Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics

Automatic Determination of Knosp Score Based on Segmentation of Anatomical Structures

Master's thesis

Bc. Filip Oplt

Study program: Medical Electronics and Bioinformatics
Specialisation: Image processing
Supervisor: MUDr. Martin Černý
Second supervisor: prof. Dr. Ing. Jan Kybic

Prague, January 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Oplt Filip**	Personal ID number:	**483697**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Cybernetics**		
Study program:	**Medical Electronics and Bioinformatics**		
Specialisation:	**Image processing**		

II. Master’s thesis details

Master’s thesis title in English:

Automatic Determination of Knosp Score Based on Segmentation of Anatomical Structures

Master’s thesis title in Czech:

Automatické určení Knosp skóre na základě segmentace anatomických struktur

Guidelines:

The Knosp classification system is widely used for scoring the severity of pituitary adenomas. Get familiar with the problem of determining the Knosp score and propose a solution to automatically classify the adenomas based on MRI scans of the brain.
1. Conduct a literature search on the Knosp classification system.
2. Design and implement a model based on geometric relationships of structures in the segmentation.
3. Design and implement a deep learning model to estimate the Knosp score from input images.
4. Evaluate the accuracy of the proposed models and compare them with respect to expert classification.

Bibliography / sources:

[1] E. Knosp, E. Steiner, K. Kitz, and C. Matula, ‘Pituitary Adenomas with Invasion of the Cavernous Sinus Space’, Neurosurgery, vol. 33, no. 4. Ovid Technologies (Wolters Kluwer Health), pp. 610–618, Oct. 1993. doi: 10.1227/00006123-199310000-00008.
[2] M. Černý et al., “Fully automated imaging protocol independent system for pituitary adenoma segmentation: a convolutional neural network—based model on sparsely annotated MRI,” Neurosurgical Review, vol. 46, no. 1. Springer Science and Business Media LLC, May 10, 2023. doi: 10.1007/s10143-023-02014-3.
[3] H. Wang, W. Zhang, S. Li, Y. Fan, M. Feng, and R. Wang, “Development and Evaluation of Deep Learning-based Automated Segmentation of Pituitary Adenoma in Clinical Task,” The Journal of Clinical Endocrinology &amp; Metabolism, vol. 106, no. 9. The Endocrine Society, pp. 2535–2546, Jun. 01, 2021. doi: 10.1210/clinem/dgab371.

Name and workplace of master’s thesis supervisor:

MUDr. Martin Černý   Ústřední vojenská nemocnice - Vojenská fakultní nemocnice Praha

Name and workplace of second master’s thesis supervisor or consultant:

prof. Dr. Ing. Jan Kybic   Biomedical imaging algorithms FEE

Date of master’s thesis assignment: **08.09.2023**   Deadline for master's thesis submission: **09.01.2024**

Assignment valid until: **16.02.2025**

MUDr. Martin Černý Supervisor’s signature	prof. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student's signature

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, 9 January 2024

....
Bc. Filip Oplt

Abstract

This thesis deals with the automatic determination of Knosp scores in magnetic resonance imaging brain scans and their segmentation masks. Knosp score is a grade in a widely used classification system for pituitary adenoma severity assessment. Its correct determination can help to stratify the risks in neurosurgical treatment. A geometric rule-based model and deep learning models are presented as a solution to this task. The available data comprise 394 training subjects and 99 test subjects. On the test dataset, the geometric model correctly classifies 79.80% of the problem's instances, and the best deep learning model exhibits an accuracy of 73.74%. Both models show a good agreement with the expert annotation with a Spearman correlation coefficient of 0.86, respectively 0.84, which is better than a previously reported inter-rater reliability of the Knosp classification system.

Keywords: image classification, medical image analysis, Knosp score, rule-based model, convolutional neural networks, deep learning

Abstrakt

Tato práce se zabývá automatickým určováním Knosp skóre ze snímků magnetické rezonance mozku a jejich segmentačních masek. Knosp skóre je stupeň na škále rozšířeného klasifikačního systému pro hodnocení závažnosti adenomu hypofýzy. Určení tohoto skóre může pomoci stratifikovat rizika při neurochirurgické léčbě. Prezentované řešení zahrnuje geometrický model založený na pravidlech a modely využívající metod hlubokého učení. Poskytnutá vstupní data obsahují 394 trénovacích subjektů a 99 testovacích subjektů. Na testovacím souboru geometrický model správně klasifikuje 79,80 % případů problému a nejlepší model hlubokého učení vykazuje přesnost 73,74 %. Spearmanův korelační koeficient 0,86, respektive 0,84 ukazuje u obou modelů ve vzathu k expertní anotaci lepší shodu, než byla u této klasifikační stupnice dříve zjištěna mezi odbornými hodnotiteli.

Klíčová slova: klasifikace obrazu, analýza medicínských obrazů, Knosp skóre, model založený na pravidlech, konvoluční neuronové sítě, hluboké učení

Acknowledgements

I would like to thank MUDr. Martin Černý for his guidance, remarks and kind attitude, which encouraged me both personally and also with the work's progress. My thanks also belong to prof. Dr. Ing. Jan Kybic, for his insights, proposed ideas and technical support, as well as to Ing. Denis Baručić for sharing his experience with me.

At the same time, I would like to express my immense gratitude towards my family for their care and support; it means a lot to me. Furthermore, I would also like to thank my friends, classmates and colleagues, who were there for me during my studies.

The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged.

List of Tables

3.1 Input datasets 22
3.2 Distribution of Knosp scores in datasets 22

4.1 Top 5 deep learning models 35
4.2 Top deep learning models by datasets 35
4.3 Comparison of geometric and deep learning model 35

List of Figures

2.1 Anatomy of the sella turcica region 4
2.2 Anatomical terminology 4
2.3 Signs of acromegaly 6
2.4 Magnetic resonance imaging: Relaxation times 8
2.5 Knosp classification system 10
2.6 Perceptron 11
2.7 Multilayer perceptron 12
2.8 Convolution in 2D image 14
2.9 Convolution with flipped filter 14
2.10 Maximum pooling 14
2.11 ResNet18 17
2.12 Popular architectures for convolutional neural networks 18

3.1 Histogram of input scans' number of slices 22
3.2 Pipeline of the geometric model 25
3.3 Example of visualized pituitary adenoma classification 25
3.4 Data augmentation examples 28

4.1 Confusion matrices for the output classifications 36

5.1 Misclassification example: edge case 38
5.2 Wrong annotation example: left side 39
5.3 Wrong annotation example: right side 39
5.4 Classification based on a predicted mask: correct example 43
5.5 Classification based on a predicted mask: failed example 44

List of Acronyms

2D 2-dimensional. 8, 14, 23, 27, 29

3D 3-dimensional. 1, 7–9, 23, 27, 29, 40

ACTH adrenocorticotropic hormone. 5, 22

ANN artificial neural network. 11

CE contrast-enhanced. 7, 21

CNN convolutional neural network. 12, 13, 15–19, 27

COM centre of mass. 24

CS cavernous sinus. 3, 9

CT computed tomography. 7

FSH follicle-stimulating hormone. 5

GH growth hormone. 5, 22

GPU graphics processing unit. 16, 34, 49

HDF Hierarchical Data Format. 34, 47

HPC high performance computing. 30, 34, 35

ICA internal carotid artery. 3, 9, 21, 23–27, 39, 41, 43, 44

ILSVRC ImageNet Large Scale Visual Recognition Challenge. 16

LH luteinizing hormone. 5

MLP multilayer perceptron. 11, 13

MRI magnetic resonance imaging. 7–9, 18, 19, 21, 23, 27, 41, 45

NIfTI Neuroimaging Informatics Technology Initiative. 21, 26, 33, 49

PA pituitary adenoma. 1, 5–9, 18, 19, 21, 23, 24, 26, 38, 39, 41, 45

PET positron emission tomography. 7

PitNET pituitary neuroendocrine tumour. 6

PNG Portable Network Graphics. 27

PRL prolactin. 5, 22

RAM random-access memory. 40

RCI Research Center for Informatics of the Czech Technical University. 34, 49

ReLU rectified linear unit. 13

RF radiofrequency. 7

RGB red, green and blue. 27

TIFF Tagged Image File Format. 27

TPU tensor processing units. 49

TSH thyroid-stimulating hormone. 5

VGG Visual Geometry Group. 16, 17

WHO World Health Organization. 5, 7

Contents

Abstract v

Abstrakt vi

Acknowledgements vii

List of Tables ix

List of Figures xi

List of Acronyms xiii

1 Motivation 1

2 Theoretical introduction 3

2.1 Pituitary gland . . . 3
2.2 Pituitary adenomas . . . 5
2.2.1 Classification of pituitary adenomas . . . 5
2.2.2 Symptoms and diagnosis . . . 6
2.3 Pituitary adenoma imaging . . . 7
2.3.1 Treatment . . . 8
2.4 Knosp classification system . . . 9
2.5 Convolutional neural networks . . . 10
2.5.1 Evolution and general concepts . . . 10
2.5.2 Building blocks . . . 13
2.5.3 Training of convolutional neural networks . . . 15
2.5.4 Overfitting . . . 16
2.5.5 Architectures for image classification . . . 16
2.5.6 Transfer learning and fine-tuning . . . 17
2.6 Deep learning and pituitary adenoma disgnostics . . . 18

3 Data and methods 21

3.1 Image dataset . . . 21
3.2 Rule-based geometric model . . . 23
3.2.1 Representation of individual objects . . . 23
3.2.2 Distinguishing the arteries . . . 23
3.2.3 Determination of the critical lines . . . 24
3.2.4 Determination of the Knosp score . . . 24
3.3 Deep learning models . . . 26

3.3.1 Preprocessing of the dataset . . . 26
3.3.2 Imbalanced dataset handling . . . 28
3.3.3 General concepts for the deep learning model . . . 29
3.3.4 Model architecture . . . 29
3.3.5 Training of the models . . . 30
3.4 Statistical evaluation . . . 31

4 Results 33
4.1 Functionality of the geometric model . . . 33
4.2 Performance of the geometric model . . . 33
4.3 Training of the deep learning models . . . 34
4.4 Performance of the best deep learning model . . . 34

5 Discussion 37
5.1 Geometry-based model . . . 37
5.2 Deep learning models . . . 40
5.2.1 Other examined approaches . . . 40
5.3 Value proposition . . . 41
5.4 Comparison with other works . . . 42
5.5 Current issues and open problems . . . 42

6 Conclusions 45

A Attachments provided 47

B Technologies used 49

Bibliography 56

Other sources 57

Chapter 1

Motivation

The diagnostic and therapeutic methods in medicine keep evolving, and the imaging domain can offer many examples of the increasing availability of diagnostic data. However, as the amount of data increases, so does the time it takes to evaluate it. Especially 3-dimensional (3D) imaging techniques require a more thorough examination to deliver insights. While the complexity of tasks in healthcare grows, the workforce remains limited, which naturally raises the demand for more automation in both data acquisition workflows and subsequent analyses.

With medicine being a rather conservative discipline, it can be problematic to incorporate automated actions among standard procedures. Despite this, suitably designed tools can advance the efficiency of work done by medical doctors and other staff in healthcare, support their decisions, reduce the time they spend on auxiliary tasks, and offer them more time for the patients.

In pituitary adenoma (PA) surgery, there are multiple tasks that require expert diagnostics. Acquiring 3D brain scans to analyse the tumours has become a common procedure in modern medicine. The analysis of the tumour includes marking it in the scan (creating a segmentation mask) and evaluating its severity. An undesirable property, which can be detected from the segmentation mask, is the tumour's invasion in the neighbouring anatomical structures. Extracting such information from the image data by automated tools can serve to provide medical doctors with aid in diagnostics.

This work contributes to the effort of creating such tools and bringing more automation to the diagnostics of PAs, with a connection to broader research. The thesis supervisors have published a paper on fully automatic segmentation of PAs and arteries of interest in their vicinity from brain scans. This thesis follows the research and explores the possibilities of using the segmentation masks (both manually and automatically created) and raw brain scans to assess the tumours' invasiveness by predicting their Knosp score: a grade in a widely used classification system for PAs.

The goal of this thesis is to develop a rule-based model implementing the Knosp classification system criteria and a model based on deep learning techniques. The performance of these models should be evaluated and compared using the available data. The proposed models can act as a proof of concept for an auxiliary diagnostic tool and can help further research in this area.

Chapter 2

Theoretical introduction

This work is dedicated to a problem from the medical domain solved by computer science methods. To overcome the gap between the two parts of this interdisciplinary task, this chapter covers the fundamentals of the related topics and should equip the reader with the required knowledge. The necessary theoretical background in both areas will be explained.

2.1 Pituitary gland

The pituitary gland, also called hypophysis, is a small endocrine organ located at the base of the brain [1], with its size being compared to a pea [2]. It is connected to the hypothalamus by the pituitary stalk (infundibulum) and lies in sella turcica, which is a saddle-shaped structure in the sphenoid bone. The surrounding space is mainly filled with the cavernous sinus (CS), a part of the venous system that serves for the outflow of blood from this brain region. Inside the CS is the cavernous part of the internal carotid artery (ICA). The ICA is a paired artery and forms a curvature within the CS. Therefore, the imaging techniques can capture up to 4 cross-sections of the ICA (two on the right and two on the left side) in some slices of this area in the coronal plane (see 2.5 for examples). (A coronal plane is a plane dividing the space of the head to the front and back half-space. This plane is orthogonal to an imaginary line running from the front of the head to the back, which is called the rostro-caudal axis.) The anatomy of this region is depicted in figure 2.1, and the anatomical planes are illustrated in figure 2.2.

There are two parts of the pituitary gland, differing in their origin and function: adenohypophysis (anterior pituitary, i.e. front lobe) and neurohypophysis (posterior pituitary, i.e. back lobe). As an endocrine organ, the pituitary gland is responsible for the production of multiple hormones. The hormones produced by adenohypophysis are [2]:

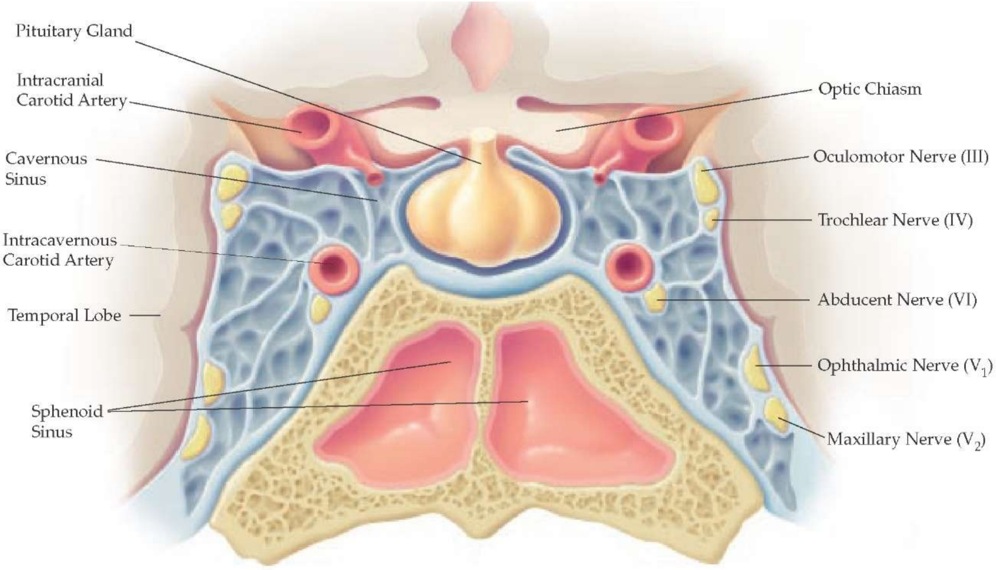


Figure 2.1:
Anatomy of the sella turcica region. Source: [3], modified

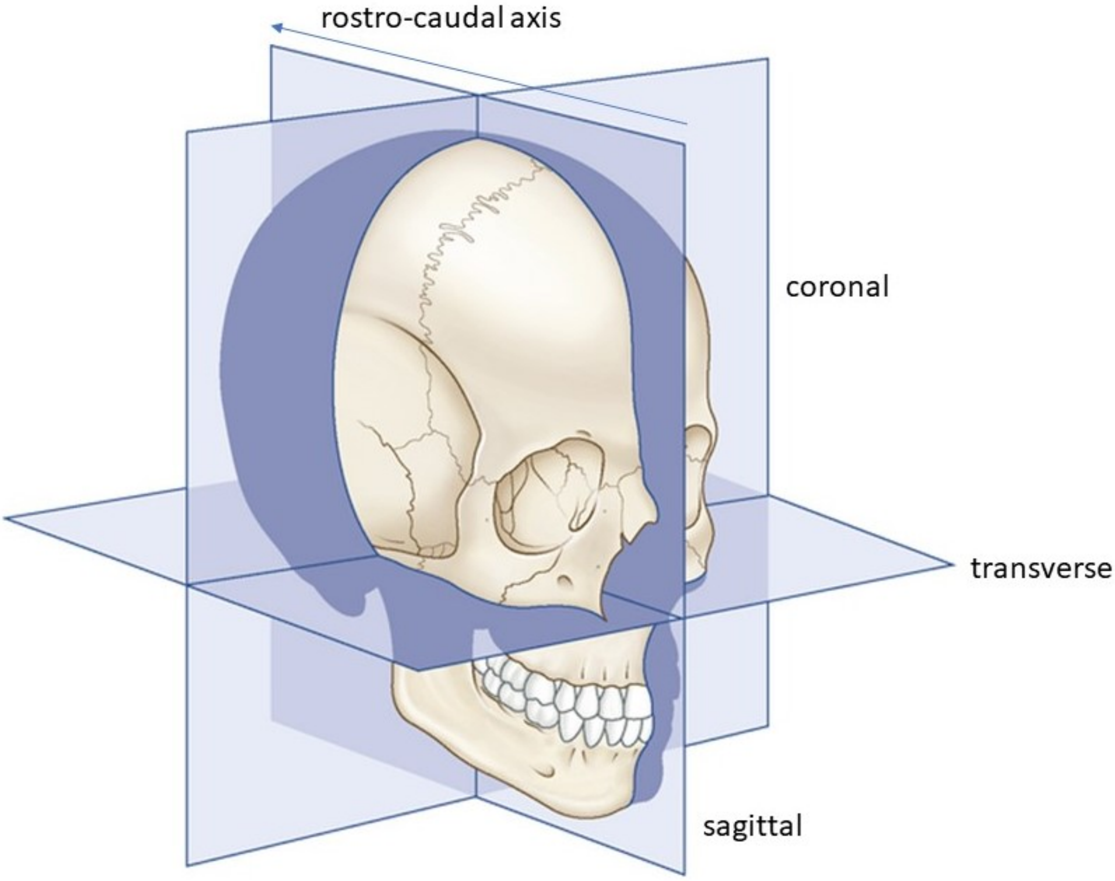


Figure 2.2:
Anatomical planes and rostro-caudal axis. Source: [4]

- adrenocorticotropic hormone (ACTH),
- follicle-stimulating hormone (FSH) and luteinizing hormone (LH), together called gonadotropins,
- growth hormone (GH),
- prolactin (PRL),
- thyroid-stimulating hormone (TSH).

The neurohypophysis produces vasopressin and oxytocin [2]. Each hormone serves as a signalling molecule, and in this way, it controls a specific bodily function or another organ's function or activity. Hence, this enumeration exhibits how vital the pituitary gland's role is. The activity of the pituitary gland itself is regulated by hormones produced in the thalamus.

2.2 Pituitary adenomas

PAs are the most frequent type of pituitary gland tumours [5], and they account for 10-15 % of all intracranial tumours [6]. Adenomas, in general, are benign tumours of glands' epitheliums: they form an abnormal mass of cells but are not malignant, and do not create metastases. The prevalence of the PAs is reported to be 37 to 116 cases in a population of 100,000 inhabitants [7][8][9][10]. In the span of a lifetime, about 10 % of people can develop a PA [11][12].

2.2.1 Classification of pituitary adenomas

The PAs can be categorized by their size into microadenomas (smaller than 1 centimetre) and macroadenomas (larger than 1 centimetre) [13][14], and each group represents approximately half of the cases [9]. Another division is to functioning and nonfunctioning adenomas, based on whether or not the adenomas release hormones [11][15][6]. The functioning PAs can be further classified into types according to the hormones they produce [14]:

- somatotroph adenomas produce GH,
- prolactinomas (also lactotroph adenomas) produce PRL,
- corticotroph adenomas produce ACTH,
- thyrotroph adenomas produce TSH,
- and gonadotroph adenomas produce FSH, LH or both.

The most often types of diagnosed PAs are non-functioning adenomas (43.0%) and prolactinomas (39.9%) [7]. Recently, prolactinomas became the most prevalent type [8][13].

The classification of pituitary tumours is regularly evaluated and updated by the World Health Organization (WHO) [16][17]. The most recent WHO classification suggests a change in nomenclature, including a new designation for PAs: now called pituitary neuroendocrine tumour (PitNET) because some of the tumours can be more harmful than the term “adenoma” indicates [17][18]. Nevertheless, for consistency with the previous research and cited sources, this text continues to use the term “pituitary adenoma”.

2.2.2 Symptoms and diagnosis

The PAs can be asymptomatic, and especially microadenomas and non-functioning adenomas can often be discovered by chance during an unrelated examination [19][11]. Manifested PAs exhibit symptoms caused by [19]:

- the enlargement of the adenoma
- or changed hormonal production.

The enlarged PAs can cause compression of neighbouring structures and lead to headaches, visual impairment (due to optic chiasm compression) and other problems [11][19].

In functioning PAs, the symptoms are determined by the hormone production. Prolactinoma may cause abnormal breast milk production, infertility, erectile dysfunction or gynecomastia (in men), amenorrhea (in women) or a decrease in libido [19][2]. Somatotroph adenomas can manifest with acromegaly or gigantism (excessive growth, illustrated in figure 2.3) together with hypertension and other problems [19][2]. Corticotroph adenomas are responsible for Cushing’s disease and are its most frequent cause [20]. Depending on the mechanism of the disorder, there may be also many other symptoms and effects present [9][12][6][21].

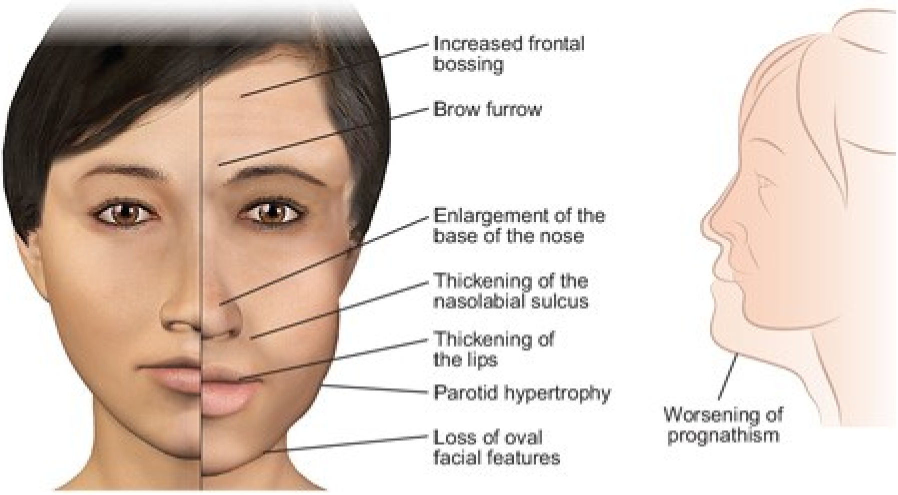

Increased frontal bossing
Brow furrow
Enlargement of the base of the nose
Thickening of the nasolabial sulcus
Thickening of the lips
Parotid hypertrophy
Loss of oval facial features
Worsening of prognathism

Figure 2.3:
Signs of acromegaly. Source: [22]

Diagnostic methods capable of detecting PAs include various biological tests (blood tests, urine tests), imaging techniques and physical examinations (typically vision testing) [23]. The tests can indicate changes caused by abnormal hormonal activity, while the imaging techniques aim at displaying anatomical changes in the affected structures. For imaging, magnetic resonance imaging (MRI), computed tomography (CT) or positron emission tomography (PET) are usually used [24][25][23], with MRI providing a better sensitivity, which is helpful for detecting microadenomas [26].

2.3 Pituitary adenoma imaging

The 5th edition of the WHO classifications [24] identifies the MRI as the most useful modality for PA diagnosis. It offers recommended strategies for specific PA types and also explains advanced MRI techniques. However, the imaging protocols are also influenced by custom procedures in hospitals, settings of medical devices, habits of the staff and other factors. Therefore, such specifics have to be taken into account, and diagnostic processes need to be aligned with the characteristics of the data used, which also holds for the design of automated tools. In this work, contrast-enhanced (CE) T1-weighted scans are used.

MRI is an imaging technique that enables acquisition of 3D scans. It utilises the magnetic properties of certain atomic nuclei in the human body. First, the patient is placed inside a strong external magnetic field, which makes the nuclei align their magnetic axis parallel or antiparallel to it. Then a radiofrequency (RF) pulse excites the nuclei: more of them move to the antiparallel state, and the nuclei align their precession spin in phase. These processes affect the measured net magnetization, which is flipped to the transversal plane and starts to rotate. Its relaxation after the RF pulse to the original state is characterized by two times, also shown in figure 2.4:

- **T1 relaxation time**, related to the recovery of the longitudinal magnetization,
- and **T2 relaxation time**, related to the time of dephasing of the spins.

To translate the measured magnetization into an image, slices in the body are selected by adding a magnetic gradient to the external magnetic field along a selected axis. A position in the slice is encoded in the frequency and phase of short signals applied between the RF pulse and the measurement of the signal [27].

There are multiple possible output images of the MRI, depending on which signal is measured. Besides the T1-weighted image (representing the T1 relaxation times), T2-weighted image and other possible sequences, there is also the option to use a contrast agent to highlight the object of interest (the lesion). In MRI, gadolinium contrast agents are used [29], which increase the intensity of the signal [30].

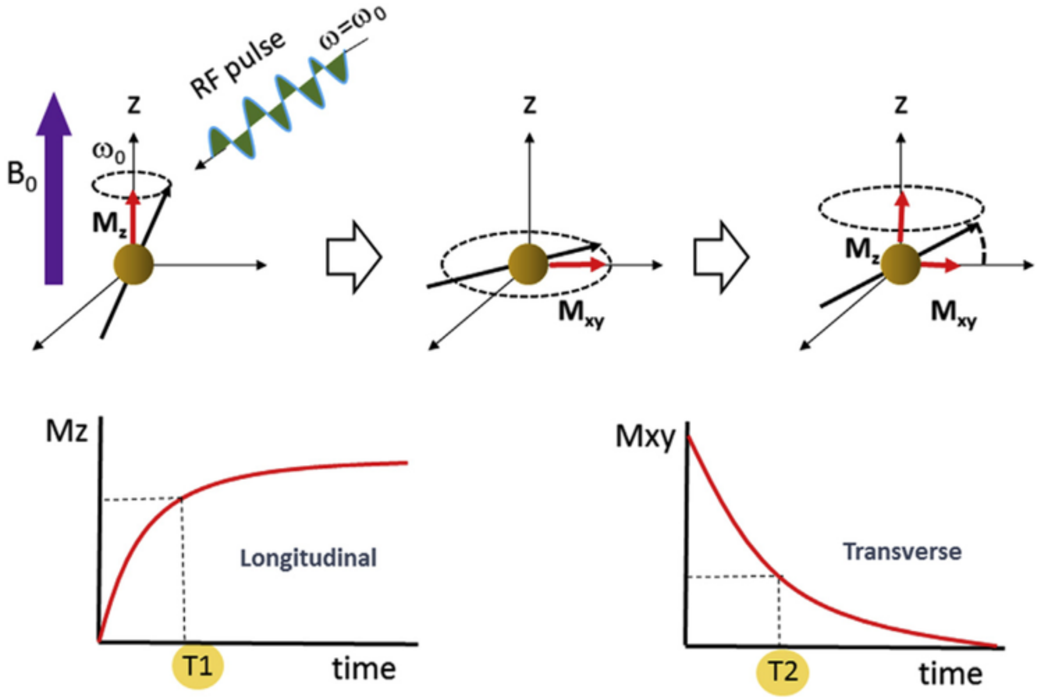


Figure 2.4:
MRI: Relaxation times. B_0 is the external magnetic field, M is the measured magnetization, decomposed into vectors along individual axes. Source: [28]

Each point in the resulting image has an intensity proportional to the intensity of the signal from the corresponding physical position. The elements of a 3D image (scan) are called voxels, similar to pixels in a 2-dimensional (2D) image.

2.3.1 Treatment

The primary treatment for most functioning PAs is surgical [15][19]. The first-line therapy is trans-sphenoidal surgery (endoscopic transnasal approach), but other options, including medication or radiation therapy (e.g., Leksell's gamma knife), are also feasible [15][31].

Medical therapy can be preferred for prolactinomas [15][19]. Asymptomatic PAs can be observed. The treatment can involve multiple modalities (pharmacotherapy, surgery, radiotherapy) and is selected in order to be most beneficial to the patient [15].

Although endoscopic trans-sphenoidal surgery is the preferred way of treatment for most functional PAs, it is still connected with possible postoperative complications and side effects, including cerebrospinal fluid leak, diabetes insipidus, postoperative nausea, bleeding, meningitis and others [32][6]. For this reason, an effort is made to predict the treatment outcomes and anticipate possible complications based on the pre-operative examinations [33][34][35].

There are multiple characteristics examined to predict the outcomes of the surgery

[34][36][37][38], but their design reflects the common goal of stratifying the risks of the treatment. Multiple classification systems have been developed to describe the PAs. The first such classification system was devised in 1976 by Hardy et al. and later modified by Wilson [34]. In 1993, Knosp et al. presented a new classification system based on the comparison of MRI images and surgical results.

2.4 Knosp classification system

The Knosp score is a number on a classification scale of 0-4 describing the PA's invasion to the CS. It describes how far the adenoma extends with respect to the neighbouring structures [39]. The classification is related to the cross-sections of the intra- and supracavernous ICA visible in the image. First, the four cross sections are identified in the image (two on the left and two on the right side). These cross-sections are then connected with a medial tangent, intercarotid line (connecting their centres), and a lateral tangent. These lines are determined separately on the left and right sides. The Knosp score is then evaluated in the following way:

- **grade 0**: the tumour does not cross any of the lines under the intracavernous ICA,
- **grade 1**: the tumour crosses the medial tangent but not the intercarotid line,
- **grade 2**: the tumour crosses the intercarotid line but not the lateral tangent,
- **grade 3**: the tumour crosses the lateral tangent but does not encapsulate the ICA,
- **grade 4**: the ICA is fully encased in the tumour.

The score is evaluated independently on the left and right sides. To capture the 3D nature of the problem, the overall score is the maximum among scores measured in all coronal slices of the scan. Grade 3 was later divided into grades 3A and 3B by Micko et al. [40] based on the position of the PA's tissue extending above (3A) or below (3B) the intracavernous ICA. The Knosp classification system is depicted in figure 2.5.

The determination of the Knosp score is usually based on a segmented MRI scan. The segmentation can be performed manually by an expert, using semi-automatic tools [41] or automatic segmentation methods. These include graph-based algorithms [42] or deep learning methods, which are actively researched (as mentioned below in section 2.6).

The disadvantage of the Knosp classification system is that it may have weak interrater reliability [43]. Recently, a new classification system, called Zurich pituitary score, was introduced [36][44]. This classification system has a better interrater agreement and can be also used for predictions of the surgery [45], but its role is not to replace the Knosp classification system. It should rather provide additional information, and both systems can be combined to achieve more accurate predictions [45], or they can be used specifically in cases where one or another performs better [36].

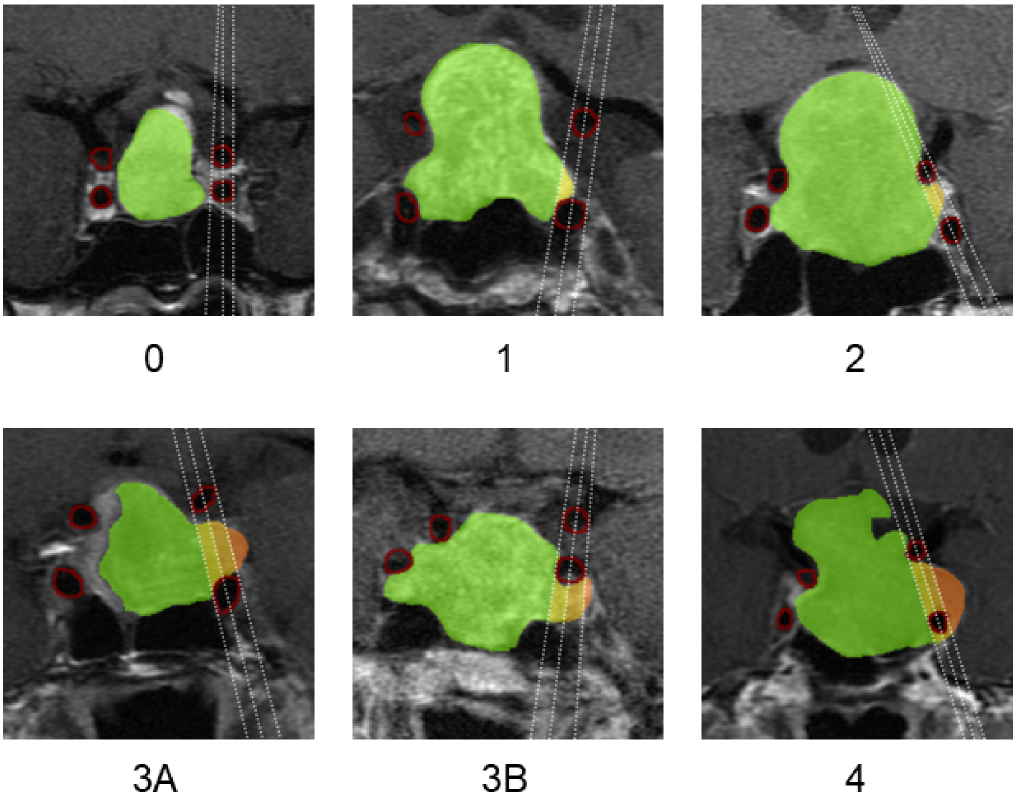


Figure 2.5:
Knosp classification system shown in examples from the input dataset.

2.5 Convolutional neural networks

Artificial neural networks (ANNs) designate complex computational models comprising individual units called neurons. Their fundamentals were established decades ago [46][47]. However, it is mainly with the massive increase in computing power that they have proved their performance in a wide range of problems and have become state-of-the-art solutions for many of them over the last decade [48]. Convolutional neural networks (CNNs) represent a specific type of ANNs with convolutional layers [49]. They allow efficient image processing, which makes them suitable for many tasks in computer vision, including image classification and segmentation. A more detailed description of CNNs will follow in the subsequent sections.

2.5.1 Evolution and general concepts

The basic concept behind ANNs can be represented by a perceptron: a linear classifier which combines multiple inputs to produce its output (figure 2.6) [46]. The inputs are weighted and summed, and the result is compared to a threshold (that can be replaced by a bias term added to the result, followed by a comparison to 0). This creates a binary decision criterion, which can be used to classify linearly separable data points. The criterion can be expressed by the following equations 2.1 and 2.2:

$$f(x) = \sum_i w_i \cdot x_i + b \tag{2.1}$$

$$y(x) = \begin{cases} 1 & \text{if } f(x) \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where x are the inputs, w are their weights, b stands for the bias term (which shifts the decision threshold towards more positive or negative values), and y represents the classification.

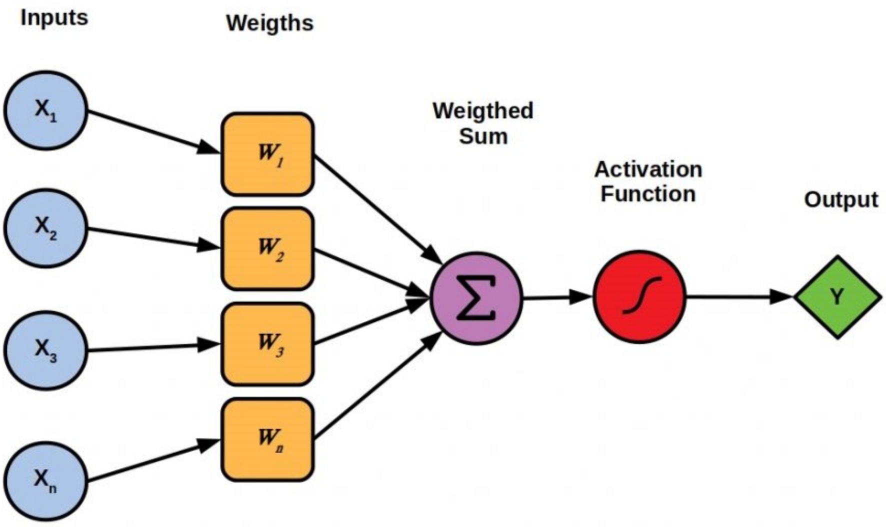


Figure 2.6:
Perceptron. Source: [50]

The inputs can be coordinates of one data point in a multidimensional space, but perceptrons can also be organized in multiple layers and outputs of one layer can be fed as inputs to the next layers. Such a structure (displayed in figure 2.7) is called a multilayer perceptron (MLP) and is a specific case of an ANN [46]. More generally, the terms “deep neural network” or “deep learning” refer to an artificial neural network (ANN) having multiple intermediate (often called hidden) layers.

The ANNs for image classification have images on their input. The image is represented by storing colour intensity values in each of its elements: pixels. Should the model learn the weights for every pixel separately, it could easily get too complex. Moreover, this representation would be prohibitive in extracting the spatial context between images’ pixels.

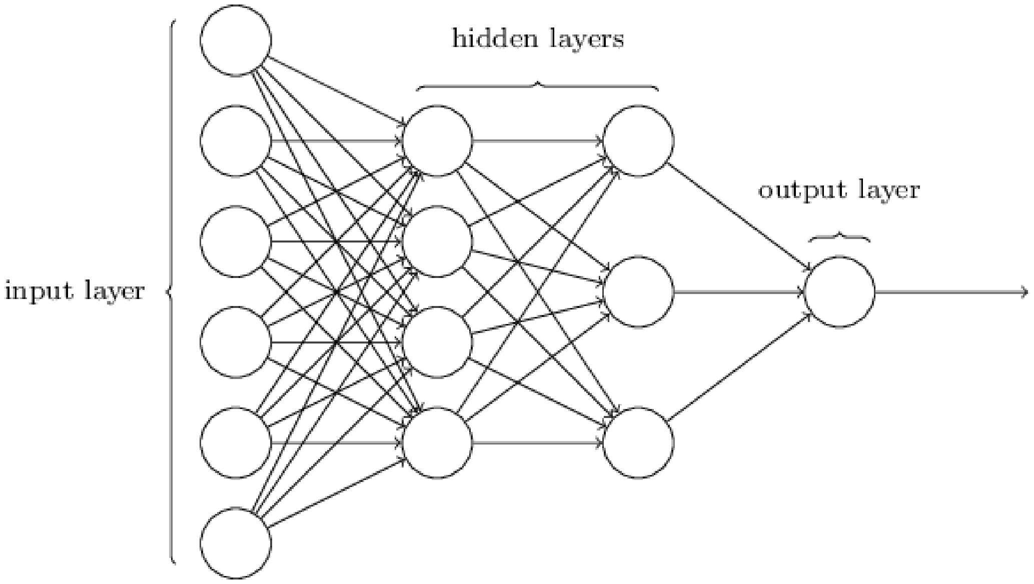


Figure 2.7:
Multilayer perceptron. Source: [51]

A convolutional neural network (CNN) employs the idea of weight sharing to reduce the computational complexity of the model. In image processing, it is common to use various filters (also called kernels) to emphasize certain structures in the images [52]. As an example, the Gaussian filter smoothens the images, and the Sobel filter is used to accentuate the edges. CNNs learn the filters that operate on the whole image (or its latent representations in the hidden layers), applying convolution as follows in equation 2.3 taken from [53] and modified:

$$I_{new}[i,j] = I * W = \sum_k \sum_l I[i-k, j-l]W[k,l] \tag{2.3}$$

where I is the input image, W is the filter, I_{new} is the output image after convolution, and i, j, k, l stand for the vertical and horizontal indices of the pixels in the images and in the convolution filter. The asterisk operator (*) denotes the convolution.

Note: For convenience, the convolution filters are often chosen with odd width and height and the middle pixel is assigned the coordinates [0, 0]. An example of such an application of the convolution operation can be seen in figure 2.8. The visualizations of convolution often simplify the representation of the filter by first flipping it both horizontally and vertically. Then, the result of the convolution is given by the sum of the element-wise multiplication of the filter values and the relevant patch of the image, as shown in figure 2.9.

On the edges of the image, the filter could extend behind the valid region of the image. This is handled by padding the area around the image (usually with 0 intensity or with some extrapolation of the image behind its edges: reflection, nearest neighbouring pixel's value, etc.) or by using only the part of the image where all the indices used in the convolution will be valid.

An image passed through a CNN is modified by the filters in subsequent layers, which creates intermediate latent representations of the input image. The shared weights (using the same filter for the whole image) ensure the efficiency of the model [54]. To allow more flexibility in the learning process, each convolutional layer can learn multiple (typically tens or hundreds) filters in parallel, similar to the MLP, which has multiple neurons in one layer.

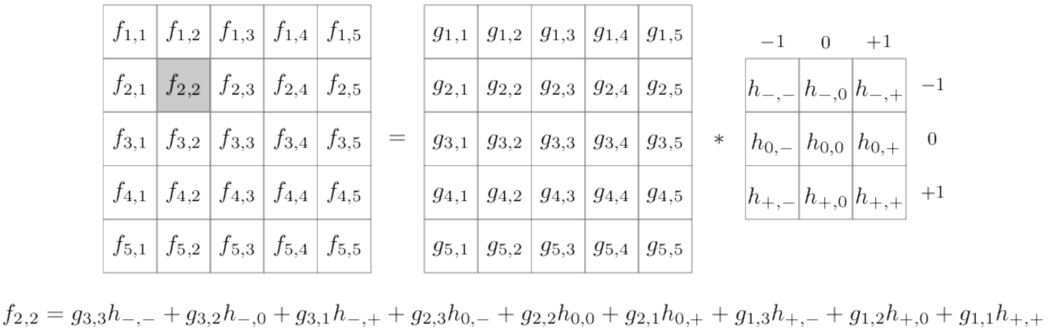
2.5.2 Building blocks

The typical architecture of a CNN for image classification consists of a convolutional part and a classification head. The convolutional part usually includes multiple convolution blocks composed of convolution layers followed by non-linear activations and finished by a pooling layer:

- **Convolution layers** learn a specified number of filters of specified size, can use stride to move the filter by more than one pixel each step when applied, and is usually followed by a non-linear activation function.
- **Activation functions** bring non-linearity to the intermediate functions applied to the image passed through the CNN. This can be thought of as thresholding the output of the operation. Commonly used activation functions include rectified linear unit (ReLU), leaky ReLU, sigmoid or softmax.
- **Pooling layers** reduce the spatial dimensions of the (latent) images by pooling multiple values together. The most common pooling layers are maximum pooling (shown in figure 2.10) or average pooling.

The convolutional part of the CNN is used to extract useful features from the input image. To classify the image, the outputs of the last convolution block are flattened and handled by the classification head. There are, again, typical layers involved in this process:

- **Flattening layer** reorganizes the outputs of the convolutional part to one long sequence.
- **Global pooling layer** can be used instead of the flattening one. It works in the same ways as a pooling layer but returns only one output for each channel of the latent image.
- **Dense layers** are layers with multiple computational units as in the MLP. It is usually followed by a non-linear activation function. The last dense layer has an output size appropriate for the classification problem. E.g., binary classification CNN can output a single output that can be thresholded to get the output class.



$$\begin{array}{|c|c|c|c|c|}\hline f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} & f_{1,5} \\ \hline f_{2,1} & f_{2,2} & f_{2,3} & f_{2,4} & f_{2,5} \\ \hline f_{3,1} & f_{3,2} & f_{3,3} & f_{3,4} & f_{3,5} \\ \hline f_{4,1} & f_{4,2} & f_{4,3} & f_{4,4} & f_{4,5} \\ \hline f_{5,1} & f_{5,2} & f_{5,3} & f_{5,4} & f_{5,5} \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|}\hline g_{1,1} & g_{1,2} & g_{1,3} & g_{1,4} & g_{1,5} \\ \hline g_{2,1} & g_{2,2} & g_{2,3} & g_{2,4} & g_{2,5} \\ \hline g_{3,1} & g_{3,2} & g_{3,3} & g_{3,4} & g_{3,5} \\ \hline g_{4,1} & g_{4,2} & g_{4,3} & g_{4,4} & g_{4,5} \\ \hline g_{5,1} & g_{5,2} & g_{5,3} & g_{5,4} & g_{5,5} \\ \hline \end{array} * \begin{array}{ccc|c} -1 & 0 & +1 & \\ \hline h_{-,-} & h_{-,0} & h_{-,+} & -1 \\ h_{0,-} & h_{0,0} & h_{0,+} & 0 \\ h_{+,-} & h_{+,0} & h_{+,+} & +1 \end{array}$$

$$f_{2,2} = g_{3,3}h_{-,-} + g_{3,2}h_{-,0} + g_{3,1}h_{-,+} + g_{2,3}h_{0,-} + g_{2,2}h_{0,0} + g_{2,1}h_{0,+} + g_{1,3}h_{+,-} + g_{1,2}h_{+,0} + g_{1,1}h_{+,+}$$

Figure 2.8:
Example of convolution in a 2D image. Source: [55]

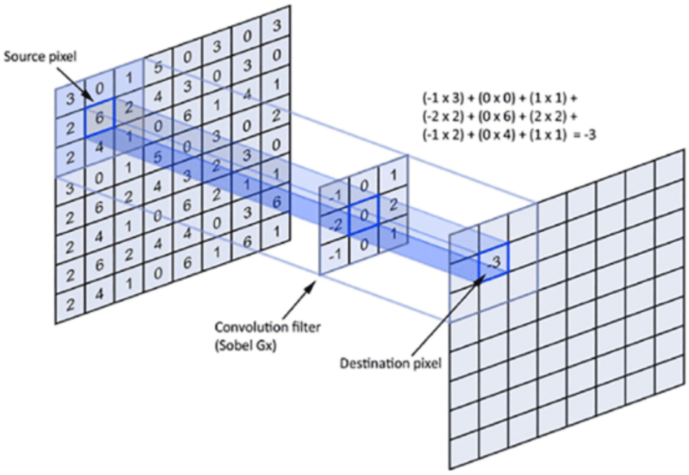
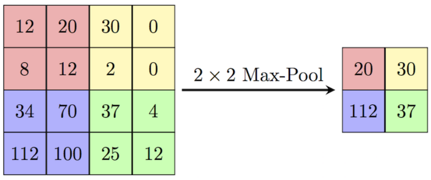

Source pixel
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(-1 x 2) + (0 x 4) + (1 x 1) = -3
Convolution filter
(Sobel Gx)
Destination pixel

Figure 2.9:
Convolution with flipped filter. Source: [56]



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Figure 2.10:
Maximum pooling. Source: [57]

The output can also list probabilities for all classes or use the one-hot encoding.

- **Dropout layers** randomly turn off a specified fraction of its inputs during the training phase. This allows the CNN to learn more reliable features because it adapts to possible missing inputs.

2.5.3 Training of convolutional neural networks

It was explained how the inputs can pass through a CNN. Passing the input through the network is called a forward pass and can be used to obtain a prediction from a trained model. However, the weights for the model are not known apriori and have to be learned in the training process. The weights are learned in the training phase from the observed data by the backpropagation algorithm.

Backpropagation updates the weights in the backward pass with respect to the error seen on the output after the forward pass during training. A loss function has to be specified according to the task performed by the model (e.g., binary classification, multiclass classification, image segmentation). Then, the gradient of the loss function can be computed on the output with respect to its inputs. It is used to update each input's parameters in order to minimize the loss function, as shown in 2.4.

$$\theta_{t+1} = \theta_t - l\frac{\partial L}{\partial \theta_t} \tag{2.4}$$

where θ_t represents the parameters at iteration point t, l denotes the learning rate (a multiplicative factor of the update term), and L stands for the loss function. The chain rule can be used to sequentially propagate the gradient to the previous layers and also update their weights [58].

The training happens sequentially in epochs. The training dataset is divided into two splits: training and validation. The inputs of the training split are grouped into batches. After the passing of one batch, the weights are updated. When all of the batches from the training split are exhausted, the validation split is used to estimate the accuracy of the intermediate state of the model without updating the weights. This step concludes one epoch [59]. The batch size affects the memory requirements and the speed of the training. It also affects the accuracy of the resulting model, but it is not clear in advance which batch size will be optimal for a given dataset and task. Usually, the training is executed multiple times with different values for batch size [60].

In the later epochs, the weights are obtained from the previous one. In the first epoch, they have to be initialized. This initialization can be random or it can use prior knowledge useful for the specific task. Symmetric initialization of the weights is undesirable because it would make all the units learn the same weights [58].

2.5.4 Overfitting

One of the common issues encountered when training the CNNs is overfitting: a situation when the model adheres to the training data and does not generalize well to new inputs. A typical sign of overfitting is good performance on the training dataset combined with poor performance on the test dataset.

There are multiple strategies to mitigate the chance of overfitting the model during training. The most popular methods to address overfitting include [61]:

- reducing the model complexity,
- using more training data or augmenting the available ones,
- applying a penalty to large weights (regularization),
- using dropout layers,
- early stopping the training.

The augmentation of an image dataset may be done by random transformations of the existing images, either in shape (zoom, rotation, flipping, shifts, perspective transformations) or intensity (brightness, contrast) [62]. The transformations need to be chosen appropriately so that they do not change the meaning of the image with respect to the task performed (e.g., counting objects and cropping one out).

2.5.5 Architectures for image classification

Image classification is a common task in computer vision. Therefore, there is a lot of research focused on this topic in deep learning, which brought great advancement, especially in the recent decade. As a result, there are many successful CNN architectures, useful for classification tasks, publicly available. Besides that, there are also datasets created for training such models (ImageNet, CIFAR-10 and more). The models can be trained on these datasets, and the weights obtained for the model can be stored and later re-used.

The performance of the CNNs is often measured with certain benchmarks. A lot of attention is also drawn to prestigious challenges, where models from various research groups compete. Results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) represent a possible benchmark for image classification. Some of the widely adapted ideas and most popular image classification CNN architectures were presented in this challenge [63]:

- In 2012, the challenge was won by Alex Krizhevsky et al. with the so-called AlexNet CNN [64]. Their solution benefited from a graphics processing unit (GPU) implementation, which made the training much faster. [65]

- In the 2014 edition, GoogLeNet and Visual Geometry Group (VGG) [66] were among the most successful entries [67]. The VGG's model became very popular. The VGG16 and VGG19 architectures belong to the most widespread models (being offered in many commonly used tools and frameworks such as PyTorch [68], TensorFlow [69], Keras [70] or MATLAB's Deep Learning toolbox [71]). The disadvantage of the VGG models is the large number of parameters used and consequently their large size.

- In 2015, ResNet [72] was introduced and achieved great success [73]. This architecture started using skip connections, which combine the outputs of some layers with their inputs by summation. This approach is efficient against the vanishing gradient problem. This problem occurs in very deep networks when the gradient is diminished in the backward pass during backpropagation, which makes the training less efficient. The skip connection increases the efficiency of training for models with a large number of layers. There are multiple varieties of the ResNet with 18 to 152 layers involved. A scheme of ResNet18 is shown in figure 2.11.

A comparison of the size and performance of selected popular CNN architectures is shown in figure 2.12.

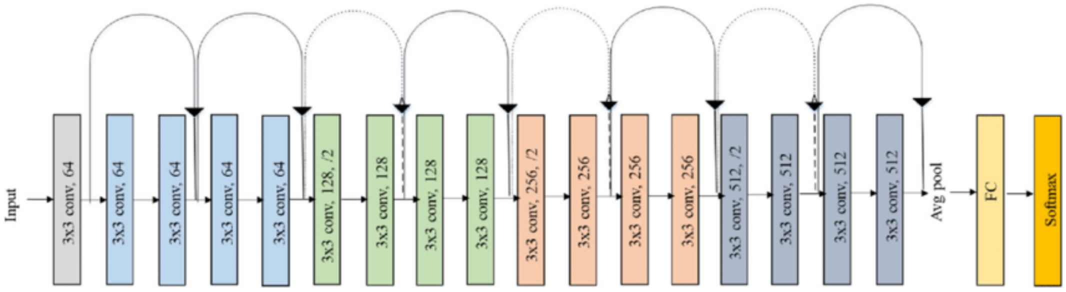


Figure 2.11:
ResNet18 architecture. Source: [74]

2.5.6 Transfer learning and fine-tuning

The practice of using a pre-trained model on a new classification task is called transfer learning. Alternatively, only the convolutional part of the CNN can be used for feature extraction, and a custom classification head can be appended. When using a pre-trained model, it can be adjusted to the new dataset by re-training it (keeping its weight for initialization) with a smaller learning rate. This approach is called fine-tuning. [76]

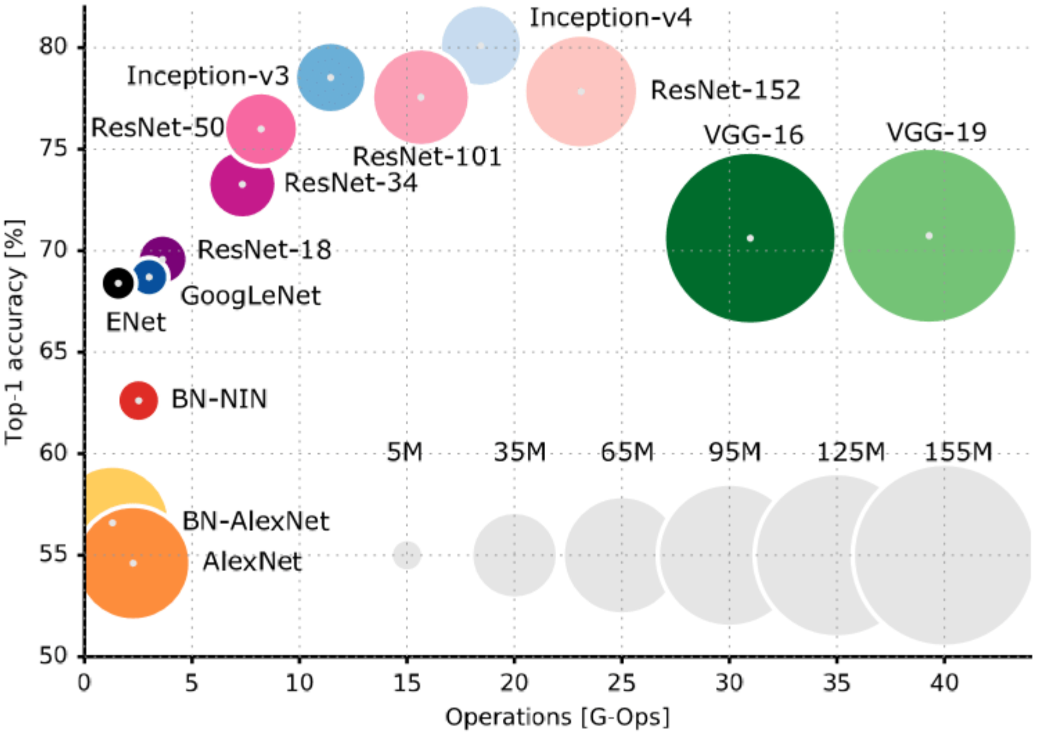


Figure 2.12:
Comparison of size and performance of popular CNN architectures. Source: [75]

Transfer learning usually helps to reduce the training times compared to training the models from scratch with random initialization. Usually, the most popular architectures are implemented in deep learning libraries and toolboxes, and their weights can be loaded, too.

2.6 Deep learning and pituitary adenoma disgnostics

In the area of PA surgery, deep learning has been applied to multiple tasks. Multiple deep-learning solutions for predicting the outcomes of surgical treatment were proposed [34][36][37][38]. The possibility of automatically segmenting the tumour in MRI was also researched.

He Wang, Wentai Zhang et al. describe a method that automatically segments the sellar region in MRI scans [77]. They also report feature extraction tools capable of predicting PA-related features, including Knosp score. These tools should be based on a deep learning model, but the authors do not specify details, as the focus of this publication is in PA segmentation. In a test dataset of 32 images with rather higher Knosp grades, they report an accuracy of 81.25%.

The thesis supervisors, MUDR. Martin Černý and prof. Dr. Ing. Jan Kybic, have published a study on another automated segmentation model in his previous research with a team of collaborators [78]. It uses a CNN-based model to automatically create segmentation masks from input MRI scans. The dataset used in this study is shared with the work presented in this thesis. The outputs of the automatic segmentation are used as a possible input dataset for the models developed in this work.

In research for available methods to automatically predict the Knosp grades, a publication by Yi Fang and He Wang was found [79]. They utilize transfer learning to classify PAs' invasiveness by the ResNet50 pre-trained model. However, while the Knosp grades are involved in the annotations, they are not the goal of the predictions. The model performs a binary classification of the PAs captured in MRI scans into the invasive and non-invasive groups.

The work of Staartjes, Serra et al. presents a deep learning model for predicting the gross total resection (defined as "the removal of all tumours" [80]) after trans-sphenoidal surgery directly, and they show that it can outperform predictors based on Knosp score and other classifications [81] in this application.

As of now, the author is not aware of any publicly available models capable of classifying Knosp grades of PAs from MRI scans or other image modalities.

Chapter 3

Data and methods

The goal of this work is to develop models for predicting Knosp scores from MRI scans. These models need to be adjusted to the formats and properties of the scans and their annotations. This chapter provides information about both the data and the developed models.

3.1 Image dataset

The dataset used for this work comes from the previous research (introduced in section 2.6). It consists of CE T1-weighted MRI brain scans annotated by subject matter experts. A total number of 493 patients' brain scans were included, with 394 of them coming from the years 2007-2018, collected retrospectively and used as a training dataset, and 99 acquired in the first half of 2022, serving as a test dataset. More detailed characteristics of the involved patients and their scans are listed in table 3.1 taken from [78]. The distribution of involved PA Knosp scores is in table 3.2. For each patient, the following files are available:

- **the brain scan**: a volumetric image centred on the sella turcica region,
- **a manually created segmentation mask**: a volumetric image of the same dimensions that assigns each voxel to one of the objects of interest (tumour, ICA) or to the background with a distinct label (number),
- **an automatically created segmentation mask**: a segmentation mask predicted by the model from the previous research,
- **annotations**: a file listing the Knosp grades for individual coronal slices of the brain scan, left and right side separately.

The volumetric images are stored in the Neuroimaging Informatics Technology Initiative (NIfTI) format.

Table 3.1: Characteristics of the available datasets. *Age* in years, *SD* = standard deviation, n = number, $\%$ = percentage.

	Training dataset	**Test dataset**
subjects, n (%)	394 (80)	99 (20)
age (mean ±SD)	53.2 ±15.6	54.2 ±14.8
male sex (%)	179 (45.4)	57 (57.6)
tumour type, n (%)		
non-functioning	242 (61.4)	59 (59.6)
GH-secreting	99 (25.1)	21 (21.2)
PRL-secreting	12 (3.0)	5 (5.1)
ACTH-secreting	39 (9.9)	14 (14.1)
plurihormonal	2 (0.5)	0 (0.0)

Table 3.2: Distribution of Knosp scores in datasets.

dataset	**side**	**grade 0**	**grade 1**	**grade 2**	**grade 3A**	**grade 3B**	**grade 4**
training	left	130	130	70	39	4	21
training	right	156	101	54	49	10	24
test	left	37	32	13	7	2	8
test	right	41	25	16	10	1	6

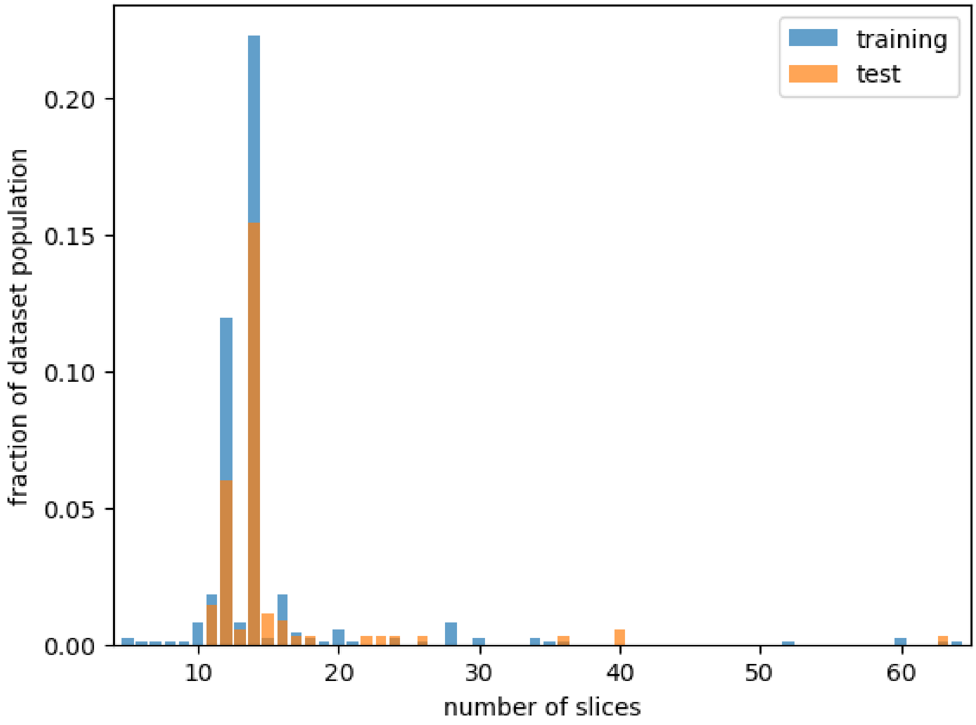


Figure 3.1:
Normalized histogram of input scans' number of slices in available datasets;
x axis (number of slices) limited to test dataset range.

As explained in section 2.4, the assignment of the Knosp grade to a tumour is based on the image view in the coronal plane. In the context of the methods used in this work, it means the following approach (if not directly stated otherwise): the scans were processed by extracting one slice in the coronal plane (a 2D grayscale image) at a time and then aggregating the results for each patient. To distinguish the whole 3D volumetric MRI image and its 2D slice in the coronal plane, the terms *scan* and *slice* will be used in the following text.

The individual brain scans were acquired by multiple devices, and therefore, also some of the scans' properties differ. Most notably, it is the spatial resolution which leads to the fact that there was a different number of slices in each scan. The distribution of the number of slices per scan is shown in figure 3.1. The spatial axes of the images will be called as *depth* for the axis orthogonal to the slices (i.e., going in the rostrocaudal direction), the terms *height* and *width* are used in the unchanged sense, referring to the vertical and horizontal axis in the slices.

3.2 Rule-based geometric model

With the segmentation masks available in the dataset, a rule-based model can be implemented to classify them, following the decision criteria specified in the Knosp classification system (2.4). Based on the geometric relationships of the objects in the segmentation mask, the appropriate scores can be devised. The following sections describe the procedure of classifying the segmentation masks with the geometric model.

3.2.1 Representation of individual objects

The segmentation masks contain dedicated labels for the PA and for the ICA cross-sections. The model uses these annotations to understand the relative positions of these objects in space to classify the score of the PA. This classification is first performed individually in each slice of the input 3D segmentation mask, and only later, the scores from all slices are aggregated to decide on the overall classification of the scan.

3.2.2 Distinguishing the arteries

In the segmentation mask (example in figure 3.2 b), we need to distinguish individual cross-sections of the ICAs. They form similarly sized, roughly circular groups of points, usually well divided in space, but they can touch each other in pairs sometimes. For this reason, pixels representing the arteries are clustered into four groups using the k-means algorithm. The clustering takes the coordinates of the pixels as its input. K-means++

initialization with 10 repetitions of the clustering procedure is used to mitigate the risk of ending up in a local minimum, not dividing the clusters (individual cross-sections) correctly.

The arteries are then assigned to the left and right intra- and supracavernous ICA based on their relative positions. For each detected cross-section, the centre of mass (COM) is computed. The COMs' coordinates are then compared, and the arteries' cross-sections are correspondingly assigned to the left and right sides and intra- and supracavernous ICA. An example of separated cross-sections of the ICA is in figure 3.2 c.

3.2.3 Determination of the critical lines

Knowing all the cross-sections of the arteries, the critical lines for the Knosp score evaluation can be determined. The following steps are performed both on the left and right sides:

1. The intercarotid line connects the COMs of the intra- and supracavernous ICA.
2. The tangents are obtained from the convex hull of both parts of the ICA.

Both cross-sections of the ICA on one side are encapsulated in a convex hull. The vertices of the convex hull form a polygon (the convex hull). On this polygon, we can check each pair of subsequent vertices. If they come from different cross-sections of the ICA (intra- versus supracavernous), the line segment between them lies on a tangent connecting the two cross-sections of the ICA. There are two such pairs of vertices, one defining the medial and one defining the lateral tangent. Which tangent is medial or lateral can be determined by the line's relative position to the COMs of the ICA.

A visualization of a convex hull of two ICA cross-sections is in figure 3.2 d, and the resulting lines found in that image are shown in figure 3.2 e.

3.2.4 Determination of the Knosp score

The next step is to identify the relative position of the adenoma to the selected lines. The score is computed separately for the left and right sides. First, every pixel's (represented by its coordinates) oriented distance from the critical lines is computed. The polarity of the computed distance determines if the pixel lies behind the line (further from the line than the centre of the PA) or not. A distance of up to 0.5 pixels behind the line is tolerated because the points lying up to 0.5 pixels far from the line are considered to be on the line, not behind.

This computation identifies definitively the points belonging to grades 0, 1 and 2. Points lying behind the lateral tangent are then further examined. If there is a hole in the

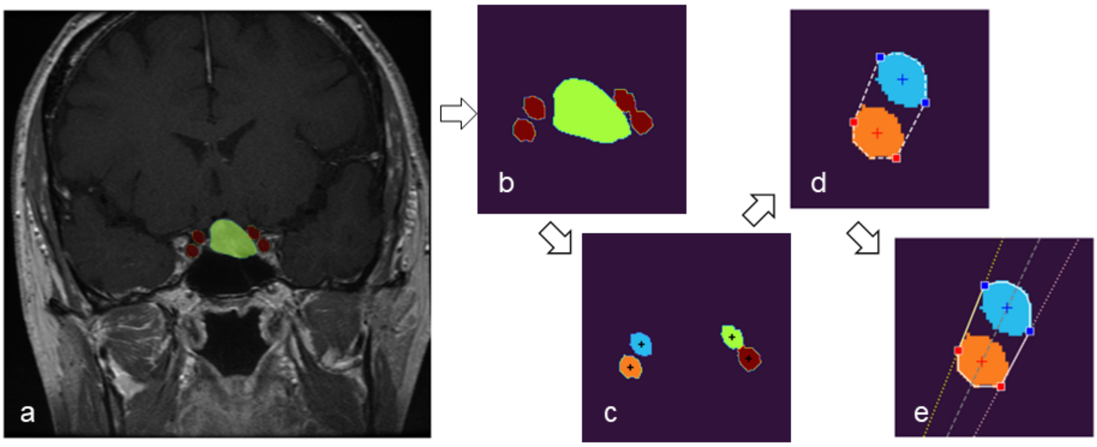


Figure 3.2:
Pipeline of the geometric model: determination of critical lines.
a) input slice, b) segmentation mask, c) separated ICAs, d) convex hull of ICAs on one side, e) critical lines determined.

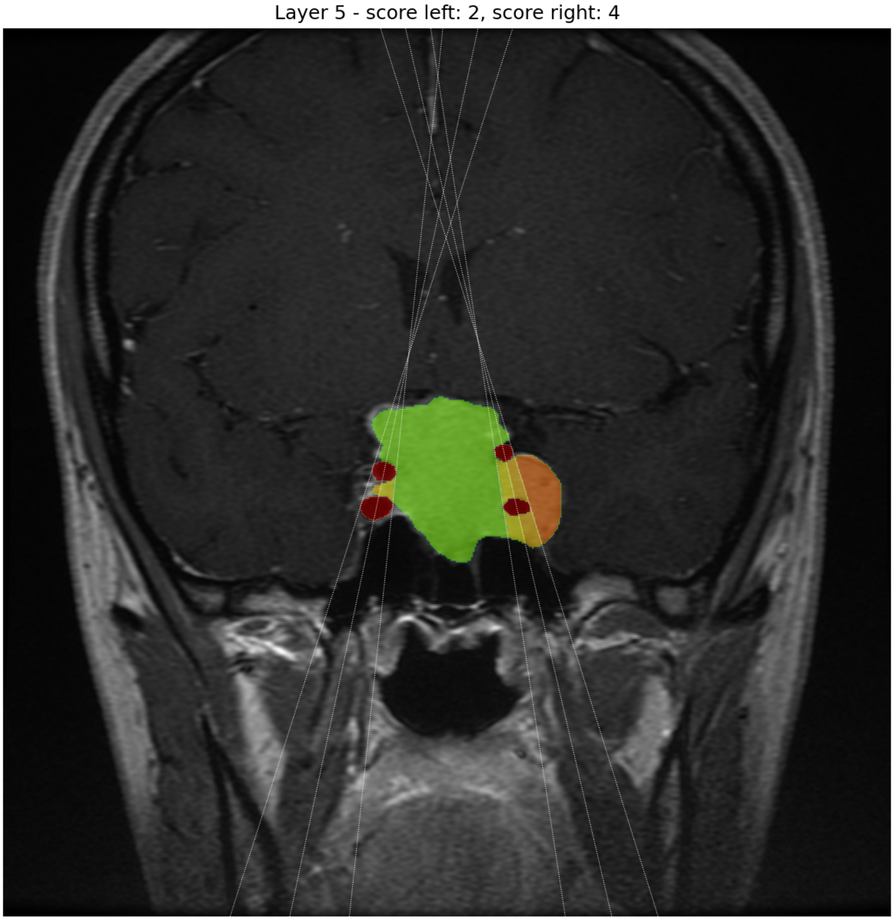


Figure 3.3:
Visualization of the output classification for a slice.

PA's body and one of the ICA's cross-sections lies in it, the PA is classified with grade 4 directly. If not, the subtypes of grade 3 are decided. If the grade 3 pixels connect with the rest of the PA above the intracavernous ICA, the score is 3A; if under, it is 3B. An example visualization of the classification is in figure 3.3.

The final score for the respective side is given by a pixel with the maximal grade on that side. The left and right scores are also stored for each slice. After processing all of them, the overall score for the patient is the maximal grade among all layers, independently for the left and right sides.

3.3 Deep learning models

In contrast with the geometry-based model, which relies on pre-set rules, the deep learning models gather and learn the decision criteria through observations of labelled data in the training phase, as explained in section 2.5.3. The available dataset for this task is relatively small (compared to general computer vision tasks), and data from individual patients differ in some parameters, which makes the development of the deep learning model challenging. To overcome these issues, the design of the proposed deep learning models starts with pre-processing the dataset to standardize the inputs. It also aims to utilise architectures that were successful in other image classification applications.

Indeed can the previously proven architectures help to reach better results, but they do not guarantee the performance of the resulting models. For this reason, the models were built in an iterative process, starting from a simpler solution and building up on top of it based on the intermediate results. The iterative enhancements led to multiple divergent strategies, out of which the most promising ones were further developed and used for the final training and evaluation.

3.3.1 Preprocessing of the dataset

The design of the methods, described in the following sections, requires a specific organization and representation of the input images, mostly in the form of slices. It is also more convenient to work with standard formats of the image files supported by commonly used image processing libraries than with the raw NIfTI format specific for medical imaging. In order to avoid preprocessing the scans repeatedly, modified versions of the input dataset were created, including:

- **Dataset of grayscale images:** Every 3D image (scans and also segmentation masks) was converted to a set of slices for each patient. The slices of scans were normalized to ensure good contrast in the images, while the segmentation masks' values were retained to keep the appropriate labels.

- **Dataset of 3-channel images:** Colourful images are commonly stored in a format that decomposes each pixel's colour to three channels: red, green and blue (RGB). Many image datasets used for training of CNNs comprise colourful images, and the architecture of such CNNs is designed to load 3-channel images on the input. The MRI images are grayscale, so the 3-channel representation of the slices was gained by also including the neighbouring slices (one from each side). This introduces more information from the spatial context, too.

- **Dataset of 3-channel slices of predicted probability segmentation objects:** The previously implemented and trained segmentation and slice selection models were used to create a 3-channel representation of each slice. The segmentation model predicts the segmentation mask as its output, but to do this, it predicts the probability of each object's label in each pixel in the previous layer of the model. The predicted probabilities of the tumour, the ICA cross-section and the normal pituitary gland in each pixel were extracted and saved as a 3-channel image. (The background probabilities were omitted.) The slices predicted as irrelevant were replaced with empty channels. In this way, another dataset was introduced, with each slice representing the predicted probability of the corresponding object in each channel.

- **Dataset of 3D images with standardized shape:** The scans did not have a unique depth, which introduces difficulties to the design of a suitable 3D deep learning model. A dataset with a standardized depth of the 3D images was created by interpolating the images to the common depth of 28 slices (taking into account the most common values of the number of slices – see in table 3.2).

These datasets store the 2D images in the Portable Network Graphics (PNG) format and the 3D images in Tagged Image File Format (TIFF). Because the scans were centred to the sella turcica region during the acquisition, all of the above datasets were also cropped in height and width to the area known to include all the important structures (194×194 pixels around the centre of each slice). Every image is also stored in "right" mode (original) and "left" mode, flipping the image horizontally. Thanks to this modification, each sample considers only one classification problem (now always on the right side of the image), which is easier for the model to learn.

3.3.2 Imbalanced dataset handling

With respect to the datasets' imbalance, the classes were weighted for the training, as mentioned above. Besides that, another modification of the training procedure used resampling of the classes either by sampling a specific number of images from each class, which serves to reduce the number of samples from the overrepresented classes or to oversample the other classes with fewer images than the required number. Eventually, the classes for grades 3A and 3B were merged before oversampling because grade 3B was the least represented one with approximately 3× fewer images than grade 4 (second smallest class, see 3.2). The following datasets were used as the input for the training:

- 1-channel grayscale slices,
- 3-channel modification of the slices,
- 3-channel slices of predicted segmentation probabilities,
- 1-channel mask slices,
- 3-channel modification of the mask slices.,

each used for training with the modifications described (remaining imbalanced, oversampling, merging 3A and 3B).

To avoid overfitting (especially when oversampling is used and the oversampled images can be seen multiple times), data augmentation was incorporated, too. It is desirable to use diverse images in the training phase so that the model can generalize more to unseen data. At the same time, it is important not to leave out any important parts of the images or to distort the images in such a way that they would lose the properties characteristic for the individual classes. To achieve this, the images were augmented using random rotation with the range of -15 to 15 degrees, random horizontal and vertical shifts up to 10% of the images' dimensions, and random zoom in/out up to 15%. The missing parts of the image after the transformation were filled with the nearest values from the edges of that image. Example images after data augmentation can be seen in figure 3.4.

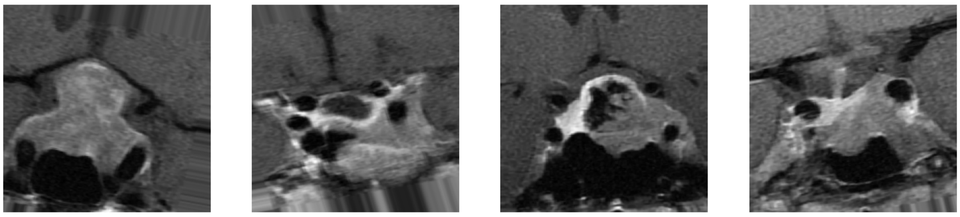


Figure 3.4:
Example of augmented images. Note that the images are slightly rotated (the vertical axis of the brain is not upright), and in the second image, the bottom part is filled with the repeated lower margin of the original image.

3.3.3 General concepts for the deep learning model

Contrary to the geometry-based model implementing the system of rules, the deep learning approach relies on inferring the patterns of correct classification from observations of the labelled training data. To offer a suitable solution to this problem, the decision for the model architecture was based on the exploration of the datasets' properties and interactive testing of potential solutions.

The most challenging aspects of the problem were given mainly by the specific nature of the datasets: The size of the dataset is rather small compared to typical deep learning applications. Moreover, the classes of the sliced datasets were heavily imbalanced. Even if the tumour in a particular scan is of grade 4, most of the slices will belong to lower grades or will not capture the tumour at all (assigned to grade 0 class, too). The higher grades were, therefore, underrepresented in the sliced datasets.

To address these challenges and avoid overfitting, the following measures were incorporated and tested:

- reducing the model size,
- using regularization techniques,
- weighing the classes and eventually resampling the dataset,
- initializing the model with pre-trained weights,
- augmenting the images,

3.3.4 Model architecture

In the effort to find a suitable model, multiple commonly used image classification models were interactively tested with the use of different modifications of the image dataset (including both 2D and 3D versions). Due to the limited size of the dataset, a rather small model was desirable: deeper models and the more complex ones with a large number of parameters tended to overfit quickly with fast improvement in accuracy on the split test of training data but no generalization on the validation split. Based on this experience, the ResNet18 architecture was selected as the base model because it offers good accuracy in image classification tasks with a lower number of parameters to be trained.

The convolutional blocks were taken, and the classification head was replaced with a custom one. On the input, an additional layer was used in the case of the grayscale dataset to convert the image to RGB format (replicating the grayscale channel). The resulting model was composed of the following parts:

1. an input lambda layer to convert the image from grayscale to RGB format (when appropriate),

2. layers of the ResNet18 network without the classification head,
3. a global average pooling layer to reduce the height and width dimensions from the convolution blocks,
4. a dense layer with 1000 units, an L2 kernel regularizer (to suppress overfitting) and ReLu activation function,
5. a dropout layer with 50% drop rate to suppress overfitting,
6. a dense layer with the number of units set to the number of classes in the dataset (i.e., typically 6, but also a variation of a dataset that merged grades 3A and 3B was tested) and softmax activation function.

Using the softmax activation function on the output, the model returns the predicted probability that a sample belongs to the corresponding classes. The class (grade) with the maximal probability was selected as the resulting prediction.

3.3.5 Training of the models

After the preparation of the datasets and classification model, the training could be performed. Still, there were many different variants of the datasets used for the training, corresponding modifications of the classification model and its settings. In the search for optimal configuration, the training of the models was automated and executed on a large scale in a parallelized fashion on a high performance computing (HPC) cluster, tracking the quality of the results by also automatically evaluating the accuracy on the test dataset. The jobs executed on the cluster differed in the image dataset used, batch size, the use of oversampling, and eventually merging grades 3A and 3B.

Following the initialization of the environment, the training dataset was loaded (with eventual modifications) and split into the training and validation parts. To avoid leaking from the training split to the validation split by using similar images, the split was performed by keeping all the images from one patient in the same split. Otherwise, similar slices (e.g., neighbouring ones with the same grades captured) could seemingly improve the perceived accuracy in the validation dataset being already observed in the training split. However, such behaviour would covertly overfit the training split, it would not generalize well, and the results on the test dataset could be much worse.

In the next step, a generator is created to load the images in batches and feed them to the model in the training phase. It includes sample normalization to set each sample's mean to 0 and standard deviation to 1. (This step is skipped in the datasets of masks because the image values in them do not represent intensity but labels.) The generator also applies the data augmentation on the training split (not on the validation split) and resizing of the images to match the expected input dimensions of the model (from 194×194 px to 224×224 px). The resizing is done so that the pre-trained weights from the ImageNet dataset can be used for the initialization of the model.

Then, the model is initialized using the selected architecture. The ImageNet weights are used for the initialization, and all the layers are set as trainable. The model is compiled with the Adam optimizer, an initial learning rate of 0.001 and categorical cross-entropy loss. Callbacks are prepared to save the model's state at the end of every epoch if the validation loss is improved, another one to reduce the learning rate by a factor of 10 if there is no improvement in the last 10 epochs and finally, an early stopping callback with the patience set to 20 epochs. The total number is set to 250 epochs, expecting that no training will take so long to converge and most of them will be finished much earlier. The training is then started, saving the history continuously.

In the case of the datasets with masks, where both the manually created and the predicted ones are available, the model is first trained on the manually created ones and then fine-tuned with the automatically predicted masks. The settings remain the same; only the initial learning rate for fine-tuning is set lower to 10^{-4}. With this approach, one trained model is saved after the first phase of the training and one after the fine-tuning. A report with information about the model used and the training's progression is automatically generated and stored for faster orientation in the results.

To evaluate the performance of every model, the training is followed by loading the testing dataset (in the same modification that was used for the training), and the classes of its images are predicted. The generator used for feeding the test images to the trained model uses neither data augmentation nor resampling.

3.4 Statistical evaluation

Statistical evaluation of the developed models is performed after the predictions are obtained. The evaluation methods are designed to compare the performance among different deep learning models and the rule-based geometric model, too. The following metrics were selected for the comparison:

- per-slice accuracy of the predictions,
- per-slice accuracy with the tolerance of ±1 grade,
- accuracy of predictions per patient,
- accuracy of predictions per patient with the tolerance of ±1 grade.

All of these metrics expect that the classification is independently done for the left and right sides, so there are two independent classification problems in each sample. The per-patient metrics are obtained after aggregating classification from all of the patient's slices. The resulting classification is the maximum grade encountered among them.

For the geometric model and the best deep learning model, a deeper analysis of the performance includes the following metrics and visualizations:

- sensitivity and specificity of the binarized classification of invasiveness (grades ≥ 3),
- Cohen's kappa score of interrater reliability,
- Spearman correlation coefficient of the predicted labels and ground truth,
- confusion matrix of the true and predicted labels.

Chapter 4

Results

This chapter summarises the performance of the developed models and provides more insights into the classification results obtained from them. The interpretation of the results is further discussed and explained with more context in chapter 5.

4.1 Functionality of the geometric model

The implemented program with the geometric model encapsulates the whole pipeline from the loading of the scans in the NIfTI format to generating predictions with supporting visualizations and annotation files. The user can control the amount of produced outputs. If the visualizations are saved, they show the segmentation masks coloured with respect to the tumour's predicted grade as an overlay on top of a grayscale slice of the scan. The borderlines are shown, too. The source codes and example outputs compose a part of the attachments.

4.2 Performance of the geometric model

The geometric model predicted the correct grade in 96.75% of the slices, and the accuracy increased to 99.10% with the tolerance of ±1 grade. Evaluating its accuracy for the aggregated scores in individual patients, the model predicted the correct grade in 158 out of 198 cases, which is 79.80%, and the prediction matched the ground truth ±1 grade in 95.45%.

The sensitivity of the geometric model on the binarized predictions of tumour invasiveness is 98.78%, while its specificity is 82.35%. Computing the interrater reliability between the geometric model and the ground truth labels, Cohen's kappa is 0.73, and the Spearman correlation coefficient is 0.86.

4.3 Training of the deep learning models

The jobs for models' training were performed on the HPC of the Research Center for Informatics of the Czech Technical University (RCI) with the search for the optimal prediction model. Table 4.1 shows the top 5 models predicting the scores from the manually created segmentation masks, and table 4.2 selects the best model for every dataset. The criterion used for the ranking of the models was the accuracy of predictions per patient; the secondary criterion in case of a match was the accuracy per patient with the ±1-grade tolerance.

The best model's training was finished after 52 epochs, which took only 6 minutes 13 seconds on a GPU computation node of the RCI cluster with NVIDIA Tesla V100 GPU. It was also the model that used the most epochs before finishing. The slowest training lasted 36 minutes and 45 seconds, and the slowest fine-tuning was finished in 9 minutes and 11 seconds.

The trained models, stored in the Hierarchical Data Format (HDF) format, including the model architecture and trained weights, are too large to be directly attached to the thesis but can be accessed from the author's online storage: [82]. The scripts for training, as well as loading the trained models and making a prediction are attached to the thesis.

4.4 Performance of the best deep learning model

The best deep learning model was the one taking 1-channel manually created segmentation masks as its input, did not used oversampling or merging of the grades 3A and 3B, and was trained with the batch size of 32 samples.

It predicted the correct grade in 95.22% of the slices, and the accuracy increased to 99.10% with the tolerance of ±1 grade. In the case of the scores aggregated for individual patients, the exact match in the ground truth and predicted grade was observed in 146 out of 198 cases, which is 73.74%, and with the relaxation of ±1-grade mismatch tolerated, the prediction matched the ground truth label in 95.45% cases.

The sensitivity of this deep learning model is 96.34%, and its specificity is 76.47%. The scores for interrater reliability between the deep learning model and the ground truth labels are 0.64 (Cohen's kappa) respectively 0.84 (Spearman correlation coefficient).

The comparison of the geometric and deep learning models is summarized in table 4.3. The confusion matrices of both models are depicted in figure 4.1.

Table 4.1: Top 5 deep learning models. *DM* = dataset modification, *ch* = number of channels of the used images, *BS* = batch size, *epoch* = epoch of the model with the smallest loss, *D* = duration of its training on the HPCcluster, *APP* = accuracy per patients, *APP±1* = *APP* with ±1 grade tolerance, *APS* = accuracy per slices, *APS±1* = *APS* with ±1 grade tolerance; accuracies shown in percent. The rows are sorted by *APP* in descending order.

DM	**ch**	**BS**	**epoch**	**D (mm:ss)**	**APP**	**APP±1**	**APS**	**APS±1**
none	1	32	52	06:13	73.74	95.45	95.22	99.10
none	3	32	29	15:46	68.18	90,40	93.14	98.93
oversampling	1	16	23	06:51	67.68	94.95	94.52	99.07
merging 3A+3B	1	32	26	05:43	66.16	94.44	93.66	99.04
oversampling	3	64	18	31:50	63.13	92.93	92.47	98.66

Table 4.2: The best performing deep learning model for every dataset. *DS* = dataset used, *ch* = number of image channels in the dataset, *BS* = batch size, *epoch* = epoch of the model with the smallest loss, *D* = duration of its training on the HPCcluster, *APP* = accuracy per patients, *APP±1* = *APP* with ±1 grade tolerance, *APS* = accuracy per slices, *APS±1* = *APS* with ±1 grade tolerance; accuracies shown in percent. The rows are sorted by *APP* in descending order. All of these models used the imbalanced datasets (no merging, no oversampling).

DS	**ch**	**BS**	**epoch**	**D (mm:ss)**	**APP**	**APP±1**	**APS**	**APS±1**
mask	1	32	52	06:13	73.74	95.45	95.22	99.10
mask	3	32	29	15:46	68.18	90.40	93.14	98.93
auto. mask	1	32	18	02:01	55.67	86.08	69.79	88.64
auto mask.	3	16	26	09:10	53.61	83.51	71.78	90.16
seg. prob.	3	32	42	18:47	46.39	82.47	66.98	88.64
raw	1	64	31	08:06	45.45	80.81	86.82	95.68

Table 4.3: Metrics compared for the geometric and the best deep learning model. *APP* = accuracy per patients, *APP±1* = *APP* with ±1 grade tolerance, *APS* = accuracy per slices, *APS±1* = *APS* with ±1 grade tolerance, *Spearman* = Spearman correlation coefficient, *Cohen* = Cohen's kappa; accuracies shown in percent; *Spearman* and *Cohen* measure the agreement with ground truth annotations.

model	**APP**	**APP±1**	**APS**	**APS±1**	**Spearman**	**Cohen**
geometric	79.80	95.45	96.75	99.10	0.86	0.73
deep learning	73.74	95.45	95.22	99.10	0.84	0.64

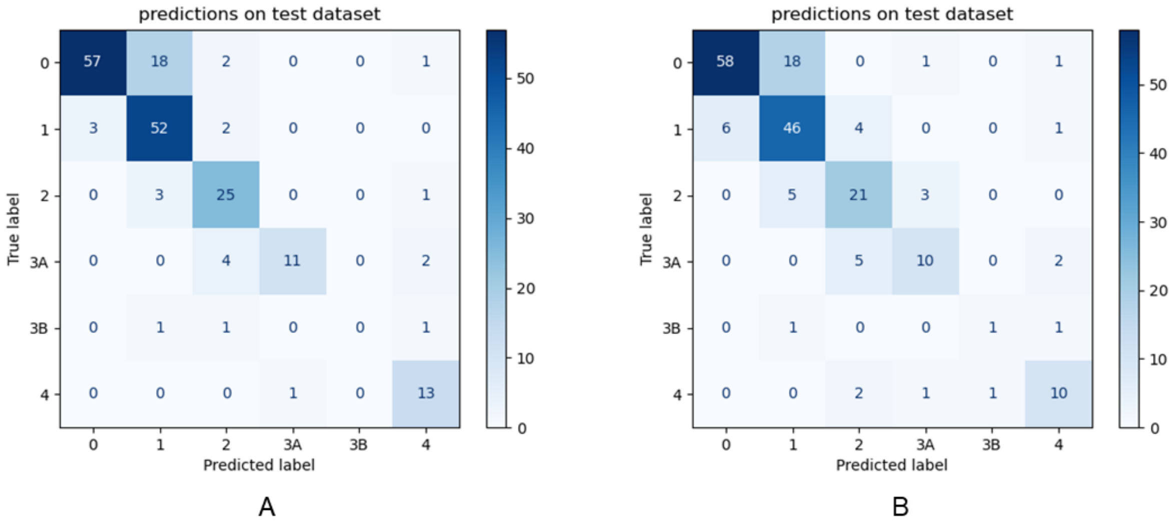


Figure 4.1:
Confusion matrices for the classification provided by the geometric model (A) and the deep learning model (B).

Chapter 5

Discussion

Models based on the implementation of geometric relationships or on learning from annotated observations and inference were introduced in chapter 3, and their performance was evaluated and stated in 4. This chapter serves to interpret the results and verify their validity, explain the inaccurate classifications, and identify the issues and open problems.

5.1 Geometry-based model

Based on the proposed metrics, the geometric model shows the best performance, classifying 79.80% of the patients' tumours correctly. The agreement with the ground truth labels (Spearman correlation coefficient: 0.86) is better than the inter-rater reliability observed in the aforementioned study (0.73) [43].

There is a remarkable difference between the percentage of correctly classified slices (96.75%) and aggregated scores for patients from the whole scans (79.80%). This is given by the fact that the grade of the tumour is evaluated as the maximal observed grade on the respective side among all slices. This means that one slice wrongly classified by a higher grade corrupts the classification of the scan as a whole, which can contain up to tens of slices. This difficulty is hard to overcome because, contrary to such a misclassification, there are also cases where the correct grade is only observed in one slice, and all the other slices capture a less invasive part of the tumour. For this reason, the selection of the maximal grade cannot be simply replaced by, for example, the average grade and is correct with respect to the definition of the classification system (section 2.4).

To confirm the classification capability of the model and explain the misclassification, a sample of wrongly classified slices was examined, with an emphasis on the most different grades and also the nearest ones. It seems that the nearest grades are usually classified differently from the ground truth because the tumour touches the borderline between two grades, as shown in figure 5.1. Based on the confusion matrix (figure 4.1 A), the most common case of misclassification is that a tumour with a ground truth label 0 is assigned to the grade 1 class. The most concerning, however, are the cases where the predicted label differs from the ground truth by multiple grades. In this case, it seems that the disagreement might be caused by an incorrect annotation. Examples of such cases are displayed in figures 5.2 and 5.3. In the large amount of slices that had to be annotated, some imprecision caused by a human factor could be expected.

An advantage of this model is that it provides, alongside the predictions, also a graphic visualization of the classification problem. This allows the user to inspect the underlying materials quickly and supports the decision if the prediction is trustworthy. Following the exploration of the edge cases leading to misclassification of neighbouring grades, an additional parameter of extent threshold was introduced to the program, too. It enables the user to adjust the condition on how far the tumour has to extend behind the critical line to be assigned to the higher grade. It can be used to relax the strictness and let more of the borderline tumours be classified to a lower grade.

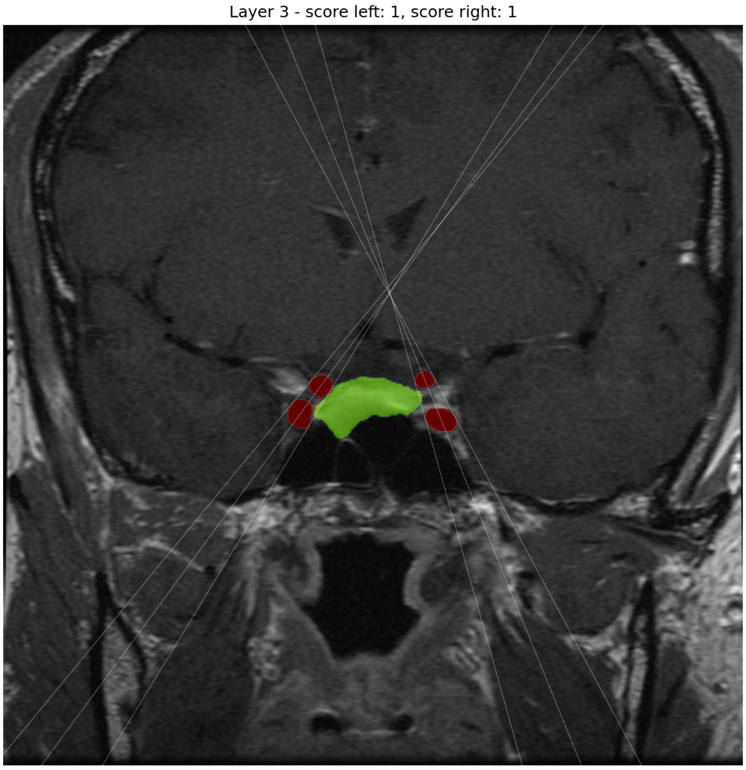


Figure 5.1:
PA with a ground truth grade 0 on the right side, classified as a grade 1.
The model detected pixels behind the critical line.

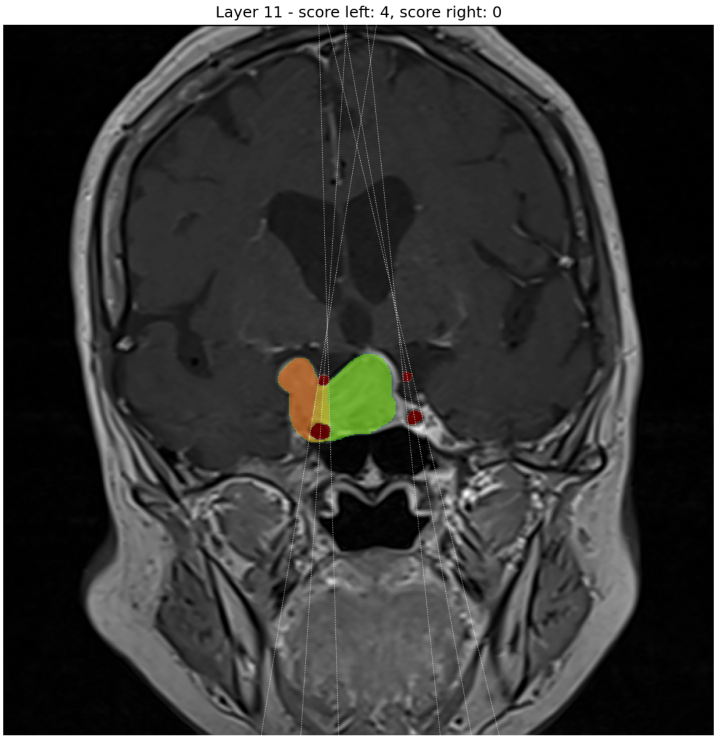


Figure 5.2:
PA with a ground truth grade 3A on the left side, classified as a grade 4.
The model detected a full encapsulation of the ICA.

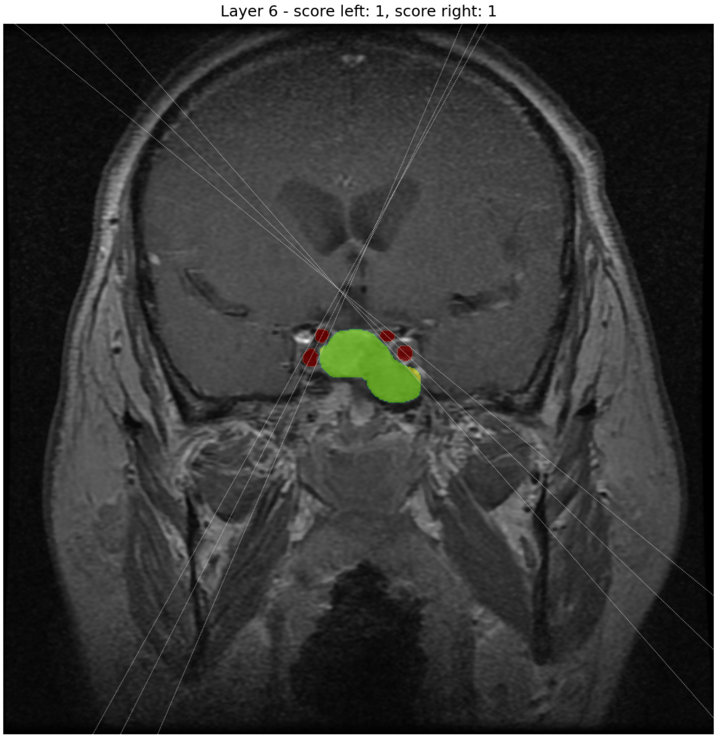


Figure 5.3:
PA with a ground truth grade 4 on the right side, classified as a grade 1.
This might be an example of a wrong label in the ground truth annotation.

5.2 Deep learning models

The deep learning approach led to a varied set of trained models enabling predictions from different modifications of the input dataset, including both manually created masks and raw scans (eventually converted to segmentation masks automatically by the available segmentation model).

The results do not reveal a predominantly more successful strategy for the training: the top-ranked models were trained with different batch size values; they also involved the imbalanced datasets alongside the oversampled ones and those with merged grades 3A and 3B.

However, the selection of the dataset modality proved to be important: the models using the manually created segmentation masks achieved results comparable to the geometric model (which uses the same modality): the best one was accurate in 73.74% of per-patient grade predictions respectively in 95.22% per-slice predictions. Also in this case, the agreement with the ground truth exhibits a better Spearman correlation coefficient (0.84) than in the mentioned study [43].

The models inferring from (either raw or automatically segmented) scans remain less successful. The most accurate one out of these predicted the correct grade in 55.67% of patient-wide classification and 69.79% of per-slice classification. This model used the segmentation masks automatically generated from the scans by the segmentation model, i.e., using only input scans and trained models.

5.2.1 Other examined approaches

Other approaches examined in this work involve an effort to predict the grades from the 3D images (described in section 3.1). This approach was particularly challenging for multiple reasons: The 3D classification is far less common than its 2D counterpart, so there are much fewer established and proven architectures, as well as datasets that would enable pre-training such models on a large scale. Likewise, the base of appropriate tools and techniques in general is less developed. The attempts to build and train a 3D CNN to classify the scans of standardized shape also demonstrated that such an approach imposes considerably higher demands on memory. The 3D nature of the images only allowed a small batch size to fit into the available random-access memory (RAM) and did not enable good enough learning. Moreover, the size of a trained model grows faster with the size of the CNN. Finally, the size of the dataset was even more limited (compared to extracting slices from the scans), which does not add to expectations of good results. For these reasons, the effort to develop a model suitable for the 3D images was discontinued.

Referring to the 2D models developed to predict the grades from raw input scans, the geometric model was also adjusted to accept them from the preprocessed datasets (section 3.1), including the masks automatically predicted from the MRI scans. This approach had a better per-slice accuracy of 88.48% (for the deep learning model predicting from the automatically generated masks, it was 69.79%) but only 38.89% accuracy on the per-patient aggregations (55.67% for the deep learning model). This shows that the deep learning model is more suitable for use with the automatically generated masks because it learns from the observations and is not bound by the rules. The geometric model assumes some characteristics that can be simply violated in these predicted masks. For example, the number and the relative positions of the ICAs are important for the geometric model, but it is a common case that there is a wrong number of ICAs present in a slice of a predicted mask or that they are inconsistent in shape. Examples of classifications for such slices are shown in figures 5.4 and 5.5.

5.3 Value proposition

The value of the developed methods and models is currently mainly in the research area, with this work being a part of ongoing research of automatic assessment of PAs, intending to publish the findings. The code will also be shared as an open-source project. The use in the clinical field is limited by multiple factors: First, data sharing in clinical use is restricted and specific to each hospital's infrastructure. The handling of patients' data is also determined by the devices used, which makes general use difficult. Finally, there are more field-specific file formats, and the codebase would need further adjustments to become compatible and fulfil relevant requirements.

In a broader sense, leaving out the emphasis on specific requirements imposed by the environment of hospitals and their information systems, the proposed solutions offer a great speed-up of the scans' evaluation. While the current diagnosis requires a trained radiologist, who needs to analyze the whole volume of the scan, one patient at a time, the models can process images from tens or hundreds of patients in seconds or minutes (depending on the computational power and if visualizations are generated).

This can also make such analysis accessible to interested non-expert users (e.g., patients) and provide a visual explanation with the plots generated by the geometric model. The models can also help standardize the form of the analysis outcomes and eliminate human-induced errors.

5.4 Comparison with other works

The results of the geometric model and the best deep learning model are comparable to the accuracy of the model mentioned in a publication by He Wang, Wentai Zwang et al. (and listed in section 2.6), which was 81.25%. The advantage of their model is that it is capable of reliably predicting the Knosp grades from the automatically created segmentation masks. However, that model lacks any closer description and is not publicly available. Also, their test dataset only involved 32 subjects, with only one of them belonging to grade 0 and two of them to grades 1 and 2. That makes their dataset's population skewed towards the higher grades, contrary to the dataset in this work. These factors make it difficult to compare the performance of the models in a general case of automatic Knosp grade classification.

5.5 Current issues and open problems

Based on the observations in the results, the space for improvement could lie in the curation of the dataset. As discussed, some of the slices might have been annotated imperfectly. The correction of such annotations could reveal a better accuracy of the geometric model, and the use of refined labels in training could also lead to improvements in the resulting deep learning models. The problem is that any such curation needs expert intervention and is time-consuming. The same holds for the eventual extension of the current dataset, which would be beneficial to the deep learning models. Compared to other classification tasks in general cases, which can use commonly available images, medical applications require specialized imaging devices and specific health conditions of the patients. This makes the medical image datasets usually rather limited.

The problem of image labelling could be partially overcome by training the deep learning models with labels provided by the geometric model, which shows good consistency. Similarly, in the joint effort of the rule-based methods and learning from observation, a more complex classifier could be designed on top of combined outputs.

In the context of the research conducted alongside this work, the segmentation model had been developed earlier and focused on the generation of segmentation masks. Future work could include extracting more information from the predictions of this model, extending the idea of the 3-channels dataset based on predicted probabilities in the segmented objects.

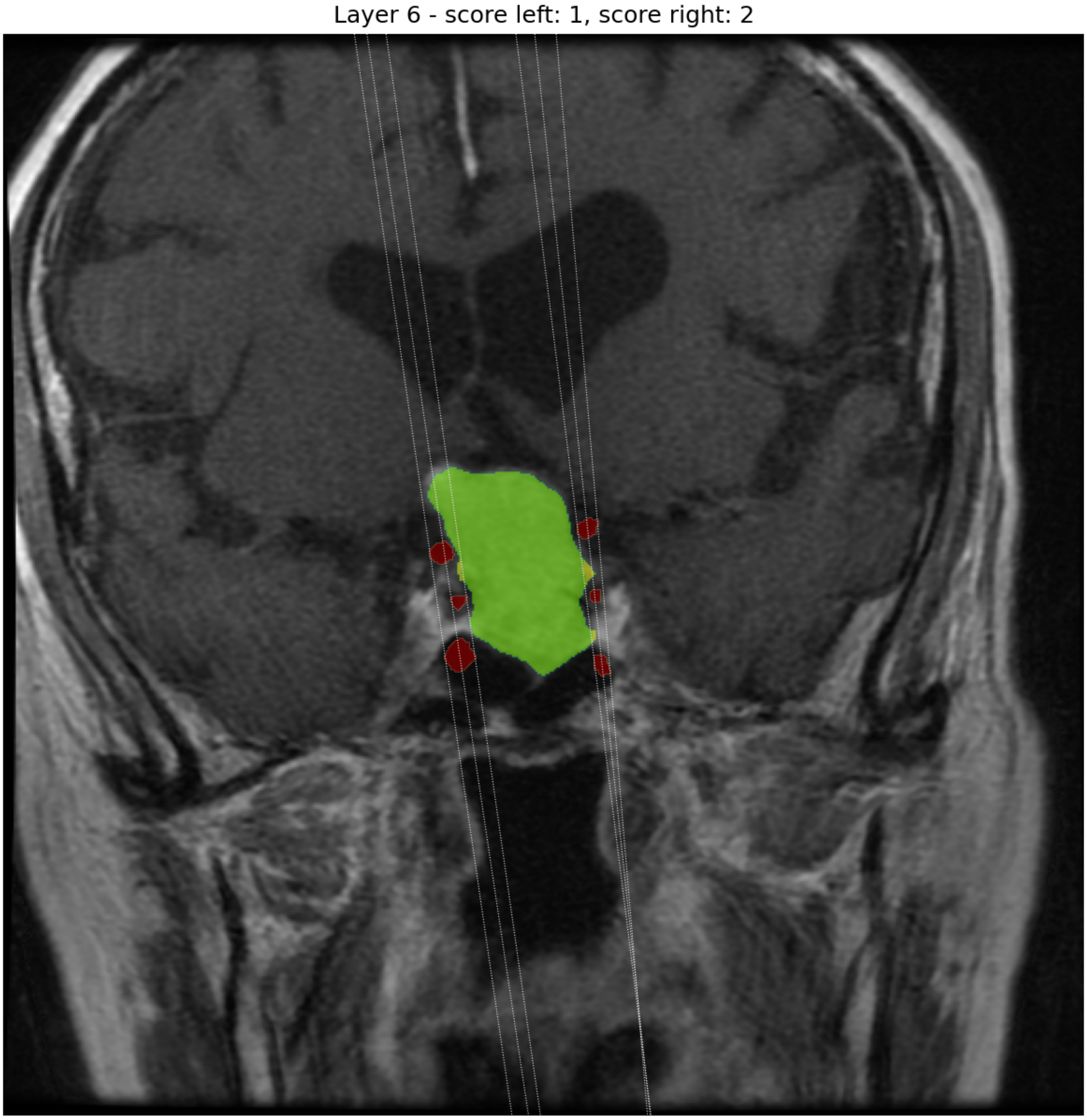


Figure 5.4:
Classification based on a predicted mask: In this case, there is a wrong number of ICA cross-sections, but it does not affect the quality of the classification thanks to the placement of the incorrectly segmented cross-sections.

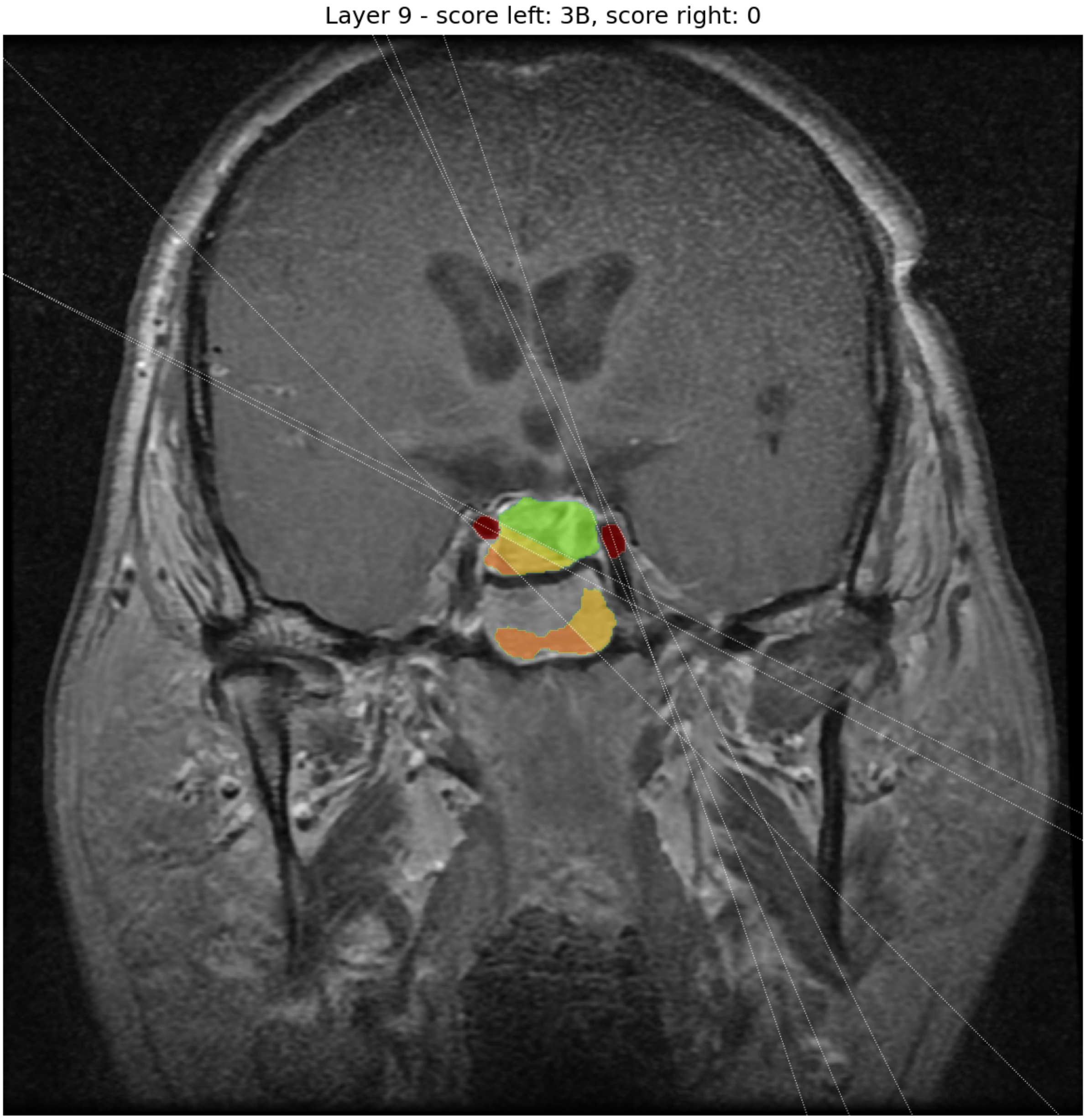


Figure 5.5:
Classification based on a predicted mask: This time, the classification fails because the assumptions of correctly segmented ICAs are not fulfilled.

Chapter 6

Conclusions

The geometric and deep learning models have been designed and implemented. The geometric model performs the best out of the proposed solutions, classifying 79.80% of whole PAs correctly (or 95.45% within the tolerance of ±1 grade) with respect to expert annotation. The best deep learning model offers an accuracy of 73.74% (95.45%). Both of these models operate on manually created segmentation masks, and both show a good agreement with the expert annotator with a Spearman correlation coefficient of 0.86, respectively 0.84. The best deep learning model capable of predicting the scores from automatically created segmentation masks was correct in 55.67 (69.79%) of cases.

Considering that there are no publicly available tools for automatic Knosp score predictions, the presented solutions can serve as proof of concept for this task. Besides the classification capability, the geometric model also offers the option to visualize the results of the classification, which demonstrates its utility as a supportive diagnostic tool, so its use is not limited to predictive tasks in research. Automated tools in medicine can provide aid with diagnostics and bring considerable time savings.

While the predictions based on the manually created masks are reliable, the models capable of predicting the scores from automatic segmentations or raw MRI brain scans will need further improvement to achieve similar accuracy. The proposed approach to reach this goal includes revising the training dataset and eventually extending it, although it is acknowledged that some of the necessary steps are not easily feasible. The data availability in the medical domain is generally more limited compared to some other areas. Other possible future steps were listed, too.

Appendix A

Attachments provided

The thesis was submitted together with the code implemented as a solution to the assignment and other files. The attached archive contains the following files and directories:

- `geom` - source code of the geometric model
    - `src` - modules
    - `main.py` - main script
    - `README.md` - instructions
    - `requirements.txt` - list of required libraries
    - `statistics.py` - postprocessing of the results
- `nn` - source codes for the deep learning approach
    - `preprocessing` - scripts for preprocessing of the dataset
    - `training` - scripts for automated training and evaluation
- `output` - example output data
    - `geom` - from the geometric model
    - `nn` - from the best deep learning model

The input dataset and trained models are too large for the attachment but can be requested from the author. Some of the best-performing models are stored in the HDF format on the author's cloud storage [82] and can be freely downloaded.

Appendix B

Technologies used

All of the presented code was developed in `Python` programming language with the use of various libraries for image processing (`scikit-image`, `SciPy`), mathematics (`numpy`), statistics (`scikit-learn`, `pandas`) and other disciplines. Handling of NIfTI files was provided by the `NiBabel` library. The deep learning models were built in `TensorFlow` and `Keras` frameworks with some pre-trained models imported from the `image-classifiers` library. The visualizations were prepared with `Matplotlib`. Other libraries were used for auxiliary tasks (`pathlib`, `tqdm`, Python standard libraries).

The geometric model was developed locally on a personal computer without any special characteristics. The deep learning models were first iteratively developed and tested in the Google Colaboratory environment with Tesla T4 tensor processing units (TPU). Later, the models were trained on a larger scale on the RCI clusters' GPU computation nodes with Tesla V100 GPU.

Other software used includes programs ITK-SNAP and 3D Slicer for the visualization of medical images.

Bibliography

[1] S. R. Emmett, N. Hill, and F. Dajas-Bailador, *Endocrinology*, Oct. 2019. DOI: 10.1093/oso/9780199694938.003.0012. [Online]. Available: http://dx.doi.org/10.1093/oso/9780199694938.003.0012.

[2] *Overview of the pituitary gland*, 2023. [Online]. Available: https://www.msdmanuals.com/home/hormonal-and-metabolic-disorders/pituitary-gland-disorders/overview-of-the-pituitary-gland.

[5] *Pituitary tumors treatment*, 2022. [Online]. Available: https://www.cancer.gov/types/pituitary/hp/pituitary-treatment-pdq.

[6] *Pituitary adenomas*, 2022. [Online]. Available: https://my.clevelandclinic.org/health/diseases/15328-pituitary-adenomas.

[7] T. T. Agustsson, T. Baldvinsdottir, J. G. Jonasson, *et al.*, *The epidemiology of pituitary adenomas in iceland, 1955–2012: A nationwide population-based study*, Nov. 2015. DOI: 10.1530/eje-15-0189. [Online]. Available: http://dx.doi.org/10.1530/eje-15-0189.

[8] A. F. Daly and A. Beckers, *The epidemiology of pituitary adenomas*, en, Sep. 2020. DOI: 10.1016/j.ecl.2020.04.002. [Online]. Available: http://dx.doi.org/10.1016/j.ecl.2020.04.002.

[9] M. E. Molitch, *Diagnosis and treatment of pituitary adenomas*, en, Feb. 2017. DOI: 10.1001/jama.2016.19699. [Online]. Available: http://dx.doi.org/10.1001/jama.2016.19699.

[10] N. Karavitaki, *Prevalence and incidence of pituitary adenomas*, en, Apr. 2012. DOI: 10.1016/j.ando.2012.03.039. [Online]. Available: http://dx.doi.org/10.1016/j.ando.2012.03.039.

[11] M. E. Molitch, *Nonfunctioning pituitary tumors*, 2014. DOI: 10.1016/b978-0-444-59602-4.00012-5. [Online]. Available: http://dx.doi.org/10.1016/B978-0-444-59602-4.00012-5.

[12] *Pituitary adenoma*, 2023. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pituitary-adenoma.

[13] S. Melmed, U. B. Kaiser, M. B. Lopes, *et al.*, *Clinical biology of the pituitary adenoma*, en, 2022. DOI: 10.1210/endrev/bnac010. [Online]. Available: http://dx.doi.org/10.1210/endrev/bnac010.

[14] K. Kovacs, E. Horvath, and S. Vidal, *Classification of pituitary adenomas*, 2001. DOI: 10.1023/a:1012945129981. [Online]. Available: http://dx.doi.org/10.1023/A:1012945129981.

[15] E. V. Varlamov, S. McCartney, and M. Fleseriu, *Functioning pituitary adenomas – current treatment options and emerging medical therapies*, en, 2019. DOI: 10.17925/ee.2019.15.1.30. [Online]. Available: http://dx.doi.org/10.17925/EE.2019.15.1.30.

[16] M. B. S. Lopes, *The 2017 world health organization classification of tumors of the pituitary gland: A summary*, en, Aug. 2017. DOI: 10.1007/s00401-017-1769-8. [Online]. Available: http://dx.doi.org/10.1007/s00401-017-1769-8.

[17] S. L. Asa, O. Mete, A. Perry, and R. Y. Osamura, *Overview of the 2022 who classification of pituitary tumors*, en, Mar. 2022. DOI: 10.1007/s12022-022-09703-7. [Online]. Available: http://dx.doi.org/10.1007/s12022-022-09703-7.

[18] H. Nishioka, *Aggressive pituitary tumors (pitnets)*, en, 2023. DOI: 10.1507/endocrj.ej23-0007. [Online]. Available: http://dx.doi.org/10.1507/endocrj.ej23-0007.

[19] *Pituitary adenoma*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK554451/.

[20] L. Guignat, G. Assie, X. Bertagna, and J. Bertherat, *Adénome corticotrope*, fr, Jan. 2009. DOI: 10.1016/j.lpm.2008.10.008. [Online]. Available: http://dx.doi.org/10.1016/j.lpm.2008.10.008.

[21] *Adenomy hypofýzy*, 2023. [Online]. Available: https://www.homolka.cz/nase-oddeleni/11635-neuroprogram/11635-neurochirurgie-nch/11751-nase-sluzby/11752-onkoneurochirurgie/adenomy-hypofyzy.

[23] *Pituitary tumors*, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pituitary-tumors/diagnosis-treatment/drc-20350553.

[24] T. Tsukamoto and Y. Miki, *Imaging of pituitary tumors: An update with the 5th who classifications—part 1. pituitary neuroendocrine tumor (pitnet)/pituitary adenoma*, en, Feb. 2023. DOI: 10.1007/s11604-023-01400-7. [Online]. Available: http://dx.doi.org/10.1007/s11604-023-01400-7.

[25] A. D. Elster, *Imaging of the sella: Anatomy and pathology*, en, Jun. 1993. DOI: 10.1016/s0887-2171(05)80079-4. [Online]. Available: http://dx.doi.org/10.1016/S0887-2171(05)80079-4.

[26] *Pituitary tumors*, 2023. [Online]. Available: https://www.msdmanuals.com/professional/neurologic-disorders/intracranial-and-spinal-tumors/pituitary-tumors?query=pituitary.

[27] R.-J. M. van Geuns, P. A. Wielopolski, H. G. de Bruin, *et al.*, *Basic principles of magnetic resonance imaging*, en, Sep. 1999. DOI: 10.1016/s0033-0620(99)70014-9. [Online]. Available: http://dx.doi.org/10.1016/S0033-0620(99)70014-9.

[29] *Gadolinium contrast agents*, 2023. [Online]. Available: https://radiopaedia.org/articles/gadolinium-contrast-agents?lang=us.

[30] *Mri sequences*, 2022. [Online]. Available: https://radiopaedia.org/articles/mri-sequences-overview.

[31] J. Ježková, V. Hána, M. Kršek, *et al.*, *Use of the leksell gamma knife in the treatment of prolactinoma patients*, en, Mar. 2009. DOI: 10.1111/j.1365-2265.2008.03384.x. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2265.2008.03384.x.

[32] T. Chowdhury, H. Prabhakar, P. Bithal, B. Schaller, and H. Dash, *Immediate post-operative complications in transsphenoidal pituitary surgery: A prospective study*, en, 2014. DOI: 10.4103/1658-354x.136424. [Online]. Available: http://dx.doi.org/10.4103/1658-354X.136424.

[33] A. T. Heffernan, J. K. Han, J. Campbell, *et al.*, *Predictive value of pituitary tumor morphology on outcomes and complications in endoscopic transsphenoidal surgery*, en, Mar. 2022. DOI: 10.1002/wjo2.16. [Online]. Available: http://dx.doi.org/10.1002/wjo2.16.

[34] M. Araujo-Castro, A. Acitores Cancela, C. Vior, E. Pascual-Corrales, and V. Rodríguez Berrocal, "Radiological Knosp, revised-Knosp, and Hardy–Wilson classifications for the prediction of surgical outcomes in the endoscopic endonasal surgery of pituitary adenomas: Study of 228 cases", *Frontiers in Oncology*, vol. 11, 2022. DOI: 10.3389/fonc.2021.807040.

[35] J. L. Sanmillán, A. Torres-Diaz, J. J. Sanchez-Fernández, *et al.*, *Radiologic predictors for extent of resection in pituitary adenoma surgery. a single-center study*, en, Dec. 2017. DOI: 10.1016/j.wneu.2017.09.017. [Online]. Available: http://dx.doi.org/10.1016/j.wneu.2017.09.017.

[36] C. Serra, V. E. Staartjes, N. Maldaner, *et al.*, "Predicting extent of resection in transsphenoidal surgery for pituitary adenoma", en, *Acta Neurochirurgica*, vol. 160, no. 11, 2255–2262, Sep. 2018. DOI: 10.1007/s00701-018-3690-x. [Online]. Available: http://dx.doi.org/10.1007/s00701-018-3690-x.

[37] N. Qiao, M. Shen, W. He, *et al.*, *Machine learning in predicting early remission in patients after surgical treatment of acromegaly: A multicenter study*, en, Oct. 2020. DOI: 10.1007/s11102-020-01086-4. [Online]. Available: http://dx.doi.org/10.1007/s11102-020-01086-4.

[38] M. Araujo-Castro, E. Pascual-Corrales, V. Martínez-Vaello, *et al.*, "Predictive model of surgical remission in acromegaly: Age, presurgical gh levels and knosp grade as the best predictors of surgical remission", en, *Journal of Endocrinological Investigation*, vol. 44, no. 1, 183–193, May 2020. DOI: 10.1007/s40618-020-01296-4. [Online]. Available: http://dx.doi.org/10.1007/s40618-020-01296-4.

[39] E. Knosp, E. Steiner, K. Kitz, and C. Matula, "Pituitary adenomas with invasion of the cavernous sinus space", en, *Neurosurgery*, vol. 33, no. 4, 610–618, Oct. 1993. DOI: 10.1227/00006123-199310000-00008. [Online]. Available: http://dx.doi.org/10.1227/00006123-199310000-00008.

[40] A. S. G. Micko, A. Wöhrer, S. Wolfsberger, and E. Knosp, "Invasion of the cavernous sinus space in pituitary adenomas: Endoscopic verification and its correlation with an mri-based classification", *Journal of Neurosurgery*, vol. 122, no. 4, 803–811, Apr. 2015. DOI: 10.3171/2014.12.jns141083. [Online]. Available: http://dx.doi.org/10.3171/2014.12.JNS141083.

[41] J. Egger, T. Kapur, C. Nimsky, and R. Kikinis, "Pituitary adenoma volumetry with 3d slicer", *PLoS ONE*, vol. 7, no. 12, 2012. DOI: 10.1371/journal.pone.0051788.

[42] J. Egger, M. H. A. Bauer, D. Kuhnt, B. Freisleben, and C. Nimsky, "Pituitary adenoma segmentation", Proceedings of Biosignal, Jul. 2010. arXiv: 1103.1778. [Online]. Available: http://arxiv.org/abs/1103.1778.

[43] M. A. Mooney, D. A. Hardesty, J. P. Sheehy, *et al.*, “Interrater and intrarater reliability of the knosp scale for pituitary adenoma grading”, *Journal of Neurosurgery*, vol. 126, no. 5, 1714–1719, May 2017. DOI: 10.3171/2016.3.jns153044. [Online]. Available: http://dx.doi.org/10.3171/2016.3.JNS153044.

[44] V. E. Staartjes, C. Serra, N. Maldaner, *et al.*, “The zurich pituitary score predicts utility of intraoperative high-field magnetic resonance imaging in transsphenoidal pituitary adenoma surgery”, en, *Acta Neurochirurgica*, vol. 161, no. 10, 2107–2115, Aug. 2019. DOI: 10.1007/s00701-019-04018-9. [Online]. Available: http://dx.doi.org/10.1007/s00701-019-04018-9.

[45] V. E. Staartjes, C. Serra, M. Zoli, *et al.*, “Multicenter external validation of the zurich pituitary score”, en, *Acta Neurochirurgica*, vol. 162, no. 6, 1287–1295, Mar. 2020. DOI: 10.1007/s00701-020-04286-w. [Online]. Available: http://dx.doi.org/10.1007/s00701-020-04286-w.

[46] I. Basheer and M Hajmeer, *Artificial neural networks: Fundamentals, computing, design, and application*, en, Dec. 2000. DOI: 10.1016/s0167-7012(00)00201-3. [Online]. Available: http://dx.doi.org/10.1016/S0167-7012(00)00201-3.

[47] X. Zhang, C. Xv, M. Shen, X. He, and W. Du, *Survey of convolutional neural network*, 2018. DOI: 10.2991/ncce-18.2018.16. [Online]. Available: http://dx.doi.org/10.2991/ncce-18.2018.16.

[48] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review”, *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017. DOI: 10.1162/neco_a_00990.

[49] K. O’Shea and R. Nash, “An introduction to convolutional neural networks”, 2015. DOI: 10.48550/ARXIV.1511.08458. [Online]. Available: https://arxiv.org/abs/1511.08458.

[52] In *Digital Image Processing*, 3rd. Springer-Verlag, 1995, pp. 100–157, ISBN: ISBN 978-3-540-59298-3.

[53] *Examples of convolutions*, 2017. [Online]. Available: https://staff.fnwi.uva.nl/r.vandenboomgaard/IPCV20162017/LectureNotes/IP/LocalOperators/convolutionExamples.html.

[54] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: Analysis, applications, and prospects”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022. DOI: 10.1109/TNNLS.2021.3084827.

[58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[59] *Image classification*, 2023. [Online]. Available: https://www.tensorflow.org/tutorials/images/classification.

[60] I. Kandel and M. Castelli, “The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset”, *ICT Express*, vol. 6, no. 4, pp. 312–315, 2020, ISSN: 2405-9595. DOI: https://doi.org/10.1016/j.icte.2020.04.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959519303455.

[61] *Overfit and underfit*, 2023. [Online]. Available: `https://www.tensorflow.org/tutorials/keras/overfit_and_underfit`.

[62] *Data augmentation*, 2023. [Online]. Available: `https://www.tensorflow.org/tutorials/images/data_augmentation`.

[63] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge", 2014. DOI: `10.48550/ARXIV.1409.0575`. [Online]. Available: `https://arxiv.org/abs/1409.0575`.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[65] *Large scale visual recognition challenge 2012 (ilsvrc2012)*, 2012. [Online]. Available: `https://image-net.org/challenges/LSVRC/2012/results.html`.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 2014. DOI: `10.48550/ARXIV.1409.1556`. [Online]. Available: `https://arxiv.org/abs/1409.1556`.

[67] *Large scale visual recognition challenge 2014 (ilsvrc2014)*, 2014. [Online]. Available: `https://image-net.org/challenges/LSVRC/2014/results.php`.

[68] *Models and pre-trained weights*, 2023. [Online]. Available: `https://pytorch.org/vision/stable/models.html`.

[69] *Module: Tf.keras.applications*, 2023. [Online]. Available: `https://www.tensorflow.org/api_docs/python/tf/keras/applications/`.

[70] *Keras applications*, 2023. [Online]. Available: `https://keras.io/api/applications/`.

[71] *Vgg-16 convolutional neural network*, 2023. [Online]. Available: `https://www.mathworks.com/help/deeplearning/ref/vgg16.html`.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", 2015. DOI: `10.48550/ARXIV.1512.03385`. [Online]. Available: `https://arxiv.org/abs/1512.03385`.

[73] *Large scale visual recognition challenge 2015 (ilsvrc2015)*, 2015. [Online]. Available: `https://image-net.org/challenges/LSVRC/2015/results.php`.

[76] *Transfer learning and fine-tuning*, 2023. [Online]. Available: `https://www.tensorflow.org/tutorials/images/transfer_learning`.

[78] M. Černý, J. Kybic, M. Májovský, *et al.*, *Fully automated imaging protocol independent system for pituitary adenoma segmentation: A convolutional neural network—based model on sparsely annotated mri*, en, May 2023. DOI: `10.1007/s10143-023-02014-3`. [Online]. Available: `http://dx.doi.org/10.1007/s10143-023-02014-3`.

[80] Q. Han, H. Liang, P. Cheng, H. Yang, and P. Zhao, *Gross total vs. subtotal resection on survival outcomes in elderly patients with high-grade glioma: A systematic review and meta-analysis*, Mar. 2020. DOI: `10.3389/fonc.2020.00151`. [Online]. Available: `http://dx.doi.org/10.3389/fonc.2020.00151`.

[81] V. E. Staartjes, C. Serra, G. Muscas, *et al.*, "Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: A pilot study", *Neurosurgical Focus*, vol. 45, no. 5, 2018. DOI: `10.3171/2018.8.focus18243`.

Other sources

[3] *Pituitary part 1*, 2023. [Online]. Available: `http://what-when-how.com/acp-medicine/pituitary-part-1/`.

[4] *Anatomy and physiology*, 2015. [Online]. Available: `https://pocketdentistry.com/4-anatomy-and-physiology/`.

[22] *Acromegaly*, 2023. [Online]. Available: `https://rmi.edu.pk/disease/acromegaly`.

[28] *Principles of nuclear magnetic resonance (nmr) technology*, 2023. [Online]. Available: `https://www.creative-proteomics.com/pronalyse/principles-of-nuclear-magnetic-resonance-nmr-technology.html`.

[50] *Perceptrons – these artificial neurons are the fundamentals of neural networks*, 2020. [Online]. Available: `https://starship-knowledge.com/neural-networks-perceptrons`.

[51] *Multilayer perceptron example*, 2020. [Online]. Available: `https://github.com/rcassani/mlp-example`.

[55] *Convolution*, 2023. [Online]. Available: `https://vincmazet.github.io/bip/filtering/convolution.html`.

[56] A. S. Almryad and H. Kutucu, *Automatic identification for field butterflies by convolutional neural networks*, en, Feb. 2020. DOI: `10.1016/j.jestch.2020.01.006`. [Online]. Available: `http://dx.doi.org/10.1016/j.jestch.2020.01.006`.

[57] *Max pooling*, 2017. [Online]. Available: `https://paperswithcode.com/method/max-pooling`.

[74] F. Ramzan, M. U. G. Khan, A. Rehmat, *et al.*, *A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks*, en, Dec. 2019. DOI: `10.1007/s10916-019-1475-2`. [Online]. Available: `http://dx.doi.org/10.1007/s10916-019-1475-2`.

[75] C. Kawatsu, F. Koss, A. Gillies, *et al.*, "Gesture recognition for robotic control using deep learning", Aug. 2017.

[82] *Trained models*, 2024. [Online]. Available: `https://drive.google.com/drive/folders/1zsRdUOhYlMdC_s_d7BZcLumWuByTpChv?usp=sharing`.