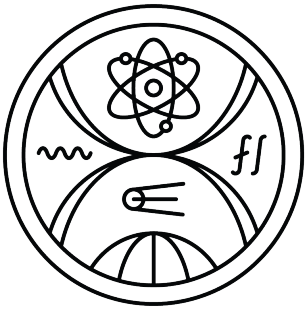
COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS

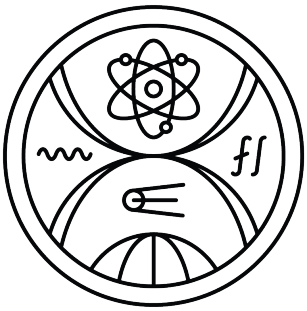


AUTOMATED EVALUATION OF THE REY-OSTERRIETH COMPLEX IMAGE TEST

Master's thesis

2026 Bc. Lucia Korbeľová

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS



AUTOMATED EVALUATION OF THE REY-OSTERRIETH COMPLEX IMAGE TEST

Master's thesis

Study program:	Applied Informatics
Branch of study:	Informatics
Department:	Department of Applied Informatics
Supervisor:	doc. RNDr. Zuzana Černeková, PhD.

Bratislava, 2026 Bc. Lucia Korbeľová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Lucia Korbeľová
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Automatizácké hodnotenie Rey-Osterriethovho komplexného obrázkového testu
Automated evaluation of the Rey-Osterrieth complete image test

Anotácia: Poruchy pamäti sú charakteristickým znakom mnohých rôznych neurologických a psychiatrických ochorení. Rey-Osterriethova komplexná figúra (ROCF) je najmodernejším hodnotiacim nástrojom neuropsychológov na celom svete na posúdenie stupňa zhoršenia neverbálnej vizuálnej pamäte. Na získanie skóre vyškolený klinický lekár kontroluje kresbu ROCF pacienta a kvantifikuje odchýlky od pôvodnej kresby.

Cieľ: Cieľom práce je vytvoriť automatizovaný systém pre hodnotenie neuropsychologického testu Rey-Osterriethovej komplexnej figúry (ROCF) s použitím metód počítačového videnia. Tento systém bude využívať pokročilé techniky počítačového videnia a umelej inteligencie na automatické rozpoznávanie a kvantifikáciu odchýlok od pôvodnej kresby. Účinnosť systému a jeho presnosť budú porovnávané s tradičnými metódami hodnotenia. Cieľom je poskytnúť efektívny nástroj na automatizované hodnotenie neuropsychologických testov, čo môže prispieť k lepšiemu diagnostikovaniu a monitorovaniu neurologických a psychiatrických porúch pamäte.

Literatúra:
Davide Di Febbo, Simona Ferrante, Marco Baratta, Matteo Luperto, Carlo Abbate, Pietro Davide Trimarchi, Fabrizio Giunco, Matteo Matteucci; A decision support system for Rey–Osterrieth complex figure evaluation; Expert Systems with Applications, Volume 213, Part C, 2023, 119226, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119226. (https://www.sciencedirect.com/science/article/pii/S0957417422022448)

Park, J.Y., Seo, E.H., Yoon, HJ. et al. Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline. Alz Res Therapy 15, 145 (2023). https://doi.org/10.1186/s13195-023-01283-w

R. O. Canham, S. L. Smith and A. M. Tyrrell, "Automated scoring of a neuropsychological test: the Rey Osterrieth complex figure," Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future, Maastricht, Netherlands, 2000, pp. 406-413 vol.2, doi: 10.1109/EURMIC.2000.874519.

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vedúci: doc. RNDr. Zuzana Černeková, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 19.10.2023

Dátum schválenia: 14.11.2023

prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

I hereby declare that I have independently completed this master's thesis on the topic 'Automated Evaluation of the Rey-Osterrieth Complex Image Test', including all its appendices and images, using the literature listed in the attached bibliography and artificial intelligence tools, under the careful supervision of my thesis advisor. I declare that I have used artificial intelligence tools in accordance with applicable legal regulations, academic rights and freedoms, ethical and moral principles, while maintaining academic integrity, and that their use is appropriately indicated in the work.

.....

Bratislava, 2026 Bc. Lucia Korbeľová

Acknowledgement

First, I would like to thank my supervisor doc. RNDr. Zuzana Černeková, PhD. for her insightful feedback and support. A special thanks to prof. Ing. Jarosav Polec, PhD for his valuable advice and insights into this topic of research. Finally, I want to extend my heartfelt gratitude to my partner and my family for their unwavering support throughout my academic journey.

Abstract

With the growing number of people suffering from dementia and other neuropsychological disorders, there is an increasing need for reliable and efficient diagnostic tools. One of the most widely used methods for assessing visual memory and cognitive abilities is the Rey–Osterrieth Complex Figure Test (ROCF). However, its manual evaluation is time-consuming and subjective, which limits its frequent use in clinical practice. For this reason, our work focuses on developing an automated system capable of evaluating ROCF drawings in a unified and objective way using modern computer vision and deep learning methods.

In this thesis, we compare two deep learning architectures, convolutional neural network ResNet18 and Swin Transformer, in a simplified classification and training setting to determine which performs better and which of five augmentation techniques brings the best results. The analysis showed that ResNet18 achieved higher accuracy and proved to be more suitable for this type of data. Using the best training configuration from this experiment, we extended our approach to a six-class classification, where each class represents an interval of six points in the ROCF scoring scale, and further optimized preprocessing techniques for this model. Finally, we proposed and tested our own design combining an encoder that extracts low-level image features with a simple neural network that performs the final classification, allowing us to explore the potential of hybrid encoder-based architectures for ROCF scoring.

Keywords: Rey-Osterrieth complex figure, deep learning, convolutional neural networks, transformer, encoder

Abstrakt

S narastajúcim počtom ľudí trpiacich následkami demencie a iných neuropsychologických porúch narastá aj potreba spoľahlivých a efektívnych diagnostických nástrojov. Jednou z najčastejšie používaných metód na hodnotenie vizuálnej pamäti a kognitívnych schopností je test Rey–Osterriethovej komplexnej figúry (ROCF). Manuálne hodnotenie tohto testu je však časovo náročné a subjektívne, čo do určitej miery obmedzuje jeho častejšie využitie v klinickej praxi. Z tohto dôvodu sa naša práca sústreďuje na vytvorenie automatizovaného systému, ktorý je schopný hodnotiť ROCF kresby jednotným a objektívnym spôsobom s využitím moderných metód počítačového videnia a hlbokého učenia.

V tejto diplomovej práci porovnávame dve architektúry hlbokého učenia – konvolučnú neurónovú sieť ResNet18 a Swin Transformer – v zjednodušenom klasifikačnom a trénovacom nastavení, aby sme určili, ktorá z nich dosahuje lepšie výsledky a ktorá z piatich augmentačných techník prináša najvyššiu presnosť. Analýza ukázala, že ResNet18 dosiahol vyššiu presnosť a potvrdil, že je vhodnejší pre tento typ dát. Pomocou najlepšej trénovacej konfigurácie z tohto experimentu sme následne rozšírili náš prístup na šesťtriednu klasifikáciu, kde každá trieda predstavuje interval šiestich bodov v ROCF skórovacej škále, a ďalej sme optimalizovali techniky predspracovania pre tento model. Nakoniec sme predostreli a otestovali náš vlastný návrh kombinujúci kóder, ktorý extrahuje nízkoúrovňové príznaky, s jednoduchou neurónovou sieťou vykonávajúcou finálnu klasifikáciu, čo nám umožnilo preskúmať potenciál hybridnej architektúry založenej na kóderoch pre automatické skórovanie ROCF.

Kľúčové slová: Rey-Osterriethova komplexná figúra, hlboké učenie, konvolúčna neuronová sieť, transformer, kóder

Contents

List of Figures

List of Tables

Terminology

Terms

- **Some term**
  Explanation of the term.

Abbreviations

- **ROCF** - Rey-Osterrieth complex figure.
- **CNN** - Convolutional neural network.
- **BQSS** - the Boston Qualitative Scoring System.
- **MCI** - Mild cognitive impairment
- **MAE** - Mean absolute error
- **ReLU** - Rectified Linear Unit
- **CV** - Cross validation.

Motivation

A world-renowned Alzheimer's disease organization estimated in its World Alzheimer Report 2015 that "46.8 million people worldwide were living with dementia in 2015. This number will almost double every 20 years, reaching 74.7 million in 2030 and 131.5 million in 2050" [13]. The main reason behind this prognosis is the increasing aging of the general population. As a result, there is a growing need for reliable diagnostic and progress assessment tools. One of the standard diagnostic methods is the pen-and-paper test known as the Rey–Osterrieth Complex Figure Test. This neuropsychological test also has countless other applications in clinical practice and in research. It serves as an effective tool for assessing visual memory, visuo-constructional abilities, fine-motor skills, and organizational and planning capabilities. Additionally, it was originally designed to monitor recovery progress during treatment of a brain injury and is often used to help diagnose various neurological and psychiatric disorders, including, but not limited to dementia, Parkinson's disease, schizophrenia, and depression.

However, over time, several systematic issues have led to a gradual decline in the frequency of ROCF usage despite its wide range of applications [3]. One of the most common problems is that administering the test and performing its subsequent evaluation takes a substantial amount of time. In addition, the scoring process is somewhat subjective, since it involves assessing the similarity between drawn and reference shapes, which can vary from one expert to another. This subjectivity introduces variability in the final scores and reduces standardization [11]. For these reasons, the need to develop an automated scoring system for the ROCF test has emerged, as a way to ensure a unified scoring procedure and significantly faster evaluation.

There have been numerous attempts by researchers to create such an automated scoring system for the ROCF test. Early efforts mainly focused on using classical computer vision algorithms and pattern recognition methods to identify specific sub-elements of ROCF drawings and then evaluate their similarity to the reference template. However, the biggest obstacle for these algorithmic approaches was the high level of variation, distortion, and rearrangement of elements in the drawings made by patients [4]. Because of this, early systems had very limited functionality and were only able to successfully recognize and score a small number of the simplest patterns in the ROCF, such as rectangles and triangles. One such system was created by Canham, Smith, and

Tyrrell in 2000 [4].

The adoption of machine learning and deep learning methods has great potential to improve the quality of software in this field, as these techniques can process complex images and automatically discover important discriminative features relevant for scoring. Although there have been several promising attempts to build automated ROCF scoring systems abroad, there is still no practical or usable software designed specifically for Slovakia that follows the Slovak standardization of ROCF test evaluation. This thesis aims to contribute toward developing such a system.

Our goal is to create an automated scoring model for ROCF drawings that is able to approximate the drawing score in six-point intervals, categorizing the images into six groups with a similar level of quality. To achieve this, we compare several preprocessing and augmentation techniques and evaluate different types of machine learning models, including standard convolutional neural networks, transformer architectures, and encoder–neural network combinations.

Chapter 1

Introduction

The aim of the following introductory chapter is to present the theoretical background of the Rey–Osterrieth complex figure test and to describe various existing methods that focus on automating the evaluation or diagnosis formulation based on Rey–Osterrieth complex figures, which have inspired the research methods used in this work. The first subchapter explains the history and various clinical applications of the ROCF Test, as well as alternative versions of complex figures that serve as equivalent or simplified alternatives to the original ROCF. Finally, it briefly introduces different types of scoring systems used to comprehensively evaluate cognitive functions of a subject, with a focus on Osterrieth's scoring system, which was used to produce the scores in our dataset. In the subchapter on existing methods, we summarize two research papers by research groups from abroad, one combining several advanced computer vision and deep learning techniques to predict patient diagnoses, and the other using a purely deep learning approach to estimate ROCF scores from drawings. Additionally, we present two recently completed master's theses from Slovakia that worked with the same image dataset used in this thesis.

1.1 Rey-Osterrieth complex figure test

Clinical neuropsychology is defined as a scientific discipline within psychology that focuses on studying the impact of neurological dysfunctions and injuries on an individual's psychological processes and behavior [2]. One of the main goals of neuropsychology is to diagnose brain dysfunctions by analyzing nonstandard and suboptimal behavior and cognitive performance [11]. For this purpose, various assessment methods are used, including 'pen-and-paper' tests, during which the subject writes or draws on paper. These types of tests are widely used because drawing and writing are complex graphomotor tasks that require visual-perceptual, motor, and coordination skills, which are highly sensitive to brain disorders affecting these abilities [11]. One of the

commonly used tests of this type is the Rey-Osterrieth complex figure test (ROCF), classified as a reproduction task in which the participant is asked to copy or recall a geometric figure presented from memory.

The Rey–Osterrieth complex figure is a standardized illustration that consists of multiple complicated geometric shapes, as shown in Figure 1.1.A [15]. The ROCF test, which is the primary focus of this thesis, can be divided into three stages, each producing a separate version of the figure drawn by the participant. The first stage, called the Copy phase, requires the individual to reproduce the ROCF image, which is presented in front of them, as accurately as possible while viewing the original. Once this drawing is completed, the template is hidden from the subjects view, and the participant is asked to draw the same figure, but this time solely from memory, this is known as the Immediate Recall phase. The final stage, Delayed Recall, takes place approximately 30 minutes later, when the subject is once again asked to recall and redraw the figure. The evaluator then scores each of the three drawings based on how much they resemble the original figure, and depending on the scoring system used, also on the participant's planning and drawing strategy [14].

The origins of the ROCF test date back to 1941, when André Rey introduced the Complex figure test as an effective method for assessing visual memory and visuospatial constructional ability of patients who suffered a brain injury, to assess the extend of their injuries. This influential work was then followed up by Paul-Alexandre Osterrieth in 1944 who standardized the way the test is performed and scoring system, which is commonly used to this day. [14]. Since its introduction, the test has been extensively applied in both research and clinical contexts.

Cognitive functions typically assessed and measured with the ROCF test in both research setting and medical practice include nonverbal and visual memory — that is, the ability to recall visual stimuli — as well as ability to concentrate on the given task, fine-motor skills, the ability to perceive and proccess the spatial relationship of objects in a image, organization and planning abilities, among others [15]. Naturally, a subject's performance in the Copy and Recall phases of the test can reveal different types of neuropsychological dysfunctions. The two Recall tasks mainly evaluate visuospatial memory, whereas the Copy task provides insight into attentional capacity, visuospatial perception required to break down the figure into its components and identify their shapes and positions, and the ability to organize and plan the drawing based on these subunits [14]. Professionally trained psychologists are able to use the results of the evaluation to recognize and diagnose various neuropsychological disorders and brain dysfunctions, including dementia, Alzheimer's disease, Parkinson's disease, attention deficit hyperactivity disorder (ADHD), eating disorders, depression, schizophrenia and the effects of traumatic brain injury [15].

1.1.1 Alternative Versions of the Rey–Osterrieth complex figure

Over time, since the original development of the Rey–Osterrieth complex figure, several alternative versions of complex figures have been designed for similar testing purposes. These variants were created either to provide figures of comparable difficulty to the original ROCF or to simplify the task for special circumstances. Examples of these versions are shown in Figure 1.1.

Initially, the reason for developing numerous figures was related to the original purpose of the ROCF (Figure 1.1.A), which was used before and after treatment of a brain injury as an assessment tool to evaluate deterioration or improvement of visuospatial capabilities. However, that caused that one patient might be asked to perform the ROCF test with the same figure multiple times, prompting the subject to perform better simply because they remember the figure from previous trials, impacting the reliability of the results.

To address this issue, researchers designed several new complex figures that are equally difficult to reconstruct and recall as the original ROCF [15]. The first one was the Taylor figure (Figure 1.1.B), introduced in 1969. However, after subsequent research it was found to be interchangeable with the ROCF in evaluation of copying ability but not when it comes to assessing nonverbal memory and organizational skills [8]. To improve equivalence, the modified Taylor figure (Figure 1.1.C) was developed, which is slightly modified version of Taylor figure with higher complexity which was proved to be truly equivalent in all aspects with ROCF. Other figures with comparable properties are the Mark figure (Figure 1.1.D) and the Medical College of Georgia Complex Figures (Figures 1.1.E–H) [15].

Furthermore, it has been shown that additional disadvantage of ROCF that it's score increases with educational level and declines with increasing age [6], which makes it a less reliable measure of cognitive abilities in elderly or low-educated populations. Therefore, various simplified versions of the ROCF were created to ensure valid assessment of these groups (Figures 1.1.I–L) [15].

1.1.2 Scoring Systems and Evaluation Methods

When it comes to the evaluation of the drawings, multiple scoring systems and methodologies have been developed to assess the Rey–Osterrieth complex figure (ROCF) and other similar complex figures. Each scoring system was designed for a specific research or clinical purpose, focusing on different cognitive processes. Nevertheless, despite the particularities of each approach, two main categories of scoring systems can be identified. The first category includes accuracy based scoring methods, such as Osterrieth's

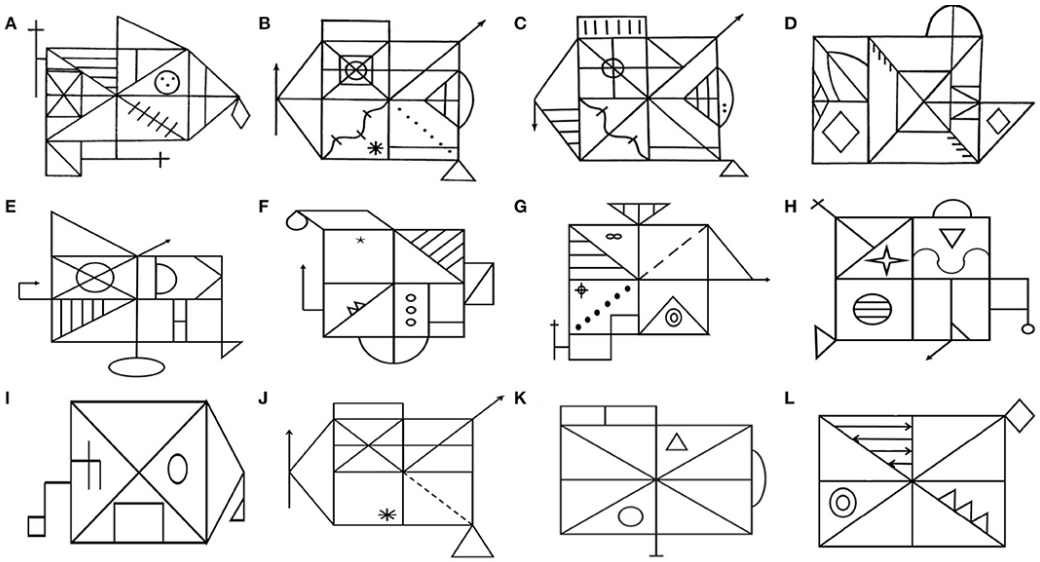


Figure 1.1: Example of the versions of Rey–Osterrieth complex figure, the original ROCF (A), the Taylor figure (B), the modified Taylor figure (C), the Mark figure (D), the Medical College of Georgia Complex Figure (E-H), the Benson figure (I), the simplified Taylor figure (J), the simplified versions of ROCF (I-L), Source: [15].

original scoring system, the Meyers scoring system, and others. The second category consists of process based scoring systems, including the Boston Qualitative Scoring System (BQSS) and the Q-Score [15].

Accuracy-based scoring systems are simpler, more straightforward, and generally more reliable, which makes them more practical for clinical use. Their main focus is to evaluate visuo-constructional abilities and nonverbal memory. The result of these methods is a quantitative score that represents how closely the subject's drawing resembles the original figure. The figure is divided into smaller logical units, and each element is scored based on its similarity to the original in terms of shape and placement relative to other components. However, a limitation of these quantitative methods is that they only consider the final static drawing, ignoring the drawing process itself. As a result, valuable information is completely lost, such as the order in which elements were drawn and how the subject organized the figure into subunits [15].

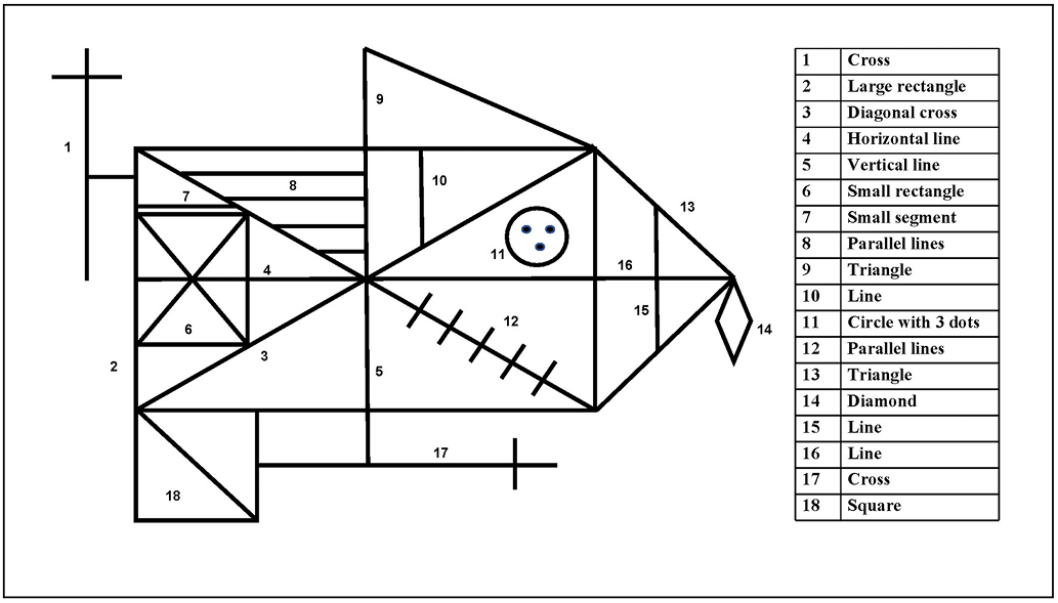
In contrast, *process-based scoring methods* were developed to address these limitations. They aim to more accurately assess cognitive and executive functions that are not captured by accuracy scoring methods, including organizational strategy, planning (drawing order), direction, and placement of elements in the figure. Although process scoring systems provide a more detailed and thorough evaluation of cognitive abilities, they are less frequently used in clinical settings because they require significant amount of time to complete and are complex to administer [15].

The BQSS is an example of a process-based scoring methodology. It produces multiple scores that reflect different aspects of executive functioning [14]. The sys-

tem evaluates not only accuracy and correctness of placement but also planning, size distortion, rotation, asymmetry, and other qualitative characteristics [15]. To capture the drawing order and assess organizational and planning abilities, multiple colored markers are often used to distinguish between different phases of the drawing process [14].

1.1.3 Osterrieth's Scoring System

Osterrieth's accuracy scoring system was developed in 1944 by Paul Osterrieth as the first scoring methodology for the Rey–Osterrieth complex figure test, and it remains commonly used today [8]. The main focus of this evaluation is to assess the correctness of placement and the geometrical accuracy of specific subunits of the ROCF drawing. As a result, it is clear that this evaluation effectively quantifies visual memory and visuo-spatial abilities, however, like other accuracy-based scoring methods, it is unable to quantitatively assess organizational or executive functions [14].



1	Cross
2	Large rectangle
3	Diagonal cross
4	Horizontal line
5	Vertical line
6	Small rectangle
7	Small segment
8	Parallel lines
9	Triangle
10	Line
11	Circle with 3 dots
12	Parallel lines
13	Triangle
14	Diamond
15	Line
16	Line
17	Cross
18	Square

Figure 1.2: Definition of 18 scoring units of Rey–Osterrieth complex figure, Source: [15].

The scoring procedure is relatively simple and can be divided into two main phases. In his original work, Osterrieth divided the Rey–Osterrieth complex figure into 18 independent units (see Figure 1.2). Each of these 18 elements is either a simple geometric shape, such as a rectangle, triangle, or line, or a logical grouping of geometric figures, such as a circle with three dots, a cross, or multiple parallel lines [15]. Therefore, the first step of the evaluation is to divide the subject's drawing into the 18 elements as defined in Figure 2. However, this process is not as straightforward as it might seem, since not all of the original 18 units are typically present in the subject's drawing. Some

elements might be missing, while others are often deformed, incomplete, or placed in unexpected positions. As a result, this division step can be somewhat subjective [15].

Subsequently, each of the 18 elements is scored independently and can receive up to two points, depending on whether the element is present in the drawing, accurately shaped, and correctly placed. Five combinations of presence, accuracy, and spatial placement are possible, and each combination corresponds to a specific score defined in Table 1.1. The element receives the full 2 points if it is present, accurately drawn, and correctly placed. Elements that are accurately shaped but misplaced, or correctly placed but have deformed shape, receive 1 point. If the element is both misshapen, but recognizable and misplaced, it receives 0.5 points. Naturally, omitted elements receive 0 points. Consequently, each individual ROCF drawing can receive a total score ranging from 0 to 36 points [15].

Score	**Accuracy**	**Placement**
2	Accurate shape	Correct placement
1	Accurate shape	Incorrect placement
1	Inaccurate shape	Correct placement
0.5	Inaccurate shape, but recognizable	Incorrect placement
0	Inaccurate shape and unrecognizable, or omitted	Incorrect placement

Table 1.1: Scoring criteria for the elements of ROCF drawing [14].

A low total score indicates impaired brain function related to visuo-constructional abilities (in the copy phase) and visual memory (in the immediate and delayed recall phases). Conversely, healthy individuals are expected to achieve higher scores, although the benchmark values also depend on age and educational background.

1.2 Existing methods

1.2.1 A decision support system for rey–osterrieth complex figure evaluation

The main goal of this study, published in 2023, was to develop a decision support system designed to assist clinicians in evaluating drawings produced during the copy phase of the Rey–Osterrieth complex figure test. Their dataset consisted of 250 reproductions of the ROCF from individuals who were either cognitively healthy, had mild cognitive impairment (MCI), or had dementia. The proposed system generated two types of outputs. First, it assigned one of four qualitative labels to each of the 18 ROCF elements, representing specific error types instead of using the standard quantitative score between 0 and 2 typical for the Osterrieth's scoring method. This reformulation simplified the scoring task into a classification problem of four classes. The labels were:

omitted (element not present), distorted (shape recognizable but inaccurate), misplaced (correct shape but wrong location), and correct. The second output represented the individual's overall diagnosis (healthy, MCI, or dementia) based on the qualitative labels predicted for each of the 18 elements.

The system's workflow consisted of three main phases. The first phase was image preprocessing, which included noise reduction using low-pass and median filtering, binarization through unimodal thresholding, and stroke enhancement using erosion, skeletonization, and dilation. The final preprocessing step involved figure standardization via image homography, which aligned the drawing's size, rotation, and proportions with the reference ROCF template using five manually selected key points of the main structure [5].

The second phase, called pattern evaluation, focused on classifying each of the 18 elements of the ROCF into one of four error categories. These elements were divided into two types of patterns, simple and complex (see Figure 1.3). Simple patterns included lines, triangles, rectangles, and other polygons. For their detection and classification, standard computer vision techniques were applied. The probabilistic Hough transform algorithm was used for line detection, while the contour detection technique was used for shape recognition. If a pattern was present in the drawing, topological analysis was then used to distinguish between the correct, distorted, and misplaced categories.

On the contrary, complex patterns could not be reliably detected using conventional computer vision methods and therefore required the use of deep learning techniques. A modified ResNet50V2 model was trained to generate simplified feature representations of the patterns in the drawings. These feature vectors were then used to train two linear support vector machine (SVM) classifiers. The first one to determine whether a pattern was omitted or present, and the second one to decide whether the detected pattern was correct or distorted. For patterns that were not omitted, topological analysis was additionally applied to determine whether they were misplaced [5].

The last phase concentrated on the diagnosis prediction of the subject based on their rocf reproduction, where the classification into one of three categories (healthy, MCI, or dementia) was based on a machine learning model trained on the 18 qualitative predictions of the ROCF elements obtained in the pattern evaluation phase. For this task, a gradient boosting decision tree algorithm (CatBoost) was used to predict the diagnosis. The system achieved an average accuracy of around 70% in classifying the error types of individual patterns. More importantly, it demonstrated strong performance in the overall diagnosis prediction, successfully distinguishing between healthy individuals and those with MCI in 89% of cases, and between healthy and dementia cases with 92% accuracy [5].

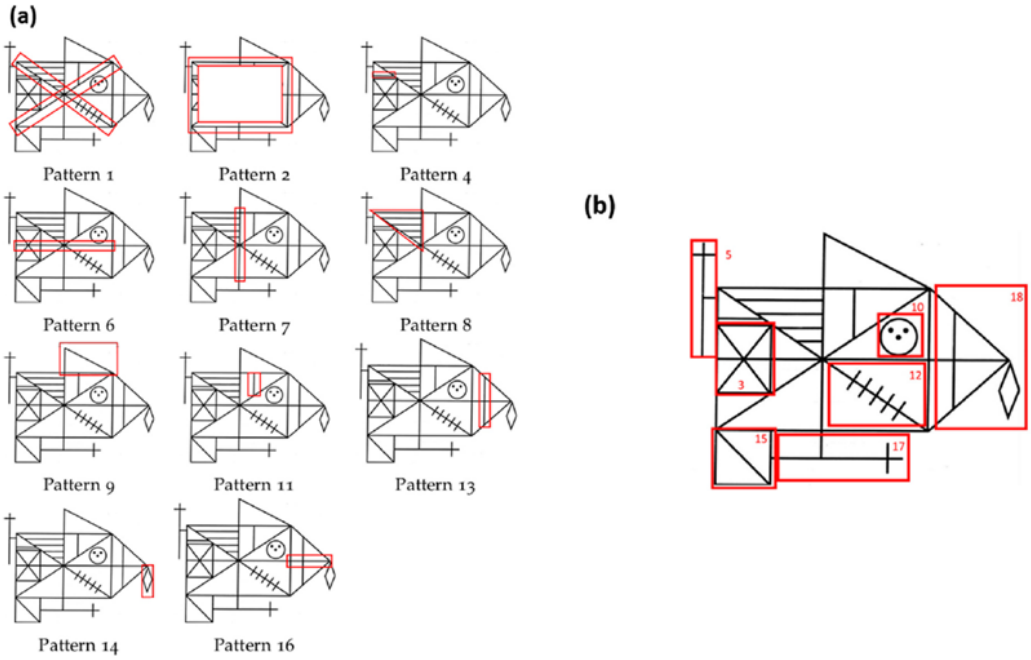


Figure 1.3: Division of 18 ROCF elements into simple and complex patterns, (A) shows simple patterns, (B) contains complex patterns, Source: [5].

1.2.2 Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline

This recent study, conducted in 2023, differs from the one described in the previous section mainly by using an almost hundred times larger dataset and a method based entirely on deep learning techniques making final prediction based on the whole drawing instead of focusing on individual elements of ROCF. Their training dataset consisted of 20,040 scanned ROCF drawings collected from 6,680 participants, each completing the copy, immediate, and delayed recall tasks. Most participants were without cognitively deficit, about one third had mild cognitive impairment, and a few hundred were diagnosed with dementia. The images were annotated by 32 clinical psychologists. Unlike the previous study, which aimed to directly predict the diagnosis from the reproduced figure, their goal was to design a system that is capable of automatically predicting the total score (ranging from 0 to 36) in the same way an expert would.

Before being fed into the deep learning model, each image was first pre-processed. Noise was reduced using median filtering, and to binarize the images the adaptive thresholding method was used. Subsequently, in order to correct rotation and detect how much the drawing is skewed they utilized the projection profile method. Finally, unnecessary background elements were cropped, the images were resized to 512 $\times$ 512 pixels, and converted to three-channel RGB format. They used DenseNet architecture pretrained on ImageNet database as the backbone model, which is a specific type

of convolutional neural network (CNN). The DenseNet architecture consisted of an one convolutional block, followed by four dense blocks, transition layers with batch normalization, a global average pooling layer, and three fully connected layers ending with a final linear activation function that produced the predicted score. The model was trained using the smooth L1 loss function [12].

They trained their DenseNet model twice using fivefold cross-validation, meaning that the model was trained five times on different subsets of the data, each time reserving one fifth of the training set for validation. The Adam optimizer was used for parameter optimization. The first model was trained on the original dataset, and its training performance was evaluated using the mean absolute error (MAE) and the coefficient of determination (R-squared). The model achieved an MAE of 1.24, which means that its predictions differed from the correct ground truth scores by about one point on average, and an R-squared value of 0.977.

Before training the second and final model, the authors identified images where the predicted scores deviated from the annotated ones by five points or more and analyzed the causes of these errors. Among them, they discovered 80 images, which had quality issues that could be corrected. Some errors were caused by poor image quality caused by scanning, others resulted from mistakes made by psychologists during scoring or were caused by accidental errors made when digitalizing the annotations. These problems were corrected to improve the quality of the dataset. Afterwards, the refined dataset was then used to train the final model using the same procedure as before. This second model outperformed the first one, achieving an improved MAE of 0.95 and an increased R-squared of 0.986 [12].

They also used an independent test set containing 150 images used for external validation of the final model. In this evaluation, they analysed how much the predicted scores of the model were in agreement with the scores assigned by five experienced human experts. The model achieved an MAE of 0.64 on the test set, and the average correlation coefficient (R-squared) between the model and each expert was 0.988, which is slightly higher than the average R-squared of 0.983 observed between two distinct human experts. These results demonstrated that the model's predictions closely approximated the scoring performance of human experts [12].

1.2.3 Automated scoring of a neuropsychological test

The following work was conducted as part of a master's thesis by Bc. Peter Iványi (2023) at the Slovak University of Technology in Bratislava, Faculty of Electrical Engineering and Information Technology. Our research is related to this thesis, since we use the same dataset and adopt the same score classification system proposed in their work. The aim of Iványi's thesis was to develop a user interface tool that could au-

tomatically classify ROCF drawings into one of four categories using a convolutional neural network. Since the dataset was relatively small and imbalanced, the author proposed dividing the drawings into four classes based on score intervals to achieve more comparable sizes of classes. The defined score intervals for each class are shown in Table 1.2.

Their proposed preprocessing procedure involved resizing the images to 500 × 500 pixels, converting them to Greyscale, and then binarizing them using the Floyd–Steinberg dithering algorithm, which maps each pixel to the most similar color within a defined palette. They designed a simple custom CNN architecture consisting of 10 convolutional layers, combined with max pooling and batch normalization layers, followed by 4 fully connected layers. The ReLU activation function was applied to introduce nonlinearity, and a final softmax layer was used for classification into four classes. The training process was repeated multiple times with identical settings to obtain the model with the highest validation accuracy. The final model achieved 72.5% accuracy on the testing dataset. Further analysis showed that the model most accurately classified images from classes 1 and 4, the lowest and highest scoring groups [9].

Class	**Score interval**	**Number of images in class**
1	0 - 14.5	240
2	15 - 22	248
3	22.5 - 30	251
4	30.5 - 36	266

Table 1.2: Division of images into four classes based on score intervals [9].

1.2.4 Classification based on Rey figures

The master's thesis authored by Bc. Miroslav Čobrda in 2024 represented a natural continuation of the previously mentioned work. It utilized the same training software and CNN architecture, which was further improved by adding dropout layers, as well as the same dataset and identical definition of four classes based on score intervals. The primary goal of the thesis was to build upon the earlier research by exploring the impact of different preprocessing methods and inclusion of data augmentation. Data augmentation was used to enlarge the dataset and introduce greater variability, by creating copy of each training image and rotating it by 5 degrees. Moreover, two preprocessing strategies were evaluated: use of greyscale images or binarized images using adaptive thresholding.

These results prompted us to attempt to discover augmentation best suited for this dataset and improve the Greyscale images with further preprocessing that would reduce noise and improve visibility of drawing stokes.

Performance results for each model variant are shown in Table 1.3. Interestingly, the results indicate that using original greyscale images without any additional preprocessing yields higher accuracy compared to binarized images. The best performance was achieved when greyscale images were combined with rotational augmentation, reaching an accuracy of 92.5%. To verify the reliability of this configuration, the same preprocessing and augmentation setup was evaluated using fivefold cross-validation and the resulting average accuracy was 88.59% [16]. These findings motivated us to attempt to discover augmentation techniques best suited for this dataset and to enhance greyscale images with additional preprocessing aimed at reducing noise and improving the visibility of drawing strokes.

Method	**Average accuracy [%]**
Adaptive thresholding without augmentation	79%
Greyscale without augmentation	84%
Greyscale with augmentation	92.5%
5-Fold CV, greyscale without augmentation	68.25%
5-Fold CV, greyscale with augmentation	88.59%

Table 1.3: Average accuracy of CNN with various preprocessing and augmentation methods, source: [16].

Chapter 2

Dataset

The images in the dataset used as the core part of this thesis were collected and evaluated as part of the Neuropsy project conducted at the Slovak institute Centrum MEMORY [7]. The Neuropsy project aimed to standardize the evaluation process of 18 neuropsychological tests for clinical use in Slovakia [10]. The images made available to us were obtained from administering the ROCF test to around 350 individuals, evaluated by nine experienced psychologists between 2016 and 2020. Most participants were instructed to draw the complex figure three times: once during the copy phase and twice from memory, after three minutes as the immediate recall and after thirty minutes as the delayed recall.

The original drawings were made with pencil or pen on white paper and were later digitized by scanning. These scans were manually cropped to include only the ROCF drawing, resulting in images of size 1500×1500 pixels, in order to remove unnecessary background and the header of the paper [9]. However, not all drawings produced within the Neuropsy project were suitable for training machine learning models due to quality issues and therefore had to be excluded from the dataset. As a result, the final dataset contains 1041 images. For each image, three types of information are available: a unique image name, the phase in which it was produced (copy, immediate recall, or delayed recall), and the final total score from evaluation.

The evaluation was done using Osterrieth's scoring system, which allows each drawing to be assigned a score in the range from 0 to 36, with a sensitivity of 0.5 points. Out of all images, 933 belong to the so-called control group, meaning they were drawn by cognitively healthy participants, while the remaining 108 drawings were created by individuals with various types of cognitive dysfunctions who were part of the clinical group.

When it comes to the statistics about the phases of the ROCF test during which the images were created, our dataset is quite balanced, with a slightly higher number of samples from the first and second phases. The exact numbers are as follows: 362 images

from the copy phase, 349 from the immediate recall phase, and 330 from the delayed recall phase (see histogram in Figure 2.1). Regarding the distribution of images across the 73 possible scores (ranging from 0 to 36 with a sensitivity of 0.5), we observed that higher scores tend to be more frequent (see histogram of score frequencies in Figure 2.2). Notably, drawings with scores up to 6 represent only 4.51% of the dataset, whereas those with scores between 30 and 36 make up 25.07%. The percentage distribution of images in six-point intervals is shown in Table 2.1.

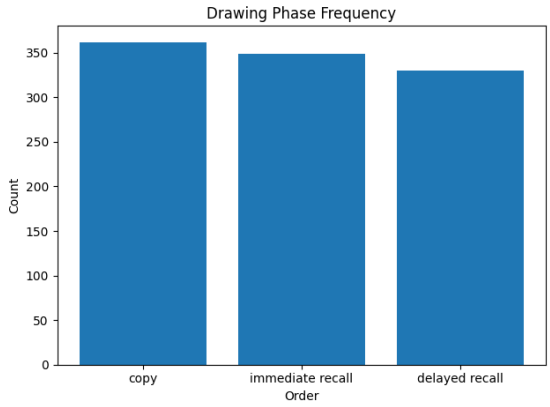


Figure 2.1: Distribution of images across the three ROCF phases.

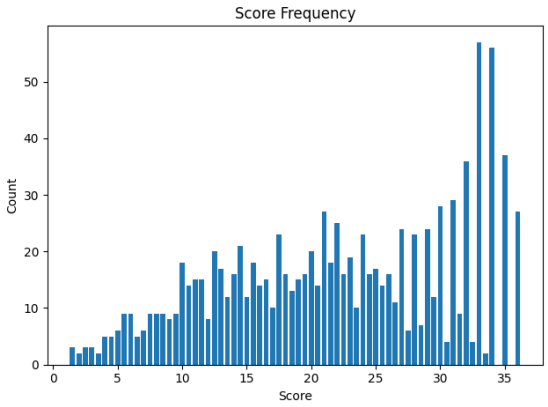


Figure 2.2: Distribution of images by total ROCF score.

Furthermore, some scores do not appear in the dataset at all: specifically 0, 0.5, 1, 34.5, and 35.5. There also seems to be a slight preference for whole-number scores, since drawings with scores with half a point (.5) represent only about 35% of the dataset. In conclusion, the distribution of scores is not evenly spread out, and there appears to be a certain bias toward higher whole-number scores.

Score interval	**Percentage of images**
$\langle 0, 6\rangle$	4.51%
$\langle 6, 12\rangle$	12.01%
$\langle 12, 18\rangle$	18.64%
$\langle 18, 24\rangle$	20.75%
$\langle 24, 30\rangle$	19.02%
$\langle 30, 36\rangle$	25.07%

Table 2.1: Percentage distribution of image scores in six-point intervals.

An example of a ROCF drawing with a high score from our dataset is shown in Figure 2.3. However, due to the nature of the ROCF task, the drawings in our dataset show a great deal of variety and can be quite challenging for computer vision processing. Common issues that appear in the dataset include lightly drawn lines that are not continuous or appear choppy (see Figure 2.4), an uneven or shaky lines caused by trembling hands, and drawings that are deformed or distorted (see Figure 2.5). In

some cases, the images contain many missing elements or so many changes that they barely resemble the original ROCF figure (see Figure 2.6).

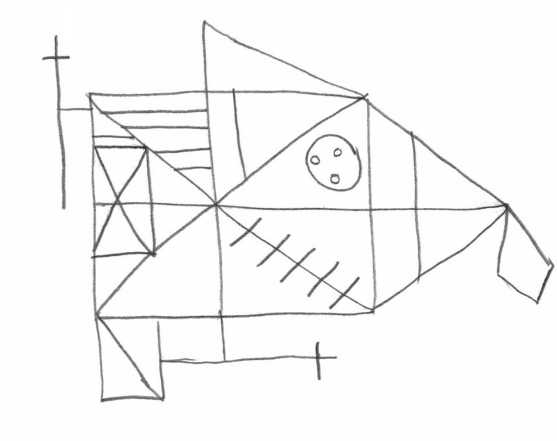


Figure 2.3: Example of a high-quality drawing with a score of 35. Source: [7]

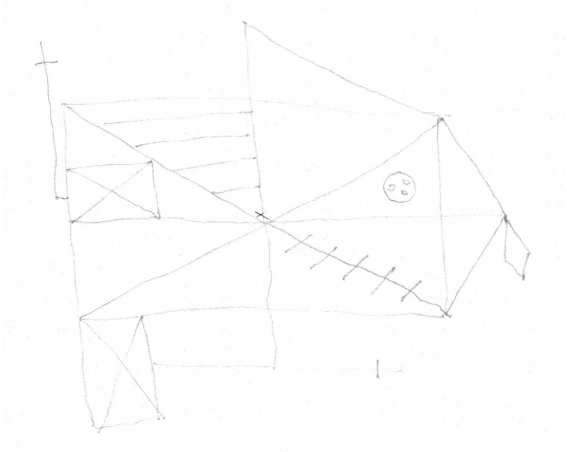


Figure 2.4: Drawing characterized by light and discontinuous lines, with a score of 24.5. Source: [7]

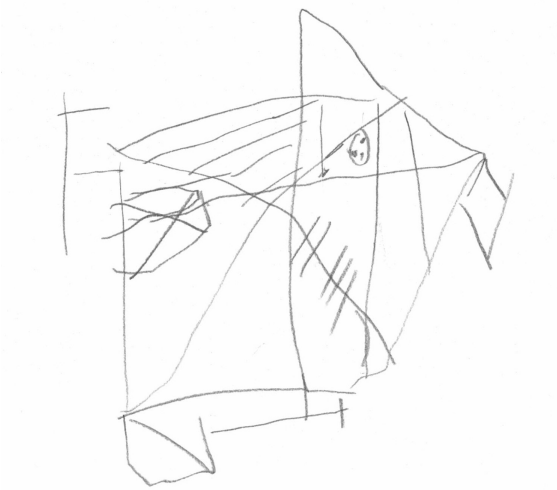


Figure 2.5: Drawing showing strong distortion of the original figure, with a score of 20.5. Source: [7]

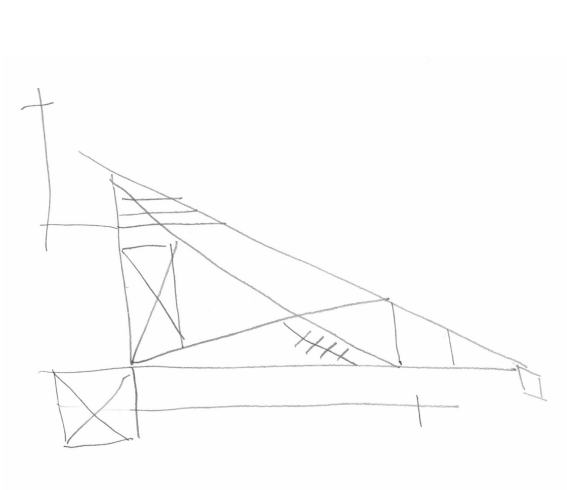


Figure 2.6: Drawing that only slightly resembles the original ROCF, with a score of 7. Source: [7]

Chapter 3

Objective and Methodology

goal of the thesis - mention every goal in the annotation from ais - subchapter for each goal what is my approach - mention article on which my methodology is based upon - my proposal of the solution

Chapter 4

Software design

short report - what inputs, outputs, UML diagrams, all main blocks/components of the pipeline - basic principles

Chapter 5

Implementation

text

Chapter 6

Research

TODO: validation, and results, graphs, images (described and explained), why we are better, false positives, false negatives, why are we not successful, changes of parameters of the system and their evaluation in a table

Draft of the research chapter

In this thesis, the dataset is divided into fixed training, validation, and test sets, each maintaining the same proportional representation of all classes. These splits are consistently used for every model to ensure fair and reproducible comparisons during training and validation.

The performance of all models is evaluated using standard metrics: recall, precision, F1 score, mean absolute error (MAE), and confusion matrix. The results were obtained using test-time augmentation to improve the accuracy.

Initially, two deep learning architectures are compared: the ResNet-18 convolutional neural network and the Swin Transformer. Both models are trained on the same training dataset and evaluated on the same validation and test datasets without any preprocessing or data augmentation. The experiments are conducted with simplified settings: input image size of 400×400 pixels and a four-class classification setup proposed By Ivanyi [9]. Their performance is compared with a baseline method proposed by Iványi et al., which uses a simple convolutional neural network.

Subsequently, both ResNet-18 and Swin Transformer models are trained again using five different data augmentation techniques. Their results are compared with the baseline method from Iványi and with the initial non-augmented versions of both models. Based on these comparisons, the best-performing architecture and augmentation strategy are selected. Preliminary results indicate that ResNet-18 achieves better performance, and the most effective augmentation technique is a combination of random translation, rotation, and adjustments of color brightness and contrast.

The best-performing configuration (ResNet-18 with optimal augmentation) will then be trained from scratch for a six-class classification task, where each class corre-

TTA metrics	Ivanyi	ResNet18	Swin Transformer
recall	72.5%	71.43%	68.57%
precision		72.14%	68.61%
F1 score		70.51%	68.25%
loss		0.9951	1.0253

Table 6.1: Test-time augmentation results on test set of ResNet18 and Swin Transformer trained without augmentation on 4-class classification and results reported by Ivanyi: [9]

sponds to a six-point interval of possible ROCF scores. The input image size will be increased to 500×500 pixels. Four variations of the ResNet-18 model will be trained:

- Without preprocessing and augmentation (serving as a new baseline ground truth method),
- With the best augmentation and no preprocessing,
- With no augmentation and with preprocessing,
- With both preprocessing and the best augmentation.

In addition, an encoder–decoder architecture will be trained on the same dataset. The encoder will be used to extract low-level feature vectors from the drawings, which will serve as input to a simple feedforward neural network performing the final classification. This encoder–neural network approach will be compared with the four ResNet-18 models trained for the six-class task.

	Augmentation type					
TTA Metric	**No**	**contrast, brightness**	**translation**	**rotation**	**crop**	**contrast, brightness, translation, rotation**
recall	71.43%	72.38%	73.33%	73.33%	50.48%	77.14%
precision	72.14%	72.69%	73.51%	73.37%	53.43%	77.61%
F1 score	70.51%	71.58%	73.17%	72.77%	42.88%	76.94%
loss	0.9951	0.8071	0.9922	0.9214	1.0312	0.6135

Table 6.2: Test-time augmentation results on test set of ResNet18 comparing multiple augmentation techniques on 4-class classification

	Augmentation type					
TTA Metric	**No**	**contrast, brightness**	**translation**	**rotation**	**crop**	**contrast, brightness, translation, rotation**
recall	68.57%	64.76%	66.67%	73.33%	24.76%	68.57%
precision	68.61%	65.41%	66.71%	73.29%	6.13%	68.01%
F1 score	68.25%	64.98%	66.57%	72.23%	9.83%	67.72%
loss	1.0253	0.987	0.7173	0.7681	2.4093	0.6958

Table 6.3: Test-time augmentation results on test set of Swin Transformer comparing multiple augmentation techniques on 4-class classification

Chapter 7

Results

interpretation of the results, with input from experts in the field, if the results are good or bad, limits of the method, what are the constraints, up to what point does it work (under what circumstances/cases does it work)

7.1 Summary

Summary text

Chapter 8

User Manual

short description of how to use the software;

Chapter 9

Conclusion

conclusion text

Bibliography

[1] [AI Summary of the article on Automated scoring of the Rey-Osterrieth Complex Figure]. (2025). Online. In: ChatGPT. Version GPT-4o. Available at: OpenAI, https://chatgpt.com/c/680d0e78-155c-8007-909a-17facee070a3. [Accessed 2025-05-12]. Task: "Summarize a scientific article on 'Automated Scoring of a Neuropsychological Test: The Rey Osterrieth Complex Figure' for use in a thesis".

[2] A.L. Benton and A.B. Sivan. Clinical neuropsychology: A brief history. *Disease-a-Month*, 53(3):142–147, 2007. Neuropsychology.

[3] Wayne Camara, Julie Nathan, and Antonio Puente. Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31:141–154, 04 2000.

[4] R.O. Canham, S.L. Smith, and A.M. Tyrrell. Automated scoring of a neuropsychological test: the rey osterrieth complex figure. In *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future*, volume 2, pages 406–413 vol.2, 2000.

[5] Davide Di Febbo, Simona Ferrante, Marco Baratta, Matteo Luperto, Carlo Abbate, Pietro Davide Trimarchi, Fabrizio Giunco, and Matteo Matteucci. A decision support system for rey–osterrieth complex figure evaluation. *Expert Systems with Applications*, 213:119226, 2023.

[6] Michèle Gagnon, Nesrine Awad, Valérie B. Mertens, and Claude Messier. Comparing the rey and taylor complex figures: A test-retest study in young and older adults. *Journal of Clinical and Experimental Neuropsychology*, 25(6):878–890, 2003.

[7] Michal Hajdúk, Petra Brandoburová, Karin Pribišová, Miroslava Abrahámová, Viera Cviková, Daniel Dančík, Sabine Gergely, Simona Krakovská, Eva Mališová, Aneta Svingerová, and Anton Heretik. *Neuropsy: štandardizácia neuropsychologickej testovej batérie na dospelej slovenskej populácii*. Univerzita Komenského v Bratislave, Bratislava, prvé vydanie edition, 2020. kolektívna monografia.

[8] Sherry Hamby, Jean Wilkins, and Neil Barry. Organizational quality on the rey-osterrieth and taylor complex figure tests: A new scoring system. *Psychological Assessment*, 5:27–33, 03 1993.

[9] Peter Iványi. Automatizované hodnotenie neuropsychologického testu. Diplomová práca, Slovenská technická univerzita v Bratislave, Fakulta elektrotechniky a informatiky, Bratislava, 2023.

[10] Centrum Memory. Projekt neuropsy. [online]. Bratislava: Centrum Memory, [cited 2025-11-03]. Available from: https://www.centrummemory.sk/projekty/neuropsy.

[11] Momina Moetesum, Moises Diaz, Uzma Masroor, Imran Siddiqi, and Gennaro Vessio. A survey of visual and procedural handwriting analysis for neuropsychological assessment. *Neural Computing and Applications*, 34, 06 2022.

[12] J. Y. Park, E. H. Seo, H. J. Yoon, and et al. Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline. *Alzheimer's Research & Therapy*, 15:145, 2023.

[13] Martin Prince, Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. World Alzheimer Report 2015.The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends. Research report, Alzheimer's Disease International, September 2015.

[14] Min-Sup Shin, Sun-Young Park, Se-Ran Park, Soon-Ho Seol, and Jun Soo Kwon. Clinical and empirical applications of the rey–osterrieth complex figure test. *Nature Protocols*, 1(2):892–899, 2006.

[15] Xiaonan Zhang, Liangliang Lv, Guowen Min, Qiuyan Wang, Yarong Zhao, and Yang Li. Overview of the complex figure test and its clinical application in neuropsychiatric disorders, including copying and recall. *Frontiers in Neurology*, 12:680474, 2021.

[16] Miroslav Čobrda. Klasifikácia na základe reyových figúr. Diplomová práca, Slovenská technická univerzita v Bratislave, Fakulta elektrotechniky a informatiky, Bratislava, 2024.