HSLU Hochschule Luzern

Machine Learning for the Prediction of Refractive Surprise after Cataract Surgery

Bachelor Thesis

B.Sc. in Computer Science

by

Boas Meier

03th of January 2023

Advisors:

Univ.-Prof. Dr. Achim Langenbucher, Dr. sc. ETH Andreas Streich

Declaration

Bachelorarbeit an der Hochschule Luzern – Informatik

Titel: Prediction of refractive surprise after cataract surgery using machine learning
Studentin/Student: Boas Meier (matriculation number: 19-875-558)
Studiengang: BSc Informatik
Abschlussjahr: 2023
Betreuungsperson: Dr. sc. ETH Andreas Streich
Expertin/Experte: Dr. Rémi Janner
Auftraggeberin/Auftraggeber: Univ.-Prof. Dr. Achim Langenbucher

Codierung / Klassifizierung der Arbeit
□ Öffentlich (Normalfall)
□ Vertraulich

Eidesstattliche Erklärung
Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt habe, alle verwendeten Quellen, Literatur und andere Hilfsmittel angegeben habe, wörtlich oder inhaltlich entnommene Stellen als solche kenntlich gemacht habe das Vertraulichkeitsinteresse des Auftraggebers wahren und die Urheberrechtsbestimmungen der Hochschule Luzern respektieren werde.

Ort / Datum, Unterschrift

Abgabe der Arbeit auf der Portfolio Datenbank
Bestätigungsvisum Studentin/Student
Ich bestätige, dass ich die Bachelorarbeit korrekt gemäss Merkblatt auf der Portfolio Datenbank ablege. Die Verantwortlichkeit sowie die Berechtigungen gebe ich ab, so dass ich keine Änderungen mehr vornehmen oder weitere Dateien hochladen kann.

Ort / Datum, Unterschrift

Abstract

This thesis aims to preoperatively predict refractive surprises, that may occur after cataract surgery, using machine learning (ML). By predicting refractive surprises, a surgeon could take preventive measures, such as choosing a intraocular lens (IOL) type that is less sensitive to refractive error. Avoiding complications would increase patient satisfaction and save additional post-operative treatments, thus also saving costs.

In total this study included 2626 eyes that underwent cataract surgery, which were split into a training set of 2363 eyes and a testing set of 263 eyes using stratified sampling. Both unsupervised learning algorithms, including principal component analysis and supervised learning algorithms, including logistic regression, decision trees, random forests, support vector machines, gradient boosted trees, and neural networks were trained to perform either regression or classification of refractive surprises. Additionally, a ML-based IOL power calculation formula was developed and compared to the Castrop and SRKT formula on the testing set.

The refractive surprise regression achieved a mean absolute error (MAE) of 0.334 ± 0.422, with an R2-Score of 0.076. The classification with a refractive surprise threshold of 0.5 dioptre (D) resulted in a precision of 0.3, a recall of 0.69, and an F1-Score of 0.42. Using a threshold of 0.25 D, the resulting metrics were 0.58, 0.52, and 0.55, respectively.

The MAE of the ML formula developed in this thesis was 0.331 ± 0.423 and the median absolute error (MedAE) was 0.269. The performance of the Castrop and SRKT formulas were as follows: Castrop MAE = 0.341 ± 0.442, MedAE = 0.275; SRKT MAE = 0.402 ± 0.515, MedAE = 0.342. T-tests with Bonferroni correction indicated significance between the ML formula and the SRKT formula (p = 0.006) but no significance between the ML formula and the Castrop formula (p = 0.7).

Acknowledgements

I would like to express my gratitude to the following individuals for their contributions to this thesis:

- Univ.-Prof. Dr. Achim Langenbucher, for enabling this work, providing the data, and sharing his knowledge in ophthalmology.
- Dr. sc. ETH Andreas Streich, for his supervision, proofreading, and extensive support throughout this work, particularly with regards to machine learning topics.
- Pia Lohri, for sharing professional slit lamp images of various cataracts with explanations.
- Dr. Remi Janner, for his helpful inputs following the intermediate presentation.
- Benjamin Haymond, for sharing his expertise in writing and proofreading some parts of this thesis.

Contents

Appendix

Listings

CHAPTER 1

Introduction

This chapter provides an introduction to cataracts and the problem of refractive surprise after surgery. The objective and significance of this study are explained, as well as who can benefit from the research findings and how.

1.1 Motivation

Vision is arguably one of the most important senses humans use to navigate and interact with their environment. However, vision affects more than one's ability to see the world clearly. Vision impairment has the potential to negatively impact almost every aspect of a person's life, and results in significant expenditures. (National Academies of Sciences et al., 2016) Globally, at least 2.2 billion people have a near or distance vision impairment. In at least 1 billion of these cases, vision impairment could have been prevented or has yet to be addressed. With 94 million cases out of this 1 billion, cataracts are the most common cause of impaired distance vision. (World Health Organization, 2021)

A cataract is the clouding of the eye's focusing lens that results in blurry vision and, if left untreated, eventually leads to vision loss. Today cataract surgery is the most common procedure performed around the world and in all of medicine. With an overall success rate of approximately 97 percent when performed in appropriate settings, it is as well the most effective procedure. (Feldman H. et al., 2022)

Due to the high success rate, patient expectations are today at an all-time high and so is the dissatisfaction in cases where vision is not restored as expected. One of the major reasons for dissatisfaction after cataract surgery is residual refractive error. (Donaldson, 2022) In cases where the intended post-operative refractive target is missed, one also speaks of a refractive surprise. This can lead to follow-up interventions up to and including replacement of the lens. (Peck et al., 2022)

So if an expected refractive surprise could be predicted in advance of the cataract surgery, the surgeon could take preventive measures, such as choosing a lens type which is less sensitive to refractive error. Avoiding complications would increase patient satisfaction and save additional post-operative treatments, thus also saving costs. Finally, this would also ease the life of a surgeon, since dealing with an unhappy patient is always an uncomfortable challenge and arguably one of the most difficult aspects of his job. (Donaldson, 2022)

Due to the large number of cataract surgeries performed each year, avoiding even a small percentage of refractive surprise will improve the vision restoration of a large group of people.

1.2 Objective

This thesis presents a machine learning (ML) approach for the preoperative prediction of refractive surprises. The research goals of this work are:

1. An overview of current research and state-of-the-art (SotA) techniques for predicting refractive surprises after cataract surgery is documented.

2. A baseline ML model using conventional algorithms for predicting whether a patient will face refractive surprises after cataract surgery is implemented and evaluated.

3. A competitive ML model based on SotA technology to predict whether a patient will face refractive surprises after cataract surgery is implemented and evaluated.

The baseline model serves as a reference point for the feasibility of the data, while the competitive model aims to set a high bar for future research in this area.

1.3 Structure of this Thesis

The structure of this thesis is mostly based on the document Aufbau WIPRO/BAA-Bericht (Hofstetter, 2020), provided by the Lucerne University of Applied Sciences and Arts (HSLU). This document determines seven chapters, each with a brief description of what that chapter should be about. What was changed is, that the main part and the appendix were split up into separate parts and the appendix is numbered using alphabetical rather than arabic numbering. The bibliography, list of figures, and list of tables are placed after the appendix, as suggested, but are not numbered.

This thesis was written in cooperation between the computer science department at the HSLU and the medical faculty at the Saarland University. The subject area lies at the intersection of ophthalmology, computer science, data science and ML. As potential readers of this work may come from these various domains, a glossary with cross-references and back-references is included to explain domain-specific technical terms. Many of these terms are also abbreviated as acronyms. In this case, a cross-reference in the text points to the list of acronyms, which in turn includes a cross-reference to the glossary.

1.4 Human Eyesight

To understand this thesis, it is necessary to have a basic understanding of how human vision works. This section provides an introduction to the concepts of optical power and refractive error, and how they are related.

1.4.1 Optical Power

The optical power of a lens is a physical quantity which measures its ability to bend light. Optical power is also referred to as refractive power and is measured in dioptre (D). 1 D is equal to 1 m^{-1} (Wikipedia, 2022a). For example, a lens with an optical power of 3 D brings two parallel light rays to focus $\frac{1}{3}$ m behind the lens, as shown in Figure 1.1. On the other hand, a lens with an optical power of -3 D diverges the light, resulting in a theoretically negative focus point at $-\frac{1}{3}$ m.

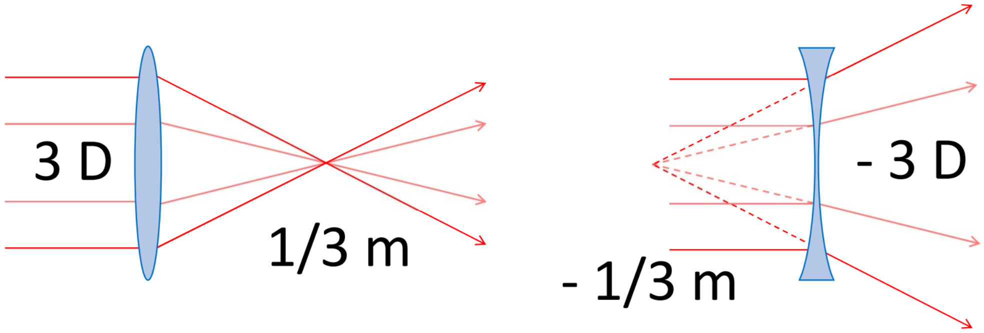


Figure 1.1: An illustration of the relationship between optical power and focal length. Converging (convex) lenses have positive optical power (left), while diverging (concav) lenses have negative optical power (right).

1.4.2 Refractive Error

The relaxed human eye typically has an optical power of around 60 D (Palanker, 2013). If the eye has no refractive error, this optical power is just right, to focus parallel rays of light directly on the retina. In this case the eye is said to have emmetropia or 20/20 vision. An eye with refractive error, on the other hand, is said to have ametropia. Types of ametropia include myopia, hyperopia and astigmatism.

Myopia, also known as short-sightedness, occurs when the optical power of the eye is too large in relation to the axial length (AL), causing the focus point to be in front of the retina rather than directly on it. This leads to difficulty seeing distant objects clearly and is corrected with a concave lens, which has negative optical power. Hyperopia, or far-sightedness, is caused by an optical power that is too small, resulting in a focus point behind the retina and difficulty seeing near objects clearly. It is corrected with convex lenses. Refractive errors in which the optical power of the eye is either too large or too small to focus light on the retina are also referred to as spherical errors.

Astigmatism, on the other hand, causes distorted or blurred vision at any distance due to rotational asymmetry in the eye's optical power. This asymmetry causes the optical power to be either too strong or too weak across one meridian, such as if the corneal curvature tends towards a cylindrical shape. Thus, astigmatism is also referred to as cylindrical error and is corrected with cylindrical and toric lenses, which refract light more in one meridian than the other. (Wikipedia, 2022c)

1.5 Cataract

Cataract is not a type of refractive error eye disorder, as discussed in section 1.4.2. It refers to the clouding of the natural crystalline lens that refracts light entering the eye onto the retina (see Figure 1.2). This cloudiness can lead to decreased vision and, if left untreated, may eventually cause blindness. Cataracts develop gradually over time, without causing pain or significant discomfort, so it may take decades before any signs of the condition are noticed. Most cataracts are caused by age-related degeneration, but there are also congenital cataracts present at birth and traumatic cataracts resulting from eye injuries. (Feldman H. et al., 2022)

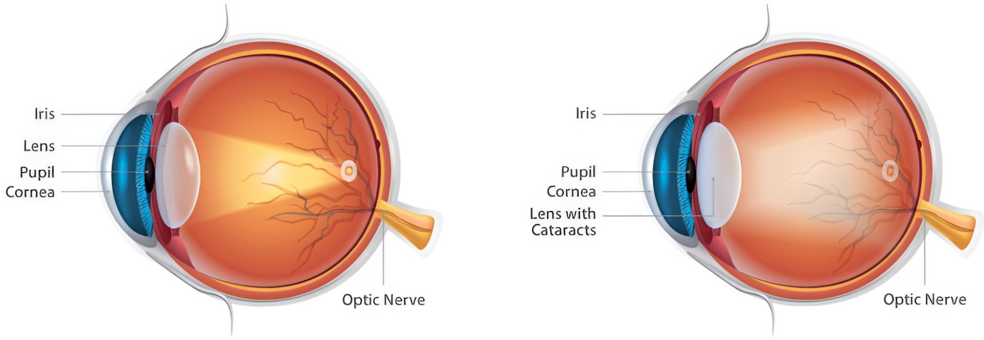


Figure 1.2: An illustration of a horizontally cut eye with normal lens (left) and a cataract lens causing distorted vision (right). (Meyer, 2022)

While all people will eventually develop age-related cataracts, research has shown that certain health, environmental, and behavioral factors can increase the risk of developing a cataract. These factors include diabetes or elevated blood sugar, smoking, alcohol consumption, exposure to ultraviolet radiation, and prior ocular surgery. (Russel, 2020a)

1.5.1 Diagnosis

Simple signs of cataract include blurred and decreased vision as well as halos. Physical findings include an opaque lens (see Figure 1.3). To fully evaluate a cataract, rule out other eye diseases, and prepare for potential surgery, various steps are taken. These include visual acuity tests, slit-lamp examinations, biometry, and refraction and intraocular pressure measurements. (Nizami et al., 2021)

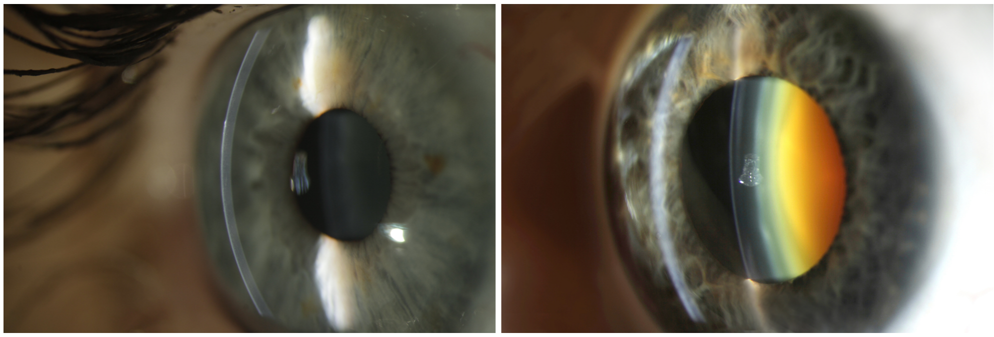


Figure 1.3: A human eye with a healthy lens (left) (Lohri, 2022c) and a lens affected by advanced nuclear cataract (right) (Lohri, 2022a). The images are taken with a slit lamp camera and mydriatic eye drops were used to widen the pupil. Both lenses are illustrated in the semi profile.

1.5.2 Treatment

Currently, there are no effective methods for preventing or treating cataracts with medication. However, early stages of cataract may be treated with corrective glasses or contact lenses. If the cataract is mature enough to interferes with daily activities, such as driving, or if visual acuity is worse than 6/24, surgery is recommended. Standards for cataract surgery are developing worldwide. First there was intracapsular cataract extraction (ICCE), then came extracapsular cataract extraction (ECCE), and today the procedure of choice is phacoemulsification (PCS). These procedures involve surgical removal of the clouded lens and implantation of an intraocular lens (IOL) (see Figure 1.4), which can restore vision in cataract patients. (Chen et al., 2021) (Nizami et al., 2021) In many cases, even a 20/20 vision, or rather emmetropia is achieved (Russel, 2020a).

1.5.3 Intraocular Lenses

Premium IOLs are being used more frequently to meet the personalized needs of patients. Multifocal IOLs[1], which are a type of premium IOLs, have been shown to be superior to traditional monofocal IOLs in terms of uncorrected distance visual acuity. Over 90 % of patients with multifocal IOLs achieve spectacle-independence for distant vision, while only 52.4-85 % of patients with monofocal IOLs do so. Additionally, 81.8-84.9 % of patients with multifocal IOLs gain both distant and near spectacle-independece, compared to 7.5-12 % of those with monofolcal IOLs. Due to these positive experiences in general and especially with multifocal IOLs, spectacle-independence is expected for either distant vision, near vision or both, in case of premium IOLs. Hence, if cataract surgery does not result in spectacle-independece, it can lead to patient dissatisfaction. Residual refractive error is the main cause for

[1] Multifocal IOLs use concentric rings of varying thickness (see Figure 1.4) to allow the eye to focus on images at all distances. The patient's brain registers the image that is most in focus based on the distance of the object. It usually takes some time for the brain to adjust to the multifocal lenses. (Dudek, 2021)

this, especially with premium IOLs, which are associated with an increased rate of visual phenomena such as glare, halos, and night vision problems that are significantly exacerbated by any refractive error. Therefore, refractive predictability has become increasingly important since the advent of premium IOLs. (Chen et al., 2021) (Peck et al., 2022)

While emmetropia is the desired outcome in most cataract cases, there are also cases where the refractive target is myopic. For example, a patient who has been short-sighted for their entire life may be unhappy if they can no longer read due to a hyperopic outcome. According to Behndig et al. (Behndig et al., 2012), the target refraction was myopic for 7.0 % of the 17'056 analyzed patients, while planned hyperopia was rare.

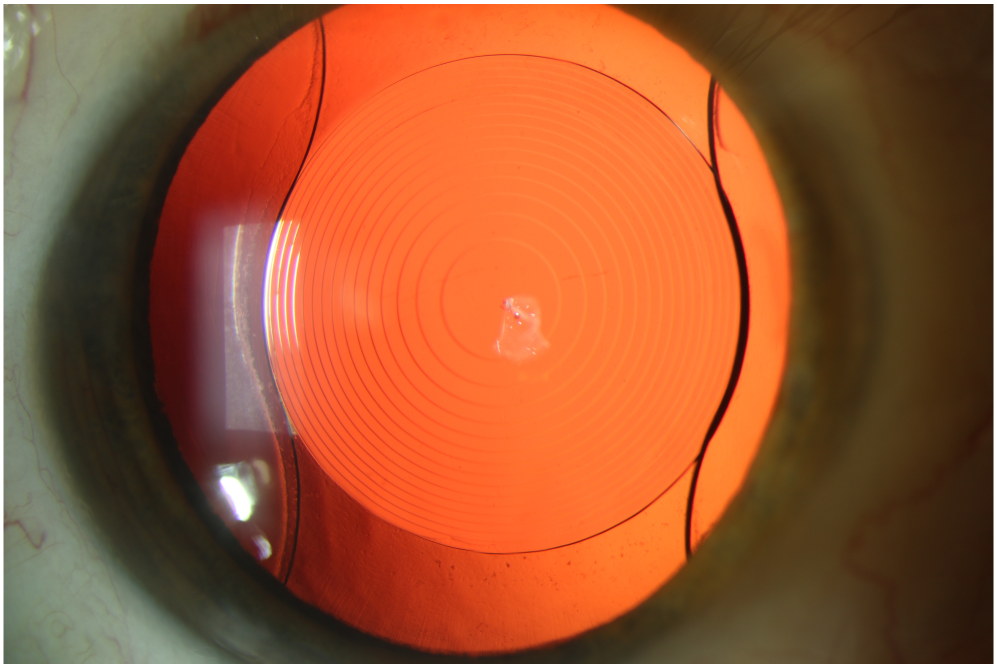


Figure 1.4: An image of an eye with an via cataract surgery implanted multifocal IOL. (Lohri, 2022b)

1.5.4 Refractive Prediction Error

As previously mentioned in Section 1.1, residual refractive error, also known as refractive prediction error, indicates by how much dioptre the post-operative refractive target is missed. If the post-operative refractive target is specified as $predSEQ$ and the effective refractive outcome is SEQ, then the refractive prediction error is calculated as follows:

$$PE = SEQ - predSEQ \tag{1.1}$$

whereby PE stands for prediction error, SEQ for spherical equivalent, and predSEQ for predicted spherical equivalent. A negative prediction error (PE) indicates a more myopic outcome than expected, while a positive PE indicates a more hyperopic outcome. How predSEQ and the corresponding IOL power is calculated, is discussed in Section 2.1.

1.5.5 Managing Refractive Surprise after Cataract Surgery

Prevention is the most effective way to manage refractive surprise. Benchmark standards for National Health Service (NHS) cataract surgery dictate that the absolute PE should be within 1.0 D for 85 % of the eyes and within 0.5 D for 55 %. (Shalchi et al., 2018) Studies on cataract surgery outcomes show that 79-94 % and 50-70 % of patients will achieve postoperative refractions within 1.0 D and 0.5 D of the intended target, respectively. (Peck et al., 2022) The ML algorithm developed in this thesis aims to further increase the number of patients within these ranges. However, if the postoperative refractive target is missed, the following options are available to the patient.

Glasses and Contacts

Not taking any action is always an option. Many refractive surprises do not require further surgery. For example, low myopia in one eye may result in monovision and the ability to read unaided. Additionally, a patient who has worn glasses their entire life may be willing to continue doing so, and some patients are comfortable wearing contact lenses. (Shalchi et al., 2018)

Surgery

For patients where spectacles are not an option, further surgery is necessary. The risks of further surgery are often greater than with the first cataract surgery, so it is important to discuss this in detail with the patient. Additionally, there are significant financial and time costs for the patient. (Peck et al., 2022)

Corneal Refractive Surgery Laser refractive surgery, such as photorefractive keratectomy (PRK) or laser-assisted in situ keratomileusis (LASIK), is a good option after refractive surprise. It can treat a wide range of refractive errors, including astigmatism. (Shalchi et al., 2018)

IOL Exchange If the source of the error and the cause of the occurrence are clear, IOL exchange may be an effective option. (Shalchi et al., 2018)

Piggyback IOL If the risk of IOL exchange is too high, a piggyback IOL may be the optimal choice. A piggyback IOL is inserted in addition to the original IOL. (Shalchi et al., 2018)

CHAPTER 2

State of Research

This chapter outlines the state of research as well as the SotA in relation to the objectives presented in Section 1.2. It also examines how other researchers have addressed similar problems, highlighting their strengths and weaknesses and comparing them to this work.

2.1 Intraocular Lens Power Calculation

Nowadays, the IOL power and the post-operative refractive outcome is calculated based on biometry of the eye, such as the AL. Numerous formulas for this have been proposed in the past. More recent formulas are generally outperforming those of prior generations in accuracy. (Melles et al., 2018) Classical and widely adopted formulas today are the Haigis (Haigis et al., 2000), Hoffer-Q (Hoffer, 1993), Holladay (Holladay et al., 1988) and SRKT formula (Retzlaff et al., 1990). As stated in Chapter 1, the ultimate goal of predicting refractive surprises should eventually help in making better preoperative predictions and thus reducing patient dissatisfaction due to PE. To achieve that goal with ML, the PE of a particular IOL power calculation formula is needed, in order to train and evaluate an algorithm. Here it seems reasonable to choose the PE of a top performing formula. Therefore, this section presents some publications of some recent and promising new formulas, studies which compared some of those formulas, as well as study guidelines for IOL power calculation. These guidelines are beneficial for this work, because IOL power calculation and the prediction of the PE of these formulas are pretty similar.4

2.1.1 Study Guidelines

Hoffer and Savini (Hoffer and Savini, 2021) proposed updated study design guidelines for IOL power calculation, which aimed to modernize the existing 2015 guidelines (Hoffer et al., 2015). Here is a direct quote from the recently published official article:

> We hope that these recommendations will help researchers improve the va-

lidity and accuracy of their studies with the ultimate goal to really improve the accuracy of IOL power calculation.

These guidelines are relevant, because in this thesis is an IOL power calculation formula based on ML developed in order to help predict the PE of existing formulas (see Section 3). The guidelines also help to design the experiments in such a way that they correspond with the current state of research. This increases the validity of the results as well as it makes them more comparable with previous and future work. The following bullet points summarize the recommendations of that new article.

- Only use samples, where the postoperative visual acuity of the patient is 20/40 or better.
- Only use one eye per patient in the test data.
- In general the sample size should be chosen as such that it is big enough to detect the effects hypothesized with reliable confidence. Based on articles published in the last decade, a sample size close to 200 for normal cataract eyes and one of at minimum 50 eyes in case of rare conditions. Such rare conditions are for example eyes which underwent LASIK or PRK.
- The age, sex and ethnicity of the study population should be reported.
- Before comparing different IOL power calculation methods, their constants must be optimized on the training data. Constant optimization is the process leading to a zero mean PE on the training data. It is required to eliminate any systematic error arising from the clinical environment, including the biometer, the surgical technique and the physical properties of the IOL. The more data is used the more accurate is the constant optimization. However, at least 100 eyes should be included to achieve reliable measurements.
- Newly published formulas should not be compared with outdated and proven inaccurate formulas such as Blinkhorst II, SRK I and SRK II regression.
- The postoperative refraction should be assessed when stable and the postoperative spherical equivalent should be measured with the highest accuracy.
- The comparison of the PE should be reported based on mean absolute error (MAE), median absolute error (MedAE) as well as standard deviation (SD). To evaluate whether there is significant difference between the formulas either a Wilcoxon matched-pairs test (two samples) or Friedman test with post hoc test (more than two samples) should be performed. In case of unpaired samples and unpaired t test (two samples) or Kruskal-Wallis test (more than two samples) should be performed. Additionally the percentage of eyes with an absolute PE within 0.25 D, 0.5 D, 0.75 D and 1.0 D should be reported as well. These percentages should be compared by Cochran's Q test.

- For a more comprehensive ranking of the accuracy of different formulas the IOL Formula Performance Index (FPI) should be used.
- If existing formulas are implemented by the researcher itself, they need to be validated by the formula author, validated against a licensed biometer or validated by another authorized source.
- All formulas used in the study must be properly referenced, including all errata.
- If ultrasound AL is necessary, only immersion should be used and never contact applanation ultrasound.
- If corneal power is used in the study, the instrument and method to obtain it, as well as the type of corneal power used, should be stated clearly.
- The software version of all instruments, programming languages and libraries should be stated clearly.
- If the anterior chamber depth (ACD) is measured from the endothelium to the lens, instead of from the corneal epithelium to the lens, it should be refered to as aqueous depth (AD).
- The term white-to-white corneal diameter (WTW) should not be used anymore. Instead the proper anatomic definition of horizontal corneal diameter (HCD) should be used.
- The IOL formula accuracy must be evaluated on unseen testdata.

2.1.2 Comparison of Modern Formulas

Melles et al. (Melles et al., 2018) analyzed a total of 18'501 eyes from 18'501 patients to evaluate popular IOL power calculation formulas (Barrett Universal II (BU-II), Haigis, Hoffer-Q, Holladay 1, Holladay 2, Olsen and SRKT). Additionally the study analyzed the extent of bias within each formula for different biometric dimensions of the eye (ACD, AL, corneal curvature, and lens thickness (LT)) that impact the predictions negatively. Results showed, that the BU-II formula was significantly more accurate than the other formulas ($P < 0.01$). The major reason for the difference between the formulas is their performance on samples where the AL is either smaller than 23 or larger than 25 mm. Inside this range all formulas gave results within 0.1 D. However, overall the BU-II appeared to have the least bias of the formulas as measured by prediction error with variations in AL, corneal power (K), ACD, and LT.

The study generally complies with the first version of study guidelines of Hoffer et al (Hoffer et al., 2015), what makes comparison with other studies easier. A limitation however is, that only two IOL models where evaluated. Thus, the results may not be generalizable to IOL models of different design.

In another study, Kane et al. (Kane et al., 2017) demonstrated as well, that BU-II has greater accuracy than other formulas. The study analyzed 3122 eyes of 3122 patients to evaluate the formulas BU-II, FullMonte, Hill-RBF, Holladay 1, and the Ladas Super Formula (Siddiqui et al., 2019). Results showed a statistically significant difference in the MAE between the five methods (P $<$ 0.001), with BU-II being the most accurate.

2.1.3 Recently published formulas

This section presents four recently published and promising new formulas. All of these showed superior performance to conventional formulas like Haigis, Hoffer-Q, Holladay 1, and SRKT. Three of the following four presented formulas are based on some sort of ML, like random forests, support vector machines (SVM), gradient boosting, Bayesian Additive Regression Trees (BART), neural networks (NN), and ensembles. They can give valuable insight, because (1) during this thesis an IOL power calculation formula was developed, (2) the data at hand for this thesis in order to predict refractive surprise is the same as the data used in these publications and (3) because no articles and papers were found about the prediction of refractive surprises with ML.

Castrop Formula

Langenbucher et al. (Langenbucher et al., 2021) proposed the Castrop formula, which is a paraxial vergence formula based on a pseudophakic model eye with 4 refractive surfaces and 3 formula constants (C, H, and R). To evaluate the performance, the Castrop formula was compared to four classical formulas (Haigis, Hoffer-Q, Holladay 1, and SRKT). The study included 1452 measurements which were split randomly into a train set (70 %, 1017 cases) and test set (30 %, 435 cases). Whereby the train set was used for constant optimization and the test set for the final evaluation. The evaluation resulted in a MAE of 0.340, 0.367, 0.417, 0.390, 0.388 for Castrop, Haigis, Hoffer-Q, Holladay, and SRKT, respectively. A Wilcoxon signed rank test with Bonferroni correction showed that the Castrop formula yields significantly better results than all other evaluated formulas. Additionally the FPI was calculated for the Castrop, Haigis, Hoffer-Q, Holladay, and SRKT formula which was 1.1284, 1.0952, 1.0624, 1.0157, and 1.0588, respectively.

Limitations of the study are that in all cases a Sensar 1 piece IOL (Johnson & Johnson, Brunswick, USA) was inserted and that it was not indicated, from how many patients the final 1452 used samples are taken. Thus, the results may not be generalizable to other IOL types and may be overly optimistic due to "both eye bias". This is because ocular measurements between bilateral eyes are more alike than between eyes of different patients. Hence, measurements of fellow eyes cannot be treated as if they were independent. To prevent this so called "both eye bias", Hoffer and Savini (Hoffer and Savini, 2021) propose to only enroll one eye per patient into IOL power calculation studies (see Section 2.1.1). Another limitation is that the newly proposed Castrop formula was not compared to other modern top performing IOL formulas such as the BU-II.

A huge advantage of the publication is, that the calculation strategy of the Castrop formula is open-source, what is not the case for many other modern formulas. This simplifies the application of the formula for other researchers and ophthalmologists, since the lens calculation can easily be automated and does not have to take place via a web interface (calc.apacrs.org, 2010), as for example when using the BU-II formula. Thus, the PE of the Castrop formula is an optimal candidate to use as target in this thesis. Not only because it is open-source and outperforms other conventional formulas, but also because its PE is already present in the data at hand (see Section 3.1).

Yamauchi Formula

Yamauchi et al. (Yamauchi et al., 2021) assessed 3331 eyes from 2010 patients to train various ML models. These models were then compared with conventional IOL power calculation formulas. Among these are the SRK/T formula, Haigis formula, Holladay 1 formula, Hoffer-Q formula, and BU-II formula. On the side of ML support vector regression (SVR), random forest regression (RFR), gradient boosting regression (GBR), and NN were assessed. With a MAE of 0.2960 the BU-II formula provided values that were significantly lower than those provided by the other formulas. The MAE on test data for the SVR, GBR, RFR, and NN were 0.2877, 0.2929, 0.2964, and 0.2891, respectively. The SVR, GBR and NN therefore had lower MAE than the BU-II formula. However, no significant difference was observed.

Not only the MAE was assessed but also the proportion of objects with errors less than 0.5 D. The BU-II formula, SVR, RFR, GBR, and NN resulted in the following proportion 81.2 %, 84.4 %, 82,4 %, 82.8 % and 82.4 %, respectively. The ML methods resulted in less errors above 0.5 D but without significant difference. The study also assessed the mean absolute prediction error categorized in short, medium, and long ALs, and no significant differences were observed among these. The authors did not give an explanation on why modern and powerful ML algorithms such as GBR and NN did only result in such a small performance improvement compared with the conventional algorithms like SVR and RFR. The NN architecture consists of three fully connected hidden layers, each of which consists of 100 neurons. Before and after the second hidden layer a dropout layer is used.

The used input features are AL, corneal curvature, ACD, LT, HCD, IOL power, and postoperative refraction as well as the predicted refraction of the SRK/T formula. Those features are measured preoperative using the IOLMaster 700, and were selected based on the GBR prediction accuracy in the training data and the calculated feature importance.

For evaluation a test set consisting of 500 samples from 500 patients was used which had no significant difference in any of the used numerical features from the training data. However, the test set only consists of samples where a YP2.2 IOL was used, although the training data consists of 12 different IOL types. Thus, the evaluation on test data only tells how good the models perform on YP2.2 implanted eyes. To really tell how good the proposed methods perform in general, further evaluation on a test set with more diverse IOL types would be needed.

Nallasamy Formula

Nallasamy et al. (Li et al., 2022) assessed a total of 6893 eyes from 5016 patients to train a stacking ensemble machine learning method. The proposed method is called the Nallasamy formula and consists of two levels. The first level consists of different ML algorithms which are trained independently to predict the postoperative refraction. The second level than takes all those outputs of the first level models as input and is then trained to make the final prediction of the postoperative refraction. However, the concrete algorithms and their parameters used in these layers were not revealed. The authors of the article also speak of some novel data augmentation methods which were utilized to generate additional synthetic training data. Like the ML algorithm those data augmentation methods were not revealed. The performance of the Nallasamy formula was compared with that of BU-II, Emmetropia Verifying Optical (EVO), Haigis, Hoffer-Q, Holladay 1, PearlDGS and SRK/T. The Nallasamy formula performed with a MAE of 0.312 and a MedAE of 0.242 on the testing set significantly better than the seven existing methods based on the paired Wilcoxon test with Bonferroni correction ($p < 0.05$). Performance of the existing methods were as follows: BU-II MAE = 0.328, MedAE = 0.256; EVO formula MAE = 0.322, MedAE = 0.251; Haigis formula MAE = 0.363, MedAE = 0.289; Hoffer Q formula MAE = 0.404, MedAE = 0.331; Holladay 1 formula MAE = 0.371, MedAE = 0.298; PearlDGS formula MAE = 0.329, MedAE = 0.258 and SRK/T formula MAE = 0.376, MedAE = 0.300.

The Nallasamy formula also resulted in a larger percentage of patients within an absolute error of 0.5 D. With a proportion of 80.2 % it achieved higher results than all other formulas and was statistically better than all except the EVO formula (79.8 %).

The performance was also compared among patients with different ALs. The AL was categorized in the following three groups: short AL (< 22,0 mm), medium AL (≥ 22.0 and ≤ 26.0 mm), long AL (> 26.0 mm). The Nallasamy formula achieved the lowest MAE and SD among all eight formulas in all 3 AL groups.

The article presents the features sex, age at surgery, power of implanted IOL, K, AL, LT, ACD, AD, astigmatism, HCD and central corneal thickness (CCT). The preoperative biometry records were obtained from Lenstar LS 900 optical biometers.

For evaluation a test set consisting of 1003 samples from 1003 patients was used. No tests were performed to proof that there is no significant difference between the two data sets. The full dataset only consists of samples where an Alcon SN60WF one-piece acrylic monofocal lens was implanted.

The Nallasamy formula did in general improve the more data was used. The trend continued even as the training set was increased from 90 % to 100 %. This indicates the potential for further improvement if more data is used to train the same model.

Bayesian Additive Regression Trees Formula

Clarke and Kapelner (Clarke and Kapelner, 2020) proposed an IOL power calculation formula based on BART. A total of 3276 eyes from 3276 patients were used to develop the model. In order to validate the formula 5-fold cross validation was used. In total the dataset included 44 possible variables per eye of which 29 are physical measurements and the remaining 15 theoretical metrics. Among the theoretical metrics are the SRK/T A-Constant, Holladay Surgeon Factor, Hoffer ACD, and the three Haigis constants, which were calculated for each lens and surgeon using the Haigis linear regression. However, some samples had missing physical measurements and hence it was not possible to calculate all the theoretical features. Thanks to missingness handling it was nevertheless possible to construct a model on samples with missing features and predicting IOL power for patients eyes where not all features were measured.

Other than the previously demonstrated formulas, the study used the difference of the true implanted IOL and the theoretical (AL adjusted) SRKT IOL power that gives the same post-operative refraction. This IOL error was then converted to refractive error using Gaussian optics, what resulted in a MAE of 0.137 D (BART), 0.278 D (Holladay 1), 0.453 D (Hill-RBF 1.0), 0.478 D (SRK/T), and 0.586 D (Hoffer-Q) with a SD of 0.242 D (BART), 0.416 D (Holladay 1), 0.569 D (Hill-RBF 1.0), 0.575 D (SRK/T), and 0.936 D (Hoffer-Q).

A limitation of the study is, that the model was only trained on samples which were accepted by the Hill-RBF calculator. Thus, results are not directly comparable with other studies, because they did as well include samples which would have been rejected by the Hill-RBF calculator.

2.2 Risk Factors for Refractive Surprises

One of the first questions that should arise in relation to the objective of this thesis is: Which factors influence the PE after cataract surgery? Another important follow up question is then: Does the data at hand contain those influencing factors? Because in order to create a ML model that correctly predicts refractive surprises, the reason for them must be present in the data. Hence, this section presents nine studies which analyzed risk factors for refractive surprise. While some studies like the one from Garay-Aramburu et al. and Lundström et al. searched broadly for a wide variety of factors, others analyzed the influence of very specific factors like biometry measuring methods, age, sex and IOL manufacturing tolerances.

Garay-Aramburu et al. (Garay-Aramburu et al., 2022) analysed a total of 1578 eyes from 1419 patients. The goal was to determine which factors increase the risk of an absolute PE greater than 1.0 D. The most significant risk factors identified were non ultrasonic biometry, previous glaucoma surgery, presence of white or hard cataract and previous visual acuity within legal blindness. Other less prominent but also significant risk factors include extreme biometric data such as axial length (AL) less than 22 mm or greater than 26 mm, ACD less than 2.5 mm, HCD less than 10 mm and the use of biometric formulas other than BU-II. The superior accuracy of the BU-II formula has already been known. Further details on performance of different IOL power calculation formulas are covered in Section 2.1.

Other studies have shown the influence of the biometry measuring method on PE as well. Moshirfar et al. (Moshirfar et al., 2019) found for example, that new swept-source ocular coherence tomography biometers are more frequently successful at measuring AL in dense cataracts, which can improve refractive outcomes. Shammas et al. (Shammas et al., 2020) evaluated the influence of segmented AL versus traditional AL on the PE. The mean PE of 595 eyes was significantly smaller in longer and shorter eyes compared with medium length eyes. Across all eyes, the mean PE was smaller as well but not significantly. Accurate biometry, with AL being one of the most critical components, is generally known as key factor in successful IOL power calculation (Norrby, 2008). There were also cases reported, in which a biometer with smeared optics repeatedly overestimated the AL, leading to a refractive surprise of +14.0 D (Carr and Gangwani, 2020).

Demographic properties of patients, like age and sex, were also found to influence the refractive outcome. Hayashi et al. (Hayashi et al., 2016) have shown, that the PE was less myopic by approximately 0.06 D per decade as age increased. The mean PE correlated positively with age ($P < 0.0001$). The study analyzed 75 eyes of 75 patients which is rather small. In another study, including 8421 eyes of 5519 patients, Zhang et al.(Zhang et al., 2021) showed, that the PE was significantly different between male and female eyes ($P < 0.0001$). Errors of male eyes skewed towards hyperopia and female eyes towards myopia. The difference between the two groups in absolute PE was not significant. However, optimization of lens constants by sex decreased the absolute PE of all five formulas. For SRK/T and Hoffer Q this reduction was significant and for Holladay, Haigis and BU-II not.

The large scale multi-centre multinational study from Lundström et al. (Lundström et al., 2018), which analysed 548'392 eyes, found multiple risk factors. Among them are poor preoperative corrected visual acuity, ocular comorbidity and previous eye surgery.

In another study, Zudans et al. (Zudans et al., 2012) showed, that IOLs available in 0.25 D increments with a labeled manufacturing tolerance of ± 0.11 D increased the percentage of patients within ± 0.25 D of the targeted refraction to a statistically significant level compared with unlabeled IOLs available in 0.50 D increments. Therefore even if the preoperative prediction of the spherical equivalent (SEQ) would be perfect, it can only be as good as the IOL manufacturing tolerance. The same goes for the prediction of the PE of a particular formula. The algorithm will probably not be able to predict the PE which is due to manufacturing tolerances of the lens. Therefore an algorithm can be considered to have fully exploited its potential, if the difference between SEQ and the sum of predicted spherical equivalent (predSEQ) and predicted prediction error (predPE) over all samples n is approximately equal to the manufacturing tolerance ϵ:

$$\frac{\sum_{i=0}^{n-1} |SEQ_i - (predSEQ_i + predPE_i)|}{n} \approx \epsilon \tag{2.1}$$

Thus, returning to the question at the beginning of this section. One important factor influencing refractive surprises is inaccurate biometry. Therefore, if the inaccurate measurements in the data are separable from the accurate ones, an ML model should be able to capture at least the PE which is due to these inaccuracies. Section 3.2 describes a concept to evaluate the degree of separability of the data. Other important factors seem to be the preoperative corrected visual acuity, sex and age. To have these features in the data will therefore probably help in predicting refractive surprises.

2.3 Imbalanced Distributions

As stated in Section 1.1, the overall success rate of cataract surgery is approximately 97 percent. Due to this fact the domain of refractive surprises can be considered moderately to extremely imbalanced. Much research has been done in the area of imbalanced data distributions, which can help in predicting refractive surprises.

Most ML algorithms tend to be biased to the most frequent class when trained on imbalanced data, leading some to ignore the minority class entirely. This is a problem because in most of the real world applications it is the minority class on which predictions are most important to the user. (Branco et al., 2015)

Branco et al. (Branco et al., 2015) provide a comprehensive overview over the state of research on how to tackle this problem of imbalanced data. A strength of this paper is, that this problem is not only tackled for classification tasks but also for regression.

In order to train a ML model properly on imbalanced data, the evaluation metric needs to take into account the users preference. In case of classification the F1-Score is recommended. In case of regression, measures such as Mean Utility and Normalized Mean Utility can be used to compare different regression models according to the users preference bias.

Branco et al. grouped the different modelling strategies for handling imbalanced domains in the following groups:

- Data Pre-processing
- Special-purpose Learning Methods
- Prediction Post-processing
- Hybrid Methods

2.3.1 Data Pre-processing

Data pre-processing approaches include solutions that pre-process the given imbalanced data set leading to a more balanced one. Existing pre-processing algorithms are re-sampling. In simple re-sampling the majority class is either under-sampled, the minority class is over-sampled or both is done simultaneously. Under-sampling can be done randomly or in more sophisticated ways in which the least important samples are dropped systematically. Also over-sampling can be done really straight forward just by duplication or in a more sophisticated way in which new data is synthesised. A popular algorithm for synthesising new data is synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002).

Another pre-processing approach is calculating class weights, which are then incorporated into the loss function. Thanks to the weights some samples are more important, when computing loss, than others and a ML algorithm is able to learn from the original distribution without being biased to the majority class. However, the drawback of this technique is that there is a risk of model overfitting.

Advantages of data pre-processing approaches are, that they are rather straight forward and there are libraries such as imbalanced-learn (Lemaître et al., 2017) which implement various re-sampling algorithms such as SMOTE. Also most of the classifiers in the scikit-learn (Pedregosa et al., 2011) library provide a class weight parameter. The classifier is then automatically configured to learn more from samples with higher weights.

2.3.2 Special-purpose Learning Methods

The special-purpose learning methods consist of solutions that modify existing algorithms to provide a better fit to the imbalanced training data. The paper describes several studies proposing adaptions of different classifiers in order to make them more sensitive to skewed data. Additionally also several studies of newly introduced ensemble techniques are presented.

All of these algorithm modification strategies have great potential. However, the implementation of those requires a deep knowledge of the selected underlying algorithm and existing implementations may not be applicable for other domains than they were created for. Thus, special-purpose learning methods are often not as straight forward to use as data pre-processing approaches.

2.3.3 Prediction Post-processing

The third category of strategies to handle imbalanced domains is prediction post-processing. Prediction post-processing is based on a probability output, that expresses the degree to which an example is a member of a class. This probability is then used to produce several models by varying the threshold for class membership.

2.3.4 Hybrid Methods

Finally the paper presents several methods in which some of the basic methods described previously are combined. Those combinations are referred to as hybrid methods. One presented hybrid method for example combines re-sampling with special purpose learning using bagging and stacking. The data set is split up into n different new data sets which include all the minority class samples and a portion of the majority class samples. Then different ML algorithms are trained on each of the n new data sets resulting in n different classifiers for each ML algorithm. Next majority voting is used to combine the classifiers trained by the same algorithm. Those aggregated outputs are then used to train a final classifier.

CHAPTER 3

Concept

This chapter provides detailed insights on how the objectives of this thesis (see Section 1.2) will be achieved, including a description of the dataset and the ML model training pipeline.

3.1 Data

The refractive prediction error dataset consists of 2626 eyes from five different studies conducted by five different surgeons and includes four types of IOLs (Vivinex, SN60WF, ZCB00, AAB00). The dataset consists of two categorical and six numerical features including biometry of the eye and the power of the implanted IOL. The categorical features are the type of IOL and the center in which the IOL was implanted. The numerical features are K, AL, CCT, ACD, LT, and IOL power (PIOL). In addition to these six numerical features, the dataset includes the SEQ and the predSEQ using different formulas, as well as the corresponding PE. The formulas included in the dataset are Castrop, Haigis, Hoffer-Q, Holladay-3, and SRKT, with constants optimized for each study

The IOLMaster 700 (Carl zeiss, Oberkochen, Germany) was used for biometry measurements in all five studies. Univ.-Prof. Dr. Achim Langenbucher prepared and provided the dataset for this thesis, which only included patients who met certain inclusion criteria, such as post-operative visual acuity of 20/25 or better and no other ocular diseases besides cataracts.

3.1.1 Data Insight

The following section provides a deeper understanding of the refractive prediction error dataset by revealing biases and important information for model training. Additionally, the existing formulas within the dataset are compared.

Ratio of Cases within Limits of Mean Absolute Prediction Error

Figure 3.1 illustrates the ratio of cases within different limits of mean absolute PE. The limits used in this analysis (0.25 D, 0.5 D, 0.75 D, 1.0 D) are recommended by current research (as discussed in section 2.1.1).The figure shows that the Castrop formula outperformed all other formulas for all four limits. Statistical tests confirmed this observation. Cochran Q tests showed, that there are significant differences between the five formulas ($p < 0.001$) and subsequent post hoc pair-wise McNemar tests with Bonferroni correction showed, that the Castrop formula resulted in a higher number of samples within each of the four limits ($p < 0.002$).

While the PE of all five formulas is present in the dataset, the primary focus is on the PE of the Castrop formula. This is because it was suggested by Univ.-Prof. Dr. Achim Langenbucher and because it has been shown to have better performance than the other formulas, as seen in Figure 3.1 and discussed in Section 2.1.2.

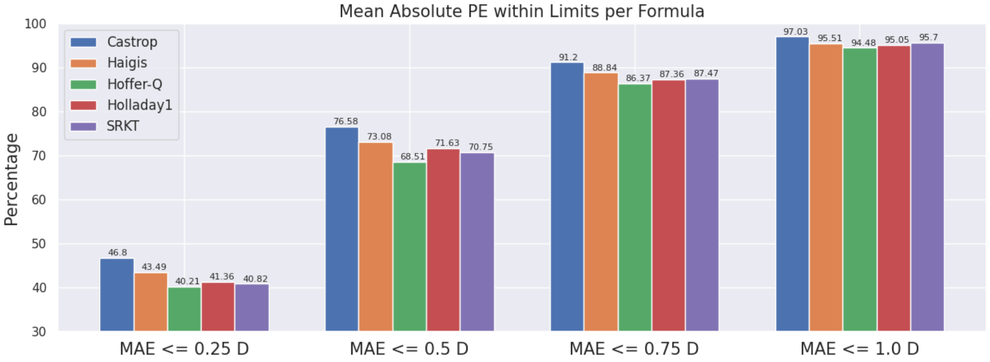


Figure 3.1: Ratio of cases within different limits of mean absolute PE for the five formulas present in the refractive prediction error dataset. It can be seen that the Castrop formula resulted in the highest number of samples within each of the four limits.

Prediction Error per Study

The refractive prediction error dataset includes data from five different studies that used four different IOL types. The IOLs were implanted at different centers. Table 3.1 summarizes the studies, indicating the IOL type, the implantation center, and the number of samples. The ZCB00 lens used in study 4 was implanted at two different centers.

Study 1, which used the Vivinex lens, had the most samples within the limits of 0.25 D, 0.5 D, 0.75 D, 1.0 D of mean absolute PE, as shown in Figure 3.2. A t-test with Bonferroni correction confirmed, that the MAE of study 1 was significantly smaller than the one of study 2 ($p < 0.001$), 4 ($p < 0.001$), and 5 ($p < 0.001$). Another t-test with Bonferroni correction reported that study 2 had a significantly smaller MAE than study 4 ($p = 0.009$) and 5 ($p < 0.001$).

Study	Lens type	Center	Count (ratio)
1	Vivinex	Castrop	588 (22.39%)
2	AAB00	Rosenheim	951 (36.21%)
3	AAB00	Castrop	54 (2.06%)
4	ZCB00	Castrop / DMEI	363 (13.82%)
5	SN60WF	DMEI	670 (25.51%)

Table 3.1: Summary of the five studies of which data are aggregated to create the refractive prediction error dataset. The table displays the IOL type, the implantation center, and the number as well as the ratio of samples.

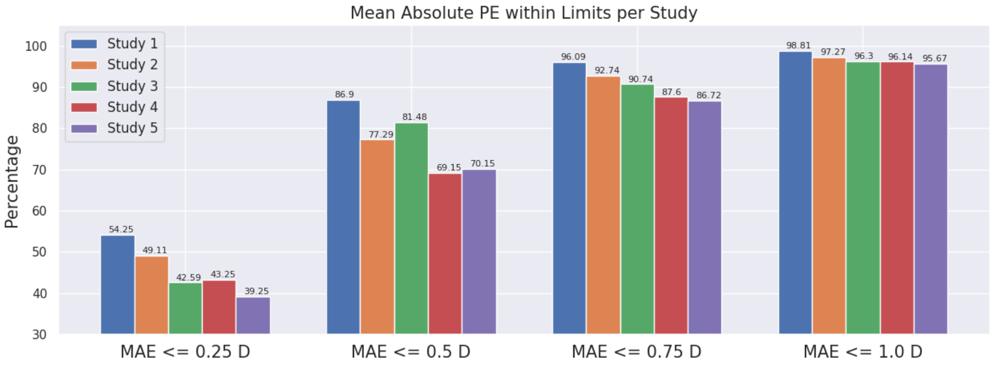


Figure 3.2: Ratio of cases within different limits of mean absolute PE for the five studies included in the refractive prediction error dataset. It is evident, that samples from study 1, in which a Vivinex IOL was used, were more likely to fall within each of the four limits.

Descriptive Statistics Benchmark of Formulas

Table 3.2 summarizes the MAE, MedAE, and SD of the PE of the different formulas included in the refractive prediction error dataset. These descriptive statistics were calculated using the entire dataset.

Formula	MAE	MSE	SD
Castrop	0.339	0.27	0.444
Haigis	0.369	0.3	0.48
Hoffer-Q	0.404	0.33	0.52
Holladay1	0.383	0.309	0.493
SRKT	0.387	0.313	0.500

Table 3.2: The MAE, MedAE, and SD for the Castrop, Haigis, Hoffer-Q, Holladay, and SRKT formulas included in the refractive prediction error dataset.

Correlation Heatmap

Figure 3.3 shows the correlation between the different features and the predSEQ as well as the PE of the different formulas. It is noteworthy that the PE of the Castrop formula (d_predSEQ_CHR1) does not correlate with any of the features. Since the Pearson correlation coefficient (Boslaugh and Watters, 2008) was used, this indicates that there is no linear relationship between these variables. Therefore, it does seem appropriate to use a linear model to predict the PE of the Castrop formula.

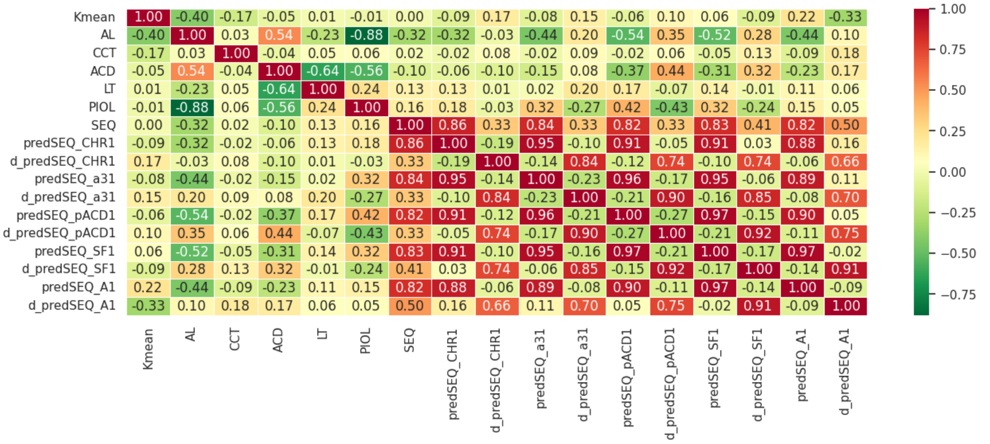


Figure 3.3: A heatmap showing the correlations between the features and targets in the refractive prediction error dataset, based on the Pearson correlation coefficient.

Plotting the features individually against the target confirms the very low correlation and does not reveal any obvious non-linear relationships (see Figure 3.4). Most of these plots appear as evenly distributed clouds of points, similar to the plot of K shown in Figure 3.4. However, the plot of predSEQ_CHR1 in the same figure shows some signs of heteroscedasticity, meaning that the variance of the target appears to increase with predSEQ_CHR1. To test this hypothesis, a Breusch-Pagan test (Breusch and Pagan, 1979) was performed. The null hypothesis of the Breusch-Pagan test is that the variance of the residuals does not depend on x. This null hypothesis was rejected ($p = 0.008$), providing sufficient evidence that heteroscedasticity is present.

In addition to the features shown in the correlation heatmap (see Figure 3.3), the radius of the cornea curvature (R) was also included in the refractive prediction error dataset. However, since R can be converted into K with the equation $K = 337.5/R$ for R in mm, these two features have a correlation of 1.0. Most of the literature reviewed in Chapter 2 reports only one of these two features. Therefore, it was decided to use only the feature K for the experiments conducted in this thesis.

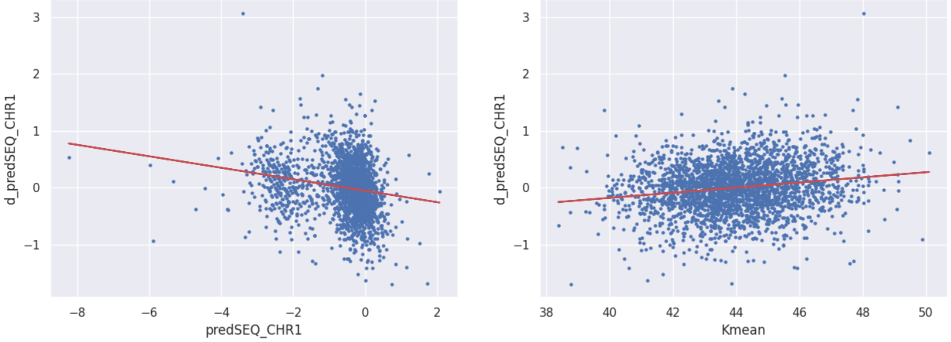


Figure 3.4: A scatter plot with a linear regression line showing the relationship between the feature predSEQ_CHR1 and the target (left) and another one showing the relationship between K and the target (right). The right plot does not show any obvious linear or non-linear relationships, but the left plot exhibits heteroscedasticity.

Descriptive Statistics

Table 3.3 presents the descriptive statistics of the numerical variables included in the refractive prediction error dataset.

N = 2626	**K in D**	**AL in mm**	**CCT in mm**	**ACD in mm**	**LT in mm**	**PIOL in D**	**SEQ in D**	**PE in D**
Mean	43.995	23.878	0.557	3.177	4.608	20.605	-0.541	0.002
SD	1.699	1.583	0.0378	0.432	0.437	4.292	0.867	0.444
Median	43.928	23.649	0.556	3.180	4.600	21	-0.340	0.008
Minimum	38.370	20.193	0.409	1.690	2.875	3	-7.700	-1.687
Maximum	50.105	31.330	0.694	4.830	6.250	34	2	3.070
Quantile 5%	41.170	21.580	0.495	2.468	3.900	12	-2.38	-0.717
Quantile 95%	46.848	26.858	0.694	3.860	5.328	27	0.5	0.712

Table 3.3: Descriptive statistics of the entire dataset, including the mean, SD, median, minimum, maximum, 5 %, and 95 % quantiles (90 % confidence intervals). The column PE in D represents the PE of the Castrop formula. Note that the CCT statistic was not computed for all 2626 samples, but only for those samples for which CCT data was available.

3.1.2 Data Quality Assessment

Data quality assessment (DQA) is the process of evaluating data scientifically and statistically to determine whether they meet the quality requirements for the intended task (Tozzi, 2021). This is especially important for real-world data samples, such as those in the refractive prediction error dataset. The DQA for the refractive prediction error dataset is described in Chapter 5.3.

3.1.3 Data Splits

To evaluate the performance of the final algorithm on new, unseen data, the dataset was split into a training set with 2363 eyes (90 % of the data) and a testing set with 263 eyes (10 % of the data). This is also known as hold-one-out cross-validation. Since the distribution of the PE is imbalanced, as seen in Figure 3.1, stratification was used for the splits to ensure that the training and testing sets have similar ratio of PE within the different benchmark limits. This is especially important for classification, as the main goal of the classifier is to detect refractive surprise, which is the minority class. Therefore, this class must be present in the testing set, for realistic results. Additionally, the splits were stratified according to the different studies, as there is a significant difference in MAE among them (see Figure 3.2). In the training set, the ratio of absolute PE within 0.25 D, 0.5 D, 0.75 D, 1.0 D and greater than 1.0 D was 46.8 %, 29.79 %, 14.6 %, 5.84 %, and 2.96 %, respectively. For the testing set, the ratios were 46.77 %, 29.66 %, 14.83 %, 5.7 %, and 3.04 %, respectively. The ratios of study 1, study 2, study 3, study 4, and study 5 in the training set were 22.39 %, 36.18 %, 2.07 %, 13.84 %, and 25.52 %, respectively. For the testing set, the ratios were 22.43 %, 36.5 %, 1.9 %, 13.69 %, and 25,48 %, respectively. Table 3.4 summarizes the training and testing set. The training set was further divided into a training and validation set with a ratio of 0.2, using the same stratification. Figure 3.5 illustrates the data splits together with the overall ML model training pipeline concept.

Column	**Training set (mean±sd)**	**Testing set (mean±sd)**	***p* Value**
Count	2363	263	n.a.
K in D	43.982±1.792	44.107±1.642	0.258
AL in mm	23.880±1.594	23.853±1.485	0.792
CCT in mm	0.557±0.038	0.559±0.039	0.489
ACD in mm	3.178±0.430	3.174±0.453	0.898
LT in mm	4.605±0.435	4.636±0.452	0.273
P_{IOL} in D	20.618±4.320	20.494±4.035	0.658
SEQ in D	-0.540±0.870	-0.546±0.843	0.914
PE in D	0.005±0.444	-0.018±0.443	0.431

Table 3.4: The summary of the training and test data, including the mean and SD. Unpaired *t*-tests showed, that there is no significant difference between the two datasets in any of the features.

3.1.4 Categorical Encoding

To use the refractive prediction error dataset for machine learning, the categorical features (type of IOL, center of implantation) must be converted to numerical vectors, so that they can be represented as coordinates in a coordinate system, which is a requirement for most ML algorithms.

There are two main ways to encode categorical features for use in ML: Label Encoding and One-Hot Encoding. Label Encoding assigns each category a unique integer value. The problem of this is, that this can create an artificial ordinal relationship between the categories if one does not actually exist and a model will be likely to learn this relationship. If there is no actual ordinal relationship between categories, this relationship should not be created artificially. One-Hot Encoding creates a separate binary column for each category, which avoids the issue of artificial ordinal relationships, but can suffer from the curse of dimensionality if there are many categories, as the representation becomes very high-dimensional and sparse (Sethi, 2020)

Since the refractive prediction error dataset only contains two categorical features, each with four or three categories, the curse of dimensionality can be considered insignificant. Furthermore, there is no ordinal relationship between these categories. Therefore, One-Hot Encoding is the appropriate choice. To avoid the dummy variable trap of One-Hot Encoding, which can cause high multicollinearity, one dummy variable will be dropped for each categorical feature.

3.1.5 Normalization

Different scaling in different features can affect the similarity of two samples. For example, a 15 % change in K with mean 43.982 will affect the similarity of two samples much more than a 15 % change in ACD with mean 3.178. This means that K is much more dominant in determining similarity. This dominance among features can distort the results of a ML algorithm, which is based on distances or similarities (Scikit-Learn, 2011). To address this, the refractive prediction error dataset should be normalized so that each features has approximately the same scaling.

There are two main normalization algorithms: Min-Max Normalization and Z-Score Normalization. Min-Max Normalization transforms the feature space into the range of [0, 1], where 1 represents the largest and 0 the smallest value. This allows for percentage interpretation, but it cannot ensure that a value larger than the maximum value in the training data will not occur. Hence, Min-Max Normalization is unusable for supervised learning. Z-Score normalization, on the other hand, transforms each feature so that it has mean 0 and a variance of 1. The disadvantage of this method is that it generates negative values and makes interpretation harder, but the advantage is, that it can be used for supervised learning and becomes more stable with larger initial datasets.

Since the refractive prediction error dataset will be used for supervised learning, Z-Score Normalization is the appropriate choice. Only the features need to be scaled, as there is only one target (the PE of the Castrop formula).

3.2 Principal Component Analysis

Principal component analysis (PCA) is an unsupervised learning algorithm in which the input data is unlabeled and the structure of the data is learned without any assistance. One common task in unsupervised learning is dimensionality reduction, and PCA is a frequently used method for this purpose. Dimensionality reduction can help with data visualization and can also address multicollinearity in the data, preparing it for supervised learning. (Kashnitsky, 2019)

The correlation heatmap (described in Section 3.3) showed that multicollinearity is present in the data. For example, the feature AL correlates with PIOL (0.88) and LT slightly correlates with ACD (-0.64). PCA can help evaluate the features that explain most of the variance and are therefore likely to be useful for predicting the Castrop PE. The PCA for the postoperative refractive error dataset is described in Chapter 5.4.

3.3 Handling Imbalanced Data

If the PE of the refractive prediction error dataset is split into multiple classes based on given thresholds, as shown in Figure 3.1, a severe skew in the class distribution is visible. Many ML algorithms are influenced by such a class imbalance in the training set, causing some to ignore the minority class entirely. This is a problem because in the refractive prediction error dataset, it is the minority class on which predictions are most important. Therefore, class imbalance will be addressed using some of the techniques discussed in Section 2.3.

3.4 Prediction of Refractive Surprise

As already stated in the original project description (see Appendix A), can refractive surprises be predicted either by classification or continuously by regression. During this work both regression models and classification models will be trained. The regression models include a dummy regressor, a support vector regressor (SVR), a decision tree regressor (DTR), a random forest regressor (RFR), and a NN. The classification models include a dummy classifier, logistic regression, a support vector classifier (SVC), a decision tree classifier (DTC), a random forest classifier (RFC), and also a NN.

3.5 Intraocular Lens Power Calculation

The ultimate goal of predicting refractive surprise is to increase performance of the Castrop formula or, in general of IOL power calculation. Another way to achieve this goal is to directly predict the postoperative SEQ based on patient characteristics and IOL power, which can help with prediction of PE. Therefore, this work will propose an IOL power calculation formula based on ML. The following models will be trained on the training set and evaluated on the validation set: dummy regressor, SVR, DTR, RFR, GBR, and NN.

3.5.1 Formula Constant Optimization

The performance of the trained models will be compared to the Castrop formula and the SRKT formula. This will provide insight into how the new formula performs against a modern high-performing formula and an older, less effective standard formula. To fairly compare the formulas, the constants of the Castrop and the SRKT formula need to be optimized separately for each study and dataset. These formulas will be implemented in Python and the constants will be optimized using the Levenberg-Marquardt algorithm with a mean squared error (MSE) loss function. Figure 3.5 illustrates the integration of constant optimization into the overall ML model training pipeline concept.

The refractive prediction error dataset already includes constants that were optimized by Univ.-Prof. Dr. Achim Langenbucher on the whole 2626 samples. These constants will be used to verify that the optimization in Python was successful. The formulas should at least perform as well with the newly optimized constants as with the existing ones. The formula implementations will also be verified. Given the same constants, the Python implementation of the formulas should yield the exact same predSEQ value as the one already present in the dataset.

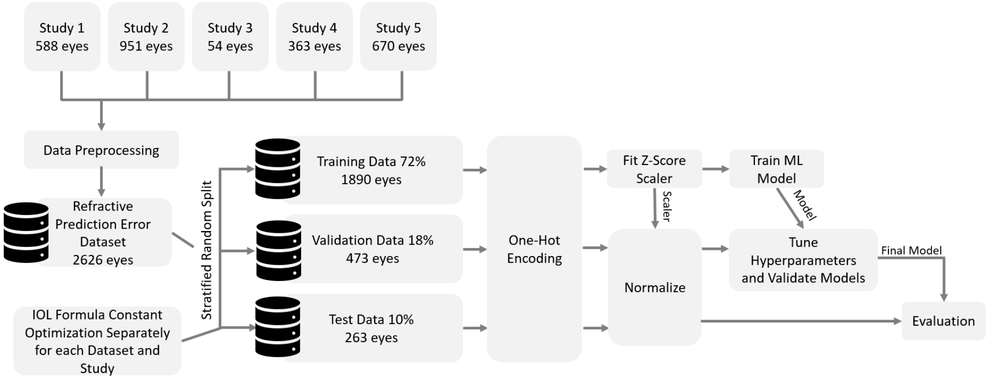


Figure 3.5: The overall ML model training pipeline concept.

3.5.2 Ensemble

The trained models for predicting refractive surprise can be useful for IOL power calculation. For example, the output of the Castrop formula and the prediction of the PE of the Castrop formula could be stacked. This concept is based on the idea, that the predSEQ of the Castrop formula should become more accurate if the predicted PE is added. However, the PE prediction will probably not be perfect, so the operation of fully leveraging the PE prediction to improve the performance of the Castrop formula may be more complex than simple addition. A second-level model can be trained on the output of the Castrop formula and the PE prediction, to learn this operation. The Castrop formula and the PE prediction form together the first-level models.

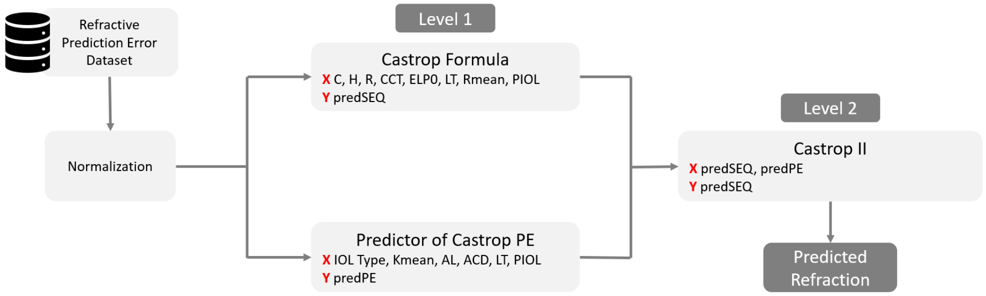


Figure 3.6: A stacking ensemble machine learning algorithm consisting of two levels. The first level consists of the Castrop formula and a model that predicts the PE of the Castrop formula. The second level is a model trained on the output of the fist-level models and predicts the SEQ.

CHAPTER 4

Methodology

This chapter describes the methodology and course of action used to carry out this project.

4.1 Project Planning and Procedure

As this project is primarily exploratory research, it was conducted iteratively and incrementally, as outlined in the original project description (see Appendix A). This section provides an overview of the individuals involved and the various methods used to manage this project.

4.1.1 Organization

The following table 4.1 shows all involved persons and their role.

Name	**E-Mail**	**Role**
Univ.-Prof. Dr. Achim Langenbucher	achim.langenbucher@uks.eu	Principal
Dr. sc. ETH Andreas Streich	andreas.streich@hslu.ch	Advisor
Dr. Rémi Janner	remi.janner@ckw.ch	External expert
Boas Meier	boas.meier@stud.hslu.ch	Bachelor candidate and author of the thesis

Table 4.1: An overview of the individuals involved in this project, including their name, email, and role.

4.1.2 Stand-up Meetings

Weekly stand-up meetings were conducted to verify progress, identify potential issues early, and gather feedback for improvement. The feedback, findings, and progress updates from these meetings are recorded in stand-up meeting minutes (see Appendix B). The progress updates were shared with the advisors in advance of each meeting.

4.1.3 Roadmap

A roadmap (see Figure 4.1) was created ath the start of the project to help verify the current state of the project. This roadmap consists of six focus points and six milestones. The milestones (see Appendix C) do not outline domain-specific results, as the outcome and direction of the project are difficult to predict. Instead, they outline a high-level ML workflow and indicate when each step should be started respectively completed. This helps to keep the project on track and avoid time pressure at the end. The roadmap was adapted several times during the project. Figure 4.1 shows the final roadmap. The initial versions and the chronological evolution of the roadmap can be traced in the stand-up meeting minutes (see Appendix B).

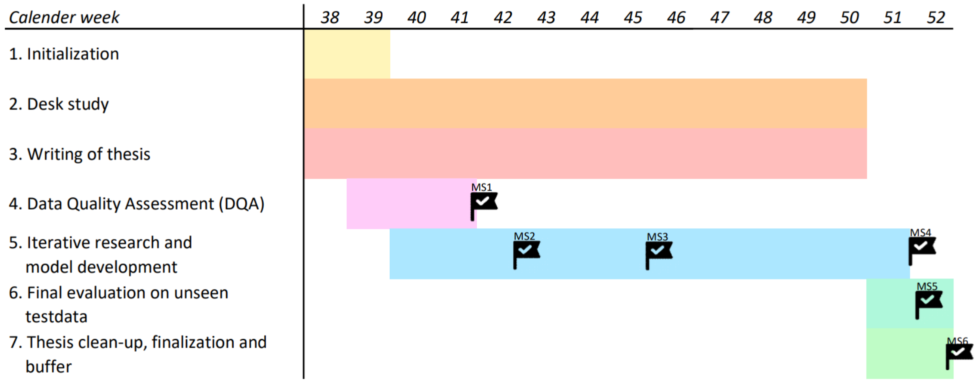


Figure 4.1: The final roadmap consists of six focus points and six milestones.

4.1.4 Risk Management

Potential risks were continuously analyzed during the project. Each stand-up meeting minutes includes in addition to the progress update and the feedback notes as well a risk update. The risk update includes newly introduced risks, updated risk scores, potential mitigations, and the current top 3 risks. Each risk has an id, title, description, likeliness L on a scale from 1 to 5, and severity S on a scale from 1 to 5, with 5 being the most likely or severe. A risk score R is calculated as:

$$R = L \cdot S \tag{4.1}$$

which is used for prioritization. Mitigation measures were defined for every risk with $R >= 10$.

Appendix D contains a complete overview of all risks, including their final risk scores before and after mitigation. Some risk scores were updated during the project.

These changes were documented and justified in the stand-up meeting minutes (see Appendix B).

4.2 Research Methodology

As seen in the roadmap (see Figure 4.1), a significant portion of the project is dedicated to iterative research and model development. To approach this part of the project systematically, a research cycle was developed that is well-suited to this project. Each iteration of this research cycle consists of five stages, as shown in Figure 4.2. This figure also illustrates some of the tools and technologies used during each stage.

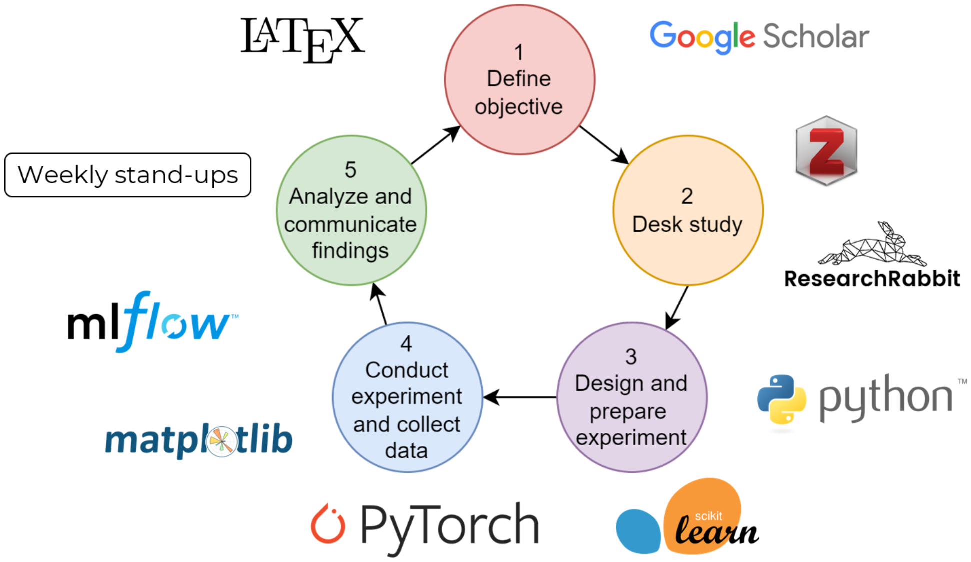


Figure 4.2: The iterative research cycle developed and used for this project consists of five stages. Each iteration begins with the definition of an objective.

A research cycle always started with the definition of an objective or hypothesis. In the second step, relevant literature was studied and any missing theory was worked up [1]. During this stage, tools such as Google Scholar, Zotero, and ResearchRabbit (ResearchRabbit, 2022) were used. ResearchRabbit was particularly helpful for displaying where relevant papers fit into the larger context, allowing for the easy listing of references and citations to find more recent work. The third stage involved the design and preparation of experiments, using tools such as the Python programming language and ML libraries such as Scikit-Learn (Pedregosa et al., 2011) for conventional ML algorithms and PyTorch (Paszke et al., 2019) for NNs. A complete list of all instruments, programming languages, and libraries used, along with their software versions, can be found in Chapter 5. The fourth stage of the research cycle involved

[1] The EyeWiki of the American Academy of Ophthalmology (EyeWiki, 2022) and the Optometrists Network (OptometristsNetwork, 2022) were mainly used to work out the theory and look up foreign words about cataracts and ophthalmology. These are considered reliable resources because the authors, the release date as well as potential reviewers and the date of the last review are known.

executing the prepared experiments and collecting the data using the Python library Mlflow (Zaharia et al., 2022). Details on the Mlflow stack can be found in Section 5.2. Finally, the findings and results of the experiments were analyzed and discussed in the weekly stand-up meetings. These discussions often yielded new hypotheses and objectives, leading to the start of the next cycle.

CHAPTER 5

Realization

This chapter leads through the realization of the refractive error prediction models and the ML-based IOL power calculation formula. It mainly follows the concept, with any deviations explicitly noted.

5.1 Tech Stack

The Python programming language with various libraries was used to execute experiments and develop the models. Figure 5.1 provides an overview of all relevant instruments, programming languages, and libraries used in this thesis, along with their specific software versions. These exact versions of the libraries were used for each experiment. If a different library was used for a specific experiment, this will be noted.

Name	**Version**	**Purpose**	**Rationale**
Python	3.10.6	Programming language	Is widely used for Machine Learning, has a rich package ecosystem and is dynamically typed.
Matplotlib	3.5.3	Creating visualizations in Python.	Oldest and most popular plotting library for Python.
Mlflow	1.29.0	Tracking of experiments.	Is open-source and has great integration with various ML libraries.
Mlxtend	0.21.0	Performing Cochran Q and Mcnemar tests.	Good documentation.
Numpy	1.23.3	Mathematics extension.	Is Very popular and needed by many other libraries.
Pandas	1.4.4	Data manipulation and analysis.	Most popular.
Pandas-profiling	3.3.0	Creating comprehensive HTML reports about datasets.	Great integration with Pandas.
Scikit-learn	1.1.2	ML library.	Is open-source and includes implementations of many conventional ML algorithms.
Scipy	1.9.1	Library for scientific and technical computing.	Needed by Scikit-learn.
PyTorch	1.13.0	ML library with great support for NN.	Widely adopted among researches. The control over the training loop allows Mlflow instrumentation.
XGBoost	1.7.2	Gradient boosting library.	Great community. Similar API as Scikit-Learn.

Table 5.1: Versions and rationale of all relevant instruments, programming languages and libraries used during this thesis.

5.2 Model Experiment Tracking

Mlflow (Zaharia et al., 2022) was used to manage all relevant aspects of the ML lifecycle in this work. The main purpose of using Mlflow was to systematically track all model experiments via the Mlflow Tracking API. The API was used to persist train, validation, and test metrics, hyperparameters, and custom tags, including information such as random seeds, used features, data split ratio, used normalization methods and everything else to enable reproducibility of the results. A relational database was used for persistence, allowing for sophisticated querying. The Mlflow Tracking API was also used to store trained models, associated pip environments, and related plots and visualizations on an FTP server. Figure 5.1 illustrates the architecture of the Mlflow stack, which was containerized using Docker and Docker Compose (Merkel, 2014). The corresponding Docker- and docker-compose-files are available via the GitLab repository (Meier, 2022), which is hosted on the HSLU Enterpriselab.

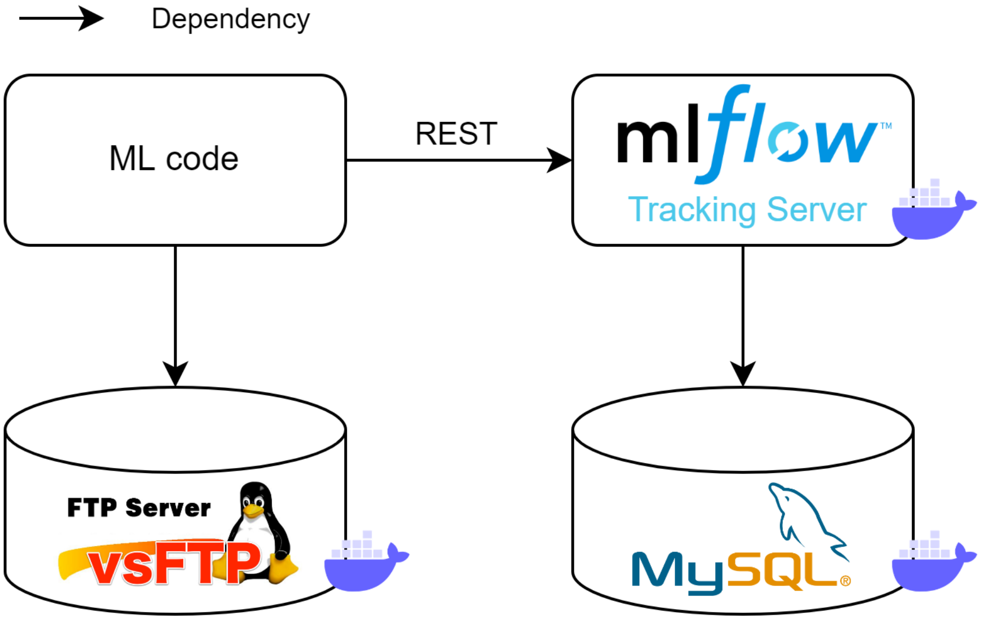


Figure 5.1: A diagram of the containerized Mlflow stack architecture for tracking model experiments. The stack includes a Mlflow server, a MySQL database (Oracle Corporation, 1995) for storing metrics and hyperparameters, and an FTP server (Evans, 2001) for storing file-based artifacts.

5.3 Data Quality Assessment

Data quality was determined based on completeness, validity, and consistency using the pandas-profiling library (Brugman, 2022). Overall, the data quality was considered very good, with only minor issues present. All the 2626 provided records were found to be of good quality, meaning that 100 % of the records are usable.

5.3.1 Completeness

Completeness was evaluated by checking every column for missing or null values and duplicates. The pandas profiling report showed that there are 951 missing CCT values. After confirming with Univ.-Prof. Dr. Achim Langenbucher, it was determined that these values were not measured by one of the surgeons. As these samples make up just over a third of the data, they cannot be dropped. Therefore, the developed ML algorithm must either be able to handle missing CCT values or the CCT feature must be dropped. Various experiments showed that the CCT feature did not affect performance of refractive surprise classification (see Section 5.6.3) or IOL power calculation (see Section 5.7.3). As a result, it was decided to drop the CCT feature.

5.3.2 Validity

To determine the validity of records, various checks were performed. First, the data types were checked to ensure that categorical features did not contain cells with numeric values and that numeric features did not contain cells with strings or categories. Additionally, the range of values was checked. The pandas profiling report did not reveal any invalid data types or formats. However, reviewing the boxplots revealed a data point with a SEQ of approx. -8 D. Upon further clarification, it was determined that this is a valid data point.

The boxplots also revealed, that most of the R1 and R2 values had the same value, while R which is defined as $R = \frac{R1+R2}{2}$ and K which is defined as

$$K = \frac{\frac{337.5}{R1} + \frac{337.5}{R2}}{2} \tag{5.1}$$

were mostly distinct. Further investigation revealed, that the features R1 and R2, representing the radius of the cornea curvature, had been overwritten by the R constants of the Castrop formula, which had the same column name. However, as it was decided that the feature R would not be used (see Section 3.1.1), no further action was taken on this issue.

5.3.3 Consistency

A data item is consistent if all representations of that item across all rows match. The DQA did not reveal any inconsistencies.

5.4 Principal Component Analysis

The principal components of the refractive prediction error dataset were evaluated using the Python package Scikit-learn. Figure 5.2 shows that the three classes $<= 0.5D$, $> 0.5D$ & $<= 1.0D$, $> 1.0D$ do not clearly separate in the space of the first three principal components. Due to the heavy overlap of the classes, it is to expect that classifiers will have difficulty distinguishing them. To determine if class separation may occur in higher dimensions, all permutations of the first five principal components were visualized as 2D scatter plots and analyzed. None of them revealed class separation, and samples with positive and negative PE did not separate either.

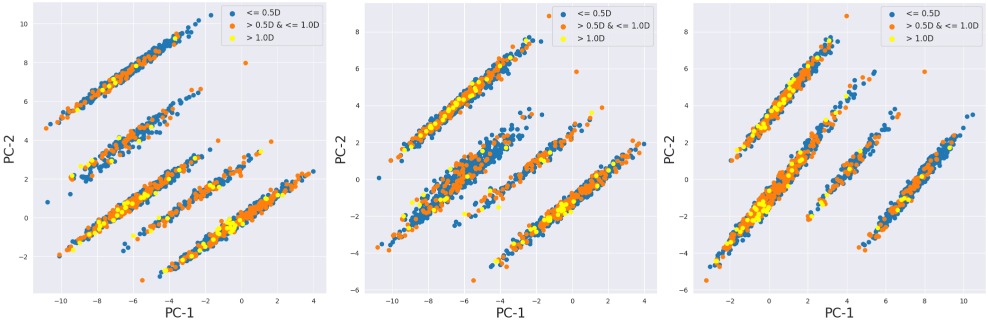


Figure 5.2: Plots showing the relationship between the first three principal components (PC-1, PC-2, and PC-3), where points are colored according to three classes indicating different severity of refractive surprise. The PCA was fitted only on samples whose absolute PE was within 0.5 D.

Although the classes did not separate, there are five clusters visible in Figure 5.2, with two of them seeming to have more points of the $> 1.0D$ (yellow) and the $> 0.5D$ & $<= 1.0D$ (orange) classes. Since the refractive prediction error dataset was aggregated from five different studies, it is likely that these five clusters represent them. To confirm this hypothesis, the points of the 2D scatter plot of the first two principal components were colored according to the five different studies (see Figure 5.3). It becomes immediately apparent that the clusters with seemingly more yellow and orange points belong to study 2 and study 5. The significant difference between the studies has already been analyzed in Section 3.1.1 and it has been shown that study 1 has a significantly smaller MAE than study 2 and 5. Thus, the observation that clusters 3 and 5 have more refractive surprise than cluster 1 can be confirmed. However, the observation that study 2 seems to have more refractive surprise than study 4 is misleading. Examining the mean absolute PE within limits per study, as discussed in Section 3.1.1, showed that study 2 has significantly less refractive surprise than study 4.

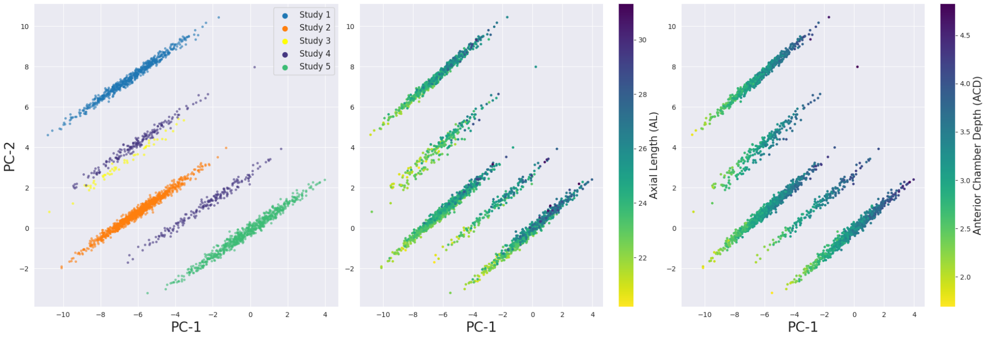


Figure 5.3: Plots showing the relationship between the first two principal components with different third variables. The third variables used are the study (left), AL (middle), and ACD (right).

However, the main conclusion that can be drawn from these PCA plots is that the study, or rather the used IOL type and the center of implantation, are the features explaining most of the variance in the refractive prediction error dataset. Plotting the importance of each feature for the first principal component, confirms this observation (see Figure 5.4), with the most important feature being DMEI, which represents a center of implantation. The second, third, and fourth most important features are AL, SN60WF, and ACD, respectively. Figure 5.3 visualizes AL and ACD as third numeric variables in the 2D scatter plot of the first two principal components.

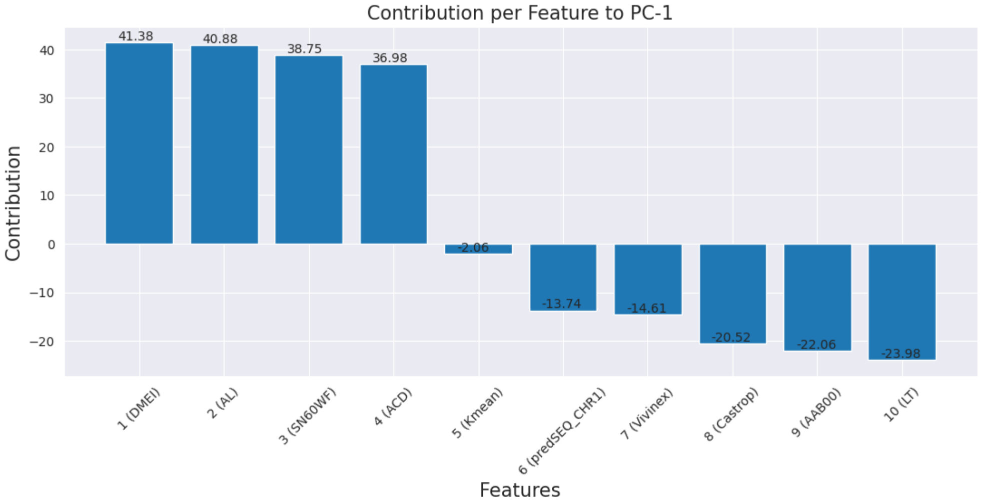


Figure 5.4: The feature importance to the first principal component. The most important feature is DMEI, followed by AL as the second most important feature.

Figure 5.5 shows that the first two principal components already explain 50.51 % of the variance of the refractive prediction error dataset, with the first principal component accounting for 28.42 %.

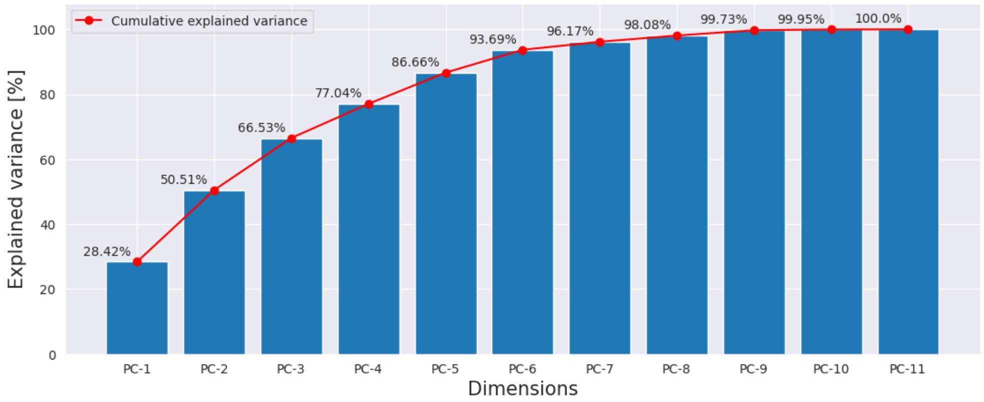


Figure 5.5: A bar plot showing the principal components with their cumulative explained variance.

5.5 Refractive Surprise Regression

This section presents the regression analysis for predicting refractive surprise after cataract surgery. Various models were compared and evaluated under similar conditions. Many experiments were conducted to find the best suited model for the regression of the Castrop PE. Different model architectures with increasing complexity were tested, as described in Section 3.4. To properly compare the different models, the same input features were used for all of them: the lens type, the implantation center, biometry data of the eye (including AL, ACD, and LT), the PIOL, and the predicted SEQ of the Castrop formula.

Dummy Regressor

First, a dummy regressor was implemented as a baseline to improve upon. This regressor simply predicts the mean of the data. The R2-Score of a model that always predicts the mean should be around 0, as it indicates how much of the variance of the target is captured by the model. The dummy regressor achieved a MAE of 0.338, a MSE of 0.206, and an R2-Score of -0.001. Table 5.2 summarizes all the training and validation results.

Decision Tree Regressor

The next model trained was a DTR. With default parameters the depth of the tree is not restricted at all resulting in perfect overfitting of the training data. To make the tree more conservative, the best model parameters were determined using k-Fold cross-validation on the combination of the training and validation sets. The parameter space to search in consisted only of the min-sample-split parameter, which determines the minimum number of samples required to split an internal node of the tree. The default value for min-sample-split is 2, allowing the tree to split a node that only consists of two samples, what can result in a separate node for each sample of the training data. The random seed was set to 42 to ensure reproducible results. For all other parameters the default value was used, as they were reasonable. The k-Fold cross-validation grid search resulted in a best min-sample-split of 689. Table 5.2 shows the achieved training and validation results.

A handy feature of decision trees is that the exact rules of the tree and the feature importance can be accessed. The most important feature was the predicted SEQ of the Castrop formula with an importance of 49.0 % and the second most important feature was ACD with 27.12 % (see Figure 5.6). It is surprising that the predicted SEQ has a relatively high importance, as it only contributed very little to the first two principal components as seen in Section 5.4.

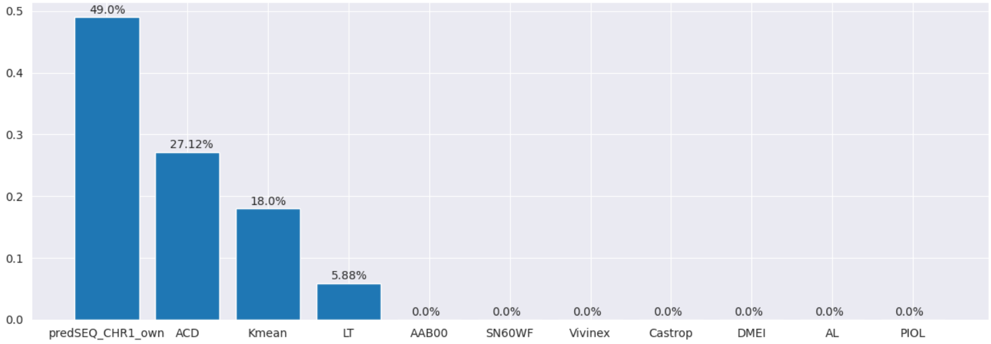


Figure 5.6: A bar plot showing the feature importance of the DTR for the Castrop PE regression.

Random Forest Regressor

Similar to the DTR the RFR overfitted the training data with default parameters. Therefore, the parameters were also tuned to make the model more conservative. Only the min-sample-split parameter was tuned, resulting in 236 being the best fit. Table 5.2 summarizes all the training and validation results. The feature importance of the RFR was very similar to that of the DTR and is not repeated here.

Support Vector Regressor

The SVR overfitted the training data as well slightly by default. Hence the squared l2 regularization, which can be configured with parameter C in Scikit-Learn, was tuned using k-Fold cross-validation and grid search. The best results were obtained with C = 0.22, where the strength of the regularization is inversely proportional to C and the default value is 1.0. Table 5.2 shows the metrics achieved on the training and validation sets.

Multi Layer Perceptron

The multi layer perceptron (MLP) is a simple deep learning architecture consisting of multiple fully connected layers. The initial model consisted of an input layer with 128 units, a hidden layer with 32 units, and a single output node, resulting in 5697 trainable parameters. The ReLU activation function was applied to the output of the input and hidden layers, and a MAE loss function was used. The AdamW algorithm was used for optimization, which is an improved version of the Adam optimizer that generalizes better (Loshchilov and Hutter, 2019). Before the depth and width of this architecture were further tuned, the batch size and learning rate were optimized. The best performance combined with the shortest training time was achieved using a batch size of 473, which corresponds to the length of the validation set, and a learning rate of 1e-3.

The performance of the initial architecture could be slightly improved by adding some dropout and adding a second hidden layer, resulting in 21’633 parameters. Figure 5.7 shows the model architecture. The MLP achieved slightly better performance on the validation set than the SVR, with less overfitting on the training set (see Table 5.2).

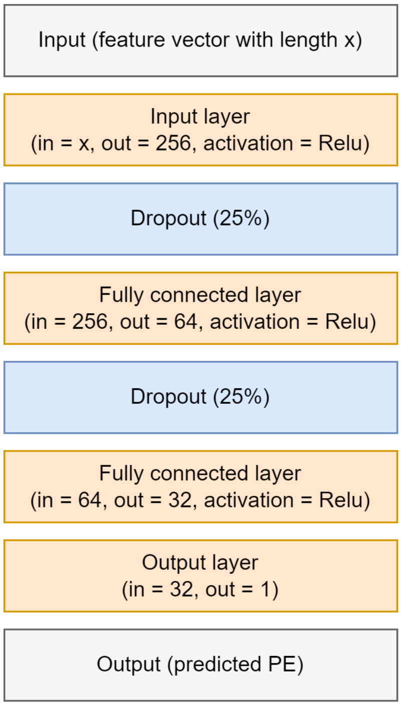

Input (feature vector with length x)
Input layer
(in = x, out = 256, activation = Relu)
Dropout (25%)
Fully connected layer
(in = 256, out = 64, activation = Relu)
Dropout (25%)
Fully connected layer
(in = 64, out = 32, activation = Relu)
Output layer
(in = 32, out = 1)
Output (predicted PE)

Figure 5.7: The model architecture of the MLP used for PE regression.

5.6 Refractive Surprise Classification

The performance of the PE regression was not satisfactory, so it was quickly switched to classification to see how ML models would perform at predicting whether the PE would be greater than a certain threshold. First, various models were explored using different thresholds and then a series of experiments were conducted to improve the model that showed the best performance during the initial exploration phase. The same features as described in Section 5.5 were used to train the following classifiers.

5.6.1 0.5 Dioptre Threshold Model Exploration

The first threshold was defined as 0.5 D. This threshold was chosen because it is a commonly used threshold in literature and because it results in only mild class imbalance, with 76.58 % of the PEs falling within 0.5 D. Therefore, if a sample has a

Dataset	Model	MAE	MSE	R2-Score
Train	Dummy	0.339	0.195	0.0
	DTR	0.325	0.177	0.095
	RFR	0.307	0.158	0.189
	SVR	0.301	0.156	0.2
	MLP	0.31	0.164	0.155
Val	Dummy	0.338	0.206	-0.001
	DTR	0.328	0.197	0.04
	RFR	0.316	0.186	0.094
	SVR	0.315	0.185	0.102
	MLP	0.31	0.18	0.122

Table 5.2: The Castrop PE regression training and validation results, including the MAE, MSE, and R2-Score values for each trained model.

PE greater than 0.5 D, it will be labeled as refractive surprise, and if not, it will be labeled as no refractive surprise.

Dummy Classifier

The first model trained was a Dummy Classifier, that always predicts the positive class, respectively classifies each sample as refractive surprise. This resulted in a recall of 1.0 and a precision of 0.237. This means that 100 % of all refractive surprises were classified correctly, but if the model predicts refractive surprise, it is only correct in 23.7 % of the time. In general, the recall can be considered the more important metric for PE prediction because it is safer to have a false positive than a false negative. It is preferable for a surgeon to take a little bit more preventive measures than too little. Table 5.3 summarizes all the training and validation results.

Logistic Regression

The next classifier trained was a logistic regression (LR). With default parameters, the precision and recall were both 0.0 with an accuracy of 0.761, which is the class ratio of cases within 0.5 D. This means that the LR always predicted the negative class, which is more dominant. This could be fixed by using the class weight parameter, which assigns different weights to the different classes, causing the classifier to consider some classes as more important. The class weights were calculated inversely proportional to class frequencies in the training data. If the y represents the list of class labels of the training data, the class weights off all classes present in y can be defined as following:

$$class\_weights = \frac{len(y)}{len(np.unique(y)) \cdot np.bincount(y)} \tag{5.2}$$

where $len(y)$ is the number of samples, $len(np.unique(y))$ is the number of classes, and $np.bincount(y)$ is a vector containing the number of each class. Using this formula on the training set resulted in a class weight of 0.652 for the negative and 2.143 for the positive class. Training the LR with these class weights resulted in an F1-Score of 0.378, which is better than before but worse than the dummy classifier (see Table 5.3).

Decision Tree Classifier

Since linear classifiers did not perform well, next a DTC was trained, which is a a non-linear model. With default parameters, the DTC also mostly ignored the minority class and heavily overfitted the training data. Therefore, the class weight parameter was provided as described in Section 5.6.1. The overfitting was addressed using similar hyperparameter tuning as described in Section 5.5. In addition to the min-sample-split also the max-tree-depth parameter was tuned and stratification was used for k-Fold cross-validation. The best performance was achieved with a max-tree-depth of 1. Examining the feature importance showed that the only feature used for classification was the boolean Vivinex, which indicates whether a Vivinex IOL was used or not. Using only this feature, the tree was able to correctly predict 87.5 % of refractive surprises but was only correct in 26.7 % of the time (see Table 5.3). The single rule the tree uses for the predictions was: if a Vivinex is used, there is no refractive surprise; otherwise, there is a refractive surprise (see Figure 5.8). As shown in Figure 3.2, the study that used the Vivinex lens had the most cases with an absolute PE within 0.5 D.

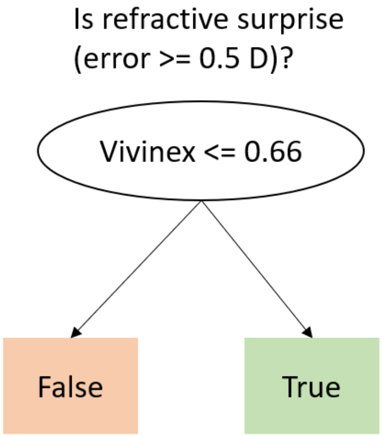


Figure 5.8: The single rule of the best DTC for PE classification with threshold 0.5 D.

Random Forest Classifier

Similar to the DTC the RFC was as well tuned using grid-search with stratified k-Fold cross-validation. To address the issue of the minority class being mostly ignored, the class weight parameter was used. The RFC achieved its best performance with a min-sample-split of 506, which resulted in more features being important than previously seen in Section 5.5. Figure 5.9 shows the feature importance of the RFC. Similar to the feature importance seen for the PE regression (see Figure 5.6), the categorical features indicating the lens type and the implantation center appear to be most important.

However, the importance is better distributed among all features and there is no feature with an importance of 0.0 %. The RFC performed worse than the DTC and only slightly better than the dummy classifier. However, no statistical tests were performed, to determine whether these differences are significant.

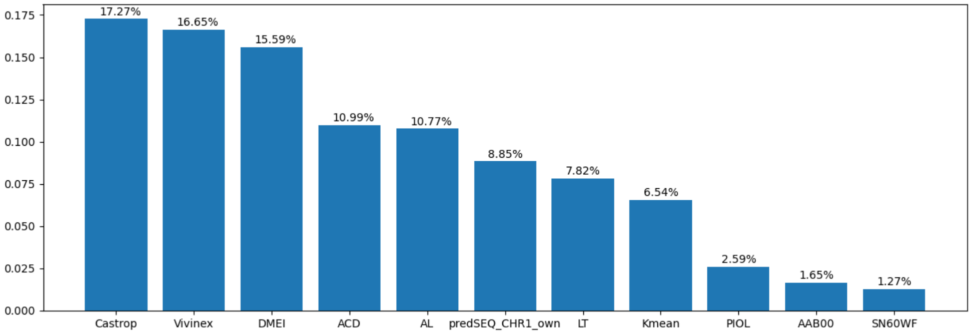


Figure 5.9: The feature importance of the RFC for the Castrop PE classification with 0.5 D threshold.

Support Vector Classifier

For the SVC the same class weight parameter was used as for the other classifiers and the l2 regularization parameter C was tuned using k-Fold corss-validation with stratification. The best performance was achieved with C = 0.1 (see Table 5.3).

Multi Layer Perceptron

The MLP was trained using the same procedure and model architecture as in Section 5.5. No activation function was applied to the output of the model because the binary cross entropy with logits loss function was used. This loss combines a sigmoid layer and the binary cross entropy loss into a single class, which is more numerically stable than using a plain sigmoid followed by a binary cross entropy loss. The log-sum-exp trick can be leveraged for numerical stability by combining the operations into one layer. (PyTorch, 2022). To prevent the minority class from being ignored due to class imbalance, the same class weight parameter was provided to the binary cross entropy with logits loss function. This multiplies the loss by the given class weight, ensuring the model learns more from the minority class. For optimization the AdamW algorithm was used. The best performance with the shortest training time was achieved using a batch size of 473 and a learning rate of 1e-3.

The MLP binary classifier performed best on the validation data, with an F1-Score of 0.416, without overfitting the training data (see Table 5.3). The model converged after about 4 epochs (see Figure 5.11). The state of the model that achieved the best validation F1-score was serialized and persisted using the Mlflow tracking stack. The F1-Score of the model, that achieved best F1-Score, could be slightly improved by post-prediction threshold tuning as described in Section 2.3.3. A threshold of 0.556 resulted in an F1-Score of 0.421 on the validation data (see Figure 5.10).

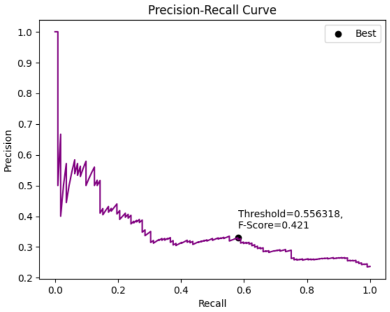


Figure 5.10: The precision-recall curve of the MLP binary classifier marked with the threshold resulting in the best F1-Score.

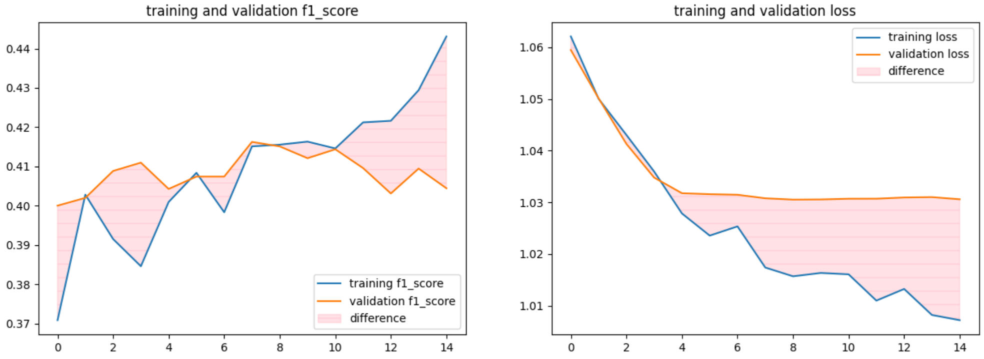


Figure 5.11: The training and validation F1-Score (left) and the training and validation loss (right) of the MLP binary classifier for Castrop PE prediction with 0.5 D threshold.

5.6.2 0.25 Dioptre Threshold Model Exploration

In the past, only thresholds of 0.5 D and 1.0 D were used as benchmarks for cataract surgery, but as the procedure has improved over time, thresholds of 0.25 D and 0.75 D have been added. The ratio of cases with an absolute PE within 0.25 D is 46.8 % for the refractive prediction error dataset. While cases outside the 0.25 D limit cannot be considered outliers, it would still be beneficial for a surgeon to know preoperatively if this is the case for a given patient. The following sections describe the classifiers that were trained and how. The same features as described in Section 5.5 were used.

Dummy Classifier

The dummy classifier achieved an F1-Score of 0.697 on the validation set by always predicting the positive class. The corresponding precision is 0.535 and the recall is 1.0

Dataset	Model	F1	Precision	Recall
Train	Dummy	0.378	0.233	1.0
	LR	0.399	0.295	0.617
	DTC	0.405	0.263	0.875
	RFC	0.437	0.332	0.639
	SVC	0.435	0.328	0.649
	MLP	0.415	0.294	0.711
Val	Dummy	0.383	0.237	1.0
	LR	0.378	0.289	0.545
	DTC	0.409	0.267	0.875
	RFC	0.389	0.3	0.554
	SVC	0.401	0.302	0.598
	MLP	0.416	0.291	0.732

Table 5.3: The Castrop PE binary classification training and validation results for a threshold of 0.5 D. For each model the F1-Score, precision, and recall of the positive class, which stands for refractive surprise, is reported.

(see Table 5.4).

Logistic Regression

The LR was tuned and trained according to section 5.6.1. The best results were achieved using the class weight parameter, and for all other parameters the default value.

Decision Tree Classifier

The DTC achieved an almost as good F1-Score as the dummy classifier, while having better precision and worse recall (see Table 5.4). The hyperparameter tuning resulted in a min-sample-split of 584. Other than with the 0.5 D threshold classification the tree is deeper and has more sophisticated rules (see Figure 5.12).

Random Forest Classifier

Hyperparameter tuning of the RFC resulted in heavy overfitting when maximizing the F1-Score. Despite the heavy overfitting, the model achieved a validation F1-Score of 0.581 (see Table 5.4). No other configuration was found that surpassed this value. When maximizing precision during stratified k-Fold cross-validation grid-search, the model resulted in an F1-Score of 0.486 with a precision of 0.594 and a recall of 0.411. A min-sample-split of 506, which gave best results for a RFC when performing 0.5

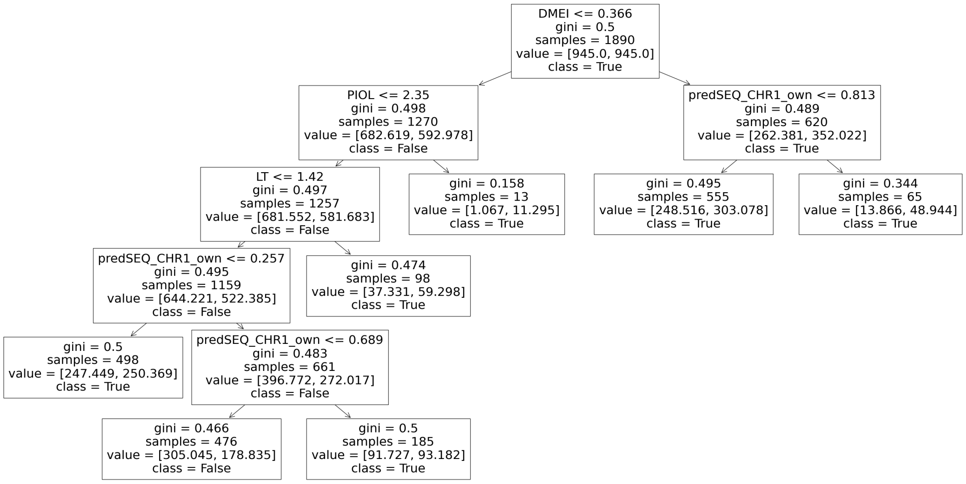


Figure 5.12: The rules of the DTC for PE classification with threshold 0.25 D.

D threshold PE classification, resulted in train and validation F1-Scores of 0.612 and 0.579, respectively.

Support Vector Classifier

For the SVC, a C parameter of 0.1 gave the best results. While the recall is low compared to the other classifiers, the SVC has second-best precision (see Table 5.4).

Multi Layer Perceptron

The model architecture, loss function, optimizer, learning rate, and batch size used to train the MLP were the same as in section 5.6.1. The model converged after about 10 epochs, and the state of the model at epoch 9 gave the best validation F1-Score, which was 0.598 (see Figure 5.14). Post-prediction threshold tuning after applying a sigmoid activation function to the logit output of the MLP classifier resulted in a threshold of 0.385 which gave an F1-Score of 0.699. However, with this threshold, the model almost always predicts the positive class. Figure 5.13 shows the precision-recall curve with the best threshold, as well as the confusion matrix of the predictions using this threshold.

Gradient Boosted Tree Classifier

Since tree models seemed to perform better than the MLP for PE prediction with a 0.25 D threshold, additional gradient boosting tree classifier (GBC)s were trained. Hyperparameter tuning with the goal of maximizing the F1-Score resulted in a GBC that always predicts the positive class, similar to the dummy classifier (see Section 5.6.2). The parameters tuned were min-child-weight, which is equivalent to min-sample-split used for DTC and RFC, alpha, which is l1 regularization, and lambda, which is l2 regularization. The optimal values achieved were 247, 0.1, and 0.001,

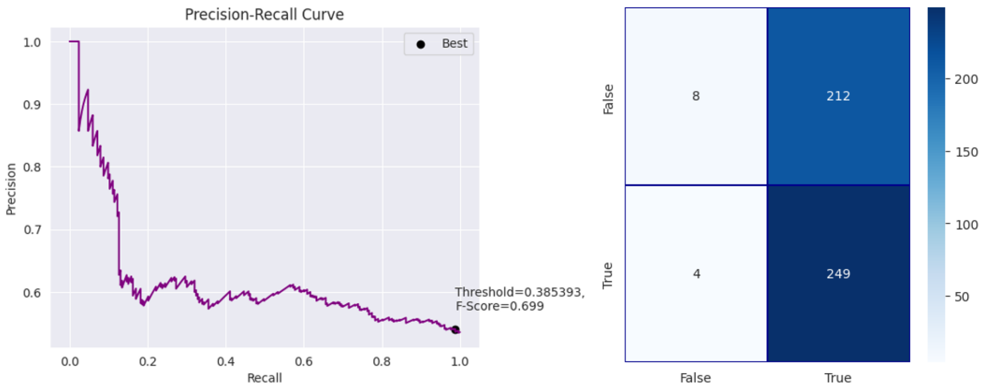


Figure 5.13: The precision-recall curve of the MLP binary classifier, with the threshold resulting in the best F1-Score marked on it (left) and the resulting confusion matrix when classifying the validation set using this post-prediction threshold (right).

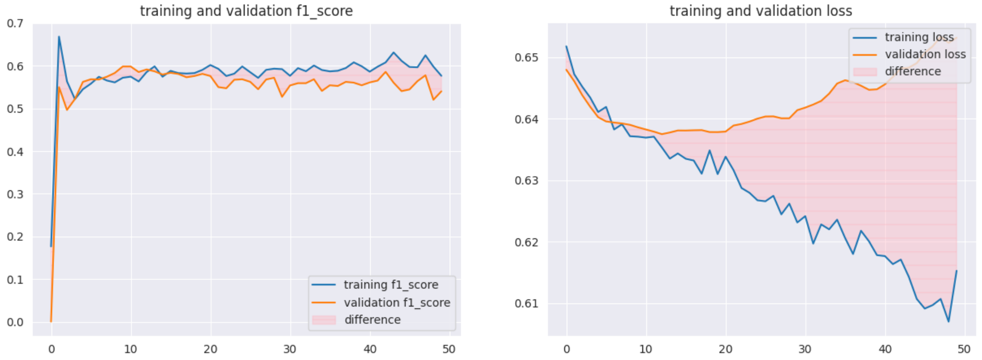


Figure 5.14: The training and validation F1-Score (left) and the training and validation loss (right) of the MLP binary classifier for Castrop PE prediction with 0.25 D threshold.

respectively. However, a model that always predicts the positive class is unusable, so the loss function for hyperparameter optimization was changed to precision. The values of min-child-weight, alpha, and lambda were 112, 0.1, and 0.1, respectively. This resulted in an F1-Score of 0.605, a precision of 0.556, and a recall of 0.664 (see Table 5.4).

5.6.3 Model Improvement Experiments

To address the class imbalance, various techniques other than the class weight parameter were experimented with, as described in Section 2.3. These techniques included SMOTE and an ensemble-based approach. For the ensemble based approach, the training set was divided into multiple datasets, each consisting of all of the minority samples and only a subset of the majority samples. These datasets were then used to train different classifiers, which were bagged and the final prediction was calculated using the logical OR operator. However, the performance of these experiments was

Dataset	Model	F1	Precision	Recall
Train	Dummy	0.694	0.531	1.0
	LR	0.554	0.596	0.517
	DTC	0.673	0.576	0.811
	RFC	1.0	1.0	1.0
	SVC	0.56	0.621	0.51
	MLP	0.572	0.608	0.541
	GBC	0.671	0.624	0.726
Val	Dummy	0.697	0.535	1.0
	LR	0.564	0.594	0.538
	DTC	0.657	0.564	0.787
	RFC	0.581	0.573	0.589
	SVC	0.549	0.605	0.502
	MLP	0.598	0.608	0.589
	GBC	0.605	0.556	0.664

Table 5.4: The Castrop PE binary classification training and validation results for a threshold of 0.25 D. For each model is the F1-Score, precision, and recall of the positive class, which stands for refractive surprise, reported.

either similar or worse.

Additionally, the influence of the CCT was analyzed. The F1-Score without the CCT feature (0.446) was higher than with the CCT feature (0.433), but the difference was not significant ($p = 0.127$).

5.7 Intraocular Lens Power Calculation

This section describes the development of an ML-based IOL power calculation formula. It begins with a description of the formula constant optimization, followed by the training and validation results of various models explored. Finally, the realization of the ensemble is outlined, along with the different experiments to improve performance.

5.7.1 Formula Constants Optimization

The Castrop and SRKT formula were implemented in Python and validated as described in Section 3.5.1. The constants of the corresponding formulas were optimized based on each dataset and study separately. The most optimal constant was selected by minimizing the MSE. The optimized constants are listed in table 5.5.

Dataset	Formula	Constant	S1 Value	S2 Value	S3 Value	S4 Value	S5 Value
Train	Castrop	C,H,R	0.298, 0.265, 0.104	0.329, -0.095, 0.247	0.354, -0.334, 0.217	0.29, 0.218, 0.323	0.326, 0.155, -0.165
	SRKT	A constant	119.281	118.998	118.866	119.44	118.875
Val	Castrop	C,H,R	0.29, 0.32, 0.051	0.333, -0.171, 0.382	0.127, 1.317, -0.773	0.507, -0.603, 0.076	0.332, 0.239, -0.268
	SRKT	A constant	119.249	118.972	118.728	119.49	118.905
Test	Castrop	C,H,R	0.151, 1.018, 0.018	0.472, -0.782, 0.246	0.184, 1.563, -1.277	0.247, 0.554, 0.202	0.204, 0.414, 0.128
	SRKT	A constant	119.31	118.867	118.739	119.439	118.797

Table 5.5: The optimized lens constants.

5.7.2 Model Exploration

Only regression models were experimented with for predicting the postoperative SEQ. Essentially, the same models were trained as for PE regression. To properly compare the different models, the same input features were used for all of them. These features included the lens type, the implantation center, the biometry data of the eye (AL, ACD, and LT), and the PIOL. Table 5.6 summarizes the training and validation metrics of all the models presented below.

Dummy Regressor

The dummy regressor always predicts the mean of the SEQ in the training set, resulting in a MAE of 0.658 and an R2-Score of 0.0.

Decision Tree Regressor

The hyperparameter tuning for the DTR resulted in a min-samples-split of 112. On the validation data, the DTR had a MAE of 0.613, only slightly better than the dummy regressor. The lens type does not seem to be as important as with the PE prediction (see Figure 5.15). The most important feature is AL, which aligns with the literature discussed in Section 2.2, as AL is one of the key components for IOL power calculation.

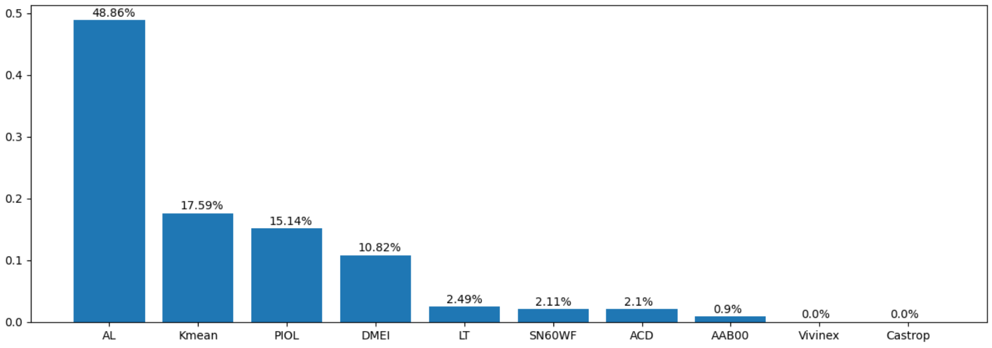


Figure 5.15: The feature importance of the DTR for the IOL power calculation.

Random Forest Regressor

Despite exhausting hyperparameter optmimization the overfitting could not be fixed for the RFR. Best results were achieved with a min-samples-split of 2, which is the default, and 200 estimators. The feature importance for the RFR looks mostly the same as the one of the DTR seen in figure 5.15.

Support Vector Regressor

Compared to the dummy regressor, DTR, and RFR, the SVR gave excellent results with a MAE of 0.335 on validation data. The SVR appears to outperform many standard IOL power formulas (see Section 3.1.1). The best results were obtained with a C value of 8.0.

Gradient Boosted Tree Regressor

GBR struggled with overfitting, similar to RFR. An exhaustive hyperparameter search resulted in the following values: min-child-weight of 1, nestimators of 122, reg-alpha of 0, and reg-lambda of 2.9. The difference between the training MAE (0.267) and validation MAE (0.411) could not be further reduced.

Multi Layer Perceptron

The MLP architecture shown in Figure 5.7 yielded the best MAE results on the validation data. A MAE loss function was used for training, with a batch size of 473 and a learning rate of 1e-3. The AdamW optimizer was used for optimization. The best MAE was achieved after the 68th training epoch, as shown in Figure 5.16.

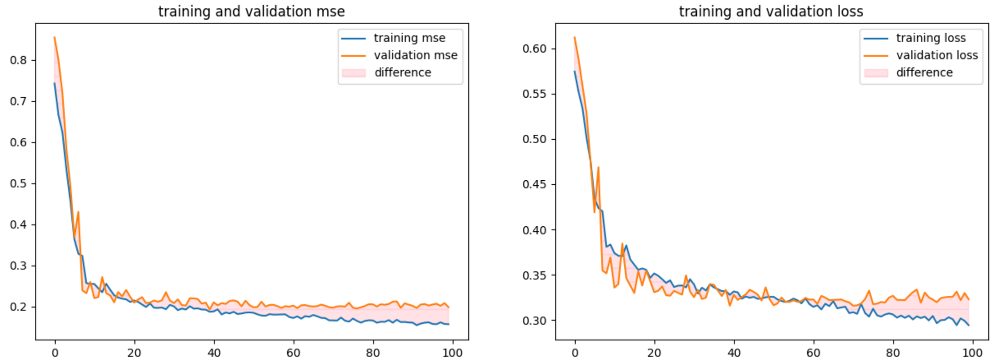


Figure 5.16: The training and validation MSE (left) and the training and validation loss (MAE) (right). The training loss appears to overtake the validation loss around epoch 50.

5.7.3 Model Improvement Experiments

The MLP model was chosen because it performed best on the validation data without overfitting training data. To improve its performance further, it was experimented

Dataset	Model	MAE	MSE	R2-Score
Train	Dummy	0.603	0.714	0.0
	DTR	0.509	0.493	0.31
	RFR	0.167	0.051	0.928
	SVR	0.278	0.14	0.803
	GBR	0.267	0.121	0.831
	MLP	0.328	0.187	0.709
Val	Dummy	0.658	0.927	-0.003
	DTR	0.613	0.772	0.164
	RFR	0.478	0.471	0.49
	SVR	0.335	0.217	0.765
	GBR	0.411	0.33	0.643
	MLP	0.316	0.193	0.752

Table 5.6: The IOL power calculation training and validation results. For each model the MAE, MSE, and R2-Score is reported.

with adding more features such as the predSEQ of the Castrop and SRKT formula and their formula constants, as well as with data augmentation.

More Features

It has been reported that adding the predSEQ values can increase performance (see Section 2.1.3). Another study reported that adding formula constants can also increase performance (see Section 2.1.3). However, adding the predSEQ of different formulas did not improve performance. Adding constants that were optimized for each study and dataset decreased performance, but constants that were optimized for the full dataset did not affect performance.

Additionally, the influence of the CCT feature was analyzed. The MAE without the CCT feature (0.321) was lower than with the CCT feature (0.329), but the difference was not significant ($p = 0.527$).

Data Augmentation

A recent study reported successful application of modern data augmentation techniques and showed that more data can increase the performance of IOL power calculation (see Section 2.1.3). Therefore, it was experimented with using synthetic minority over-sampling technique for regression with gaussian noise (SMOGN) (Branco et al., 2017). The open-source Python SMOGN library implemented by Kunz was used for this. Since SMOGN has difficulties synthesising one-hot encoded categorical features, the library randomly samples these features. This led to impossible synthetic samples, such as Vivinex and SN60WF being true simultaneously. To avoid such inconsistencies, samples for each study were synthesized separately, synthesizing only numerical features and manually setting the categorical ones according to the rest of the samples of the particular study. This also ensures that the variance from different studies is not mixed. The newly synthesized samples were then shuffled into the existing training set. The synthetic samples were only synthesized from the training set and the validation set was not modified at all. This increased the length of the training set from 1890 to 2673. Figure 5.17 shows the training and validation loss. Compared to the training and validation loss without synthetic data (see Figure 5.16), it is apparent that the loss converges slower using synthetic data. However, the performance did not change significantly ($p = 0.496$).

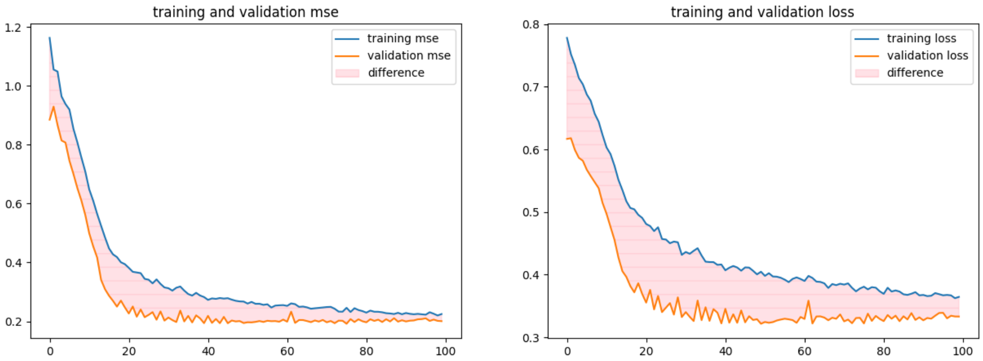


Figure 5.17: The training and validation MSE (left) and the training and validation loss (MAE) (right) of MLP training on a training set containing synthetic samples.

Ensemble

It was experimented with using the PE regression to improve the performance of the Castrop formula, as described in Section 3.5.2. Different models with increasing complexity and more additional features for the second-level model were tried. The best performance with the least overfitting was achieved using the MLP architecture shown in Figure 5.7, with the predSEQ of the Castrop formula and the MLP PE regression as input. With a MAE of 0.311, this ensemble performed better than the MLP from Section 5.7.2, but the differences were not significant ($p = 0.961$).

CHAPTER 6

Evaluation

The best model for PE regression (see Section 5.5), PE classification (see Section 5.6), and IOL power calculation (see Section 5.7) were evaluated on the truly unseen test data.

6.1 Prediction Error Regression

For the evaluation of the Castrop PE regression the metrics MAE, MSE, R2-Score, and SD were used. Table 6.1 shows the results on the truly unseen test set described in Section 3.1.3. The distribution of errors over the true values shows, that the model predicts a value inside the range of ± 0.25 D most of the time.

MAE	**MedAE**	**R2-Score**	**SD**
0.334	0.283	0.076	0.422

Table 6.1: The performance of the Castrop PE regression with respect to the metrics MAE, MedAE, R2-Score, and SD.

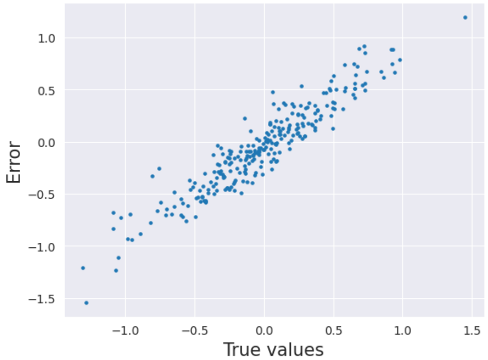


Figure 6.1: The distribution of the Castrop PE regression error over the true values.

6.2 Prediction Error Classification

Table 6.2 shows the results obtained by both the 0.5 D as well as 0.25 D threshold classification. For the 0.5 D model, the tuned threshold from Section 5.6.1 was used. For the 0.25 D model, the default threshold of 0.5 was used, since the tuned threshold resulted in always predicting the positive class on the validation set (see Section 5.6.2).

Figures 6.2 and 6.3 show that the classification models predict samples over the full range of the Castrop PE incorrectly. It does not seem to matter whether a sample is near the threshold or far from it, indicating that the models cannot distinguish refractive surprises from normal cases based on the features present in the refractive prediction error dataset.

Class	Precision	Recall	F1	M-Pre	M-Rec	M-F1
0 (0.5 D)	0.84	0.51	0.63	0.57	0.60	0.53
1 (0.5 D)	0.3	0.69	0.42			
0 (0.25 D)	0.51	0.57	0.54	0.55	0.55	0.54
1 (0.25 D)	0.58	0.52	0.55			

Table 6.2: The performance of the Castrop PE binary classification on truly unseen test data. The performance on classification with 0.5 D as well as 0.25 D threshold is reported with respect to the metrics precision, recall, F1-Score for each class and the respective macro-averaged score over both classes. The negative class (0) represents no refractive surprise, while the positive class (1) stands for refractive surprise.

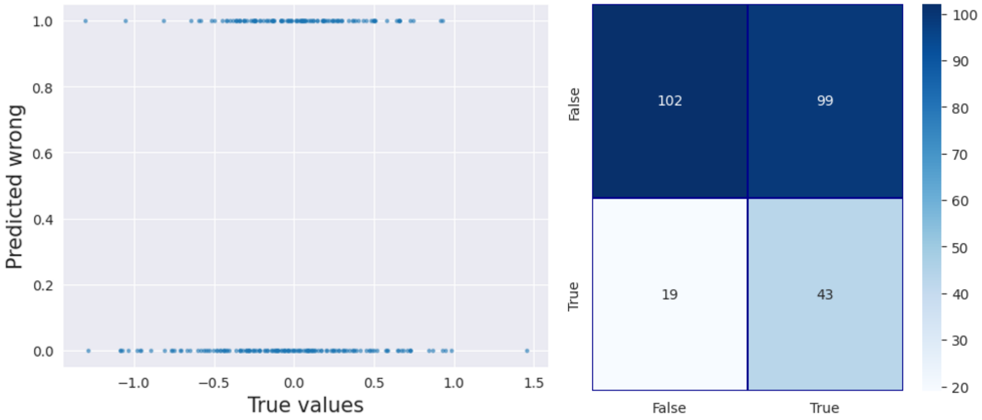


Figure 6.2: A scatter plot showing which values of Castrop PE of the test set were classified incorrectly by the 0.5 D model (left) and the confusion matrix of its predictions (right).

6.3 Intraocular Lens Power Calculation

The performance of the best IOL power calculation formula was compared to the Castrop and the SRKT formulas based on the recommendations in the literature discussed in Section 2.1.1. Figure 6.4 shows the number of absolute PEs within the limits of 0.25, 0.5, 0.75, and 1.0 D. The ML-based formula proposed in this work

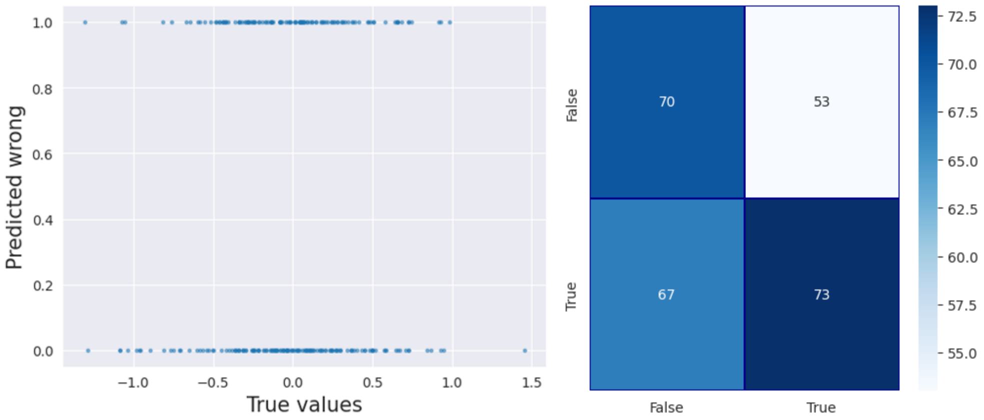


Figure 6.3: A scatter plot showing which values of Castrop PE of the test set were classified incorrectly by the 0.25 D model (left) and the confusion matrix of its predictions (right).

performed slightly worse within 0.25 D but slightly better within 0.75 D and 1.0 D than the Castrop formula, and outperformed the SRKT formula on all thresholds. Cochran-Q tests with Bonferroni correction showed significant differences among these three formulas within the limits of 0.5 D and 0.75 D ($p <= 0.001$). Subsequent post-hoc McNemar tests with Bonferroni correction reported that the Ensemble significantly outperformed the SRKT formula within the limits of 0.5 D and 0.75 D ($p <= 0.002$). However, no significant differences were detected between the Castrop formula and the ensemble ($p >= 0.083$).

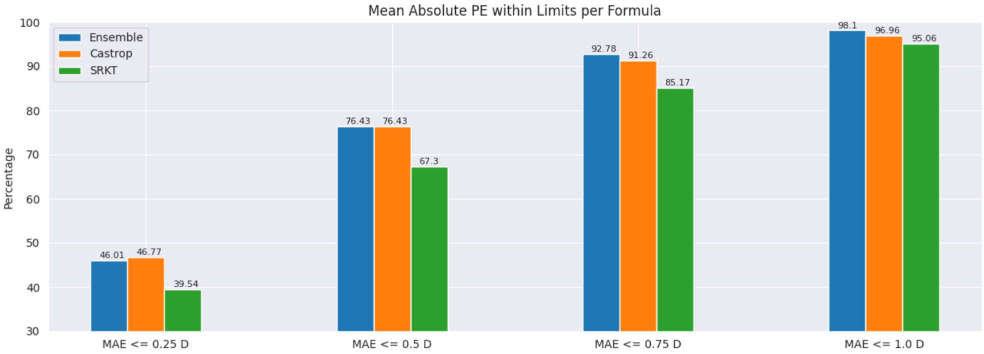


Figure 6.4: Ratio of predictions within the limits of the mean absolute PE for the ensemble ML formula developed in this work and two existing formulas on the test set.

In addition, the formulas were compared with respect to the MAE, MedAE, SD, and FPI. The MAE, MedAE, and SD were smaller for the ensemble on both the validation and testing sets (see Table 6.3). T-tests with Bonferroni correction did not detect significance between the ensemble and the Castrop formula ($p = 0.669$) but did detect significance between the ensemble and SRKT formula ($p = 0.006$).

Dataset	**Formula**	**MAE**	**MedAE**	**SD**	**FPI**
Val	Ensemble	0.311	0.251	0.427	1.061
	MLP	0.312	0.238	0.428	1.087
	Castrop	0.338	0.268	0.453	0.861
	SRKT	0.387	0.284	0.525	0.914
Test	Ensemble	0.331	0.269	0.423	0.925
	Castrop	0.341	0.275	0.442	1.009
	SRKT	0.402	0.342	0.515	0.631

Table 6.3: IOL power calculation formula validation and test results, including the MAE, MedAE, SD, and FPI for each formula.

The regression error plots of the three different formulas appear similar (see Figure 6.5). The PE of all the formulas seems to become more hyperopic with increasing SEQ. For example, none of the three formulas got any of the samples right where the SEQ is around 1.0 D, while, all of them got samples right that have a SEQ of -1.0 D or greater. This may be related to the fact that planned hyperopia is more rare than planned myopia, as discussed in Section 1.5.3. Thus, there are fewer samples where the SEQ is hyperopic and the existing formulas were more optimized for samples were the SEQ is more myopic. The same is true for the ML-based formula developed in this work. There were too few samples with a hyperopic SEQ to properly learn patterns from, so performance on these hyperopic cases may improve if the model could see more of them. However, this is just a hypothesis, and still needs to be proven.

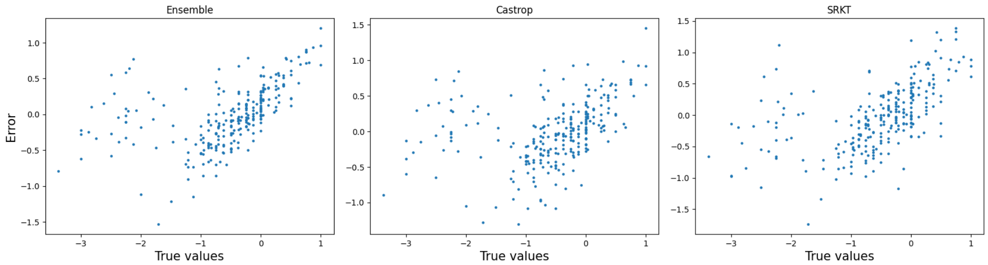


Figure 6.5: The distribution of regression error over the true SEQ values of the ensemble proposed in this work (left), the Castrop formula (middle), and the SRKT formula (right).

CHAPTER 7

Conclusion

This chapter discusses limitations, strengths, unsolved problems, and further ideas of this thesis.

7.1 Limitations of this Thesis

The refractive error prediction dataset this thesis did not include a patient ID, making it impossible to determine which samples belong to the same patient or to only enroll one eye per patient as recommended in the literature. As a result, the results for both the PE prediction as well as the IOL power calculation may be overoptimistic.

Furthermore, this thesis does not report the demographics of the study population, such as age, sex, and ethnicity, which can significantly affect the results. While it is not possible to definitely determine any ethical bias, it must be assumed that the models trained in this study are biased in all three respects.

7.2 Strengths of this Thesis

This thesis demonstrates that predicting the PE of the Castrop IOL power calculation formula using ML is difficult based on the available features. PCA showed that the IOL type used and the center of implantation explain most of the variance of the Castrop PE, but that refractive surprises and normal cases heavily overlap. Supervised learning also produced similar results, with both regression as well as classification models struggling to identify any patterns. Thus, it seems like that the cause for refractive surprises is not present in the data.

However, this thesis proposes a stacking ensemble ML architecture that aims to improve the performance of existing IOL power calculation formulas through PE regression. It also demonstrates that conventional ML models such as SVM can outperform various standard IOL power calculation formulas. Additionally, it shows that NNs have the potential to outperform modern top-performing IOL power calculation formulas.

7.3 Outlook

In the future, the performance of both PE prediction and IOL power calculation could certainly be improved with more data and features. A desk study showed that preoperative visual acuity, sex, and age, which were not included in the data for this task, are all correlated with refractive surprise to some extent. It would be interesting to examine the impact of these additional features on the performance of the developed models. Additionally, a gold standard benchmark for PE prediction and IOL power calculation could be established, making comparisons between such studies more reliable.

The problem of limited and imbalanced data could be addressed through data augmentation. Data augmentation experiments conducted during this thesis did not significantly impact performance. However, recent developments in synthesizing tabular data using transformer-based architectures (Borisov et al., 2022) may produce better results when applied.

Finally, it may also be worthwhile to experiment with some recent deep learning architectures for tabular data (Hollmann et al., 2022), which were not covered in this thesis. Although deep learning has not yet fully overtaken tree-based models for tabular data (Grinsztajn et al., 2022), these architectures may have the potential to further increase performance for PE prediction and IOL power calculation.

Appendix

APPENDIX A

Original Project Description

Following the original project description in German as created in JointCreate (jointcreate.com, 2022).

A.1 Ausgangslage und Problemstellung

Die Katarakt oder Grauer Star ist eine Trübung der Augenlinse, die zu einem langsam fortschreitenden Verlust der Sehschärfe führt. Die Katarakt tritt bei ca. jeder sechsten Person über 40 Jahren auf. Bei der Kataraktoperation wird die trübe Linse durch ein künstliches Implantat ersetzt. Die Kataraktoperation ist die heute weltweit am häufigsten durchgeführte Operation. Alleine in Deutschland werden rund 950.000 Eingriffe pro Jahr durchgeführt. Die exzellente Erfolgsquote und Zufriedenheit der Patienten im Allgemeinen täuscht darüber hinweg, dass bei wenigen Prozent der Patienten refraktive Überraschungen oder Komplikationen auftreten, so dass hier ggf. Folgeeingriffe nötig sind bis hin zur Explantation / zum Ersatz der Linse. Oft werden dabei Premiumlinsen mit Zusatzfunktionen gegen deutlich besser tolerierte Standardlinsen ausgetauscht.

Viele der refraktiven Überraschungen sollten vermeidbar sein, wenn entsprechende Screeningverfahren vorhanden wären welche die vor dem Eingriff erhobenen biometrischen Daten sowie patientencharakteristische Größen abgleichen und dem Operateur eine Warnung an die Hand geben, z.B. von Premiumlinsen (multifokale oder torische Linsen) abzusehen.

A.2 Datenmaterial

Zur Verfügung stehen einige Tausend vor einer Kataraktoperation erhobene biometrische Messungen (mit dem IOLMaster700 der Firma Carl-Zeiss-Meditec, vollständige Datensätze), patientencharakteristische Daten wie das Alter und Geschlecht, der Brechwert und Typ der implantierten Linse, sowie das refraktive Ergebnis nach der Operation. Der Brechwert der zu implantierenden Linse bzw. die zu erwartenden Refraktion nach dem Eingriff können mit den biometrischen Größen abgeschätzt werden, so dass die Abweichung der tatsächlich gemessenen Refraktion von der vorhergesagten Refraktion als "Refraktionsüberraschung" definiert ist.

A.3 Ziel der Arbeit und erwartete Resultate

In dieser Arbeit soll ein Machine Learning Verfahren entwickelt werden, mit dem die Refraktionsüberraschung vorhergesagt werden kann. Dabei ist sowohl eine Vorhersage in Form einer Klassifizierung (deutliche/mittlere/geringe Abweichung in Richtung Myopie/Hyperopie) oder auch kontinuierliche Vorhersage (Regression) möglich.

Abgegeben werden soll ein Bericht mit State-of-the-Art, Konzept, Ansätzen, der/den entwickelten Methoden und einer robusten Evaluation, sowie lauffähiger, kommentierter Programmcode.

A.4 Gewünschte Methoden, Vorgehen

Bei diesem Projekt handelt es sich um explorative Forschung, das in einem iterativen, inkrementellen Ansatz umgesetzt werden soll. Die/der Student:in soll den aktuellen Stand des Projektes und die nächsten Schritte in regelmässigen Absprachen mit dem Betreuer besprechen, um Feedback zu sammeln und sich zu verbessern. Dabei gilt es, den Fokus auf der Entwicklung eines Vorhersage-Algorithmus zu halten und diesen im Sinne einer Machbarkeitsstudie zu entwickeln und zu testen.

Darüber hinaus müssen Risiken so früh wie möglich gesammelt, verfolgt und gemindert werden, um zu überprüfen, ob einige Risiken ein Hindernis für das Projekt darstellen.

A.5 Kreativität, Varianten, Innovation

Das Projektziel ist bewusst offen formuliert und lässt viel Raum für eigene Kreativität.

Die Auswahl und Umsetzung des geeigneten Projektvorgehens ist Teil der Projektaufgabe und liegt grundsätzlich in der Verantwortung der/s Student:in. In regelmäßigen Treffen mit dem Betreuer werden der aktuelle Stand und die nächsten Schritte besprochen.

Der Betreuer soll regelmässig bis einen Tag vor dem geplanten Treffen schriftlich (max. 1 Seite) über den aktuellen Stand informiert werden:

- Welche Arbeiten wurden im letzten Berichtszeitraum durchgeführt, welche Arbeiten sind für die nächste Periode geplant
- Stand der Arbeiten (Soll-Ist-Vergleich mit Planung), ggf. Begründung von Abweichungen
- Top-3-Risiken inklusive geplanter Maßnahmen

Die Architektur soll so einfach wie möglich gehalten werden. Bezüglich der Programmiersprache ist der/die Student:in frei; es sollen jedoch wenn möglich und sinnvoll vorhandene Open-Source-Bibliotheken (wie sklearn, pytorch, ...) wiederverwendet werden, um das Ziel effizient zu erreichen.

APPENDIX B

Minutes

Stand-Up Meeting Minutes

Biweekly stand-up meetings on the progress of this bachelor thesis together with Univ.-Prof. Dr. Achim Langenbucher and Dr. sc. ETH Andreas Streich.

Kick-off Meeting – 15.09.2022, Virtual

Agenda

1. Introduction
2. Discussion of problem
3. Discussion of data
4. Discussion of organization
5. Risk update

Discussion of problem

- The cataract surgery is the third most operation on humans (~950’000 operations per year).
- Most common lens is one with 21 diopters.
- Post operative refractive surprise: The patient’s sight is much worse than expected. This happens in approx. 2% of the cases.
- Since the surgery is successful most of the times it is really devastating for the 2% in which it’s not. The patients then need a lot of care and need to hear that the doctor will fix it.
- The dissatisfaction of patients with refractive surprises tended to increase in the past.
- The Scheimpflug Method is often used to calculate the lens and the expected refraction after the cataract surgery.

- There are 3 levels of desirable outcomes. This thesis focuses mostly on point one. Point two and three are more of the long-term vision:
    - Predict based on the OCT data of the patient whether an aftertreatment will be needed.
    - Predict which specific aftertreatment will be needed if one will be needed.
    - Directly predict the optimal lens which should be used in the cataract surgery with AI so that never an aftertreatment will be needed.

- Other potential outcomes could be:
    - Evaluate based on the algorithm which IOL calculation method is the most error prone.
    - Evaluate based on the algorithm which lens is the most error prone.

Discussion of data

- Because the refractive surprise only happens in approx. 2% of the cases the dataset will be skewed. → Anomaly detection.
- The data consists of:
    - 12-13 numeric distance data of the eye (cornea -> iris -> lens -> retina). E.g., bending of the cornea, horizontal diameter of the cornea, etc.
    - Which lens was implanted.
    - Label -> how big was the refractive surprise.

- The datasets consist of approx. 1000 rows.
- The threshold between good and bad refractive surprises is to determine.
- The output of the algorithm can be a class or a continuous numeric value e.g., in percent.
- The data is in the range of MB.
- Dr. Achim Langenbucher and Dr. Andreas Streich think it should be possible to predict refractive surprises with the available data and technology, but it is also possible, that the dataset does not contain the variance and no "positive" results can be delivered.

Discussion of organization

- Sync meeting per teams every two week on Friday 14:00 o'clock with a preceding progress, risk and next steps update per mail.
- The intermediate presentation should take place early, e.g., week 7 or 8.
- Dr. Andreas Streich offers to read 30 pages of the thesis to provide feedback.
- This bachelor thesis is an exploratory research project. Therefore, it is important that every decision is documented comprehensible!

SW02 – 26.09.2022, Virtual

What has been done

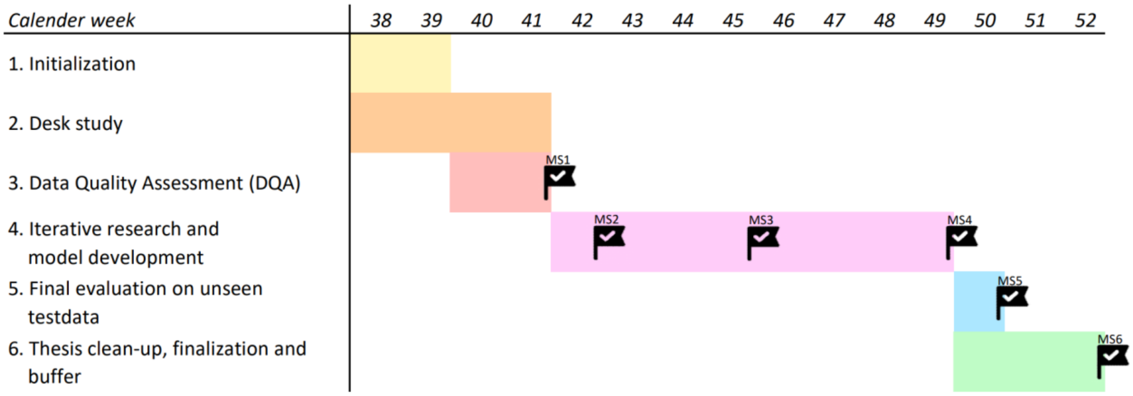
- Initialization of project: definition of roadmap and milestones, risk analysis, formulating goal, setting up thesis document.
  - ➔ Thesis will be written in English
  - ➔ HSLU GitLab will be used as VCS
- Desk study: no papers on predicting refractive error after cataract surgery were found yet.
  - ➔ Zotero and ResearchRabbit are used for managing papers and articles.

Next steps

- Finalization of initialization.
- Desk study.
- Start with data quality assessment if some first data will be available.

Progress update

Following is the created initial project roadmap which is separated in six stages and six milestones. Future progress updates will be done based on this roadmap.



ID	*Title*	*Description*	*Deadline*
MS1	Data quality	The data is visualized, analyzed, and cleaned up. Each step and potential change in the data is documented clearly and comprehensible.	2022-10-16
MS2	Model baseline	The dataset is split up into a train, validation, and test set. A simple first baseline model is fully trained and validated.	2022-10-23
MS3	Intermediate presentation	The presentation of the intermediate results to the advisors and experts is completed.	KW45 (SW08)
MS4	Final model	The final version of the model is trained and ready for evaluation on unseen test data.	2022-12-11
MS5	Evaluation	The evaluation of the final model on truly unseen test data is completed.	2022-12-18
MS6	Submission	The thesis has been submitted.	2023-01-03

Risk update

The following four new risks were identified:

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R1	Illness or accident	Absence due to illness or accident.	2	5	10
R2	Part-time studies	Absence due to part-time studies of the author.	3	3	9
R3	Lacking variance	A lack of predictive capacity (variance) in the training data. Whereby it is not possible to build a ML model with satisfying performance.	3	4	12
R4	Lacking ophthalmology know-how	It is hard to comprehend the model output since the data cannot be interpreted due to lacking ophthalmology know-how.	4	3	12

Mitigation:

ID	*Mitigation*	*Likeliness*	*Severity*	*Risk Score*
R1	The deadline can be shifted backwards in case of illness or an accident.	2	3	6
R3	Use of techniques like model prediction distribution, normality check, model explainability, … to prematurely realize that the data may be the problem. Thereby the amount of time wasted on experiments with other hyperparameters and models can be limited and instead be invested in the search for more suitable data.	3	2	6
R4	Univ-Prov. Dr. Achim Langenbucher offered his support in case of questions.	2	3	6

Top 3 Risks (sorted descending):

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R2	Part-time studies	Absence due to part-time studies of the author.	3	3	9
R1	Illness or Accident	Absence due to illness or accident.	2	3	6
R3	Lacking variance	A lack of predictive capacity (variance) in the training data. Whereby it is not possible to build a ML model with satisfying performance.	3	2	6

Questions

- Is there a possibility to get more data?
    - ➔ Yes, there is the possibility to get more labeled data probably also from other institution around Europe. The limiting factor are the labels.

Other potential interesting features:

- Gender
    - ➔ will be added
- Does the patient have other eye diseases besides cataract, e.g., keratoconus which could lead to an abnormal keratometry?
    - ➔ LVC = 3 or 4: Means keratoconus. But in the data at hand are only patients with LVC = 0 which means nothing.
- Has the patient already had refractive surgeries in the past?
    - ➔ LVC = 1:
- Stage of the cataract (severity)
    - ➔ This information is implicitly available through IOL power (LT).
- Data for the diagnosis of cataracts (OCT?), e.g., type of cataract
    - ➔ Could be provided for a very small number of patients (~200 samples)

Ethical considerations:

- Which medical institution provides the dataset?
    - ➔ Mainly Saarland University Hospital. But it's possible to get data from other medical institutions.
- Which population around the world occurs the most in the data at hand?
    - ➔ Mostly people from Europe.
- Gender

Feedback

- Roadmap is too sequential. Desk study and writing of the thesis should be done iterative incremental over the whole project and not only in the first four weeks or rather in the last three weeks.
- The stages of the roadmap should more represent focus blocks than hard sequential separated work packages.
- For each read paper a few sentences about the most interesting results/findings should be written down.

- Keywords to find related work: refractive prediction error (cataract).
- R3 should be rephrased to lack of predictive capacity.
- The second and third goal are very similar -> reformulate second goal (second goal should be the baseline).

SW03 – 07.10.2022, Virtual

Agenda

- Discuss progress and risk update.
- Presentation of data quality assessment results.
- Presentation of first experiments with basic ML algorithms.

What has been done

- Analyzing dataset.
- Data quality assessment.
- First experiments with basic ML algorithms.
- Study cataract theory and writing introduction.

Next steps

- Finalize DQA to earn MS1 -> PCA.
- Finalize baseline model to earn MS2
- Continue with iterative research and model development.

Progress update

- The roadmap has been updated according to the feedback of sw02.
- The start of focus point 5 has been shifted from 42 to 40.
- Project is on track. There are no impediments why MS1 and MS2 should not be accomplished.

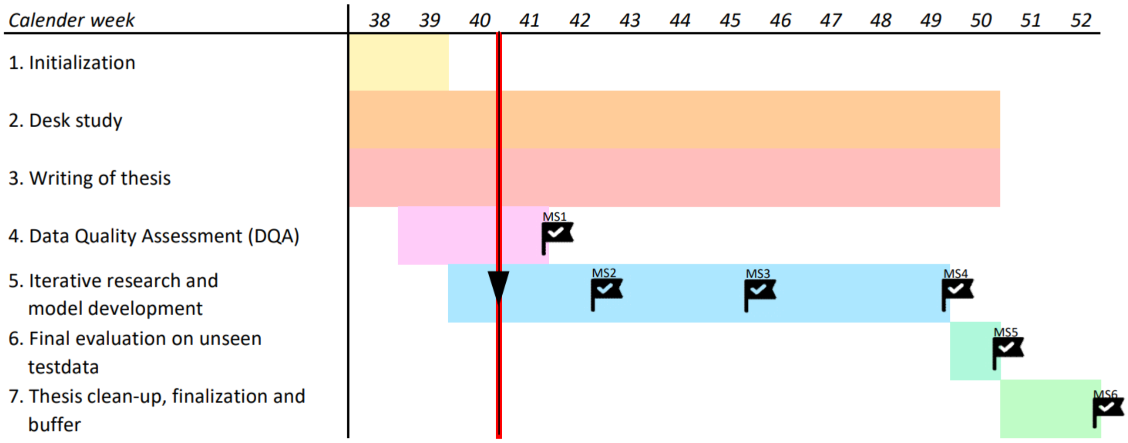


Figure 1: The red line indicates where the project should be and the black triangle where the project currently is.

Risk update

- Reformulated R3 according to feedback from SW02.

- Increased severity of R2 from 3 to 4 because the time schedule is tight and more absence due to the job of the author would be hard to compensate.

The following new risk was identified:

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R5	No access to relevant articles	Since the department of computer science at the HSLU is a technical institution, it is not subscribed to medical journals. Whereby the author of this thesis has no free access to relevant articles.	4	3	12

The following risks were updated:

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R2	Part-time studies	Absence due to part-time studies of the author.	3	4	12
R3	Lack of predictive capacity	A lack of predictive capacity in the training data. Whereby it is not possible to build a ML model with satisfying performance.	3	2	6

Mitigation:

ID	*Mitigation*	*Likeliness*	*Severity*	*Risk Score*
R2	The boss and colleagues at work of the author were informed about the BAA. No additional effort will be requested from the author, and it is possible to temporarily decrease the pensum. Tracking of work done.	1	4	4
R5	TODO (Maybe access via Achim?)	2	3	6

Top 3 Risks (sorted descending):

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R1	Illness or Accident	Absence due to illness or accident.	2	3	6
R3	Lack of predictive capacity	A lack of predictive capacity in the training data. Whereby it is not possible to build a ML model with satisfying performance.	3	2	6
R4	Lacking ophthalmology know-how	It is hard to comprehend the model output since the data cannot be interpreted due to lacking ophthalmology know-how.	2	3	6

Questions

- What is SEQ in the data?

- What is R1, R2 in the data?
- Is K keratometry?
- Which column represents the post operative refractive target?
- Which column represents the type of IOL?
- What is the difference between negative and positive prediction error?
- Some patients have rather high visus e.g., 1.6 but still needed a cataract surgery?
- Patient 131 has visus of 20?
- Patient 49 has visus of 10?
- Why is Barret Universal II not present in the data?
- Can we append the date of birth or rather the age of the patients to the data?

Feedback

- Null values in the dataset should not be replaced by e.g., median but be removed.
- Do scatter plots of regression error to see on which samples the error is biggest.
- Use feature importance to see which features are more important.

SW05 – 21.09.2022, Virtual

Agenda

- Discuss progress and risk update
- Presentation of percentage within 0.5D and 1.0D of all four errors.
- Presentation of baseline results and Mlflow tracking stack.

What has been done

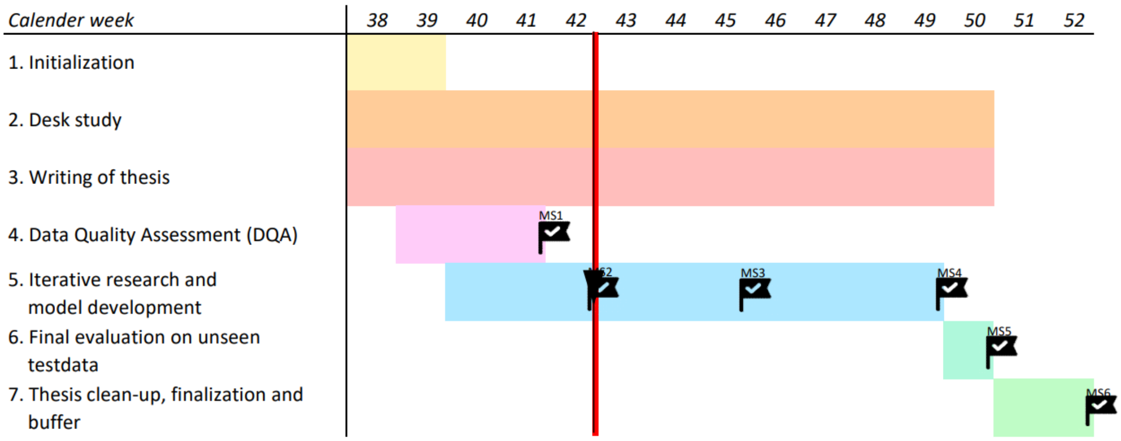
- Setting up Mlflow Tracking Stack to keep track of experiments, hyperparameter, metrics, etc.
- Finalize baseline experiments: plotting of regression error and feature importance.
    - Regression [R2-Score]
        - KNR: 0.086 (cross validation even worsened the score)
        - SVR: 0.056
    - Classification [F1-Score]
        - DTC: 0.3
        - LRC: 0.244

Next steps

- Experiment with neural networks.
- Experiment with down-sampling and upweighting.

Progress update

- Nothing changed on the roadmap.
- Project is on track.



Risk update

- No new risks identified, and no risk scores updated. The Top 3 is still the same as in SW03.

Questions

- Where would you put Mlflow Tracking in the thesis? Chapter 5 (realization), chapter 3 (concept) or chapter 4 (methodology)?

Feedback

- Validate hypothesis that with k-fold cross validation the validation performance decreases with an increasing k.
- Try to plot and cluster the date to evaluate whether outlier detection is the better approach. E.g., plot the most important features in a scatter plot and colorize the records with the highest CAS error. If all the colored points don’t form a group this is an indicator, that outlier detection probably will work better.
- Beware of overfitting. There is high variance in train and validation metrics in the baseline experiments conducted so far.

SW07 – 04.11.2022, Virtual

Agenda

- Short summary what has been done so far
- Presentation of PCA results
- Presentation of MLP results

What has been done

- Experiment with by patient splits, k-fold cross validation and manual splits
- Principal component analysis
- Desk study
- Experiment with simple neural networks (multilayer perceptron)

Next steps

- Analyze new data

- Evaluate the impact of additional data on performance of already done experiments
- Experiment with the package imbalanced learn
- Experiment with XGBoost Decision Trees.

Progress update

- Project is on track and there are no impediments.

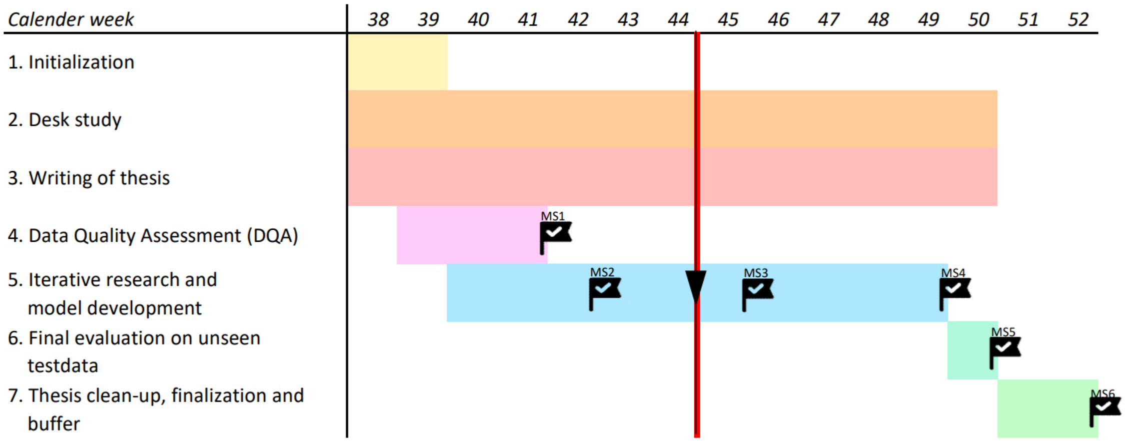


Figure 2: The red line indicates where the project should be and the black triangle where the project currently is.

Risk update

- No new risks identified, and no risk scores updated. The Top 3 is still the same as in SW03.

Questions

Feedback

- Plot PC-1 with PC-3, PC-2 with PC-3. Sometimes this reveals clusters.
- Dimension of principal components should be normalized, e.g., mean of 0 and standard deviation of 1.
- Check why the sum of the contribution per feature to principal components is greater 1
- Reduce class threshold to 0.25D to have less imbalanced data.
- Check if the used MLP really is nonlinear.
- Use dropout and regularization to tackle overfitting in MLP.
- Analyze, why the validation performance of the MLP has such high volatility. Maybe observe some weights of the MLP during training. Probably it is because some batches contain no outliers, but others do.

SW10 – 25.11.2022, Virtual

What has been done

- Experiment with re-sampling (under-sampling, over-sampling, SMOTE) -> did not result in significant better results than those already achieved with balanced class weight parameter in scikit-learn library.
- Study of two papers which do IOL power calculation with ML.
- Study of paper which proposes study guidelines for IOL power calculation.

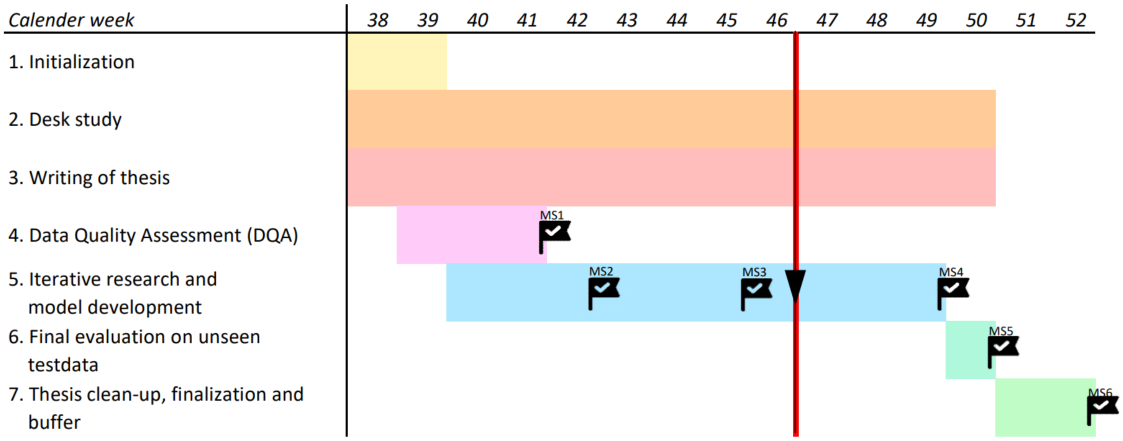
- Study of reasons and risk factors for refractive prediction error.
- Study of best practices for ML with imbalanced data.

Next steps

- Continue with incorporation of feedback from intermediate presentation.
- Directly predict the refractive outcome from the data (IOL power calculation).
- Use the directly predicted refractive outcome as feature for prediction of refractive surprises and vice-versa.

Progress update

- The project is on track so far.
- If the requested data (patient id, age, gender, preoperative visus) will be delivered late, MS4 and MS5 will probably be shifted back by one week, so that MS5 will be accomplished by the end of week 51.



Risk update

The following new risk was identified:

ID	*Title*	*Description*	*Likeliness*	*Severity*	*Risk Score*
R6	No related work	There is no previous work, which tries to predict refractive surprises with ML.	4	2	8

However, after discussing the risk with the advisors after the intermediate presentation the severity of R6 was adjusted to 1, resulting in a rather small risk score of 4. The reason for decreasing the severity was, that the scope, in which was searched for previous work, was opened and potential papers were analyzed with more creativity. This resulted in various papers which can be built up on.

Top 3 Risks (sorted descending):

- Since the risk score of R6 was updated to 4, the Top 3 is still the same as in SW03.

Feedback

- Second eye refinement does work but is not practical because you need to wait between the surgery of the two eyes at least for one month to have stable results on the first eye, which then can be incorporated in the second eye. But waiting for one month is not an option.

SW12 – 09.12.2022, Virtual

Agenda

What has been done

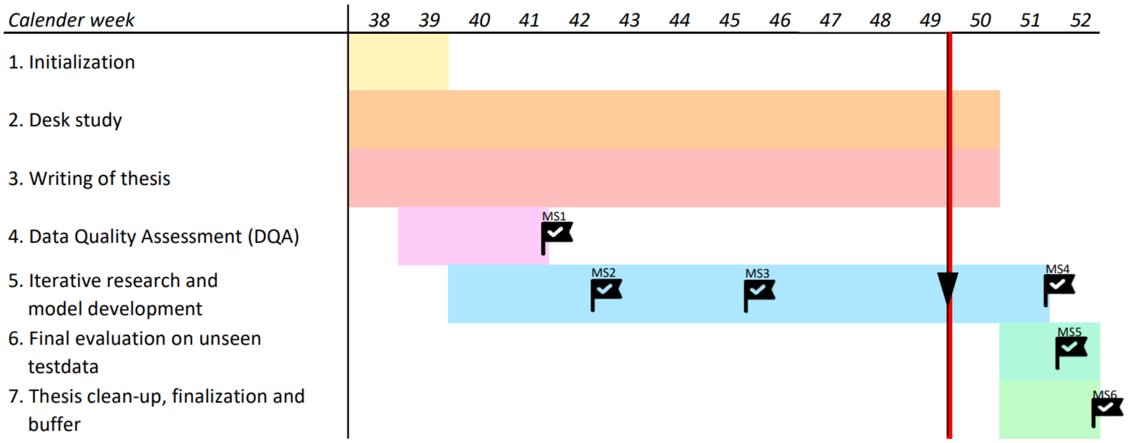
- Clean up chapter 1 and 2 of thesis.
- Continue writing chapter 3 concept and chapter 4 methodology.

Next steps

- Writing thesis.
- Continue with experiments outlined in SW10 as soon as patient id will be delivered.

Progress update

- MS2 missed. Shifted back by two weeks.
- MS5 not yet missed but as well shifted back by a bit more than one week.
- Project is still on track, since the time which could not be invested in experiments due to the missing patient IDs could be invested in writing of thesis and study of literature. However, since the web abstract needs to be delivered by the beginning of week 52, no more such compensations are possible, because the evaluation needs to be finished by then. Therefore, the delivery of the patient IDs until Thursday evening of week 50 is critical. Otherwise, the experiments will be conducted without by patient splits.



Risk update

- No new risks identified. Top 3 risks still the same as in SW10.

Questions

- What are R1, R2 and Rmean features? Why is R1 and R2 constant per study?
- Is a SE of -8 D valid?
- If its valid, should this sample nevertheless be discarded for training of ML algorithms?
- From how many surgeons do the data come?

SW13 – 15.09.2022, Virtual

Agenda

What has been done

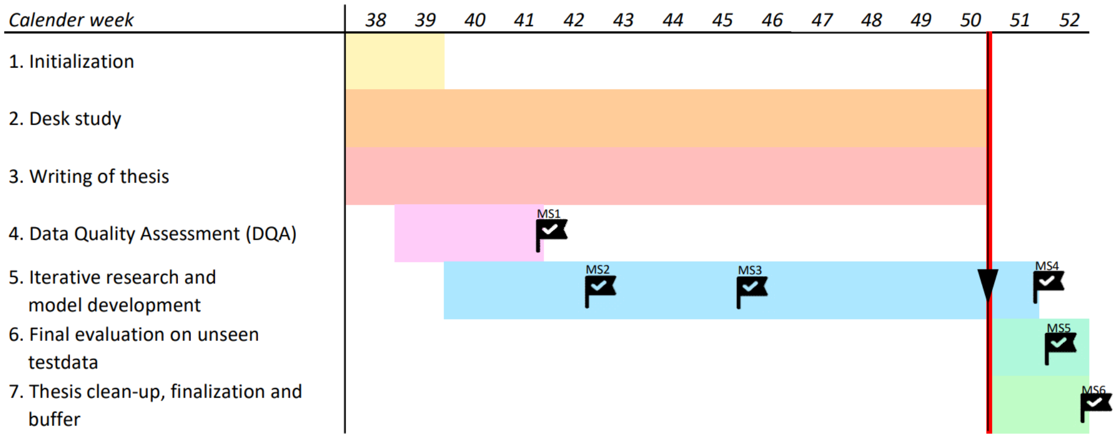
- Experiment with neural networks. Classification with 0.5 threshold reached M-F1 of 0.6 and F1 of 0.626.
- Little convergence could be achieved with neural network regression. MAE of 0.306.

Next steps

- Implement constant optimization for SRKT and Castrop formula
- Continue with experiments outlined in SW10 without splits by patient.
  - Directly predict the refractive outcome from the data (IOL power calculation).
  - Use the directly predicted refractive outcome as feature for prediction of refractive surprises and vice-versa.

Progress update

- Project on track.



Risk update

- No new risks identified. Top 3 risks still the same as in SW10.

Feedback

- Plot precision-recall curve.
- Analyze why loss function is overfitting but metric not. Try out more epochs. -> This was due to the linear LR scheduling. With constant LR this phenomenon disappears.
- Do not use linear learning rate scheduler.
- Stratify splits additionally also based on lens type.

SW14 – 23.12.2022, Virtual

What has been done

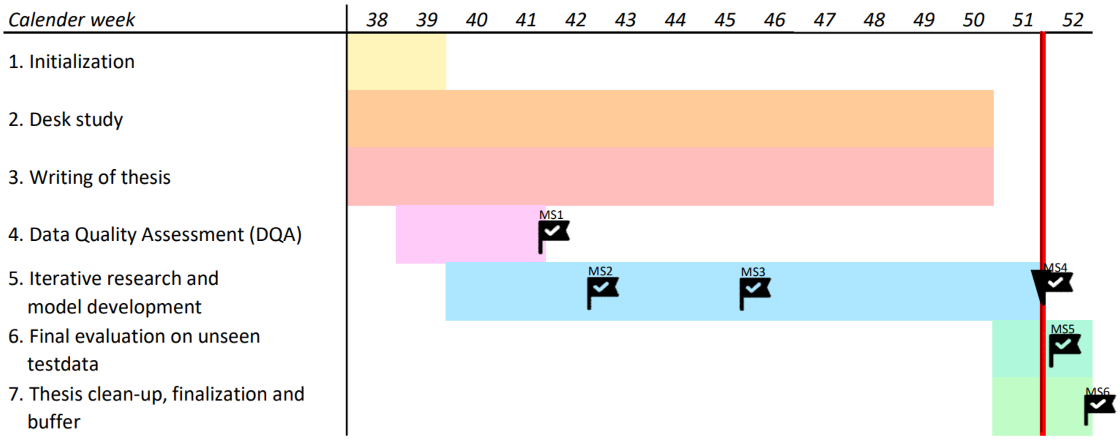
- Threshold tuning for PE classification. For classification of PE above 0.25 D a F1-Score of 0.7 was achieved, with a threshold of 0.31.
- Implemented and tested Castrop and SRKT formulas according to the Matlab script.
- Implemented constant optimization for Castrop and SRKT. The new constants optimized via Python Levenberg Marquardt implementation gave better results for Study 3, 4, and 5. However, the difference is not significant (p-Value > 0.05).
- Implemented IOL power calculation models (SVR, XGBoostRegressor, MLP). MLP yielded significantly better results than Castrop on training and validation set (Mc-Nemar test p-Value < 0.05).

Next steps

- Write Web-Abstract.
- Try to improve IOL power calculation with TabTransformers.
- Try to improve IOL power calculation with modern transformer-based data augmentation.
- See if IOL power calculation model can be used to improve prediction of PE.

Progress update

- Project on track. Evaluation scripts are set up. MS4 and MS5 will be reached by end of week 51. Therefore, the focus in week 52 can be set to thesis clean-up and finalization.



Risk update

- No new risks identified. Top 3 risks still the same as in SW10.

Feedback

- Overfitting of XGBoost in IOL power calculation can potentially be fixed with better hyperparameter tuning.
- Plot precision over threshold and recall over threshold for classification metrics, to see the influence of different thresholds on these metrics.

APPENDIX C

Milestones

In Table C.1 all defined milestones are listed, including an ID, a title, a description, and a deadline.

ID	Title	Description	Deadline
MS1	Data quality	The data is visualized, analyzed and cleaned up. Each step and potential change in the data is documented clearly and comprehensible.	2022-10-16
MS2	Model baseline	The dataset is split up into a train, validation and test set. A simple first baseline model is fully trained and validated.	2022-10-23
MS3	Intermediate presentation	The presentation of the intermediate results to the advisors and experts is completed.	KW45
MS4	Final model	The final version of the model is trained and ready for evaluation on unseen test data.	2022-12-11
MS5	Evaluation	The evaluation of the final model on truly unseen test data is completed.	2022-12-18
MS6	Submission	The thesis has been submitted.	2023-01-03

Table C.1: The project roadmap consists of six milestones. Each milestone has an ID, a title, a description, and a deadline.

APPENDIX D

Risks

Table D.1 lists all risks which came up during the project and Tables D.2 and D.3 presents their unmitigated respectively mitigated risk scores. Table D.4 describes the mitigation measures taken and how they impacted the risk score.

ID	Title	Description
R1	Illness or accident	Absence due to illness or accident.
R2	Part-time studies	Absence due to part-time studies of the author.
R3	Lack of predictive capacity	A lack of predictive capacity in the training data. Whereby it is not possible to build a ML model with satisfying performance.
R4	Lacking ophthalmology know-how	It is hard to comprehend the model output since the data can not be interpreted due to lacking ophthalmology know-how.
R5	No access to relevant articles	Since the department of computer science at the HSLU is a technical institution, it is not subscribed to medical journals. Whereby the author of this thesis has no free access to relevant articles.
R6	No related work	There is no previous work, which tries to predict refractive surprises with ML.

Table D.1: All risks which came up during the project, including ID, title, and description.

ID	Likeliness	Severity	Risk Score
R1	2	5	10
R2	3	4	12
R3	3	4	12
R4	4	3	12
R5	4	3	12
R6	4	2	8

Table D.2: The unmitigated risk scores of all risks.

ID	Likeliness	Severity	Risk Score
R1	2	3	6
R2	3	3	4
R3	3	2	6
R4	2	3	6
R5	2	3	6
R6	4	2	8

Table D.3: The mitigated risk scores of all risks.

ID	Title	Mitigation
R1	Illness or accident	The deadline can be shifted backwards In case of illness or an accident. *Update:* S 5 → 3
R2	Part time studies	The boss and colleagues at work of the author were informed about the BAA. No additional effort will be requested from the author and it is possible to temporarily decrease the pensum. Tracking of work done. *Update:* L 3 → 1
R3	Lack of predictive capacity	Use of techniques like model prediction distribution, normality check, model explainability, . . . to prematurely realize that the data may be the problem. Thereby the amount of time wasted on experiments with other hyperparameters and models can be limited and instead be invested in the search for more suitable data. *Update:* S 4 → 2
R4	Lacking ophthalmology know-how	Univ-Prov. Dr. Achim Langenbucher offered his support in case of questions. *Update:* L 4 → 2
R5	No access to relevant articles	Univ-Prov. Dr. Achim Langenbucher offered his support in case of questions. *Update:* L 4 → 2

Table D.4: Mitigation measures that have been taken for risks with an overall risk score greater than 10 and the resulting risk update.

Listings

List of Figures

List of Tables

List of Equations

Acronyms

ACD anterior chamber depth 10, 12–14, 19, 25, 26, 38–40, 50, 83

AD aqueous depth 10, 13

AL axial length 3, 8, 10, 12–15, 19, 26, 38, 39, 50, 83

BART Bayesian Additive Regression Trees 11, 14

BU-II Barrett Universal II 10–13, 15

CCT central corneal thickness 13, 19, 23, 36, 49, 52, 85

D dioptre 3, 7, 9, 10, 12–15, 20, 23, 24, 36, 37, 41–49, 54–57, 83–86

DQA data quality assessment 23, 36

DTC decision tree classifier 26, 43, 44, 46, 47, 83

DTR decision tree regressor 26, 27, 39, 40, 50, 51, 83

ECCE extracapsular cataract extraction 5, 91, 92

EVO Emmetropia Verifying Optical 13

FPI IOL Formula Performance Index 10, 11, 56, 57, 86

GBC gradient boosting tree classifier 47

GBR gradient boosting regression 12, 27, 51

HCD horizontal corneal diameter 10, 12–14

HSLU Lucerne University of Applied Sciences and Arts 2, 92

ICCE intracapsular cataract extraction 5, 91, 92

IOL intraocular lens 5–15, 19–21, 24, 26, 27, 33, 36, 38, 43, 49–52, 54, 55, 58, 59, 82, 83, 85, 89, 91, 92

K corneal power 10, 13, 19, 22, 23, 25, 36, 82, 87

LASIK laser-assisted in situ keratomileusis 7, 9

LR logistic regression 42, 43, 46

LT lens thickness 10, 12, 13, 19, 26, 39, 50

MAE mean absolute error 9, 11–14, 20, 21, 24, 37, 39, 41, 42, 50–54, 56, 57, 83–86

MedAE median absolute error 9, 13, 21, 54, 56, 57, 85, 86

ML machine learning 2, 7–9, 11–14, 16–19, 24–27, 30, 31, 33–36, 41, 49, 55–58, 82, 84

MLP multi layer perceptron 41, 44, 45, 47, 48, 51, 53, 83, 84

MSE mean squared error 27, 39, 42, 49, 51–54, 83–85

NHS National Health Service 7

NN neural networks 11, 12, 26, 27, 31, 34, 58

PCA principal component analysis 26, 37, 38, 58, 83

PCS phacoemulsification 5, 92

PE prediction error 6–9, 12, 14–16, 19–28, 37, 39–50, 53–59, 82–86

PIOL IOL power 19, 26, 39, 50

predPE predicted prediction error 15

predSEQ predicted spherical equivalent 15, 19, 22, 27, 52, 53

PRK photorefractive keratectomy 7, 9

R radius of the cornea curvature 22, 36

RFC random forest classifier 26, 43, 44, 46, 47, 83

RFR random forest regressor 26, 27, 40, 51

RFR random forest regression 12

SD standard deviation 9, 13, 14, 21, 23, 24, 54, 56, 57, 85, 86

SEQ spherical equivalent 15, 19, 26, 28, 36, 39, 40, 50, 57, 82, 84

SMOGN synthetic minority over-sampling technique for regression with gaussian noise 52

SMOTE synthetic minority over-sampling technique 17, 48

SotA state-of-the-art 2, 8

SVC support vector classifier 26, 44, 47

SVM support vector machines 11, 58

SVR support vector regressor 26, 27, 40, 41, 51

SVR support vector regression 12

WTW white-to-white corneal diameter 10

Glossary

20/20 vision Visual acuity is measured in a numeric fraction, such as 20/20 or 20/40. The top number represents the distance from the chart (20 feet), and the bottom number represents the distance at which the average person with normal eyesight can correctly read the same line. (Russel, 2020b) 3, 5, 91

ametropia An eye that has refractive error is said to have ametropia or be ametropic. (Russel, 2021) 3

aphakic Aphakic means an eye without a lens. (Porter, 2021) 92

astigmatism Astigmatism is a type of refractive error that causes distorted vision, usually at all distances. It occurs when the curvature of the cornea, the front of the eye, is irregularly shaped. (Russel, 2020c) 3, 4, 7, 13

bagging Bagging or bootstrap aggregating is an ensemble machine learning method that involves training the same algorithm many times by using different subsets sampled from the training data. The final output prediction it then averaged across the predictions of all the sub models. (Wen and Hughes, 2020) 18

dioptre The unit of measurement of optical power. Optical power is a physical quantity which describes the degree to which for example a lens bends the light. (Wikipedia, 2022a) 3, 6, 88

emmetropia Emmetropia is the clinical term used by eye doctors to describe a person with perfect vision, also known as 20/20 vision. (Russel, 2021) 3, 5, 6

extracapsular cataract extraction ECCE leaves the posterior capsule of the lens intact, removing the nucleus and cortex of the lens. Therefore a smaller incision than with ICCE is needed. The removed lens is then replaced by a posterior chamber IOL. (Yorston and McGavin, 2009) 5, 88

GitLab GitLab is an open-core DevOps software package that combines the ability to develop, secure, and operate software in a single application. (GitLab Inc., 2022) 35

heteroscedasticity In statistics, heteroscedasticity happens when the standard deviation of y is not constant in x. The tell-tale sign upon visual inspection is, that y tends to fan out with an increasing x. (Hayes, 2022) 22, 23, 82

HSLU Enterpriselab The HSLU Enterpriselab provides IT resources for computer science and research at the HSLU. (HSLU Enterpriselab, 2006) 35

hyperopia Hyperopia, also known as farsightedness or long-sightedness, causes near objects or images to appear blurry. 3, 15, 57

intracapsular cataract extraction ICCE is one of the first methods for lens removal. Traditionally, it involved removal of the complete intact lens through a large incision measuring 12-14 mm. In earlier years the eyes were left aphakic. In the late 1970 it has been superseded by the back then more modern ECCE mostly due to the lower rates of postoperative posterior segment complications. (Yanoff, 2019) 5, 88

legal blindness A person is considered legally blind if her visual acuity is 20/200 or worse. (WebMD Editorial, 2020) 14

monovision Monovision involves one eye, usually the dominant eye, being corrected for distance viewing, and the other eye being corrected for near viewing, allowing one to see clearly at any distance. (Boyd, 2018) 7

myopia Myopia, also known as nearsightedness or shortsightedness, causes distant objects to appear blurry or out of focus. 3, 7, 15, 57

ocular comorbidity A combination of different eye disorder which exist simultaneously. (Pinazo-Durán et al., 2016) 15

ophthalmology Ophthalmology is a surgical subspeciality within medicine that deals with the diagnosis and treatment of eye disease. (Wikipedia, 2022b) 2

phacoemulsification PCS is a modern cataract surgery method. During this surgery, a tiny probe is inserted into the side of the cornea, through a small incision, usually about 2.8 mm. The probe emits ultrasound waves that soften and break up the lens to enable easy suctioning from the eye. (Benítez Martínez et al., 2021) 5, 89

pseudophakic Pseudophakia translates from the Latin to mean false lens. The term refers to the implanting of an IOL to replace a natural lens. An eye is therefore pseudophakic if its natural lens is removed and replaced by an IOL. (Huizen and Griff, 2017) 11

stacking Stacking is an ensemble machine learning method that is concerned with combining heterogeneous machine learning models using another machine learning model. (Wen and Hughes, 2020) 18

Bibliography

Behndig, A., Montan, P., Stenevi, U., Kugelberg, M., Zetterström, C., & Lundström, M. (2012). Aiming for emmetropia after cataract surgery: Swedish national cataract register study. *Journal of Cataract and Refractive Surgery*, *38*(7), 1181–1186. https://doi.org/10.1016/j.jcrs.2012.02.035 (see p. 6)

Benítez Martínez, M., Baeza Moyano, D., & González-Lezcano, R. A. (2021). Phacoemulsification: Proposals for improvement in its application. *Healthcare (Basel, Switzerland)*, *9*(11), 1603. https://doi.org/10.3390/healthcare9111603 (see p. 92)

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2022, October 12). Language models are realistic tabular data generators. https://doi.org/10.48550/arXiv.2210.06280. (see p. 59)

Boslaugh, S., & Watters, P. A. (2008, July). *7. the pearson correlation coefficient*. Retrieved October 22, 2022, from https://learning.oreilly.com/library/view/statistics-in-a/9781449361129/ch07.html. (see p. 22)

Boyd, K. (2018, May 7). *What is monovision (or blended vision)?* [American academy of ophthalmology]. Retrieved October 6, 2022, from https://www.aao.org/eye-health/treatments/what-is-monovision-blended-vision. (see p. 92)

Branco, P., Torgo, L., & Ribeiro, R. (2015, May 13). A survey of predictive modelling under imbalanced distributions. https://doi.org/10.48550/arXiv.1505.01658. (see p. 16)

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: A pre-processing approach for imbalanced regression [ISSN: 2640-3498]. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 36–50. Retrieved January 1, 2023, from https://proceedings.mlr.press/v74/branco17a.html (see p. 52)

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation [Publisher: [Wiley, Econometric Society]]. *Econometrica*, *47*(5), 1287–1294. https://doi.org/10.2307/1911963 (see p. 22)

Brugman, S. (2022). *Pandas-profiling: Exploratory data analysis for python* [Version: 3.3.0]. Retrieved October 21, 2022, from https://github.com/pandas-profiling/pandas-profiling. (see p. 35) calc.apacrs.org. (2010). *Barrett universal II formula v1.05*. Retrieved November 19, 2022, from https://calc.apacrs.org/barrett_universal2105/. (see p. 12)

Carr, F., & Gangwani, V. (2020). Refractive surprise after cataract surgery secondary to smeared optics of swept-source optical coherence tomography biometer: A case report. *BMC ophthalmology*, *20*(1), 352. https://doi.org/10.1186/s12886-020-01629-0 (see p. 15)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953 (see p. 17)

Chen, X., Xu, J., Chen, X., & Yao, K. (2021). Cataract: Advances in surgery and whether surgery remains the only treatment in future. *Advances in Ophthalmology Practice and Research*, *1*(1), 100008. https://doi.org/10.1016/j.aopr.2021.100008 (see pp. 5, 6)

Clarke, G. P., & Kapelner, A. (2020). The bayesian additive regression trees formula for safe machine learning-based intraocular lens predictions. *Frontiers in Big Data*, *3*. Retrieved November 19, 2022, from https://www.frontiersin.org/articles/10.3389/fdata.2020.572134 (see p. 14)

Donaldson, K. E. (2022). Tips for dealing with unhappy refractive cataract surgery patient. *Current Ophthalmology Reports*, *10*(1), 1–4. https://doi.org/10.1007/s40135-022-00282-8 (see pp. 1, 2)

Dudek, L. (2021, April 15). *Multifocal and monofocal IOLs: Visual performance comparison*. Retrieved December 9, 2022, from https://www.heartoftexaseye.com/blog/multifocal-vs-monofocal-iols/. (see p. 5)

Evans, C. (2001). *Vsftp*. https://security.appspot.com/vsftpd.html. (see p. 35)

EyeWiki. (2022). *EyeWiki*. Retrieved October 7, 2022, from https://eyewiki.org/Main_Page. (see p. 31)

Feldman H., B., Heersink, S., & Alpa S., P. (2022, July 7). *Cataract - EyeWiki*. Retrieved October 1, 2022, from https://eyewiki.aao.org/Cataract. (see pp. 1, 4)

Garay-Aramburu, G., Bergado-Mijangos, R., Irizar-Amilleta, R., Saez-Espejo, B., Serrano-Zurbitu, L., Arakama-Alustiza, J., Gutiérrez-Soto, M., Ojanguren-Zugazaga, M. E., Areitio-Garcia, L., & Molpeceres-Uriszar, A. (2022). Risk factors for predicted refractive error after cataract surgery in clinical practice. retrospective observational study. *Archivos de la Sociedad Española de Oftalmología (English Edition)*, *97*(3), 140–148. https://doi.org/10.1016/j.oftale.2022.02.004 (see p. 14)

GitLab Inc. (2022). *The one DevOps platform*. Retrieved October 21, 2022, from https://about.gitlab.com/. (see p. 91)

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022, July 18). Why do tree-based models still outperform deep learning on tabular data? https://doi.org/10.48550/arXiv.2207.08815. (see p. 59)

Haigis, W., Lege, B., Miller, N., & Schneider, B. (2000). Comparison of immersion ultrasound biometry and partial coherence interferometry for intraocular lens calculation according to haigis. *Graefe's Archive for Clinical and Experimental Ophthalmology = Albrecht Von Graefes Archiv Fur Klinische Und Experimentelle Ophthalmologie*, *238*(9), 765–773. https://doi.org/10.1007/s004170000188 (see p. 8)

Hayashi, K., Ogawa, S., Yoshida, M., & Yoshimura, K. (2016). Influence of patient age on intraocular lens power prediction error [Publisher: Elsevier]. *American Journal of Ophthalmology*, *170*, 232–237. https://doi.org/10.1016/j.ajo.2016.08.016 (see p. 15)

Hayes, A. (2022, April 20). *Heteroscedasticity definition: Simple meaning and types explained* [Investopedia]. Retrieved October 22, 2022, from https://www.investopedia.com/terms/h/heteroskedasticity.asp. (see p. 92)

Hoffer, K. J. (1993). The hoffer q formula: A comparison of theoretic and regression formulas. *Journal of Cataract & Refractive Surgery*, *19*(6), 700–712. https://doi.org/10.1016/S0886-3350(13)80338-0 (see p. 8)

Hoffer, K. J., Aramberri, J., Haigis, W., Olsen, T., Savini, G., Shammas, H. J., & Bentow, S. (2015). Protocols for studies of intraocular lens formula accuracy. *American Journal of Ophthalmology*, *160*(3), 403–405.e1. https://doi.org/10.1016/j.ajo.2015.05.029 (see pp. 8, 10)

Hoffer, K. J., & Savini, G. (2021). Update on intraocular lens power calculation study protocols: The better way to design and report clinical trials [Publisher: Elsevier]. *Ophthalmology*, *128*(11), e115–e120. https://doi.org/10.1016/j.ophtha.2020.07.005 (see pp. 8, 11)

Hofstetter, J. (2020). Aufbau wipro/baa-bericht (see p. 2).

Holladay, J. T., Musgrove, K. H., Prager, T. C., Lewis, J. W., Chandler, T. Y., & Ruiz, R. S. (1988). A three-part system for refining intraocular lens power calculations. *Journal of Cataract & Refractive Surgery*, *14*(1), 17–24. https://doi.org/10.1016/S0886-3350(88)80059-2 (see p. 8)

Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2022, November 29). TabPFN: A transformer that solves small tabular classification problems in a second. https://doi.org/10.48550/arXiv.2207.01848. (see p. 59)

HSLU Enterpriselab. (2006). *Enterprise lab*. Retrieved October 21, 2022, from https://eportal.enterpriselab.ch/. (see p. 92)

Huizen, J., & Griff, M. (2017, October 13). *Pseudophakia (IOL): Definition, signs you may need them, and types*. Retrieved November 26, 2022, from https://www.medicalnewstoday.com/articles/319685. (see p. 92)

jointcreate.com. (2022, June 22). *Original project description*. Retrieved September 21, 2022, from https://home.jointcreate.com/de_ch/ventures/542/. (see p. 61)

Kane, J. X., Van Heerden, A., Atik, A., & Petsoglou, C. (2017). Accuracy of 3 new methods for intraocular lens power selection. *Journal of Cataract & Refractive Surgery*, *43*(3), 333–339. https://doi.org/10.1016/j.jcrs.2016.12.021 (see p. 10)

Kashnitsky, Y. (2019). *Topic 7. unsupervised learning: PCA and clustering*. Retrieved October 28, 2022, from https://kaggle.com/code/kashnitsky/topic-7-unsupervised-learning-pca-and-clustering. (see p. 26)

Langenbucher, A., Szentmáry, N., Cayless, A., Weisensee, J., Fabian, E., Wendelstein, J., & Hoffmann, P. (2021). Considerations on the castrop formula for calculation of intraocular lens power [Publisher: Public Library of Science]. *PLOS ONE*, *16*(6), e0252102. https://doi.org/10.1371/journal.pone.0252102 (see p. 11)

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html (see p. 17)

Li, T., Stein, J., & Nallasamy, N. (2022). Evaluation of the nallasamy formula: A stacking ensemble machine learning method for refraction prediction in cataract surgery [Publisher: BMJ Publishing Group Ltd Section: Clinical science]. *British Journal of Ophthalmology*. https://doi.org/10.1136/bjophthalmol-2021-320599 (see p. 13)

Lohri, P. (2022a). *An image of an eye with a lens affected by an advanced nuclear cataract. eye clinic - lucerne cantonal hospital, switzerland.* (see p. 5).

Lohri, P. (2022b). *An image of an eye with an via cataract surgery implanted multifocal iol. eye clinic - lucerne cantonal hospital, switzerland.* (see p. 6).

Lohri, P. (2022c). *An image of an eye with healthy lens. eye clinic - lucerne cantonal hospital, switzerland.* (see p. 5).

Loshchilov, I., & Hutter, F. (2019, January 4). Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.1711.05101. (see p. 41)

Lundström, M., Dickman, M., Henry, Y., Manning, S., Rosen, P., Tassignon, M.-J., Young, D., & Stenevi, U. (2018). Risk factors for refractive error after cataract surgery: Analysis of 282.811 cataract extractions reported to the european registry of quality outcomes for cataract and refractive surgery. *Journal of Cataract & Refractive Surgery*, *44*(4), 447–452. https://doi.org/10.1016/j.jcrs.2018.01.031 (see p. 15)

Meier, B. (2022). *Baa-hs22 / refractive-error-predicter · GitLab*. Retrieved October 21, 2022, from https://gitlab.enterpriselab.ch/baa-hs22/refractive-error-predicter. (see p. 35)

Melles, R. B., Holladay, J. T., & Chang, W. J. (2018). Accuracy of intraocular lens calculation formulas [Publisher: Elsevier]. *Ophthalmology*, *125*(2), 169–178. https://doi.org/10.1016/j.ophtha.2017.08.027 (see pp. 8, 10)

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2:2 (see p. 35).

Meyer, F. (2022, June 7). *Cataract awareness month*. Retrieved October 2, 2022, from https://eyesage.net/cataract-awareness-month-2021/. (see p. 4)

Moshirfar, M., Buckner, B., Ronquillo, Y. C., & Hofstedt, D. (2019). Biometry in cataract surgery: A review of the current literature. *Current Opinion in Ophthalmology*, *30*(1), 9–12. https://doi.org/10.1097/ICU.0000000000000536 (see p. 15)

National Academies of Sciences, E., Division, H., Medicine, Practice, B. o. P. H., Health, P., Health, C. o. P. H. A. t. R. V. I., Eye, P., Welp, A., Woodbury, R. B., McCoy, M. A., & Teutsch, S. M. (2016, September 15). *The impact of vision loss* [Publication Title: Making Eye Health a Population Health Imperative: Vision for Tomorrow]. National Academies Press (US). Retrieved December 9, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK402367/. (see p. 1)

Nizami, A. A., Gulani, A. C., & Redmond, S. B. (2021). Cataract (nursing). *StatPearls Publishing, Treasure Island (FL)*. http://europepmc.org/article/NBK/NBK568765 (see pp. 4, 5)

Norrby, S. (2008). Sources of error in intraocular lens power calculation. *Journal of Cataract & Refractive Surgery*, *34*(3), 368–376. https://doi.org/10.1016/j.jcrs.2007.10.031 (see p. 15)

OptometristsNetwork. (2022). *Home* [Optometrists.org]. Retrieved October 7, 2022, from https://www.optometrists.org/. (see p. 31)

Oracle Corporation. (1995, May 23). *Mysql*. https://www.mysql.com/. (see p. 35)

Palanker, D. (2013, October 28). *Optical properties of the eye - american academy of ophthalmology*. Retrieved November 25, 2022, from https://www.aao.org/munnerlyn-laser-surgery-center/optical-properties-of-eye. (see p. 3)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library, 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (see p. 31)

Peck, T., Brad H., F., & Alpa S., P. (2022, May 8). *Refractive error after cataract surgery - EyeWiki*. Retrieved October 1, 2022, from https://eyewiki.aao.org/Refractive_Error_After_Cataract_Surgery. (see pp. 1, 6, 7)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved October 21, 2022, from https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (see pp. 17, 31)

Pinazo-Durán, M. D., Zanón-Moreno, V., García-Medina, J. J., Arévalo, J. F., Gallego-Pinazo, R., & Nucci, C. (2016). Eclectic ocular comorbidities and systemic diseases with eye involvement: A review. *BioMed Research International*, *2016*, 6215745. https://doi.org/10.1155/2016/6215745 (see p. 92)

Porter, D. (2021, December 10). *What is aphakia?* [American academy of ophthalmology]. Retrieved October 2, 2022, from https://www.aao.org/eye-health/diseases/what-is-aphakia. (see p. 91)

PyTorch. (2022). *BCEWithLogitsLoss — PyTorch 1.13 documentation*. Retrieved December 30, 2022, from https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html. (see p. 44)

ResearchRabbit. (2022). *ResearchRabbit* [ResearchRabbit]. Retrieved October 7, 2022, from https://www.researchrabbit.ai. (see p. 31)

Retzlaff, J. A., Sanders, D. R., & Kraff, M. C. (1990). Development of the SRK/t intraocular lens implant power calculation formula. *Journal of Cataract & Refractive Surgery*, *16*(3), 333–340. https://doi.org/10.1016/S0886-3350(13)80705-5 (see p. 8)

Russel, L. (2020a, May 27). *What are cataracts?* [Optometrists.org]. Retrieved October 1, 2022, from https://www.optometrists.org/general-practice-optometry/guide-to-eye-conditions/guide-to-cataracts/cataracts/. (see pp. 4, 5)

Russel, L. (2020b, August 23). *What is a visual acuity test?* [Optometrists.org]. Retrieved October 1, 2022, from https://www.optometrists.org/general-practice-optometry/guide-to-eye-exams/eye-exams/what-is-a-visual-acuity-test/. (see p. 91)

Russel, L. (2020c, April 29). *What is astigmatism?* [Optometrists.org]. Retrieved October 1, 2022, from https://www.optometrists.org/childrens-vision/guide-to-pediatric-eye-conditions/what-is-astigmatism/. (see p. 91)

Russel, L. (2021, August 29). *Emmetropia* [Optometrists.org]. Retrieved October 1, 2022, from https://www.optometrists.org/general-practice-optometry/guide-to-eye-exams/eye-exams/emmetropia/. (see p. 91)

Scikit-Learn. (2011). *6.3. preprocessing data* [Scikit-learn]. Retrieved December 28, 2022, from https://scikit-learn/stable/modules/preprocessing.html. (see p. 25)

Sethi, A. (2020, March 5). *Categorical encoding — one hot encoding vs label encoding* [Analytics vidhya]. Retrieved December 28, 2022, from https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/. (see p. 25)

Shalchi, Z., Restori, M., Flanagan, D., & Watson, M. (2018). Managing refractive surprise, 2 (see p. 7).

Shammas, H. J., Shammas, M. C., Jivrajka, R. V., Cooke, D. L., & Potvin, R. (2020). Effects on IOL power calculation and expected clinical outcomes of axial length measurements based on multiple vs single refractive indices. *Clinical Ophthalmology (Auckland, N.Z.)*, *14*, 1511–1519. https://doi.org/10.2147/OPTH.S256851 (see p. 15)

Siddiqui, A. A., Juthani, V., Kang, J., & Chuck, R. S. (2019). The future of intraocular lens calculations: Ladas super formula [Number: 3 Publisher: AME Publishing

Company]. *Annals of Eye Science*, *4*(3), 19–19. https://doi.org/10.21037/aes.2019.04.02 (see p. 11)

Tozzi, C. (2021, February 19). *The data quality assessment: Does your data measure up?* [Precisely]. Retrieved October 20, 2022, from https://www.precisely.com/blog/data-quality/does-your-data-measure-up-assess-data-quality. (see p. 23)

WebMD Editorial. (2020). *What does it mean to be legally blind?* [WebMD]. Retrieved December 2, 2022, from https://www.webmd.com/eye-health/legally-blind-meaning. (see p. 92)

Wen, L., & Hughes, M. (2020). Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques [Number: 10 Publisher: Multidisciplinary Digital Publishing Institute]. *Remote Sensing*, *12*(10), 1683. https://doi.org/10.3390/rs12101683 (see pp. 91, 92)

Wikipedia. (2022a, November 19). Dioptre [Page Version ID: 1122813313]. In *Wikipedia*. Retrieved November 25, 2022, from https://en.wikipedia.org/w/index.php?title=Dioptre&oldid=1122813313. (see pp. 3, 91)

Wikipedia. (2022b, October 3). Ophthalmology [Page Version ID: 1113862586]. In *Wikipedia*. Retrieved October 7, 2022, from https://en.wikipedia.org/w/index.php?title=Ophthalmology&oldid=1113862586. (see p. 92)

Wikipedia. (2022c, August 23). Refractive error [Page Version ID: 1106092892]. In *Wikipedia*. Retrieved November 25, 2022, from https://en.wikipedia.org/w/index.php?title=Refractive_error&oldid=1106092892. (see p. 4)

World Health Organization. (2021, October 14). *Vision impairment and blindness*. Retrieved October 1, 2022, from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. (see p. 1)

Yamauchi, T., Tabuchi, H., Takase, K., & Masumoto, H. (2021). Use of a machine learning method in predicting refraction after cataract surgery. *Journal of Clinical Medicine*, *10*(5), 1103. https://doi.org/10.3390/jcm10051103 (see p. 12)

Yanoff, M. (2019). Intracapsular cataract extraction - indications for lens surgery/indications for application of different lens surgery techniques. Retrieved October 2, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/intracapsular-cataract-extraction (see p. 92)

Yorston, D. H., & McGavin, M. (2009). Extracapsular cataract extraction - ophthalmology in the tropics and subtropics. Retrieved October 2, 2022, from https://www.sciencedirect.com/topics/nursing-and-health-professions/extracapsular-cataract-extraction (see p. 91)

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Xie, F., & Zumar, C. (2022). Accelerating the machine learning lifecycle with MLflow, 7. Retrieved October

21, 2022, from https://cs.stanford.edu/~matei/papers/2018/ieee_mlflow.pdf (see pp. 32, 35)

Zhang, Y., Li, T., Reddy, A., & Nallasamy, N. (2021). Gender differences in refraction prediction error of five formulas for cataract surgery. *BMC Ophthalmology*, *21*(1), 183. https://doi.org/10.1186/s12886-021-01950-2 (see p. 15)

Zudans, J. V., Desai, N. R., & Trattler, W. B. (2012). Comparison of prediction error: Labeled versus unlabeled intraocular lens manufacturing tolerance. *Journal of Cataract and Refractive Surgery*, *38*(3), 394–402. https://doi.org/10.1016/j.jcrs.2011.08.044 (see p. 15)