Poznan University of Technology
Faculty of Computing and Telecommunications


Doctoral dissertation

# APPLICATION OF MACHINE LEARNING TECHNIQUES IN GLAUCOMA DETECTION AND CONTROL

Hubert Świerczyński

Supervisor: Szymon Szczęsny, Ph. D., Dr. Habil.
Auxiliary supervisor: Cezary Mazurek, Ph. D.

Poznań 2024

# Acknowledgements

First, I would like to thank my supervisor, Szymon Szczęsny for support during the research.
I would also like to thank my auxiliary supervisor, Cezary Mazurek for support during the research.
I would like to thank Robert Wasilewicz for sharing medical knowledge and ideas.
I would like to thank Juliusz Pukacki for support in data collection and management.

# Abstract

Glaucoma is a group of eye diseases which result in damage to the optic nerve. The disease affects many millions of patients worldwide and it is a substantial public health challenge. There is no single examination that can efficiently detect different types of glaucoma. Diagnostic routine includes diverse tests and requires significant level of experience.

Current approach for detecting glaucomatous structural damage is based mainly on optical coherence tomography. While it provides high-quality tomograms of anatomical structure of the eye, we can't use the images to track dynamic changes in this complex system which is influenced by many external factors that vary over time.

New diagnostic options independent of standard imaging techniques became available in the recent years as use of wearable medical devices is growing in many fields of healthcare. Triggerfish (Sensimed) device is based on a contact lens sensor with embedded strain gauge that measures ocular volume changes during 24-hour session. Triggerfish measurements are related to intraocular pressure changes and it can record low-amplitude ocular pulsations related to the heart rate.

This thesis considers application of machine learning techniques for analysis of data acquired using Triggerfish contact lens sensor and devices for continuous monitoring of cardiovascular system properties. Overview of basic machine learning concepts is provided before presenting the results of the research. Development of new machine learning models for glaucoma diagnosis is described thoroughly. Predictive performance of the models was estimated using cross-validation and many relevant metrics. Heart monitoring data associated with Triggerfish measurements can be used to more accurately detect glaucoma. We can improve predictive performance of the models by including measurements of the corneal biomechanical properties (e.g. corneal hysteresis).

System for the support of glaucoma diagnosis and control is proposed. It enables application of predictive models based on multi-sensor data and clinical measurements. Services for data sharing, management and visualization facilitate clinical decision-making and collaborative research.

This thesis also refers to the aspects of personalized medicine and the concept of transdisicplinarity.

# Streszczenie

Jaskra to grupa postępujących neuropatii nerwu wzrokowego, przy jednoczesnych morfologicznych zmianach warstwy włókien nerwowych siatkówki. Wiele milionów pacjentów na całym świecie choruje na jaskrę, a jej wykrywanie i leczenie jest poważnym wyzwaniem dla systemu opieki zdrowotnej. Nie zaproponowano dotychczas uniwersalnej, skutecznej metody diagnozowania różnych typów jaskry. Procedura diagnostyczna obejmuje wiele badań i wymaga znacznego doświadczenia klinicznego.

Aktualne metody wykrywania zmian patologicznych związanych z jaskrą bazują przede wszystkim na optycznej tomografii koherencyjnej (OCT). Tomogramy anatomicznych struktur oka nie umożliwiają jednak obserwacji dynamicznych zmian w funkcjonowaniu tego złożonego układu, na który oddziałuje wiele zewnętrznych czynników. Wraz z wprowadzeniem przenośnych urządzeń monitorujących stan pacjenta, pojawiły się w ostatnich latach nowe opcje diagnostyczne, niezależne od standardowych technik obrazowania stosowanych w okulistyce. Urządzenie Triggerfish (Sensimed) wykorzystuje soczewkę kontaktową z wbudowanym tensometrem oporowym do pomiaru zmian kształtu rogówki w ciągu doby. Względne zmiany objętości gałki ocznej zarejestrowane tą metodą umożliwiają ocenę zmian ciśnienia wewnątrz gałki ocznej.

W rozprawie opisano zastosowanie technik uczenia maszynowego w analizie danych zarejestrowanych za pomocą sensora Triggerfish i urządzeń monitorujących w sposób ciągły parametry układu sercowo-naczyniowego. Rozprawa zawiera też przegląd podstawowych koncepcji i algorytmów uczenia maszynowego. Szczegółowo opisano proces transformacji danych i właściwości zaproponowanych modeli predykcyjnych. Przedstawiono porównanie oszacowania efektywności predykcyjnej modeli z uwzględnieniem wielu kryteriów oceny (z zastosowaniem m.in. walidacji krzyżowej). Wykazano, ze uwzględnienie pomiarów właściwości biomechanicznych gałki ocznej (np. histereza rogówki) poprawia efektywność predykcyjną modeli.

Zaprezentowano koncepcję systemu wspomagającego diagnostykę jaskry, który wykorzystuje opracowane modele predykcyjne bazujące na danych sensorycznych i klinicznych. Omówiono scenariusze wspierania współpracy badawczej i podejmowania decyzji klinicznych z wykorzystaniem udostępniania, wizualizacji i eksploracji danych.

Perspektywa wdrożeniowa zaproponowanych rozwiązań została przedstawiona z uwzględnieniem założeń medycyny personalizowanej i badań transdyscyplinarnych.

# Preface

Some ideas and results presented in this dissertation have appeared previously in the following items:

A1. Hubert Świerczyński, Juliusz Pukacki, Szymon Szczęsny, Cezary Mazurek and Robert Wasilewicz. Application of machine learning techniques in GlaucomAI system for glaucoma diagnosis and collaborative research support, 2024 (under review)

A2. Hubert Świerczyński, Juliusz Pukacki, Szymon Szczęsny, Cezary Mazurek and Robert Wasilewicz. Sensor data analysis and development of machine learning models for detection of glaucoma. *Biomedical Signal Processing and Control*, 86:105350, 2023. Elsevier

C1. Robert Wasilewicz, Cezary Mazurek, Juliusz Pukacki and Hubert Świerczyński. GlaucomaAI - the first in classs, intelligent decision support system for glaucomatous optic neuropathy risk diagnostics. *10th World Glaucoma Congress*, Abstract Book, pp. 10-11, Rome, Italy, June 28 - July 1 2023. World Glaucoma Association

P1. Robert Wasilewicz, Cezary Mazurek, Juliusz Pukacki and Hubert Świerczyński. Method for creating a predictive model for predicting glaucoma risk in a subject; Method for determining glaucoma risk in a subject using such predictive model; Device for predicting glaucoma risk in a subject. Patent no. 7278323, *Japan Patent Office*, Certificate of patent registration date: May 11 2023

# Selected abbreviations

| | |
|---|---|
| AL | axial length. 32 |
| AUC | Area under the ROC curve. 23 |
| CCT | Central corneal thickness. 31 |
| CDSS | Clinical decision support systems. 5 |
| CH | corneal hysteresis. 31 |
| CLS | contact lens sensor. 27 |
| CRF | Corneal resistance factor. 31 |
| CV | Cross-validation. 20 |
| GAT | Goldmann applanation tonometry. 31 |
| GBM | Gradient Boosting Machine. 16 |
| IOP | Intraocular pressure. 30 |
| ML | machine learning. 12 |
| NTG | normal tension glaucoma. 8 |
| OCT | Optical coherence tomography. 8 |
| OH | ocular hypertension. 31 |
| PCA | Principal component analysis. 74 |
| POAG | Primary open-angle glaucoma. 8 |
| SOMNOtouch NIBP | Somnomedics noninvasive continuous blood pressure monitor. 29 |
| TF | Triggerfish. 28 |

# Contents

# Chapter 1

# Introduction

The presented research was conducted as part of Applied Ph.D. Programme supported by Ministry of Science and Higher Education (DWD/4/24/2020).

## 1.1 Motivation

Glaucoma is a worldwide vision threatening disease that is a significant public health challenge. Substantial efforts has been made to improve noninvasive diagnostic methods based on retinal fundus images or optical coherence tomography (OCT) [1]. Alternative diagnostic options have become available in the recent years [2, 3]. Eye and cardiac sensors can continuously record data during 24-hour session. Acquired sensor data can be processed to build patient's diagnostic profile that provides insights independent of imaging techniques commonly used in ophthalmology [4]. There is a consensus that Triggerfish measurements are related to the changes of intraocular pressure and properties of such relation were investigated and analysed [5, 6]. Triggerfish contact lens sensor (CLS) can also record low-amplitude ocular pulsations [7] related to the heart rate with good accuracy in a majority of eyes [8]. Biomechanical properties of the eye have influence on the recorded CLS signal values [9], therefore such factors should be taken into consideration in the analysis of the CLS output. In addition, cardiovascular system properties have impact on ocular blood flow [10]. Increasing availability of sensor based devices enables observation of subtle interactions of cardiovascular system and eye function during the whole day [11].

Deep convolutional neural networks can perform optic disc, cup and retinal vessel segmentation and enable objective quantification of the optic nerve head changes [12]. Over the last years OCT technology has been significantly refined and the cost of single examination decreased. It is now commonly used in diagnosing glaucoma and other retinal diseases.

Although OCT provide high-quality tomograms of anatomical structure of the eye, we can't use the images to observe dynamic changes in this complex system which is

affected by many external factors that vary over time.

Clinical routine for glaucoma detection includes diverse tests and requires extensible experience. Intraocular pressure (IOP) measurement is one of the most important assessment criteria in the diagnosis and management of glaucoma patients. Conventional tonometric techniques only allow IOP to be measured several times a day at the clinic. Such sparse data don't provide enough information for detailed assessment of the eye as it responses to the factors related to the patient's activities (e.g. stress) and normal circadian biorhythm (e.g body position). Using Triggerfish CLS it is possible to record series of measurements every 5 minutes during the whole day (including the sleep time). Data acquired by continuous monitoring of the physiological signals can be essential in development of reliable diagnostic methods and management standards for the disease.

Application of machine learning techniques for analysis of Triggerfish CLS record and cardiac sensor data can lead to accurate assessment of the eye in early glaucoma stages and potentially allow more precise control of the condition.

## 1.2 Goals and scope of the thesis

The thesis of this research can be summarized as follows:

*It is possible to efficiently support glaucoma diagnosis and control using machine learning techniques for analysis of Trigerfish CLS and cardiac sensor data supplemented with selected clinical measurements of the eye.*

Development of analytical tools and application of machine learning models for diagnosis of glaucoma can be seen as implementation of personalized (or precision) medicine premise assuming that individual patient data can be used to more precisely detect or treat a disease.

Selection of appropriate statistical methods and machine learning algorithms (ML) is required to perform valuable assessment of a single case or specific groups of cases. In medicine it is important to understand how input data affect prediction of a ML model. Consequently, this research mainly considers interpretable methods where the output can be explained in terms closely related to the basic data properties and common clinical concepts.

In practice, software system which supports efficient application of ML models may address the requirements arising in diagnostic and analytic scenarios. Specific functions for data sharing and analysis can be provided to facilitate adoption of the system in the medical community. Exploratory data analysis approach can be used for comprehensive characterization of the available data properties (including data recorded by Triggerfish CLS and devices for continuous monitoring of cardiovascular system parameters). Ultimately, the retrieved data-based knowledge combined with the experience in the field of ophthalmology enables clinical hypothesis evaluation and refinement.

## 1.3 Key contributions

This thesis describes the following key contributions:

- Implementation of raw data processing workflow that quantifies relationship of Triggerfish CLS and cardiac sensor data in time intervals defined according to the physiological circadian cycle properties. It can be used to characterize patient subgroups and may contribute to understanding the pathogenesis and progression of the disease.

- Development of efficient predictive ML models for glaucoma diagnosis support that involve Triggerfish CLS and cardiac sensor data. Consideration of functional properties of the eye enables accurate assessment of conditions in early glaucoma stages. Models supplemented with measurements of corneal biomechanical properties have better performance metrics.

- Conception and initial implementation of the software system for glaucoma diagnosis support based on ML models involving sensor data and selected clinical measurements of the eye properties. Provision of data visualization functions facilitates identification of specific features related to the course of disease and prognosis.

- Proposal of collaborative research scenarios for the eye doctors and data scientists with reference to the aspects of transdisciplinarity and personalized medicine.

# Chapter 2

# Clinical decision support and personalized medicine

This chapter briefly describes concept of personalized medicine, clinical decision support and provides examples of application of ML methods in medicine.

## 2.1 The role of ML in personalized medicine

The aim of personalized (also known as precision) medicine is to provide a more precise approach for diagnosis and treatment of disease. It harnesses innovative methods to characterize a specific case on the basis of its comprehensive data profile (including genomic and environmental information). ML and statistical techniques are used in many fields of modern medicine. Radiomics is quantitative analysis of medical images involving deep learning algorithms [13] and advanced geometrical and topological methods [14]. Bioinformatics is using assembled DNA and RNA sequencing data for assessment of disease at the cellular level.

Integration of clinical data acquired by multiple diagnostic tools can facilitate initial case classification and the assessment of selected treatment. Quantitative features extracted from different imaging modalities output can be used in combination with other patient information to improve patient management and clinical resource allocation. Challenges related to the integration of heterogeneous health data sources also include sharing and privacy issues [15].

Within the field of glaucoma detection and control there are several approaches that can be recognized as individualized or personalized [16]. One of the aims of personalized or precision medicine is identification of patient subgroups that have specific characteristic related to diagnosis or management of a disease. There are reports (describing mainly single nucleotide polymorphisms) on genetic variants associated with risk and progression of particular glaucoma types [17, 18]. Such preliminary studies may contribute to

understanding the pathogenesis of the disease, but in clinical practice it seems useful to identify subgroups according to the available data such as correlations of Triggerfish and cardiac signal in particular time intervals [19], eye biomechanical properties [20] and selected clinical data of the patient.

Interpretability of predictive models is an important issue in application of the models developed in research process. Identification of domain-specific set of constrains that make reasoning process understandable is essential in medicine. General properties of the models involving e.g. relative attribute importance can be provided for the users to help adequate model choice and comparison of results. Prediction of an interpretable model for an individual case can be explained in terms known for the users, providing description in the form of inference rules or relevant quantitative summary.

Generally, there is no tradeoff between accuracy and interpretability when we consider the full process of turning data into knowledge that can be used in practice. Interpretability is useful for troubleshooting or comprehensive asessment of the models, which leads to better accuracy, not worse. We should expect both performance metric and interpretability metric to be iteratively refined in the multi-step process of model development [21]. Interpretability is also a key element of trust for ML models as users can decide whether predictions are reliable for specific cases and general model properties are consistent with expected characteristics.

## 2.2 Clinical decision support

Clinical decision support systems (CDSS) are intended to improve healthcare delivery by providing essential clinical knowledge, patient information and other health data relevant to decision-making [22]. Currently CDSS are software systems with web interface that can be connected to data sources such as electronic health records (EHR) and repositories of diagnostic data.

CDSS can be based on the set of reference information (knowledge base). Clinical structured information is processed by algorithms in the system inference engine to produce rules that can be used in decision-making for an individual case. This type of CDSS implements evidence-based guideline recommendations that address both preventive practices and management of disease. GLIDES (Guidelines Into Decision Support) is an example of such a system that has been created at Yale University School of Medicine (USA). It provided clinical decision support for pediatric asthma and obesity management using recommendations of National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC) [23].

ML-based (or AI-based) CDS systems don't use structured knowledge base or any explicit reference information representation. These systems facilitate decision-making by using ML models designed for handling specific clinical scenarios.

ML models for multiple disease detection based on standard laboratory tests can facilitate initial patient assessment [24]. Predictive performance of these models is relatively decent compared to standard routine in clinical environment, especially when time allocated for patient is limited. Application of the models can prevent the oversight of important diagnostic factors when the range of clinical data available for patient is large. Implementation of comprehensive visualization and data management methods in the system enable embedding ML model results in the extended context of the patient's data [25].

Combined ML and rule-based approach also has been proposed for EHR data processing in CDSS (e.g. for detection of patient allergies [26]).

CDSS can increase compliance with specific guidelines and standards in clinical environment. Important requirements and suggestions can be provided as alerts or reminders for CDSS users. Tracking and controlling functions can reduce risk of errors. CDSS can enable efficient scheduling and selection of diagnostic path for a patient. CDSS also can include reporting services for clinical documentation maintenance.

Potential pitfalls of CDSS include negative impact on users and clinical process due to inappropriate design of some services, limited customization or introduction of too many constraints in a dynamic healthcare environment. Initial cost of CDSS deployment may be relatively high as it usually encompasses implementation of organizational changes in the clinic, user training and engineering support.

# Chapter 3

# Glaucoma

This chapter provides basic knowledge about the eye and glaucoma. Application of ML techniques in diagnosis and control of the disease was also described.

## 3.1 Basic medical knowledge

Advances in understanding anatomical structure of the eye and its physiology facilitate development of new diagnostic methods and therapeutic standards in ophthalmology.

Eyes dynamically adapt to changing external conditions. Light enters the eye through the clear cornea. The amount of light is controlled by the circular pupil located in the center of the thin iris. The lens which lies behind can change its shape to focus the light onto the light sensitive tissue called retina. Retina contains photoreceptor cells that absorb the photons and finally produce electrical impulses that are transmitted via the fibres of the optic nerve to the visual cortex in the brain.

The aqueous humour is a transparent fluid inside the front part of the eyeball. Intraocular pressure (IOP) depends on the balance between aqueous humor production and its drainage through the trabecular meshwork and ciliary muscle. Ocular aqueous humor is produced continuously in the ciliary processes of the ciliary body to supply nutrients to the lens, cornea and avascular tissues. It flushes away their metabolic waste products, provides stabilization of the ocular structure and regulation of the homeostasis of eye tissues [27]. Diverse pathological conditions affecting IOP can develop when the balance between inflow and outflow is disturbed. Understanding the complex mechanisms that regulate aqueous humor circulation is essential for better diagnosis and control of glaucoma.

Glaucoma is a group of eye diseases that lead to damage of the optic nerve. This neurodegenerative disorder is characterized by progressive loss of retinal ganglion cells and optic nerve axons. It is the second leading cause of blindness worldwide (after cataract) and the most frequent cause of irreversible vision loss. Prevalence of glaucoma

was investigated in many studies for different populations [28, 29]. It is estimated within the range 1%-4% for European populations.

Primary open-angle glaucoma (POAG) is the most common type of glaucoma. POAG is classified into high tension glaucoma (HTG) and normal tension glaucoma (NTG). Elevated intraocular pressure is the main feature in HTG, whereas in NTG, the IOP value is within the normal range (IOP $\leq$ 21 mm Hg) for the population.

Main risk factors for glaucoma include older age, family history of the disease and high myopia. Comprehensive eye examination is necessary to detect glaucoma early. Diagnostic routine includes diverse tests and requires significant clinical experience. IOP measurement is one of the most important assessment criteria in the diagnosis and management of glaucoma patients [29].

Optical coherence tomography (OCT) is common diagnostic approach for detecting glaucomatous structural damage [30]. OCT enables objective quantification of optic nerve head changes in glaucoma [31]. Analysis of retinal images taken by high-resolution fundus camera is another diagnostic method [32].

While development of techniques for image data processing is quite intensive there are other diagnostic options that became available in recent years [33]. Eye and cardiac sensors continuously register data that can be processed to build patient's diagnostic profile that provides insights of the eye and cardiovascular system interactions. It is the main research issue considered in this thesis.

Automated perimetry is also used for visual field testing. It enables monitoring of disease progression.

Treatment in glaucoma is usually focused on reduction of IOP with various drugs or minimally invasive surgical procedures. Emerging stem cell therapies are aiming at restoring function of the eye by reconstructing ocular tissue. Trabecular meshwork stem cells can be used in regenerative or protective treatment. Preliminary experimental results seem promising but there are many challenges related to the stem cell delivery, integration and safety [34].

## 3.2 ML in glaucoma diagnosis and control

Introduction of new diagnostic devices and advances in ML techniques enable progress in ophthalmology and fundamental eye research.

Optical coherence tomography is noninvasive and safe method of examining soft tissue. Since 1991 OCT has demonstrated its applicability in detailed, cross-sectional visualization of the eye's structure. It allows a qualitative assessment of tissue features and detection of pathological changes. An OCT tomogram is a cross-sectional image representing the optical reflectance properties of the examined tissues.

Annotation of 3D structures requires analysis of a series of cross-sections, that can include several hundred images (depending on the device and scanning protocol). It is not possible to make extensive analysis of such amount of data manually in a limited time (what is common in clinical practice). Automatic parameterization of the retinal images can support understanding the effects of structural changes in the eye on vision quality. Application of ML algorithms for image segmentation and comprehensive analysis of OCT data allows quick assessment of the features relevant to many eye pathologies such as diabetic retinopathy (DR), age-related macular degeneration (AMD) and glaucoma. Currently it is possible to determine thickness of the retinal nerve fiber layer (RNFL), structure of the optic nerve head (ONH) and morphology of vascular network of the retina [35].

Figure 3.1 shows sample tomograms of the ONH. ONH cupping is the common clinical feature of glaucoma. Prelaminar cupping of the ONH surface is characterized by progressive loss of the prelaminar neural tissues, which results in the increase of the depth and width of the cup (thus increasing the cup to disk ratio). In most cases, glaucoma leads to damage and remodeling of the laminar connective tissues and progressive loss of retinal ganglion cell (RGC) axons [36]. Figure 3.2 shows topografic thickness information of the RNFL for the same eyes as in figure 3.1.

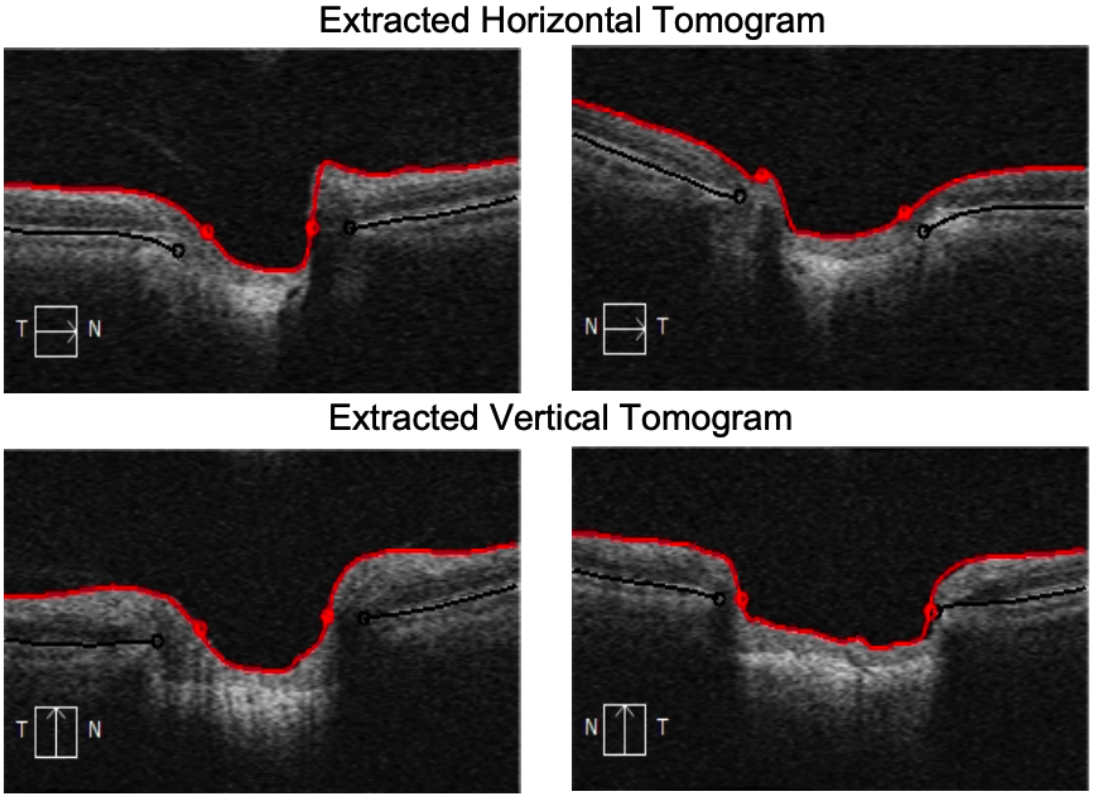


FIGURE 3.1: Sample OCT tomograms of the optic nerve head (ONH). Healthy eye images on the left. Images of the eye with glaucomatous optic neuropathy on the right (increased average cup to disc ratio). Image courtesy of Robert Wasilewicz.

Many advanced deep learning architectures have been introduced in the last years [37]. Quality of the OCT imaging output has improved (quantified by the histogram-based maximum tissue contrast index (mTCI), image resolution or other metrics). Inherent issues affecting the quality of OCT data output can be technology-based (noise, acquisition errors) and biology-based (eye movement, heterogeneous tissue reflectivity, shadows related to the blood flow).

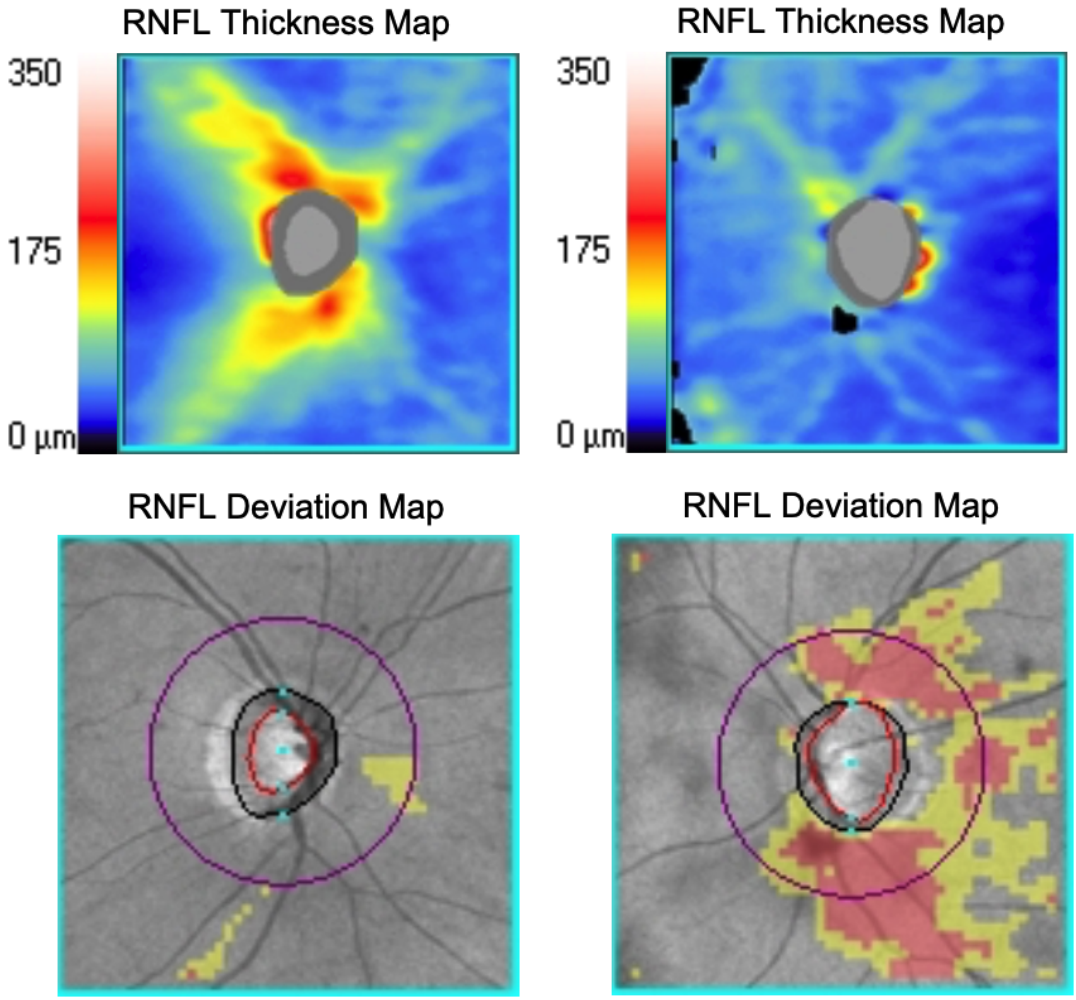


FIGURE 3.2: Sample OCT imaging output providing topographic thickness information of the retinal nerve fiber layer (RNFL). Healthy eye images on the left. Images of the eye with glaucomatous optic neuropathy on the right (retinal nerve fiber layer assymetry and decreased average thickness). Image courtesy of Robert Wasilewicz.

Convolutional neural networks (CNN) have been used for automated segmentation and classification of OCT images of cases with suspected glaucomatous neuropathy [38]. CNN are feed forward (i.e. without recurrent connections) neural networks applied for solving computer vision task in many fields of medicine. CNN is a sequence of layers that are specifically useful for transformation of the input raw images (composed of many channels). Convolutional layer is the basic building block of CNN. This layer computes dot product of the filters (kernels) and local regions of the input volume. It is

called feature (or activation) map. Pooling layer is usually placed after a convolutional layer. It performs downsampling operation on the previous layer volume to reduce the amount of parameters, which can prevent overfitting of the network. Average and max pooling of squared region are the common operations in this layer. Transition to fully-connected layers at some point is a standard network pattern for CNN architecture. The last fully-connected layers are used to handle complex hierarchical relations in feature representation produced by the previous layers. Objectives such as class scores are finally determined. Rectified linear unit (ReLU) is the common activation function used in CNN. It is a simple, nonlinear function that will output the input directly if it is positive, otherwise, it will output 0. It allows use of backpropagation for efficient network training.

Fully connected networks (FCN) have been also proposed for image segmentation and classification. This architecture uses CNN to extract image features and performs transposed convolution to produce output image that has the same dimensions as the input and can represent predicted class for the input pixels [37].

Deep learning models provide the highest overall performance for segmenting the different structures of the optic nerve head [38]. Common network architectures include U-Net based on FCN [39] and deep CNN such as VGGNet with reduced set of network parameters [40].

Deep learning algorithms require a large amount of training data [41]. Several data augmentation methods have been proposed for efficient generation of new artificial instances that are similar to the instances from the available clinical datasets [42].

Predictive models of long-term glaucoma progression based on initial ONH structural features, IOP and selected clinical data have been proposed for POAG cases. RNFL thinning can be predicted using random forest model with lamina cribrosa (LC) curvature index (considered as indicator of LC deformation related to the degree of mechanical strain on the LC) and IOP as the most important attributes [43].

Genome-wide association studies have revealed many genetic variants associated with variation in IOP. CNN have been used to investigate impact of gene variations on the trabecular meshwork cells morphology. Cell painting protocol for multiple fluorescent channels was applied to generate images suitable for morphological profiling. Differences in cellular morphology quantified by CNN indicated significant effect of gene knockout (e.g. LTBP2, BCAS3) for overall morphological variation or individual organelles (e.g. ANAPC1) [44]. This approach enables analysis of the complex genetic background related to the development of glaucoma.

# Chapter 4

# Machine learning basic concepts

In this chapter, basic concepts related to the scope of the thesis are introduced. It provides an overview of machine learning (ML) algorithms, model performance metrics and remarks on the interpretability and explanations for predictive models.

## 4.1 Overview

In 1950 Alan Turing published a famous paper [45] in which he asked if machines can think and proposed a test that can be used to distinguish between a (digital) computer and human. Since the mid-20th century, we have seen significant progress in the fields related to computational intelligence.

ML is a broad research field that lies at the intersection of the artificial intelligence (AI) and data science. Data science considers data collection, organization and analysis in order to extract important data properties or knowledge. AI concerns problems that seem to require intelligence when solved by humans and uses computational methods to find reliable solutions. ML algorithms can learn from experience in the form of observational data or interactions with an environment. We train or use ML algorithm for available data instead of explicitly writing a code to handle all possible cases or patterns (which is often infeasible or impractical).

Suppose $x_1, x_2, ..., x_p$ are independent variables (called predictors or features) and y is an output variable (called response). We assume that relationship between $X = (x_1, x_2, ..., x_p)$ and y can be written in the general form as y=$f(X) + \varepsilon$, where $f$ is some fixed but unknown function of X, and $\varepsilon$ is random error term (called noise). Noise $\varepsilon$ is independent of X, has zero mean and variance $Var(\varepsilon)$. ML refers to methods for estimating $f$ and evaluating of the results. In order to predict output $\hat{y}$ for given features X we generate $\hat{f}$ which is an estimate for $f$, i.e. $\hat{y} = \hat{f}(X)$. In general the accuracy (quality) of prediction $\hat{y}$ depends on reducible error and irreducible error. Inaccuracy related to the fact that $\hat{f}$ will not be a perfect estimate of $f$ will introduce some error. It can be reduced as it is possible to improve accuracy of $\hat{f}$ by using the most suitable

ML technique to estimate $f$. There will be other unmeasured or unknown variables that contribute to y, including measurement error. Irreducible error is an upper bound on the accuracy of prediction for y and is generally unknown in practice [46].

Other important questions in ML are related to the properties of predictive models in the context of inference, i.e. relationship between X and y. These questions refer to the association of predictors with response and adequate representation and interpretation of this association.

In supervised learning approach we consider set of observations (containing both the inputs and outputs) called training set. In the process known as learning by example difference of reference output and predicted value i.e. $y_i - \hat{f}(X_i)$ for the observations is used to build a general model of input and output relationship.

In the approach known as unsupervised learning we don't have reference output and there is no direct measure of success. For the most of unsupervised algorithms it is difficult to assess quality of the results. Unsupervised techniques are used to characterize specific patterns in big datasets or summarize properties of groups of similar objects.

## 4.2 Regression

One of the common problems in ML is prediction of continuous output variable on the basis of a set of continuous input variables (features).

Linear regression is a standard statistical method based on the assumption of linear relations in dataset. For a vector of inputs $X^T = (x_1, x_2, ..., x_p)$ we predict output y using the following model

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} x_j \hat{\beta}_j \tag{4.1}$$

If we include $\hat{\beta}_0$ called intercept as the first element in the column vector of coefficients $\hat{\beta}$ and include $x_0 = 1$ as the first element in X, it can be written as a product

$$\hat{y} = X^T \hat{\beta} \tag{4.2}$$

There are many different methods of fitting linear model to a given dataset, but the most common is the least squares method. We choose the coefficients of $\beta$ to minimize the residual sum of squares (RSS) which is a quadratic function of the parameters

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - X_i^T \beta)^2 \tag{4.3}$$

Let X be N x (p+1) matrix which contains N input vectors as rows, and vector Y contains relevant outputs. Final solution can be written as

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4.4}$$

providing $X^T X$ is invertible (nonsingular).

## 4.3 Classification

In classification we predict discrete class labels (i.e. qualitative output) on the basis of a set of continuous input variables (features). Regression and classification problems have a lot in common, and both can be viewed as approximation of a function. If we have more than two distinct class labels defined it is called multiclass classification. Binary classification is a specific type of the problem where only two distinct class labels are defined. This thesis refers mainly to the binary classification algorithms.

## 4.4 Algorithms

Wide range of ML algorithms have been proposed for analysis of various data types. There is no single approach that outperforms all others across all possible datasets. In practice, one of the most challenging tasks in ML is selecting the best algorithm for a given dataset. This section contains overview of the algorithms that were applied to analysis of the experimental data.

### Logistic regression

Logistic function can be applied to model relationship of X and probability $p(X) = P(y = 1|X)$ i.e. conditional probability that response equals 1 given the predictors. The fitted logistic curve has the following form:

$$p(X) = \frac{e^{\beta_0 + X^T\beta}}{1 + e^{\beta_0 + X^T\beta}} \tag{4.5}$$

It can be written as

$$ln(\frac{p(X)}{1 - p(X)}) = \beta_0 + X^T\beta \tag{4.6}$$

Decision boundary for this model is the set of points $\{X : \beta_0 + X^T\beta = 0\}$ which is a hyperplane.

Logistic regression model is usually fitted by maximum likelihood method. Likelihood is probability density function of a given data seen as a function of the parameters of a model. It can be written as

$$l(\beta_0, \beta) = \prod_{i:y_i=1} p(X_i) \prod_{j:y_j=0} (1 - p(X_j)) \tag{4.7}$$

Plot of the sample sigmoidal curve is shown in the figure 4.1.

### Naive Bayes

Naive Bayes algorithm is based on the simple/naive assumption of independence of the predictors [47]. For an input vector $X = (x_1, ..., x_p)$, m class labels $C_1, ..., C_m$ using

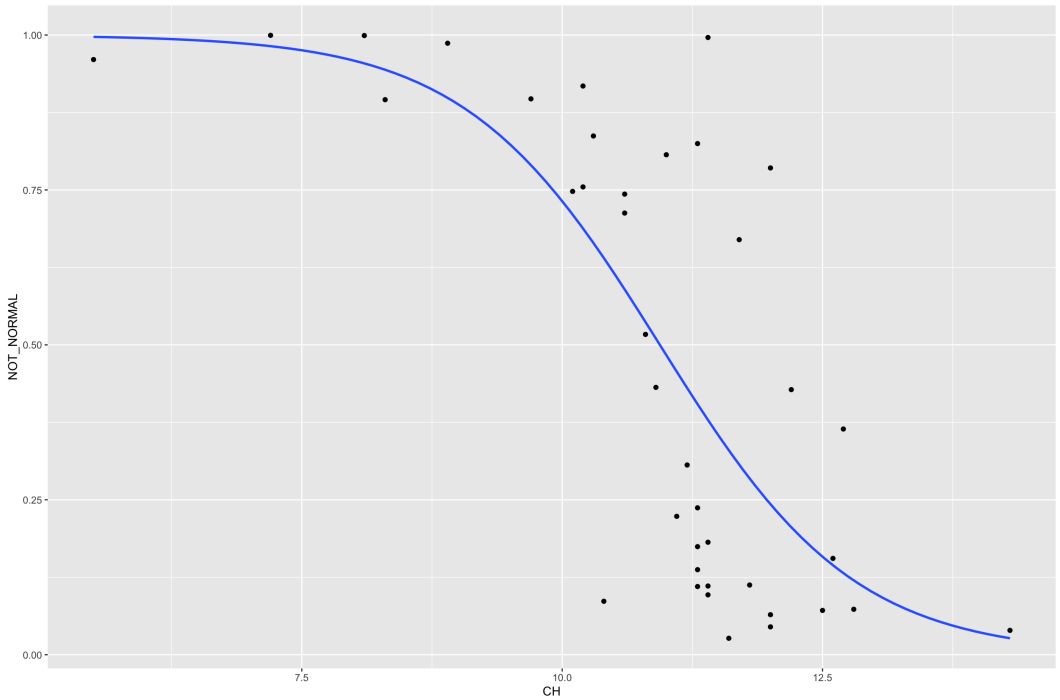


FIGURE 4.1: Example of the fitted logistic regression curve (blue) for glaucoma predictive model based on one continuous input variable (CH). Points (black) represent predictions of complex logistic regression model involving CH, CRF and Triggerfish data for the relevant CH values.

Bayes theorem we can write:

$$p(C_i|X) = \frac{p(C_i)p(X|C_i)}{p(X)} \tag{4.8}$$

We try to determine class label of X, p(X) is the same for each class and can be skipped. Let C(X) means selection of class with the highest probability for X:

$$C(X) = \arg\max_i \frac{p(C_i)p(X|C_i)}{p(X)} = \arg\max_i p(C_i)p(X|C_i) \tag{4.9}$$

Prior probability $p(C_i)$ refers to the assignment of $C_i$ class label without any specific condition. Application of the chain rule for conditional probability and assumed independence of the predictors give

$$p(X|C_i) = \prod_{j=1}^{p} p(x_j|C_i) \tag{4.10}$$

Finally

$$C(X) = \arg\max_i p(C_i) \prod_{j=1}^{p} p(x_j|C_i) \tag{4.11}$$

where $p(C_i)$ and $p(x_j|C_i)$ are easy to calculate for the given input data.

## Decision tree

Decision tree is a binary branching structure used in methods for solving classification and regression problems. Each node in the tree involves simple feature comparison against

some value (i.e. logic conditon for a feature). Predictor space is divided into a number of distinct and non-overlapping regions. It is non-parametric approach (i.e. without assumptions on data distribution) provided that no constraint on maximum tree depth is set. Many techniques such as pruning are used to limit complexity of the resulting tree by testing sequence of subtrees generated according to diverse strategies in order to reduce prediction error. It is possible to improve prediction accuracy by combining a large number of trees, with some loss in interpretability of the results. This led to ensemble methods like random forest or gradient boosting.

## Random forest

High variance is one of the disadvantages of standard decision trees: small change in the input data can result in different sequence of splits and different prediction result. Bagging (also called bootstrap aggregation) is a technique for reducing variance that is particularly useful for decision trees. In bagging we generate B training sets using random sampling with replacement for a given input data. We draw a fixed number of samples for each of the training sets. Subsequently we build the relevant separate trees using only a randomly selected subset of features for the each split in tree. Final prediction $\hat{f}$ for new observation x is by obtained by averaging (in regression) output from the all trees (similarly by majority voting in classification):

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) \tag{4.12}$$

Low correlation among the trees in the ensemble is possible due to the high variance of a single decision tree. Error rate for the ensemble is usually considerably lower than for any of the trees.

## Gradient Boosting Machine (GBM)

Boosting is yet another technique that can improve performance of decision trees. In GBM trees are grown sequentially, each tree is built using information from the previously grown trees [48]. For the each step new tree is fitted to residuals of the ensemble. This tree is added to the model to update the residuals. The shrinkage parameter $0 < \lambda \leq 1$ controls learning rate and improves generalization properties of the final model. Recurrence relation for subsequent approximations of GBM model can be written as:

$$\hat{y}_m = \hat{y}_{m-1} + \lambda\Delta_m(X), \tag{4.13}$$

where X is a matrix of observations and $\Delta_m$ is a tree built so as to minimize differentiable loss function using residual vector. Usually the trees are small (i.e. number of splits is limited). Typical values of $\lambda$ are small and a large number of trees is required to achieve good performance.

## Clustering

Clustering refers to unsupervised learning techniques for finding clusters or subsets in a dataset. Dataset is divided into distinct clusters so that the observations within each cluster are more similar to each other than compared with observations assigned to different clusters. Choice of the relevant measure of distance or dissimilarity between observations is the basis of reliable cluster analysis. Properties of distance function d defined for a set M containing observations are determined by the axioms of metric space:

1. Positivity:
$$d(x, y) \geq 0 \tag{4.14}$$
2. Identity:
$$d(x, y) = 0 \iff x = y \tag{4.15}$$
3. Symmetry:
$$d(x, y) = d(y, x) \tag{4.16}$$
4. Triangle inequality:
$$d(x, y) \leq d(x, z) + d(z, y) \text{ for all x, y, z} \in M \tag{4.17}$$

In the thesis we only consider M which is a set of vectors of fixed length. Euclidean distance is the common metric defined as

$$d_E(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \tag{4.18}$$

It is related to the shortest straight line between points in space. Many other distances are used, e.g. Manhattan (taxicab) distance based on paths along the rectangular grid:

$$d_T(x, y) = \sum_{i=1}^{p} |x_i - y_i| \tag{4.19}$$

Clustering can be used for data reduction of large datasets (for processing or visualization) or outlier detection (finding possible errors).

### K-means

K-means is a simple and efficient algorithm for partitioning a dataset into specified number of clusters. It is a centroid-based method in which each cluster is represented by a central vector that doesn't have to be a member of the dataset. K-means uses iterative refinement approach and is intended for squared Euclidean distance. Observations are randomly assigned to the K clusters $C_1, ..., C_K$ at the beginning. These assignments

are updated in each step using measure $W(C_l)$ of the amount by which the observations within a cluster differ from each other such that total within-cluster variation is minimized:

$$\underset{C_1,\ldots,C_K}{\arg\min} \sum_{l=1}^{K} W(C_l) \tag{4.20}$$

K-means algorithm finds local optimum, therefore it should be run multiple times with different random initial cluster assignments. Finally we select solution for which the objective is smallest. Many methods for determining the relevant/adequate number of clusters in a given dataset have been proposed (e.g. elbow heuristic based on the explained variance as a function of K).

#### Hierarchical clustering

Hierarchical clustering is a combinatorial algorithm processing the observations without any direct reference to an underlying probability distribution. Connectivity-based clustering produces hierarchical representation (i.e. ordered sequence of groupings) in which the clusters at each level of the hierarchy are created by merging clusters at the nearest preceding level. Two groups with the smallest intergroup dissimilarity are selected for merge at the next level in the most common agglomerative strategy. Distance between clusters X and Y is defined as the linkage criterion D(X,Y). The following main linkage types are used:

- Single linkage is the smallest distance between the observations from X and Y (minimal intercluster dissimilarity):

$$D(X,Y) = min_{x\in X, y\in Y} d(x,y) \tag{4.21}$$

- Complete linkage is the largest distance between the observations from X and Y (maximal intercluster dissimilarity):

$$D(X,Y) = max_{x\in X, y\in Y} d(x,y) \tag{4.22}$$

- Average linkage is based on the average distance between the observations from X and Y (mean intercluster dissimilarity):

$$D(X,Y) = \frac{1}{|X|\,|Y|} \sum_{x\in X} \sum_{y\in Y} d(x,y) \tag{4.23}$$

## 4.5 Feature selection

Feature (or attribute) subset selection is the process of choosing a subset of features according to some requirements (or criteria). Main aim of feature selection is usually

improvement of model performance. Another aim is selection of the most important features to facilitate interpretation of model and understanding of properties of the available data. Feature selection can also reduce complexity of model as some redundant features can be removed. Feature selection problem is particularly important in bioinformatics (for datasets where the number of features is greater than the number of observations).

There are $2^p$ possible subsets for the given p features that have to be considered in order to solve the feature selection problem. It is computationally infeasible to directly check all candidate sets for large p values. Many efficient methods have been proposed for choosing the optimal subset. Wrapper methods use output of ML model to find solution (e.g. recursive feature elimination). Filter methods determine subset independently of ML algorithms, on the basis of statistical or information related properties of the features (e.g. minimum redundancy maximum relevance or $\chi^2$ test for categorical features). Embedded methods use internal properties of specific ML algorithm (e.g. lasso or elastic net).

## Stepwise feature selection

Stepwise feature selection is one of the common wrapper methods. Forward stepwise selection begins with an empty set of features. One feature is added to the set in subsequent iteration steps. Feature that mostly improves performance of a model for the enhanced set is chosen at each step (we compare feature sets of equal size using e.g. residual sum of squares for regression). At the end we build models for the generated subsets and select one that has the best performance estimate (using e.g. cross-validation). Backward stepwise selection begins with the set containing all features. In the next steps we remove the least useful feature (regarding model performance). This method gives results close to the forward selection. There are hybrid approaches where features are added sequentially to the model and some features that no longer improve model performance can be removed at each step.

## Ridge regression and LASSO

Ridge regression and LASSO (least absolute selection and shrinkage operator) are techniques that allow shrinking coefficients or regularization of a fitted model by introducing penalties. Ridge regression is similar to the least squares fitting method but includes additional penalty term that has impact on the final coefficient estimates. Ridge regression coefficients minimize the following expression

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2 \tag{4.24}$$

As tuning parameter $\lambda \geq 0$ increases, the coefficients estimates are shrunken towards 0. For $\lambda$ equal 0 coefficients are the same as in the least squares method. Increasing the

value of $\lambda$ will not result ultimately in setting any coefficient to 0. In LASSO the penalty term in the above expression is replaced by $\lambda \sum_{j=1}^{p} |\beta_j|$. This change results in different solution: LASSO performs feature selection as some coefficients are set exactly to 0 for sufficiently large $\lambda$ values. Optimal value of $\lambda$ can be determined using cross-validation for a grid of $\lambda$ values.

## 4.6 ML model performance evaluation

Assessment of prediction capability of a ML model is very important in practice, since it defines the scope of its application. Generalization error is especially significant, as it refers to the prediction error for independent test dataset and selected loss (cost) function. Training and test data need to be drawn from the same distribution to achieve reliable results of model assessment. Training error is the average loss over the training dataset. It is not a good estimate of the test error as training error decreases when we increase complexity (also called capacity) of the model. More complex model can adapt to complicated patterns in the training dataset and fit to some irrelevant properties or noise. Eventually such overfitting leads to poor generalization performance of the model. On the other hand, if complexity of the model is too little, it generalizes poorly because it can't fit to the all relevant patterns in the dataset.

Expression for the expected prediction error of regression estimate $\hat{f}$ of f at an input point $X = x_0$ and squared loss can be written as

$$E[(y - \hat{f}(x_0))^2 | X = x_0] = Var(\hat{f}(x_0)) + Bias^2(\hat{f}(x_0)) + Var(\varepsilon) \tag{4.25}$$

This expression (known as bias-variance decomposition) refers to the mean squared error tested at $x_0$ and f estimated multiple times using a large number of training sets. $Var(\hat{f}(x_0))$ term quantifies variability of $\hat{f}$ in reference to its mean. Squared bias term is related to the error of approximation of a complicated real function f by a simpler model. It is the amount by which the average of estimate $\hat{f}$ differs from the true/reference value $f(x_0)$ The last term is related to the irreducible error $\varepsilon$ (noise) for f.

Bias-variance tradeoff is an important challenge regarding any supervised learning technique as we are trying to balance the bias and variance of a model while we have limited knowledge of all properties of the problem and incomplete data resources.

### 4.6.1 Cross-validation

Cross-validation (CV) is a resampling method used for estimation of the test error for predictive ML algorithms. In the 1970s publications it was referred to as a statistical method for assessment of the quality of any data-derived quantity [49]. It can be applied in model selection scenario to find the optimal model parameters (or complexity).

It is also a standard approach in model assessment scenario when we are estimating generalization performance of the selected predictive model and availability of data is limited.

In K-fold cross-validation the available dataset is randomly sorted at the beginning. Then it is splitted into K non-overlapping subsets (folds) of equal size. In K consecutive steps, a fold selected in the current iteration is used as a validation set and the data from the remaining K-1 folds is used as a training set. At the end we compute average estimate for the results:

$$CV_K(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}_{\chi(i)}(x_i)), \tag{4.26}$$

where $\chi : \{1, ..., N\} \mapsto \{1, ..., K\}$ is an indexing function that returns index of the fold assigned for the observation, $\hat{f}_{\chi(i)}$ is the function fitted for the fold relevant to i-th observation, L is loss function or metric.

In the approach known as leave-one-out cross-validation we have K=N and validation set contains only one observation whereas training set contains the remaining data. It can be computationally expensive compared to the K-fold approach since we have to build N models. It also can have higher variance because training sets are similar with each other [46]. Optimal value of K depends on the size of available dataset and its distribution (usually K close to 10 is selected).

### 4.6.2 Evaluation metrics

We can assign predicted binary classification label (0/1 or negative/positive) for an observation to the one of four categories: true positive (TP) for correctly predicted positive label, true negative (TN) for correctly predicted negative label, false positive (FP) for incorrectly predicted label of negative observation and false negative (FN) for incorrectly predicted label of positive observation. Confusion matrix (2x2) is convenient representation of such information for a predictive model.

Class-specific performance is commonly used in life sciences (especially in medicine). Binary classification model usually returns (raw) numerical value for an observation. Suitable classification (discrimination) threshold $t \in (0, 1)$ is needed to assign a relevant class label for the predicted probability. It is determined depending on implementation scenario, domain knowledge and the cost/risk associated with the specific classification errors. The following threshold-dependent metrics are commonly used in evaluation of binary classification models:

- Sensitivity (recall, true positive rate):

$$\frac{|TP|}{|TP| + |FN|} \tag{4.27}$$

- Specificity (true negative rate):

$$\frac{|TN|}{|TN| + |FP|} \tag{4.28}$$

- Precision (positive predictive value):

$$\frac{|TP|}{|TP| + |FP|} \tag{4.29}$$

- False positive rate (fall-out):

$$\frac{|FP|}{|TN| + |FP|} \tag{4.30}$$

- Accuracy:

$$\frac{|TP| + |TN|}{|TP| + |FN| + |TN| + |FP|} \tag{4.31}$$

Sensitivity and specificity are usually reported together (as precision and recall). Accuracy is inappropriate for imbalanced datasets (as it is high for a classifier that assigns majority class for the all observations).

F-score is a metric that combines precision and recall using weighted harmonic mean (with equal importance/weights for $F_1$):

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} \tag{4.32}$$

General F-score for parameter $\beta$ is defined by

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{4.33}$$

Real number $\beta > 0$ is chosen such that recall is considered $\beta$ times as important as precision.

Matthews correlation coefficient (MCC or $\phi$ coefficient) is a metric that summarizes the properties of confusion matrix for model. It is defined as

$$MCC = \frac{|TP| \cdot |TN| - |FP| \cdot |FN|}{\sqrt{(|TP| + |FP|)(|TP| + |FN|)(|TN| + |FP|)(|TN| + |FN|)}} \tag{4.34}$$

It is related to Pearson correlation coefficient of the predicted labels (0/1) and reference (actual) labels for observations. MCC equals 1 for perfect classifier, 0 for random class assignment and -1 for the entirely inverted classification output.

Classification threshold is determined as a result of partly subjective decision related to the application scope of model and assessment of the cost of wrong label assignment. It may lead to misleading conclusions regarding model properties. Predictive models commonly return raw probability value and class distribution can be imbalanced. Moreover, categorizing continuous outcomes of model results in information loss (e.g. small change in predicted probability for an observation can change its class label) and makes their interpretation difficult [50]. Many metrics not based on classification threshold have been proposed for estimation of performance of the models.

### Brier score

Brier score (squared loss) is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{4.35}$$

where $y_i$ is reference (true) label (i.e. 0/1), $\hat{y}_i \in [0,1]$ is predicted value for the i-th observation. For regression ($y_i, \hat{y}_i \in R$) problems it is known as mean squared error (MSE).

If we set classification threshold $t \in (0,1)$ we can replace $\hat{y}_i$ with $\hat{y}_i(t)$ that depends on t in the parameterized formula for BS(t). We assign $\hat{y}_i(t) = 0$ when $\hat{y}_i \leq t$ and $\hat{y}_i(t) = 1$ otherwise. Then $BS(t) = \frac{1}{N}(|FP| + |FN|)$. It gives $1 - BS(t) = \frac{1}{N}(|TP| + |TN|)$ which is accuracy.

### AUC

Receiver operating characteristics (ROC) graph shows (true positive rate) and FPR (false positive rate) pairs for the range of possible classification thresholds. Diagonal line (y=x) is the ROC plot for a model that randomly assigns labels (we can say it has no useful information related to classification).

Area under the ROC curve (AUC) metric is numerical (scalar) representation of model performance based on the approximation of area under the ROC curve. AUC equals 1 for perfect model, 0.5 for random class assignment and 0 for the entirely inverted classification output. AUC of a model is equivalent to the probability that the model will rank a randomly chosen positive observation higher than a randomly chosen negative observation [51]. We can use AUC metric when a model returns uncalibrated scores or class distribution is imbalanced (as ROC depends on TPR related to positive class and FPR related to negative class and these are proportions separately computed for the each class).

### Logistic loss

Logistic loss (log loss) estimates how close predicted values (uncalibrated probabilities) are to the actual values. It increases exponentially as the difference gets larger, equals 0 for perfect predictions and is defined as

$$log\,loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)), \tag{4.36}$$

where N is the count of observation dataset, $y_i$ is reference (actual) value, $\hat{y}_i$ is the predicted value for the i-th observation. This metric is used mainly for fitting ML models.

## 4.7 Explanations for predictive ML models

Complex predictive ML models built for large datasets can involve a big number of variables and parameters. It may be impossible to understand how input variables affect prediction of a model. Moreover, complex models may in many cases not work as well as it is assumed. IBM Watson for Oncology system was delivering unsafe and inaccurate recommendations [52]. Predictive performance of Google Flu Trends model deteriorated due to data drift over time [53].

Many explainable artificial intelligence (XAI) techniques have been proposed to address the issues of model prediction validation and justification. We would like to know how strong the evidence is that supports a prediction of model. Also we would like to understand which variables affect prediction and to what extent [54].

Model can be interpretable by design as linear models, rule-based models or basic classification trees. However predictive performance reduction may be the cost of direct model interpretability. We can explain predictions of complex (or black-box) models by using adequate simplifications or approximations. Many model-specific techniques for model properties exploration has been proposed. For example, there are methods for measuring variable importance in generalized linear models, random forests and neural networks. Model-agnostic explanations seem particularly useful as new ML algorithms are constantly being developed.

### Break-down plots

Break-down plot is a model-agnostic method that can be applied to any predictive model that returns numeric value. It estimates effect of explanatory variables on prediction of a model.

We can assume that prediction of a model is an approximation of the expected value of dependent variable y given values of explanatory variables X. Break-down profile quantifies contribution of an explanatory variable to model prediction by calculating the shift in the expected value of y, when the values of other variables are fixed [55]. The plots are compact and easy to understand. Basic version is suitable for the additive attributions and depends on the ordering of explanatory variables. There is also extended version of break-down profile for models with interactions (when the effect of an explanatory variable depends on the values of other variables) [54].

### Shapley Additive Explanations

Shapley Additive Explanations (SHAP) use Shapley values developed in cooperative game theory. This method is based on averaging the value of attribution for a variable

over all possible orderings. Subsampling can be used to efficiently compute SHAP for a large number of variables. Shapley values have the following properties:

- (additivity) if a model m is sum of two models g and h, then the Shapley value for m is sum of Shapley values for g and h
- if a given explanatory variable doesn't contribute to any prediction for any set of remaining explanatory variables, then its Shapley value is equal to 0
- (local accuracy) the sum of Shapley values match the difference between prediction and a baseline (expected) value

Standard SHAP methods can incorrectly estimate contributions of correlated variables. Asymmetric Shapley technique has been proposed for such input data.

## LIME

Local Interpretable Model-agnostic Explanations (LIME) method is useful for sparse (with a limited number of variables) explanation of predictive models with a very large number of input variables [56]. Complex (black-box) model f is locally approximated around a given instance (case) $x_*$ by a simple (interpretable) model. We minimize loss function L to find optimal approximation $\hat{g}$:

$$\hat{g} = \arg\min_{g \in G} L(f, g, v(x_*)) + \Omega(g), \tag{4.37}$$

where g belongs to a class G containing interpretable models (e.g. linear models or decision trees), $v(x_*)$ is neighbourhood of approximation, L measures difference between models f and g in the neighbourhood, $\Omega(g)$ is penalty for complexity of model g.

Black-box model f is defined on high-dimensional space, whereas glass-box model is defined on low-dimensional space of explanatory variables. In the case of tabular input data continuous variables can be discretized to obtain interpretable categorical and combination of categories can be used for categorical variables. New, artificial data points are often required to build a glass-box model as there may not be enough points in high-dimensional input dataset which is usually very sparse. New data can be generated using various perturbations of the instance of interest (e.g. adding Gaussian noise to continuous variables).

Combining the results of various techniques for instance-level explanation can provide additional insights and more detailed view of the predictive model properties. The scope of application of a particular method depends on the data characteristics, specifically it is determined by relationship between the input variables [54].

The methods presented in this section can also be used in many tasks in the iterative modelling process:

- Model refinement/debugging: investigation of the reasons for incorrect predictions for selected cases may provide hints for model improvement
- Domain-specific validation: user can consider a model as reliable or plausible when the influence of the explanatory variables on model predictions is consistent with expectations based on the domain knowledge.
- Model selection: if overall performance of models is similar, we may use explanation techniques to select one of the candidates by the examination of instance (case) predictions.
- Extraction of new knowledge: if domain knowledge is limited or unavailable then we can extract valid information by the assessment of dependencies between model explanatory variables.

# Chapter 5

# Sensor data and clinical data

This chapter refers to the basic types of patient data involved in the research and relevant measurement methods.

## 5.1 Triggerfish contact lens sensor

The doctor is able to perform eye measurements regarding the IOP only at single point in time (when patient visits clinic or hospital). This does not provide complete information for reliable assessment as the eye changes during the day in response to the factors related to the patient's activities (e.g. stress) and normal circadian biorhythm (e.g body position).

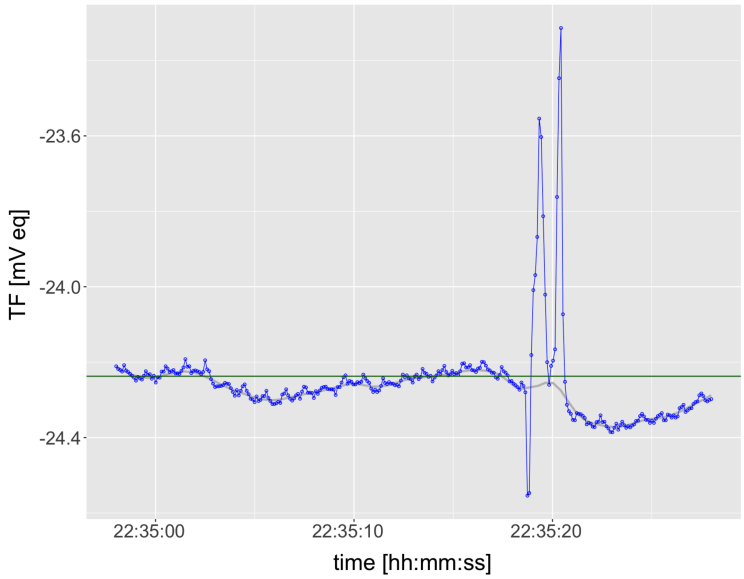


FIGURE 5.1: Sample series of Triggerfish sensor measurements (one burst). Two peaks recorded after 20 seconds from the begining of the burst are probably related to the eye blink.

Soft disposable silicone contact lens sensor can record circumferential changes at the corneoscleral area. Triggerfish (Sensimed) device [8] is based on a contact lens sensor (CLS) with embedded strain gauge that measures ocular volume changes during 24-hour

session. Series of measurements is recorded every 5 minutes in the units of milllivolt equivalents with sampling rate of 10 samples/s (each series called burst has 300 values). Adhesive antenna is placed around the eye. It wirelessly receives the information from the CLS. The data is transmitted via a thin cable from antenna to the portable recorder. The recorded data is transferred via Bluetooth interface to the computer at the end of the recording session (usually at doctor's office). Triggerfish device can be used during the whole day, assuming normal daily activity not affecting external antenna around the eye. Triggerfish is safe and generally well tolerated by patients. Side effects are rather rare and include dry eye or eye irritation. CLS is supplied in a pre-packed sterile delivery unit. It has been designated as a single use device. Figure 5.1 shows measurements of the one series, where low-amplitude ocular pulse can be noticed. Triggerfish record for 24-hour session is presented in Figure 5.2. There is a consensus that Triggerfish (TF) measurements are related to IOP changes and properties of such relation were investigated and analysed [5, 6]. CLS can also record low-amplitude ocular pulsations [7] related to the heart rate with good accuracy in a majority of eyes [8]. Biomechanical properties of the eye have influence on CLS signal values [9], therefore such factors should be taken into consideration in the analysis of the CLS output. As the CLS is placed on the eye surface for many hours it affects corneal surface and may change measurements especially at the end of a long recording session. Such inaccuracy of the output is not exactly known and depends on many factors such as biomechanical properties of the cornea.

Sensimed is developing novel pressure measuring contact lens sensor for direct monitoring of IOP in mm Hg (standard units for tonometry) instead of relative units used by Triggerfish. This CLS is based on a Micro-Electro-Mechanical System (MEMS) [57] that acquires data with sampling rate of 51 samples/s (series of measurements is recorded every 3 minutes). Current bioengineering research involves soft contact lenses designed to release medication into the eye over an extended period of time. They can be used to treat some ocular conditions including glaucoma [58]. Future research challenges encompass contact lenses with integrated eye monitoring and therapeutic capabilities (i.e. controlled drug release dependent on continuous CLS measurements).

## 5.2 Noninvasive continuous blood pressure monitoring and related physiological parameters

Continuous monitoring of arterial blood pressure is important for accurate assessment and control of many conditions. Simple devices based on the cuff inflated at equal time intervals don't provide sufficiently precise results (especially during sleep time). Currently some wearable devices are able to continuously monitor cardiovascular sys-

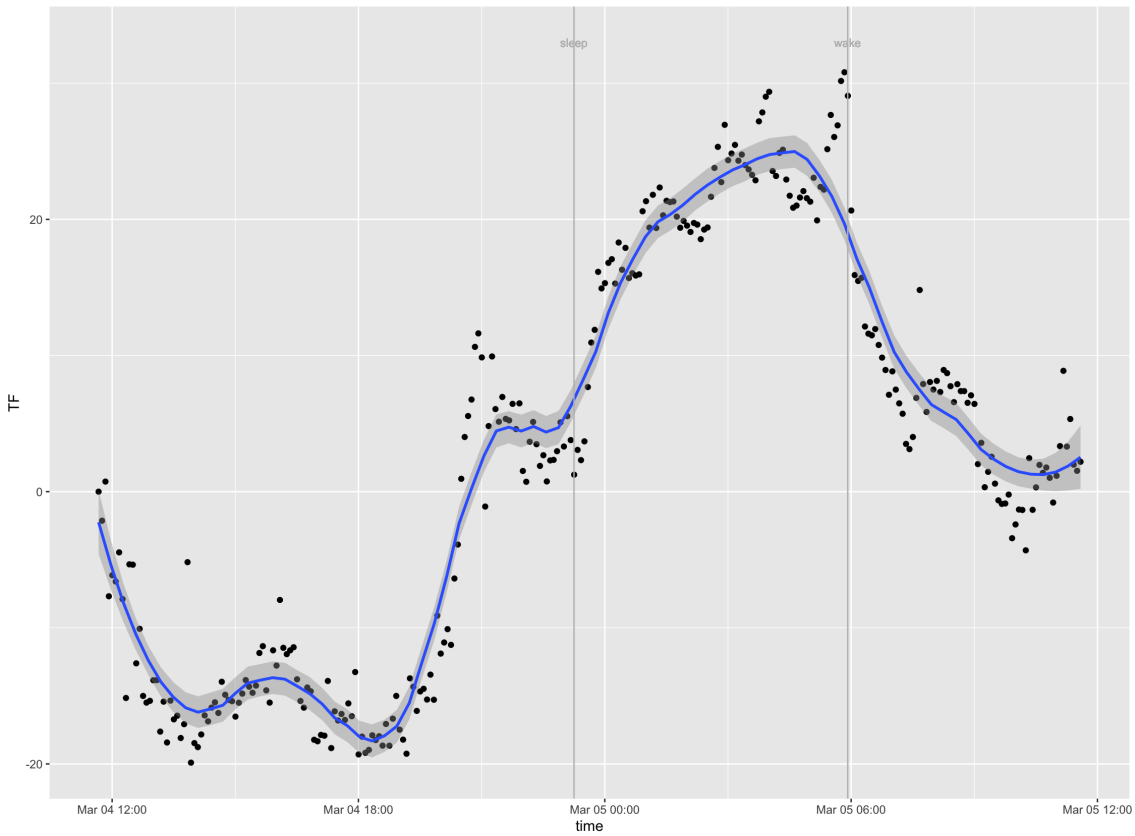


FIGURE 5.2: 24-hour session Triggerfish CLS output for a normal (healthy) patient. Each black point represents median of one TF series (one burst) in the units of milllivolt equivalents [mV eq]. Two grey vertical lines mark the beginning and the end of sleep period. Smoothed blue line is the loess (locally weighted polynomial regression) approximation of TF.

tem parameters with accuracy comparable to the professional ambulatory equipment. Measurement method can be based on the piezoelectric sensors encapsulated with flexible silicone, providing soft and conformal contact between the sensors and the artery located skin region. It allows accurate conversion of local deformation of the sensor caused by the expansion or contraction of the artery into electrical output [59]. Other methods involve photoplethysmography (PPG), ultrasound wall tracking, bioimpedance or capacitive sensors [60]. Cardiovascular system properties have impact on ocular blood flow [10]. Increasing availability of sensor based devices enables observation of subtle interactions of cardiovascular system and eye function during the whole day [11].

Somnomedics noninvasive continuous blood pressure monitor (SOMNOtouch NIBP) [61] is a cuff-less, compact device (see plot of the values recorded for the one burst in the Figure 5.3). It measures blood pressure beat-to-beat, on the basis of pulse transit time (PTT), which is determined by a 3-lead ECG (electrocardiogram) and the PPG from the finger clip. PTT is the time required for the pulse wave to propagate along the vessel wall between two defined points. In SOMNOtouch NIBP it is the distance from the left ventricle of the heart (defined by the R peak of the ECG) to the finger tip (determined by PPG). It uses complex algorithm for estimation of the following parameters:

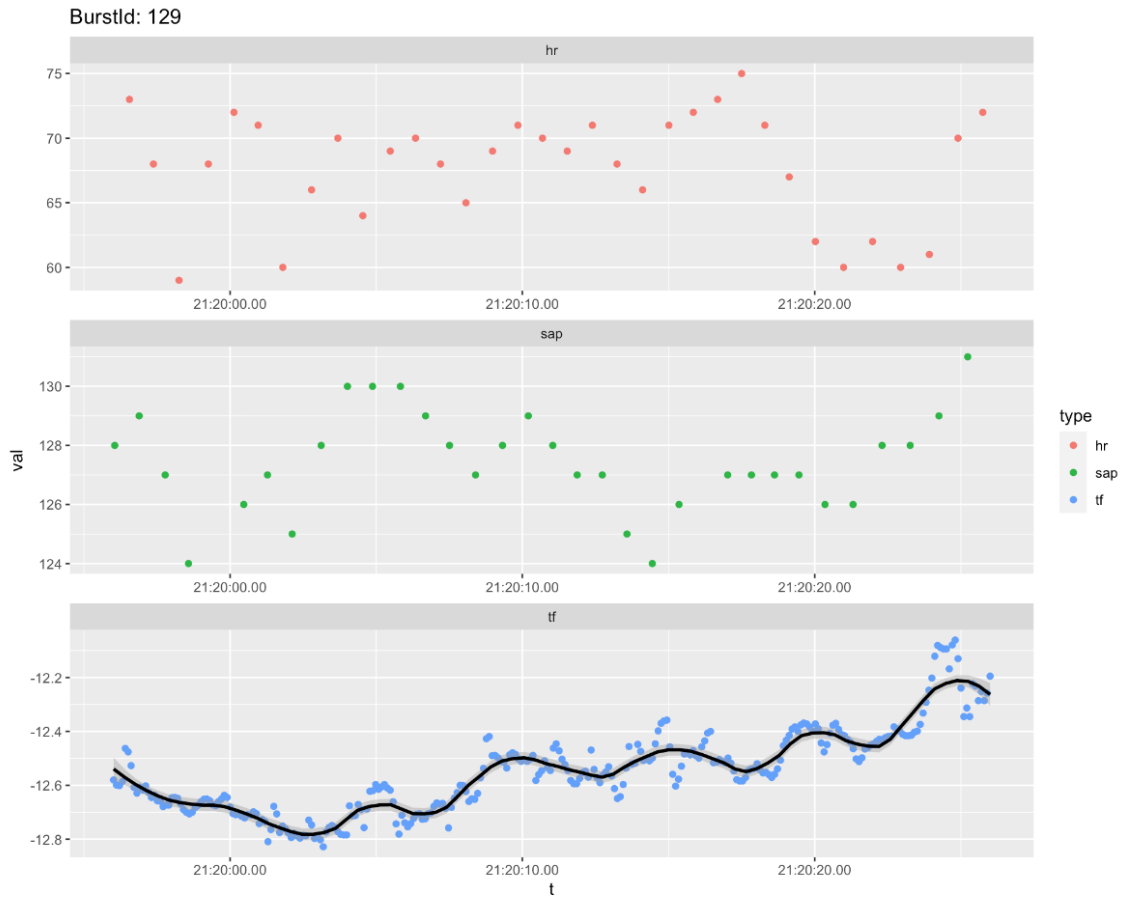


FIGURE 5.3: Plot of the TF joined with cardiac sensor signal for the one sample burst (about 30 seconds). Initial times of Triggerfish and SOMNOtouch NIBP are synchronized.

- blood pressure (SAP: systolic arterial pressure, DAP: diastolic arterial pressure, MAP: mean arterial pressure), reported in mm Hg
- heart rate (HR), reported in beats per minute
- blood oxygen saturation ($SpO_2$), reported in percent

Initial calibration at the beginning of the session is required for the device (independent measurement of blood pressure). Although SOMNOtouch NIBP and similar devices have undergone preliminary evaluation and clinical tests, they are not used on a large scale. In some cases, measurement inaccuracy can be higher than expected therefore further clinical assessment is needed [62].

## 5.3 Clinical data

### 5.3.1 Intraocular pressure

Intraocular pressure (IOP) is measured in millimeters of mercury (mm Hg). IOP is the main risk factor for glaucoma and its assessment is important in eye conditions related

to the ocular hypertension (OH) when the IOP is higher than normal without detectable glaucomatous neuropathy.

Goldmann applanation tonometry (GAT) is a standard technique for measurement of IOP. It is based on Imbert–Fick law that can be used for description of axially symmetric ellipsoidal elastic shell loaded with internal pressure and flattened over a small surface [63]. The ratio of the external force and the area of the flattened cornea surface is the measure of IOP. GAT results depend on central corneal thickness (CCT) and axial radius of the external corneal curvature. Area of the applanation surface is controlled during measurement. Appropriate calibration (correction) values for these parameters are required to calculate the final IOP value [63].

Dynamic contour tonometry (DCT) is a new technique for noninvasive IOP measurement. DCT is based on a miniature piezoresistive pressure sensor embedded within a tonometer tip that approximates shape of the corneal surface when the pressures on both sides are equal. It is less dependent on the effect of individual biomechanical properties of the cornea, as is the case in applanation tonometry. It measures IOP in a continuous way and provides a pressure curve that is synchronous with the cardiac cycle for a period of several seconds [7]. DCT is independent noninvasive method for IOP assessment as correction formulas for GAT are not uniformly accepted or validated.

Another common technique for IOP measurement is non-contact tonometry. It is useful for screening tests. Non-contact tonometer uses a jet of air that flattens external surface of the cornea. Light that is reflected from the central cornea is used for indirect determination of the IOP.

### 5.3.2 Properties of the cornea

Hysteresis is a measure of the viscoelastic response of the cornea affected by the external force. Cornea has elastic properties (associated with reversible deformation) and time-dependent viscous resistance to an applied force. Reichert ocular response analyzer (ORA) is a device for measuring corneal compensated IOP and biomechanical properties of the cornea that uses brief pulse of air to perturb its surface. Data recorded by the ORA when the air pressure is affecting the outer cornea can be used to build the model of the viscoelastic corneal surface [64]. Surface of the cornea reflects infrared light which is recorded by the detector. The reflected light is maximally aligned with the detector when the cornea undergoes applanation (is flattened). Loading (P1) and unloading (P2) applanation pressures are different and P1 is higher than P2. Difference between P1 and P2 is reported as corneal hysteresis (CH) in mm Hg. Corneal resistance factor (CRF) measured by ORA also depends on the values of P1 and P2. It can be considered an indicator of the overall resistance of the cornea and is reported in mm Hg.

Central corneal thickness (CCT) is important parameter in refractive surgery and

assessment of the conditions related to ocular hypertension. It is usually reported in micrometers (μm). Measurement techniques of CCT include ultrasound pachymetry, optical coherence tomography (OCT) and corneal topography. CTT measurements are not directly comparable for different types of devices [65].

### 5.3.3 Other data

The following data were also collected for the patients:

- axial length (AL) is the distance from the corneal surface to the retinal pigment epithelium (in optical measurements). Large AL (e.g. related to myopia) can have influence on biomechanical properties of the eye and may lead to progressive thinning of some of the retinal layers.
- other patient's data: age, sex, underlying medical conditions (e.g. diabetes).

# Chapter 6

# Development of machine learning models for glaucoma detection

This chapter is based on the results which were published in the article written by the author (see [A2]).

Development and application of machine learning models in the field of ophthalmology focused on glaucoma can be seen as implementation of personalized medicine [66] premise assuming that individual patient data can be used to more precisely detect or treat a disease. Application of wearable medical devices is growing in many fields of healthcare. Data acquired by continuous monitoring of the physiological signals can be essential in development of reliable diagnostic methods and management standards for disease. Identification and evaluation of the relationship between time series recorded by multiple sensors can be a way to better understanding the nature of condition. The great majority of existing papers on use of AI in the field of glaucoma detection is related to deep learning methods (mainly convolutional neural networks) for structural image analysis (fundus photos or OCT) [30, 31]. Previous research in using ML for sensor data is limited to the assessment of models involving Triggerfish signal. Investigation of relationship between eye sensor signal and cardiac activity can result in refinement of detection models and facilitate automatic diagnostic decision support for the disease.

Research hypothesis in this study assumes that heart monitoring data associated with Triggerfish measurements can be used to more accurately detect glaucoma. We continue investigation on influence of cardiovascular system on 24-hour ocular volume changes measured with CLS [67] where correlation of Triggerfish and cardiac sensor data was examined for healthy and POAG cases. Earlier studies concerned daily biorhythms of the eyeball volume changes and cardiovascular system functional properties depending on diagnosis [68]. This thesis reports results for predictive models involving wide range of sensor data based attributes. Particularly, it presents new approach for division of recording period according to physiological circadian cycle properties. Main innovative

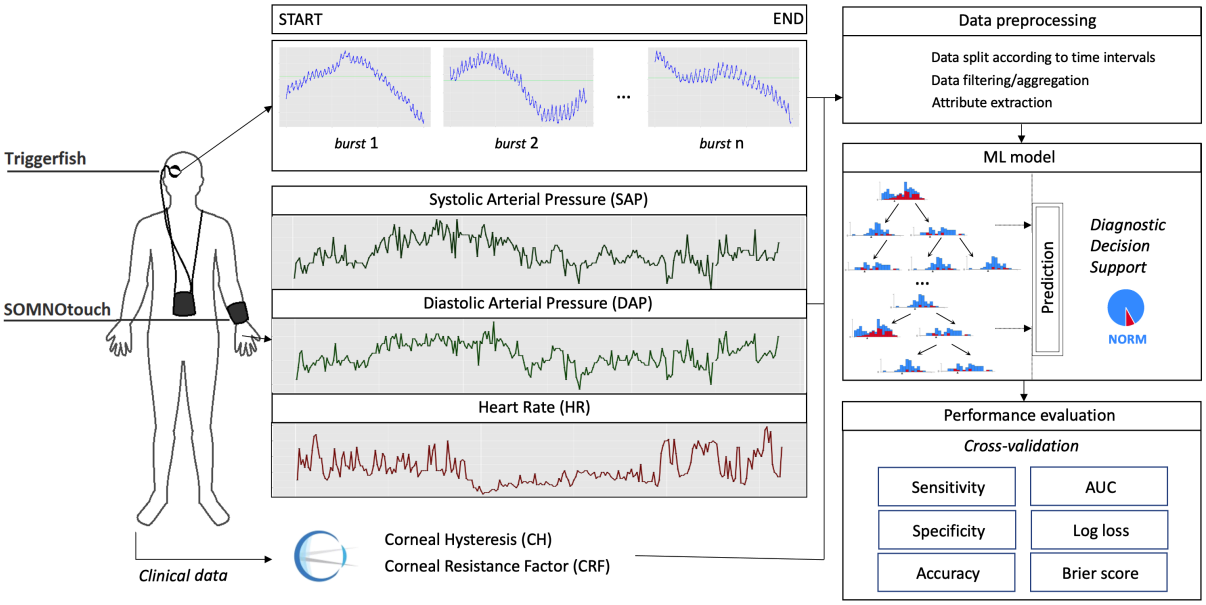


FIGURE 6.1: Overview of data flow in development and application of machine learning models for glaucoma detection and control.

aspect of the research is incorporation of data from cardiovascular system monitoring device that is aligned to the eye sensor signal. We also show that models supplemented with measurements of corneal biomechanical properties have better performance metrics.

General schema of the proposed method is shown in Figure 7.2. Diagram components refer to data acquisition (24-hour recording session for Triggerfish and SOMNOtouch devices), data preprocessing (involving feature/attribute extraction), model performance evaluation and application of the model in diagnostic decision support.

The key contributions of the research can be summarised as follows:

1. Machine learning approach involving Triggerfish CLS record and cardiac data considers functional properties of the eye that can lead to accurate assessment of conditions in early glaucoma stages and potentially allow more precise control of the disease. For comparison, deep learning techniques involving OCT imaging data detect structural changes which are typically related to progression of glaucoma [30, 31].

2. New models presented in the thesis don't rely on IOP value that can be registered (e.g. by Goldmann applanation tonometer) no more than few times a day. Instead we consider Triggerfish and cardiac sensor data based attributes for 24-hour monitoring session (that is essential e.g. in characterization of normal tension glaucoma cases).

3. Techniques applied for processing of sensor data and clinical data results in improved performance of classification models that can be included in diagnosis sup-

port system for glaucoma. Relation of Triggerfish and cardiac signal in selected time intervals which is considered in the study can be used to identify/characterize patient subgroups and may contribute to understanding the pathogenesis and progression of the disease.

## 6.1 Input data

Input data was collected at [Wasilewicz] Eye Clinic in Poznań. Main input dataset contains 105 cases (67 females and 38 males). According to the diagnosis it includes 45 high tension glaucoma (POAG/HTG), 21 normal tension glaucoma (POAG/NTG) and 39 control/healthy (NORM) cases. Prevalence of glaucoma in different populations was investigated in many studies [29] and is estimated within the range 1%-4%. Prevalence and characteristics of glaucoma depends on many factors (e.g. age) and the minority class involves common diagnosis types that are labeled in the thesis as NOT_NORM (also as glaucomatous neuropathy). Standard random undersampling approach was applied during data collection process to obtain closer case count of the both classes from the dataset. Randomly selected cases in the majority class (NORM) were skipped. The resulting distribution is more balanced and seems appropriate for the chosen binary classification algorithms.

Initial intraocular pressure (IOP) was measured by Goldmann applanation tonometer [63] before application of Triggerfish contact lens sensor (CLS). Basic biomechanical parameters of the cornea i.e. corneal hysteresis (CH) [69] and corneal resistance factor (CRF) were measured additionally. Systolic/diastolic arterial pressure (SAP/DAP) and heart rate (HR) were being recorded continuously during 24-hour period by SOMNOtouch NIBP (see description of the devices and measurement methods in 5). Table 6.1 summarises the basic attributes of the input dataset. 24-hour CLS record (containing 288 bursts) is available for the each case. Triggerfish and SOMNOtouch NIBP device internal times were synchronized to enable derivation of reciprocal relation of the acquired signals.

| attribute | min | Q.25 | median | Q.75 | max | mean | sd |
|---|---|---|---|---|---|---|---|
| age [years] | 22.00 | 46.0 | 59.0 | 69.0 | 86.0 | 56.4 | 14.32 |
| initial IOP [mm Hg] | 10.00 | 15.0 | 18.0 | 21.0 | 52.0 | 18.6 | 5.71 |
| CH [mm Hg] | 6.30 | 9.0 | 10.3 | 11.4 | 14.8 | 10.2 | 1.73 |
| CRF [mm Hg] | 7.50 | 10.2 | 11.2 | 12.4 | 16.8 | 11.3 | 1.78 |
| systolic [mm Hg] in *SLEEP_WAKE* interval | 84.00 | 107.0 | 121.0 | 130.1 | 154.0 | 119.8 | 16.59 |
| diastolic [mm Hg] in *SLEEP_WAKE* interval | 34.50 | 64.5 | 71.0 | 81.0 | 106.0 | 72.5 | 11.78 |
| heart rate [bpm] in *SLEEP_WAKE* interval | 45.00 | 57.0 | 61.6 | 67.0 | 82.0 | 61.8 | 6.92 |

TABLE 6.1: General statistical summary of the input dataset selected properties. Lower and upper quartile (Q.25 and Q.75) were included.

## 6.2 Data preprocessing

Sensor raw signals are transformed in order to get data suitable for feature (attribute) construction which is the next step in the development of ML models considered in this chapter. We compute separate median value for the each raw signal in time intervals overlapping Triggerfish series (bursts). This low-dimensional representation was proposed earlier [70] and captures the underlying characteristics of the high-dimensional input data. Median TF dispersion during 24-hour session is shown in Figure 6.2. Scatter plot left panel's NORM group contains healthy and ocular hypertension (OH) cases, right panel's NOT_NORM (glaucomatous neuropathy) group contains POAG cases. Scatter plot was generated for almost 300 cases using *ggplot2* library in R (median TF black points were plotted partially transparent to reduce effect of overlapping).

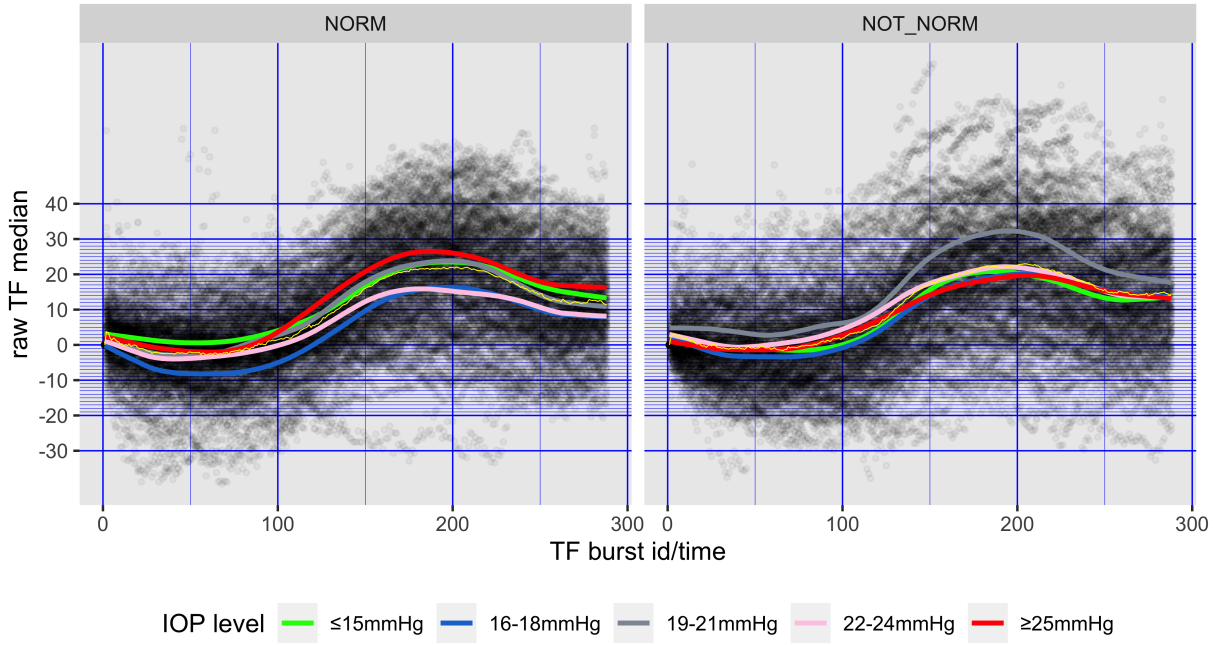


FIGURE 6.2: Comparison of 24-hour TF variability in NORM (healthy) and NOT_NORM (POAG cases) group. Each black point represents median of one TF series (one burst). Additionally, scatterplot contains smoothed color lines which are TF approximation generated by the loess function (locally weighted polynomial regression) for the enumerated initial IOP ranges.

According to the protocol proposed by Robert Wasilewicz and on the basis of the physiological circadian cycle properties [71], 24-hour session is divided into the consecutive time intervals. Main division contains the following base time points: start of the recording session (START), begin of the main/night sleep period (SLEEP), end of the main/night sleep period (WAKE), end of the recording session (END). SLEEP and WAKE time points were determined from Triggerfish record, where sleep period is considered generally as time interval without eye blinks. Table 6.2 shows main time intervals used for attribute generation.

We can derive attributes that provide characteristics of Triggerfish, cardiac sensor

| interval name | time range |
|---|---|
| *START_TP1* | *START* until *SLEEP*-5h |
| *TP1_SLEEP* | *SLEEP*-5h until *SLEEP* |
| *SLEEP_TP2* | *SLEEP* until *SLEEP*+2h |
| *TP2_WAKE* | *SLEEP*+2h until *WAKE* |
| *WAKE_TP3* | *WAKE* until *WAKE*+2h |
| *TP3_END* | *WAKE*+2h until *END* |

TABLE 6.2: Main time intervals range for 24-hour session.

signal or relation of CLS and cardiac sensor signal record for the each time interval [19]. Median value for each burst measurements was generated in the earlier transformation. The following attributes were defined for the input sensor data:

**1.** Sum of Triggerfish (TF) in the interval. Constant TF value between consecutive bursts was assumed (sum).

**2.** Slope (in radians) of linear regression line for TF fitted using standard least squares method in the interval (slope).

**3.** Amplitude of signal in the interval i.e. difference of the maximal value and the minimal value (ampl). We also computed modified amplitude for cardiac signal (SAP, DAP or HR) as 95 % quantile minus the minimal value (flat_ampl).

**4.** Sum of the numerical approximation of TF second derivative in the interval (sec_deriv_integral). It estimates total change of TF variability rate in the interval.

**5.** Correlation coefficient of TF and cardiac signal (SAP, DAP or HR) in the interval (cor). We also divided correlation value range at points {-0.65,-0.25,0.25,0.65} and mapped these intervals onto five integer levels (cor_level).

**6.** Aggregated measures (summary statistics in the form of mean and sum of numerical approximation of the second derivative) for linear convolution of TF and cardiac signal (SAP, DAP or HR) in the interval (conv).

In 1. and 2. we computed sum and slope for raw TF values and TF scaled into [0,20] range separately for each case ($TF^s$).

In 5. we generated Spearman's rank correlation coefficient to determine monotonic relation of TF and cardiac signal. It is less sensitive to outliers than standard Pearson's coefficient. Figure 6.3 shows heatmap generated by *pheatmap* library in R for SLEEP_WAKE interval. Clusters of cases with strong negative (red) or strong positive (green) correlation coefficient of TF and arterial pressure can be found for this time interval. Heatmaps and clustering techniques can be used to explore patterns in correlations generated for different time intervals listed in Table 6.2.

In 6. we use *convolve* function from *stats* library in R to compute linear convolution values (setting type="open" in the function call). We considered convolutional attributes only for short (no longer than 2 hours) intervals starting at SLEEP time point. Generally, TF curve increases substantially in this period, after the patient changes their position from vertical to horizontal [72].

We also took account of alternative time intervals for the main division defined in Table 6.2. Any shift (in hours) of interval begin or end is specified in squared brackets at the end of the attribute description (e.g. wake_[tp3+3h] means WAKE until WAKE+5h range). Short interval starting at SLEEP time point and spanning 15 bursts (75 minutes) is labelled as b15.

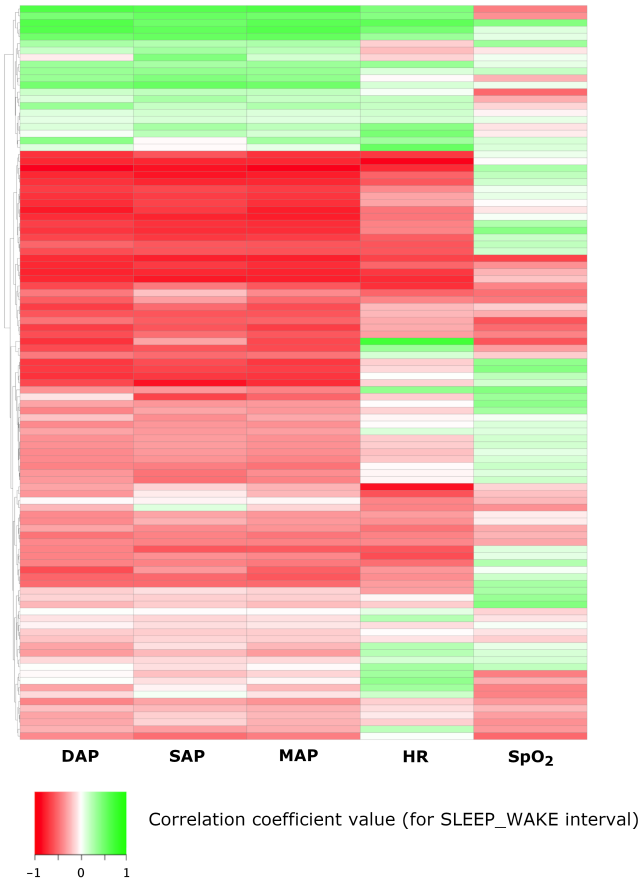


FIGURE 6.3: Correlations of TF and cardiac sensor signal: DAP, SAP, MAP (mean arterial pressure), HR, $SpO_2$ (peripheral oxygen saturation) in the main/night sleep period (SLEEP_WAKE). Cases (rows) in the heatmap are ordered according to the hierarchical clustering results for euclidean distance measure.

### 6.2.1 Detection of peaks in Triggerfish CLS signal

Triggerfish signal contains high peaks that are mainly related to the eye blinks and considered a noise in most analytical scenarios. Median value for each TF burst measurements is generated during initial transformation of the input data. This low-dimensional representation of TF signal sufficiently reduces impact of the high peaks and captures

important properties of the signal. Most of the predictive ML models presented in the thesis involve features based on median TF values.

However in some analytical scenarios we need high-dimensional input data without peaks (e.g. in spectral analysis). Additionally, continuous TF increase over long time interval (i.e. large shift of TF level) can be related to shift of CLS on the eye surface. It also can be considered as an undesirable effect. Basic peak detection procedure has been implemented by the author in R language using *find_peaks* function from ggpmisc package. The procedure is based on the variation of TF loess approximation and TF derivative (slope of a tangent at a given point).

Figure 6.4 and 6.5 show the output of the procedure of selection of TF fragments (ranges) without peaks and large level shifts. Blue horizontal line represents median of the burst. Original TF measurements are plotted as blue circles connected by auxiliary yellow line. Black crosses represent TF values left after removing peaks and large level shifts. Auxiliary smooth green curve is the loess approximation of TF (without high peaks) shifted by its standard deviation (sd) value.

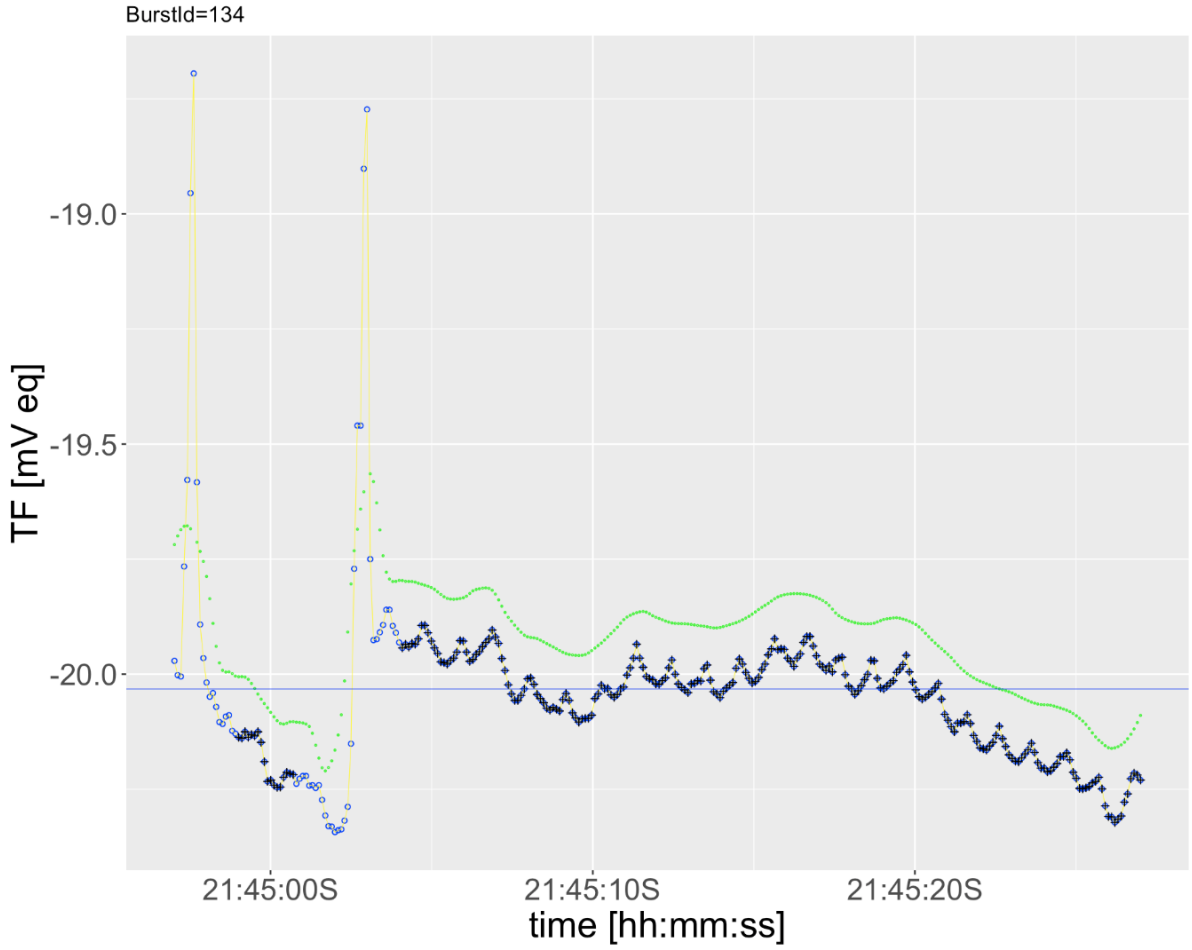


FIGURE 6.4: TF fragments (ranges) without sharp peaks and large level shifts marked with black crosses. The burst contains two sharp peaks around 21:44:57 and 21:45:03.

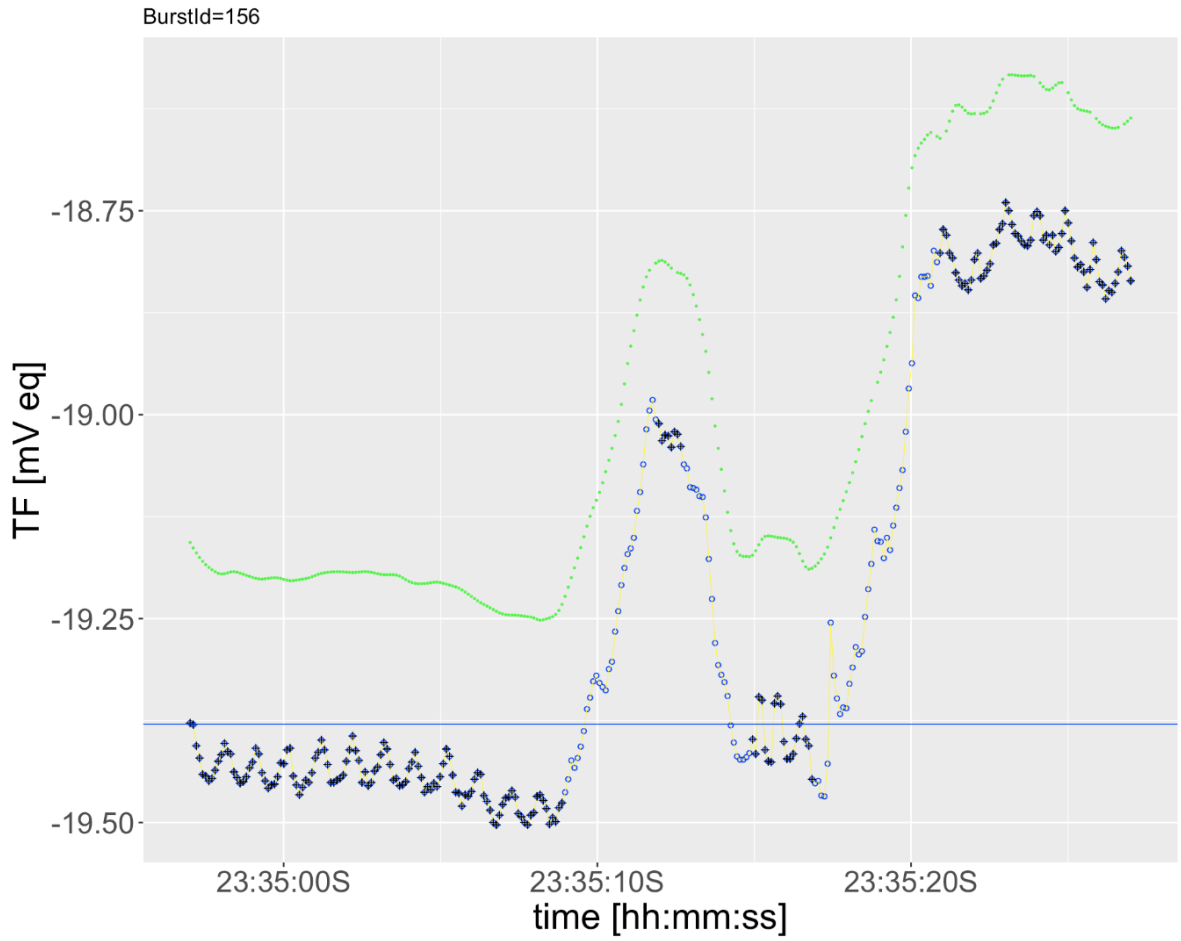


FIGURE 6.5: TF fragments (ranges) without sharp peaks and large level shifts marked with black crosses. The burst contains one rounded peak in the middle and continuous TF increase can be seen from 23:35:17 to 23:35:21.

### 6.2.2 Spectral analysis of sensor data

We performed basic spectral analysis of TF and cardiac sensor signal. We generated periodograms to investigate important frequency components in the signals. Periodogram is an estimator of the spectral density computed using the discrete Fourier transform (DFT) of a signal. It represents distribution of signal energy in the frequency domain [73].

Figure 6.6 shows raw TF and HR measurements for a one burst (time interval of 30 s). Heart rate is directly related to the ocular blood flow and has influence on the changes in TF signal.

Large peaks were removed from the raw TF signal. Missing TF measurements were substituted with the median of a burst. Linear trend was removed initially from the data. Periodograms of raw TF were generated for sleep_wake interval. Peaks around 1 Hz are related to the heart rate. We computed frequency of the main components and maximum and mean power in the relevant peak clusters. Preliminary evaluation of the selected ML models enhanced with these numerical features didn't show improvement

of classification performance.

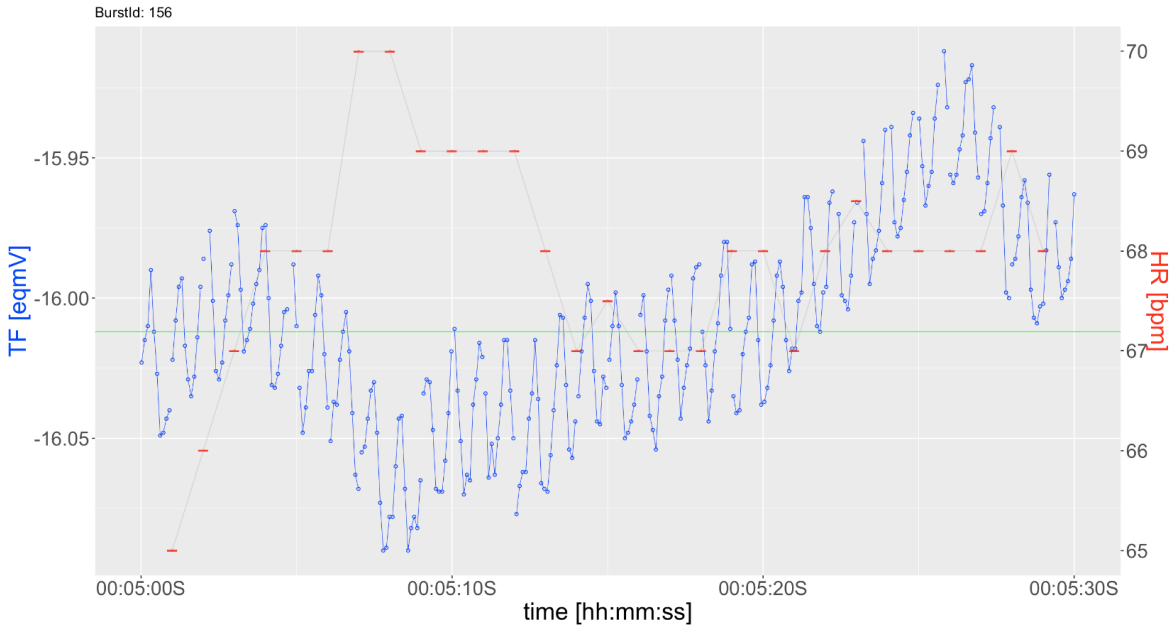


FIGURE 6.6: Plot of raw TF (blue circles) and HR (red dashes) values for one sample burst. Sampling rate for TF is 10 samples/s and 1 sample/s for HR. Range of the values is indicated on the axes.

## 6.3 Predictive ML models for glaucoma detection

### 6.3.1 Algorithms

There are many ML algorithms that can solve binary classification problem (see description of the algorithms in 4.4). Here we consider assigning NORM label for healthy and NOT_NORM for the other cases (diagnosed as POAG/HTG or POAG/NTG) in our dataset. Given the size of the dataset and data distribution two basic ML algorithms were selected: logistic regression and random forest which are common methods in medical data analysis. Logistic regression is based on properties of sigmoidal curve used for modelling the probability of belonging to a class. Logistic regression is fast algorithm and produces results that are quite easy to interpret. It is a linear classification algorithm i.e. its decision boundaries are linear. Random forest is an ensemble learning method that generalizes standard decision trees. It efficiently handles non-linear relations and outlier values. We build and test models with H2O framework which is open source, distributed and scalable ML and predictive analytics environment [74]. H2O Flow web interface was used for model evaluation and visualization of the results.

Determining the best subset of attributes in order to achieve high prediction performance of the models requires a considerable computational cost due to the large number of possible candidates. For initial attribute selection we used LASSO regularization (least absolute selection and shrinkage operator) [46] and RSM (Random Subspace Method)

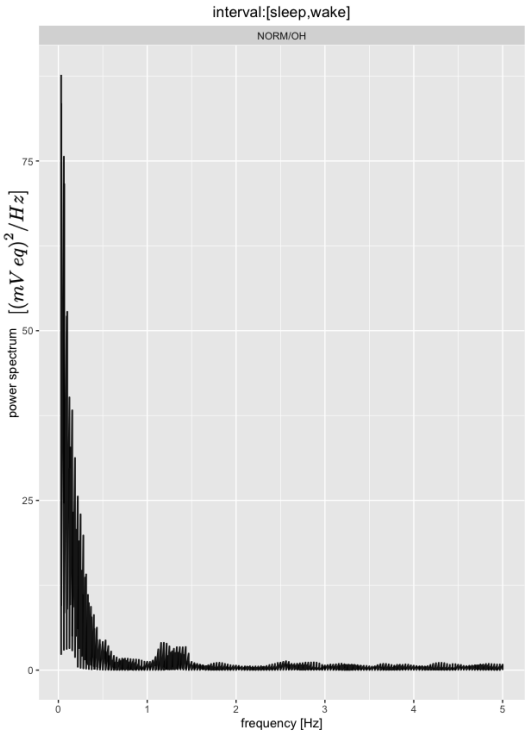


FIGURE 6.7: Periodogram of raw TF for a normal case.

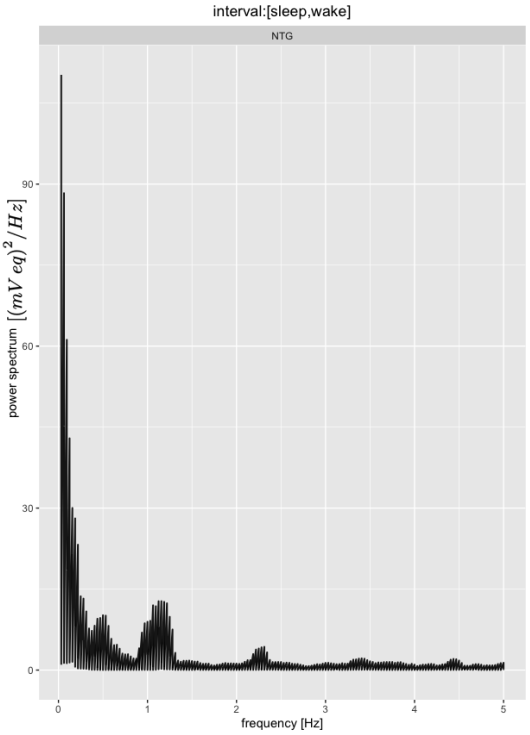


FIGURE 6.8: Periodogram of raw TF for a NTG case.

[75] technique. LASSO enables selection of a small number of attributes by penalizing coefficients magnitude of the fitted linear models. We used LASSO regularization available in H2O for generalized linear models (GLM). Ranking of attributes in RSM is based on fitting linear models on small randomly chosen subsets of attributes. Final subset of the attributes is chosen using information criteria or validation set. We have chosen parallel version of RSM implemented in regRSM package for R.

### 6.3.2 Evaluation

We applied cross-validation (CV) resampling procedure to estimate model prediction performance. In standard k-fold CV input dataset is randomly divided into k equally sized subsets. In each of the k steps we subsequently use one subset as validation/test data and the remaining subsets as training data. Prediction performance (for the selected metrics) is computed for test data in each step and average of the results is the final estimation. We have chosen CV fold size equal to 8 considering given input dataset count (105 cases). Additionally, full CV procedure was repeated 50 times to assess variance of the estimation for the each model.

Quality of the results of binary classification models can be evaluated with many different metrics. The following metrics were computed:

**1.** Accuracy which is the ratio of correctly classified cases to the all cases count.

**2.** Brier score (BS) given by the formula $BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where N is the all cases count, $y_i$ is the actual value (assigned to the reference class), $\hat{y}_i$ is the predicted value (probability) for the case.

**3.** Logistic loss (log loss) estimates how close predicted values (uncalibrated probabilities) are to the actual values and increases exponentially as the difference gets larger. $log\,loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i))$, where N is the all cases count, $y_i$ is the actual value (assigned to the reference class), $\hat{y}_i$ is the predicted value (probability) for the case.

**4.** Area under the ROC curve (AUC) is an estimate based on properties of the plot containing value pairs of TPR (true positive rate: probability of positive prediction for truly positive case) and FPR (false positive rate: probability of positive prediction for truly negative case) for different classification threshold values.

For the selected models we reported classification threshold dependent metrics such as sensitivity, specificity, precision, negative predictive value (NPV) and F-score.

Additionally, assessment of the results was performed using SMOTE: Synthetic Minority Over-sampling Technique [76] for the input dataset to oversample NORM class. We generated balanced set containing 70 NORM and 69 NOT_NORM cases (i.e. 139 total). AUC and accuracy of the selected models for synthetic data are similar or higher than the cross-validation results for the original data with a slightly greater variance. Results for the original dataset were reported in the tables.

### 6.3.3 Models involving sensor data

Existing papers on application of ML techniques to Triggerfish data analysis provide results for models involving only CLS signal based attributes, optionally with IOP included [3]. In this section we present models for input data that includes TF and cardiac signal derived attributes. In Table 6.3 we enumerated attributes for the models with highest AUC value estimated in CV repeated routine. We evaluated diagnostic performance metrics for the logistic regression (LR) and random forest (RF) algorithms. We didn't perform full hyperparameter optimization/tuning for random forest so it is possible that some RF predictive performance results can be slightly refined. We built two versions of model: one involving cardiac signal derived attributes and the shorter one with TF derived attributes only to compare each other performance estimates.

Model s5 (LR) achieved the best performance with mean AUC of 0.74±0.02, Brier score at 0.20±0.01 and accuracy at 0.71±0.01. It is based on TF slope, sum, sec_deriv_integral attributes and cardiac attributes (correlation and amplitude related ones). Model s4 has the same TF attributes as s5 and no additional cardiac attributes. AUC estimate for s5 is higher respectively for LR and RF algorithms compared to s4, and

| id | type | attributes |
|---|---|---|
| s1.LR<br>s1.RF | logistic regression<br>random forest | slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h],<br>sum_TF$^s$_wake_[tp3+3h], sec_deriv_integral_TF_wake_[tp3+3h],<br>cor_SAP_sleep_wake, conv_sec_deriv_integral_HR_sleep_b15, cor_HR_tp2_wake |
| s2.LR<br>s2.RF | logistic regression<br>random forest | slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h],<br>sum_TF$^s$_wake_[tp3+3h], sec_deriv_integral_TF_wake_[tp3+3h] |
| s3.LR<br>s3.RF | logistic regression<br>random forest | slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h],<br>sec_deriv_integral_TF_[wake-4h]_wake, sum_TF$^s$_[wake-4h]_wake,<br>sum_TF$^s$_wake_[tp3+3h], ampl_HR_sleep_b15,<br>cor_SAP_sleep_wake, cor_HR_tp2_wake |
| s4.LR<br>s4.RF | logistic regression<br>random forest | slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h],<br>sec_deriv_integral_TF_[wake-4h]_wake, sum_TF$^s$_[wake-4h]_wake,<br>sum_TF$^s$_wake_[tp3+3h] |
| s5.LR<br>s5.RF | logistic regression<br>random forest | slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h],<br>sec_deriv_integral_TF_[wake-4h]_wake, sum_TF$^s$_[wake-4h]_wake,<br>sum_TF$^s$_wake_[tp3+3h], ampl_HR_start_sleep,<br>flat_ampl_DAP_sleep_wake, cor_level_HR_tp2_wake |

TABLE 6.3: Summary of the attributes selected for the models based on sensor data only.

the difference (0.04 for LR) is above the standard deviation range. This comparison shows that additional cardiac attributes joined with TF data can improve predictive performance of model involving only sensor data.

| id | type | AUC | log loss | Brier score |
|---|---|---|---|---|
| s1.LR | logistic regression | 0.66±0.02 | 0.65±0.02 | 0.23±0.01 |
| s1.RF | random forest | 0.69±0.03 | 0.62±0.01 | 0.21±0.01 |
| s2.LR | logistic regression | 0.67±0.02 | 0.63±0.02 | 0.22±0.01 |
| s2.RF | random forest | 0.67±0.03 | 0.61±0.01 | 0.21±0.01 |
| s3.LR | logistic regression | 0.71±0.02 | 0.63±0.02 | 0.21±0.01 |
| s3.RF | random forest | 0.70±0.02 | 0.61±0.01 | 0.21±0.01 |
| s4.LR | logistic regression | 0.70±0.02 | 0.61±0.02 | 0.21±0.01 |
| s4.RF | random forest | 0.68±0.02 | 0.61±0.01 | 0.21±0.01 |
| s5.LR | logistic regression | 0.74±0.02 | 0.59±0.02 | 0.20±0.01 |
| s5.RF | random forest | 0.71±0.02 | 0.61±0.01 | 0.21±0.01 |

TABLE 6.4: Performance metrics for the models defined in Table 6.3.

### 6.3.4 Models involving sensor data and clinical data

IOP is recognized as the important risk factor in glaucoma onset and progression. It has been shown that IOP combined with attributes based on CLS record can improve accuracy of classification models for POAG and NORM cases [3]. Detailed role of IOP level and its influence on optic nerve head in glaucoma is not fully explained so far. There is OH group, where high ocular pressure doesn't result in disease progression. On the other hand, NTG group contains cases with alterations of the optic nerve despite normal IOP level. In this section we focus on the alternative attributes that can replace or

complement IOP in classification models. General relationship between IOP and corneal biomechanical properties was investigated [77] but its complex nature is not fully explained across the spectrum of glaucoma at the present time. Overall, IOP measured by Goldmann applanation tonometer is affected by eye biomechanical properties quantified by CH and CRF [78]. We have included these measurements in extended attribute set for our models.

| id | type | attributes |
|---|---|---|
| m1.LR | logistic regression | IOP |
| m2.LR | logistic regression | IOP, CRF |
| m3.LR | logistic regression | CH |
| m4.LR | logistic regression | CH, CRF |
| m5.LR<br>m5.RF | logistic regression<br>random forest | IOP, CRF, slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h], sum_TF$^s$_wake_[tp3+3h], cor_HR_tp2_wake, cor_SAP_tp3_end |
| m6.LR<br>m6.RF | logistic regression<br>random forest | IOP, CRF, slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h], sum_TF$^s$_wake_[tp3+3h] |
| m7.LR<br>m7.RF | logistic regression<br>random forest | CH, CRF, slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h], sum_TF$^s$_wake_[tp3+3h], cor_HR_tp2_wake |
| m8.LR<br>m8.RF | logistic regression<br>random forest | CH, CRF, slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h], sum_TF$^s$_wake_[tp3+3h] |
| m9.LR<br>m9.RF | logistic regression<br>random forest | CH, CRF, slope_TF$^s$_tp1_sleep, sec_deriv_integral_TF_sleep_[tp2+1h], sum_TF$^s$_wake_[tp3+3h], ampl_HR_start_sleep, flat_ampl_DAP_sleep_wake, cor_level_HR_tp2_wake |

TABLE 6.5: Summary of the attributes selected for the models with clinical data. Simple models (m1 to m4) added as a baseline reference.

| id | type | AUC | log loss | Brier score |
|---|---|---|---|---|
| m1.LR | logistic regression | 0.65±0.01 | 0.62±0.01 | 0.22±0.01 |
| m2.LR | logistic regression | 0.80±0.01 | 0.53±0.01 | 0.18±0.01 |
| m3.LR | logistic regression | 0.79±0.01 | 0.54±0.01 | 0.18±0.01 |
| m4.LR | logistic regression | 0.79±0.01 | 0.56±0.01 | 0.18±0.01 |
| m5.LR | logistic regression | 0.86±0.01 | 0.48±0.02 | 0.15±0.01 |
| m5.RF | random forest | 0.76±0.02 | 0.59±0.01 | 0.20±0.01 |
| m6.LR | logistic regression | 0.85±0.01 | 0.48±0.02 | 0.16±0.01 |
| m6.RF | random forest | 0.76±0.02 | 0.57±0.01 | 0.20±0.01 |
| m7.LR | logistic regression | 0.83±0.01 | 0.50±0.02 | 0.17±0.01 |
| m7.RF | random forest | 0.84±0.01 | 0.53±0.01 | 0.18±0.01 |
| m8.LR | logistic regression | 0.82±0.01 | 0.52±0.02 | 0.17±0.01 |
| m8.RF | random forest | 0.84±0.01 | 0.52±0.01 | 0.17±0.01 |
| m9.LR | logistic regression | 0.87±0.01 | 0.46±0.02 | 0.15±0.01 |
| m9.RF | random forest | 0.85±0.01 | 0.51±0.01 | 0.17±0.01 |
| m9.XGB | XGBoost | 0.84±0.02 | 0.54±0.04 | 0.17±0.01 |
| m9.GBM | GBM | 0.86±0.01 | 0.46±0.02 | 0.15±0.01 |
| m9.NB | naive Bayes | 0.80±0.01 | 0.58±0.02 | 0.18±0.01 |

TABLE 6.6: Performance metrics for the models defined in Table 6.5.

Table 6.5 enumerates attributes of the models involving clinical data (i.e. IOP, CH and CRF). We built m1 - m4 LR models as baseline reference for the extended models.

Model m5 (LR) with IOP achieved the best performance with mean AUC of 0.86±0.01, Brier score at 0.15±0.01 and accuracy at 0.81±0.01. Model m6 contains the same attributes except cardiac derived ones which were left out. It has slightly lower AUC estimate for LR and the recorded difference with m5 is within the standard deviation range. Model m7 is based only on sensor data and corneal biomechanical measurements. It has mean AUC of 0.84±0.01, Brier score at 0.18±0.01 and accuracy at 0.78±0.01. Model m8 is based on m7 attributes without cardiac derived ones. It has mean AUC estimates similar to m7. AUC estimate for m7 (RF) is significantly greater by 0.04 than AUC for the best baseline model (m2) based on clinical data only. Therefore sensor data derived attributes can be regarded as complementary to CH and CRF measurements. Also IOP joined with biomechanical corneal properties and sensor attributes lead to improved predictive performance as in (LR) model m5 or m6.

Finally we check the models supplemented with cardiac attributes describing SAP, DAP, HR range summary (amplitude, flat_amplitude) in the main time intervals. Model m9 (LR) has mean AUC of 0.87±0.01 that is higher than AUC for m8 model and the corresponding difference is above the computed standard deviation range. Mean accuracy estimate for m9 is 0.81±0.01. Training and mean cross-validation ROC curve for m9 (LR) is shown in Figure 6.9 and 6.10. All ROC curves generated in cross-validation for the selected random division of the input data set are shown in Figure 6.11. This extended cardiac based attribute set joined with TF data and CH, CRF yields improved estimate of logistic loss and Brier score for LR. Summary of m9 metrics depending on classification threshold is shown in Table 6.7. Table 6.8 lists hyperparameters with optimal values determined for the m9 model.

Additionally we assessed predictive performance of several other ML algorithms for m9 attribute set. XGBoost mean AUC is 0.84±0.02, Brier score is 0.17±0.01 and accuracy is 0.79±0.02. Gradient Boosting Machine (GBM) mean AUC is 0.86±0.01, Brier score is 0.15±0.01 and accuracy is 0.80±0.02. Naive Bayes classifier performance is weaker than the other algorithms: it has mean AUC of 0.80±0.01 and accuracy at 0.78±0.02. Performance of GBM model is close to the estimate of LR.

Figure 6.12 shows relative attribute importance for the m9 (LR) model. In the bar chart of standardized coefficient magnitudes for logistic regression we can see that highest values are assigned to CH, sum_TF$^s$_wake_[tp3+3h], cor_level_HR_tp2_wake, ampl_HR_start_sleep. Figure 6.13 shows supplementary SHAP (Shapley additive explanations) summary plot of attribute contribution for m9 (RF) model. CH, flat_ampl_DAP_sleep_wake and sum_TF$^s$_wake_[tp3+3h] have the highest rank. Approximation of attribute importance for the models using Shapley values depends on the available dataset size and data distribution [79]. Sum under TF curve in the time interval of length 5h

beginning at WAKE time point seems to be the most important attribute related to Triggerfish for the input dataset.

Specific eye tissue properties are different for healthy and glaucoma positive cases and can have impact on biomechanical phenomena leading to glaucomatous process [3]. It is possible that value of attribute sum_TF$^s$_wake_[tp3+3h] (associated with change of ocular volume and shape of the eye surface) can be intermediately related to such tissue properties as it refers to time interval after the end of sleep, considerably long time since application of Triggerfish lens having an influence on the eye surface. Autoregulatory capacity of the eye requires adequate ocular blood flow which depends on cardiovascular system efficiency. Attributes related to heart rate (ampl_HR_start_sleep) and diastolic arterial pressure (flat_ampl_DAP_sleep_wake) can quantitatively approximate some properties of such mechanism. Ratio of nocturnal and daytime level of blood pressure association with POAG progression was investigated in [10]. Case-specific relationship between TF and HR in nighttime interval is quantified by cor_level_HR_tp2_wake attribute.

| metric name | metric value for LR | metric value for GBM |
|---|---|---|
| F1 | 0.86 | 0.85 |
| F2 | 0.88 | 0.87 |
| F0point5 | 0.83 | 0.83 |
| accuracy | 0.81 | 0.80 |
| precision | 0.82 | 0.82 |
| sensitivity (recall) | 0.89 | 0.88 |
| specificity | 0.68 | 0.68 |
| NPV (negative predictive value) | 0.80 | 0.78 |

TABLE 6.7: Summary of m9 (logistic regression and GBM) model performance metrics.

### 6.3.5 Models for the extended input dataset

This section reports results for the selected ML models built using extended input dataset. Methods described earlier in this chapter were applied in development and evaluation of these predictive models.

#### Input data

Input data for the research was collected at [Wasilewicz] Eye Clinic in Poznań. Input dataset contains 138 cases. It has been supplemented with the new ones since the results presented in previous sections were published in [19]. Some minor corrections were introduced in the dataset (including removal of one case with low quality SAP/DAP/HR data). The following diagnosis labels were assigned to the cases:

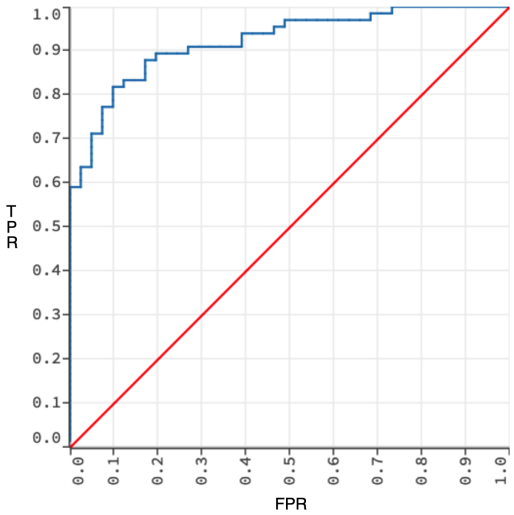


FIGURE 6.9: Training dataset ROC curve of the model m9.LR (logistic regression). Training dataset AUC is estimated at 0.92, relevant accuracy at 0.86.

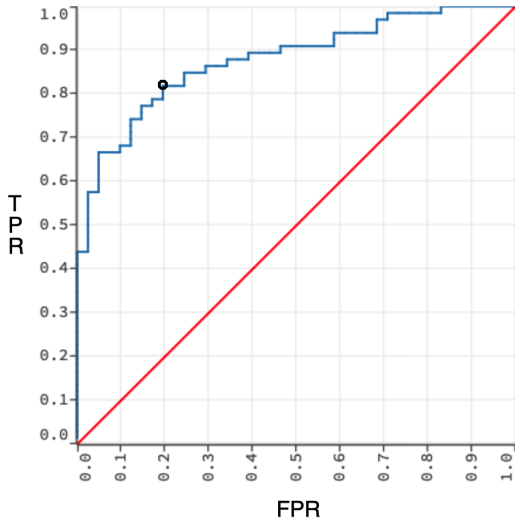


FIGURE 6.10: Mean cross-validation ROC curve for the m9.LR model (logistic regression). Black point on the curve represents (FPR,TPR) value pair for the maximal accuracy classification threshold. Accuracy of the model is estimated at 0.81±0.01.

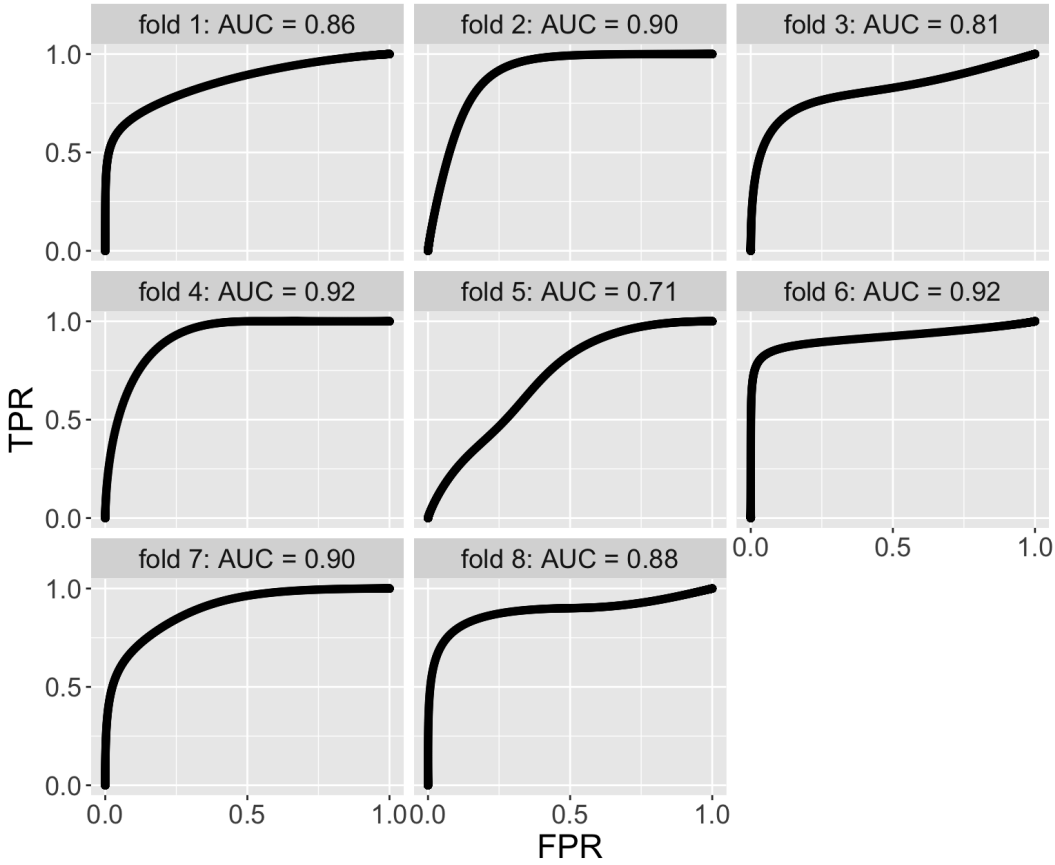


FIGURE 6.11: ROC curves for the all cross-validation folds in selected random division of the input data set (m9.LR model).

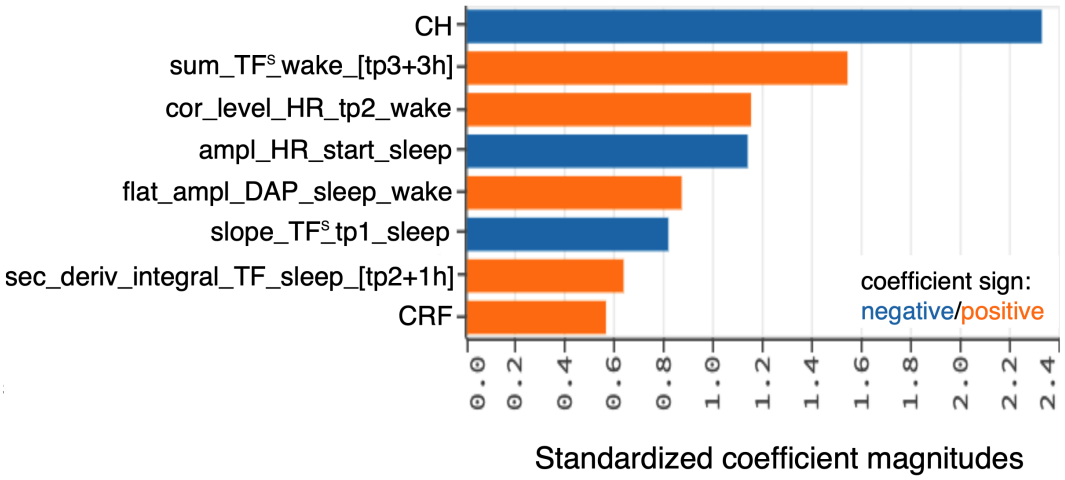


FIGURE 6.12: Attribute importance summary for the m9.LR (logistic regression) model. Blue color of bar refers to the negative sign of coefficient.

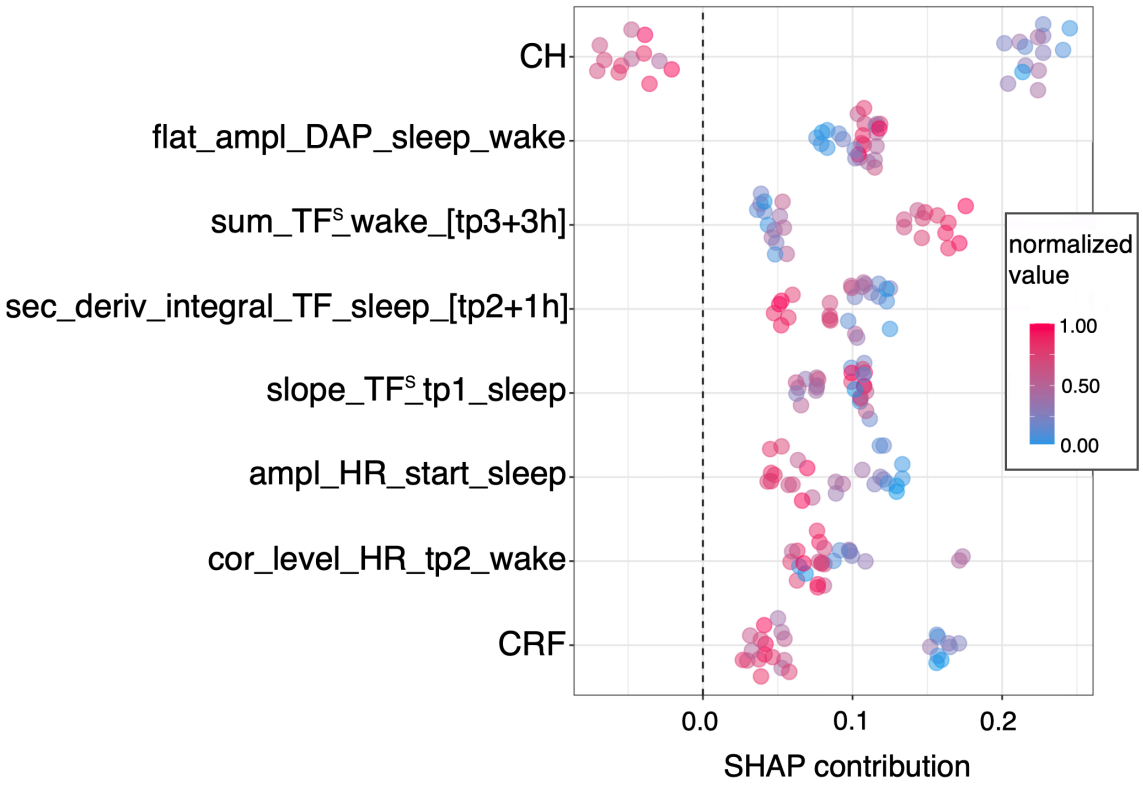


FIGURE 6.13: SHAP summary plot for the m9.RF (random forest) model shows the contribution of its attributes ranked by importance.

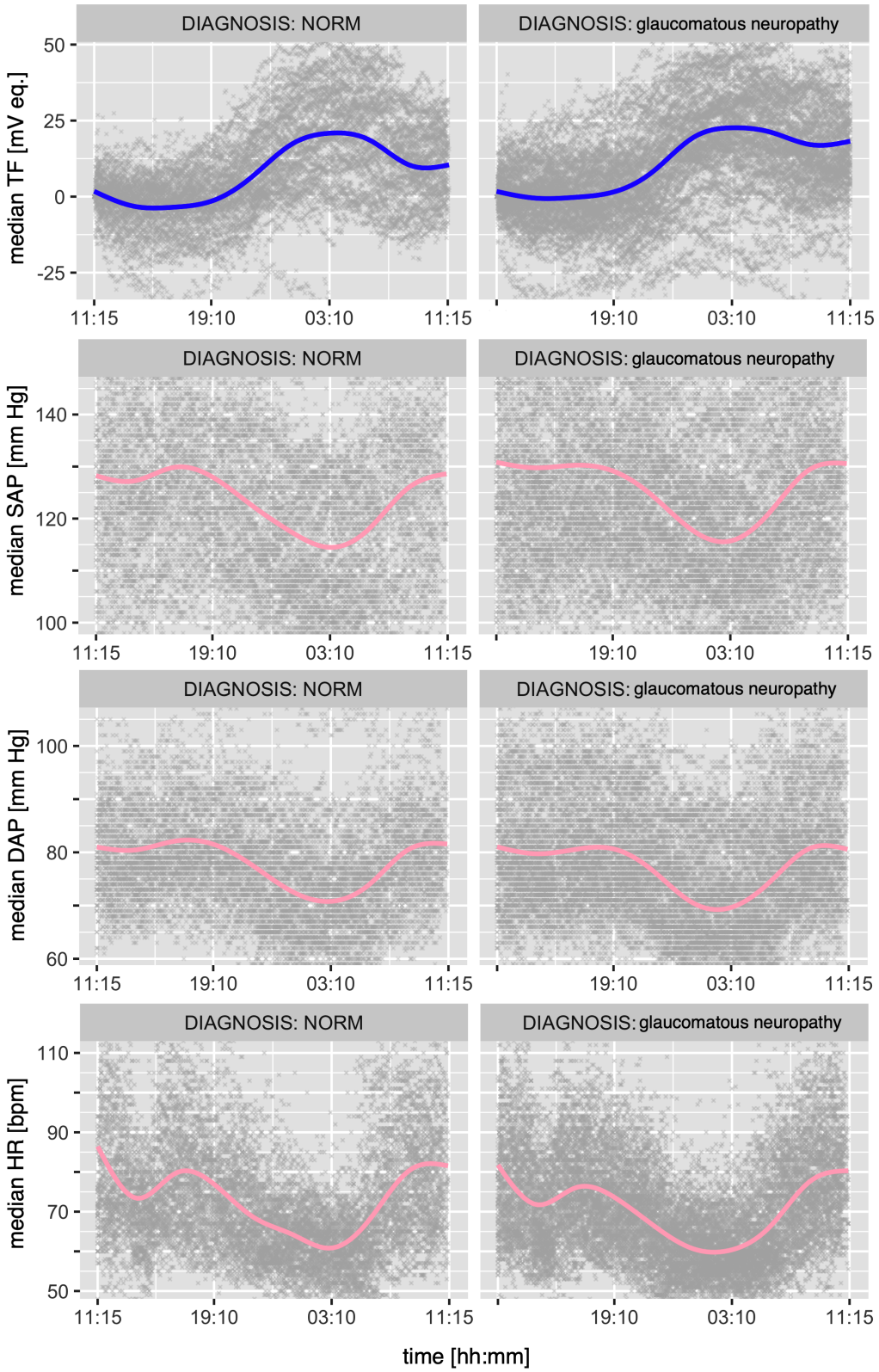


FIGURE 6.14: 24-hour overview of attributes in the input dataset for healthy (NORM) and POAG cases (glaucomatous neuropathy). Each point represents median of the values within one series (one burst). Scatterplots contain smoothed color lines which are attribute approximation generated by loess function (locally weighted polynomial regression) for TF [mV equivalents], SAP/DAP [mm Hg], HR [beats per minute].

| hyperparameter | value | description |
|---|---|---|
| *random forest* | | |
| ntrees | 180 | number of trees |
| max_depth | 5 | maximum depth to which each tree will be built |
| min_rows | 11 | minimum number of observations for a leaf required to split |
| nbins | 20 | maximum number of bins included in the histogram for determining the split of the attribute space |
| sample_rate | 0.67 | row sampling rate set to improve generalization and reduce validation error |
| *logistic regression* | | |
| lambda | 0 | LASSO regularization penalty |
| *XGBoost* | | |
| ntrees | 130 | number of trees |
| learn_rate (eta) | 0.27 | specifies shrinkage of the feature weights after each boosting step |
| *GBM* | | |
| ntrees | 160 | number of trees |
| learn_rate (eta) | 0.09 | specifies shrinkage of the feature weights after each boosting step |

TABLE 6.8: Main hyperparameters of the machine learning models. Optimized values for the m9 model.

- 50 high tension glaucoma (POAG/HTG)
- 30 normal tension glaucoma (POAG/NTG)
- 58 control/healthy (NORM)

Figure 6.14 shows 24-h overview of the sensor data from the extended input dataset.

### Predictive models

Table 6.9 enumerates attributes for the models with highest AUC value estimated in repeated 10-fold CV routine for the input dataset. Model $G_0$ is based only on TF and cardiac data derived attributes. Additional cardiac attributes joined with TF data can improve predictive performance of model. Quantification of the reciprocal relation of TF and cardiac signal (e.g. correlation coefficient) seems particularly valuable. Model $G_1$ is based on sensor data supplemented with measurements of corneal biomechanical properties (CH, CRF) [77]. Such models can be a tool suitable for glaucoma detection regardless of direct IOP measurements. Sensor data derived attributes are complementary to CH and CRF with the highest mean AUC of $0.88 \pm 0.01$ for $G_1$ (see Figure 6.15

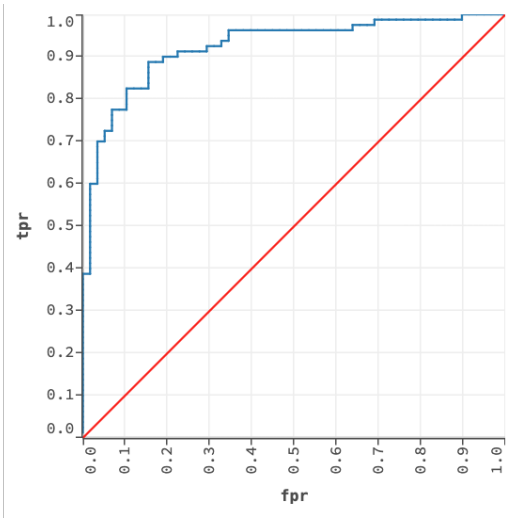


FIGURE 6.15: Training dataset ROC curve of the model $G_1$ (logistic regression). Training dataset AUC is estimated at 0.92.

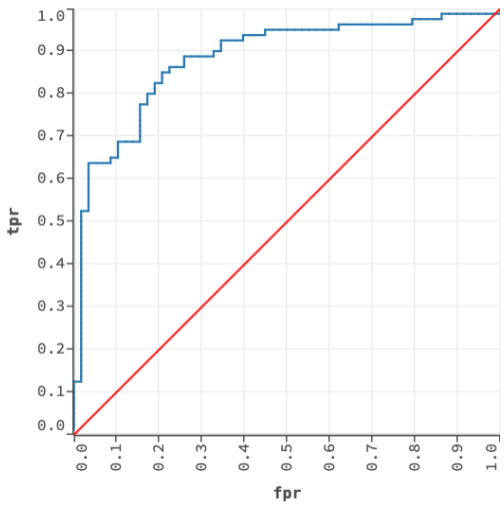


FIGURE 6.16: Mean cross-validation ROC curve of the model $G_1$ (logistic regression). AUC of the model is estimated at 0.88±0.01.

and 6.16). Best performance metrics estimated for the logistic regression model (see Table 6.10) are similar to the results reported for the earlier (initial) input dataset [19].

| Model id | Attributes |
|---|---|
| $G_0$ | slope_TF$^s$_[sleep-5h]_sleep, sec_deriv_integral_TF$^s$_sleep_[sleep+3h], sum_TF$^s$_[wake-4h]_wake, sum_TF$^s$_wake_[wake+5h], ampl_HR_start_sleep, flat_ampl_DAP_sleep_wake, cor_level_HR_[sleep+2h]_wake |
| $G_1$ | CH, CRF, slope_TF$^s$_[sleep-5h]_sleep, sec_deriv_integral_TF_sleep_[sleep+3h], sum_TF$^s$_wake_[wake+5h], ampl_HR_start_sleep, flat_ampl_DAP_sleep_wake, cor_level_HR_[sleep+2h]_wake |

TABLE 6.9: Summary of the attributes selected for the model $G_0$ involving only sensor data derived attributes and $G_1$ with corneal biomechanical measurements (CH and CRF).

| Model id | Type | AUC | Brier score | Accuracy |
|---|---|---|---|---|
| $G_0$ | logistic regression | 0.77±0.01 | 0.20±0.01 | 0.73±0.01 |
| | XGBoost | 0.70±0.02 | 0.22±0.01 | 0.70±0.01 |
| | naive Bayes | 0.68±0.02 | 0.23±0.01 | 0.68±0.01 |
| $G_1$ | logistic regression | 0.88±0.01 | 0.14±0.01 | 0.83±0.01 |
| | XGBoost | 0.85±0.01 | 0.15±0.01 | 0.81±0.01 |
| | naive Bayes | 0.84±0.01 | 0.16±0.01 | 0.82±0.01 |

TABLE 6.10: Estimation of performance metrics for the models from Table 6.9.

## 6.4 Conclusions

Glaucoma detection and control requires acquisition of data from multiple sources including standard measurements of eyeball properties (using e.g. Goldmann applanation

tonometry), imaging techniques, Triggerfish CLS and devices for monitoring cardiovascular system properties. Currently IOP is the main identified risk factor [80] for glaucoma which can be modified. We focused our study on TF joined with cardiac data derived attributes that can be candidate for additional modifiable risk factors in extended ML models. We also evaluated models involving CH, CRF and sensor data based attributes without IOP. Such models can be a tool suitable for glaucoma detection independently of direct IOP measurements. Sensor data derived attributes are complementary to CH and CRF with the highest mean AUC of 0.87 for m9 (or 0.88 for $G_1$). Cardiac data derived attributes are also complementary to TF based attributes as mean AUC estimate is higher for extended models built for TF joined with cardiac data based attributes. Predictive performance metrics improvement is noticeable for such extended models based on sensor data only and for models based on sensor data along with CH, CRF measurements.

Comprehensive patient's data profile which involves Triggerfish and cardiac sensor record can provide new insights into glaucoma detection and control. Such integrated approach for monitoring of the disease can also facilitate advances in research focused on understanding basic pathological mechanisms related to the retinal function [66].

Machine learning models based on joint CH and sensor data are important part of the study. Previous publications reported results for either CH or sensor data [3, 81]. Performance estimates for our extended models that include CH measurements are better than results for the models based only on corneal biomechanical properties or the models based only on Triggerfish derived data. Such extension of the attribute set of the models can be considered as an advantage. Important aspect of the work is introduction of cardiac data derived attributes that can be regarded as prospective modifiable risk factors of glaucoma and element in patient profiling or treatment recommendation.

# Chapter 7

# System for glaucoma diagnosis and collaborative research support

This chapter is based on the results which were presented in the article written by the author (see [A1]) and included in the description of invention in the patent application (see [P1]). This chapter also contains some ideas and the results presented at the 10th World Glaucoma Congress (see [C1]).

Diagnostic support services were designed and implemented by the author using R, Python and Java programming languages.

## 7.1 System overview

Triggerfish is a relatively new device and there are no extensive software tools for supporting complete analysis of its data output. Current cost of a single examination (disposable contact lens) is relatively high, can be many times higher than OCT scan (see description of OCT imaging technique in 3.2) and there are no specific multi-sensor data based clinical protocols ready to apply in glaucoma detection and treatment. The majority of ML-based systems for glaucoma diagnosis is intended for structural analysis of image data [16] and the range of implemented functions is usually limited. One of the few systems that handle clinical data is Maggelan which supports diagnosis using OCT extracted features and selected clinical data including parameters based on visual field (VF) test and IOP. However this system can't use any sensor data for case assessment [82].

System designed for management and analysis of the sensor data can address many of the aforementioned issues in the context of glaucoma diagnosis. Deployment of the system will increase availability of the data that is typically stored in isolated/closed repositories. It will encourage research on the personalized approaches focused on the multi-sensor data and the exchange of new ideas.

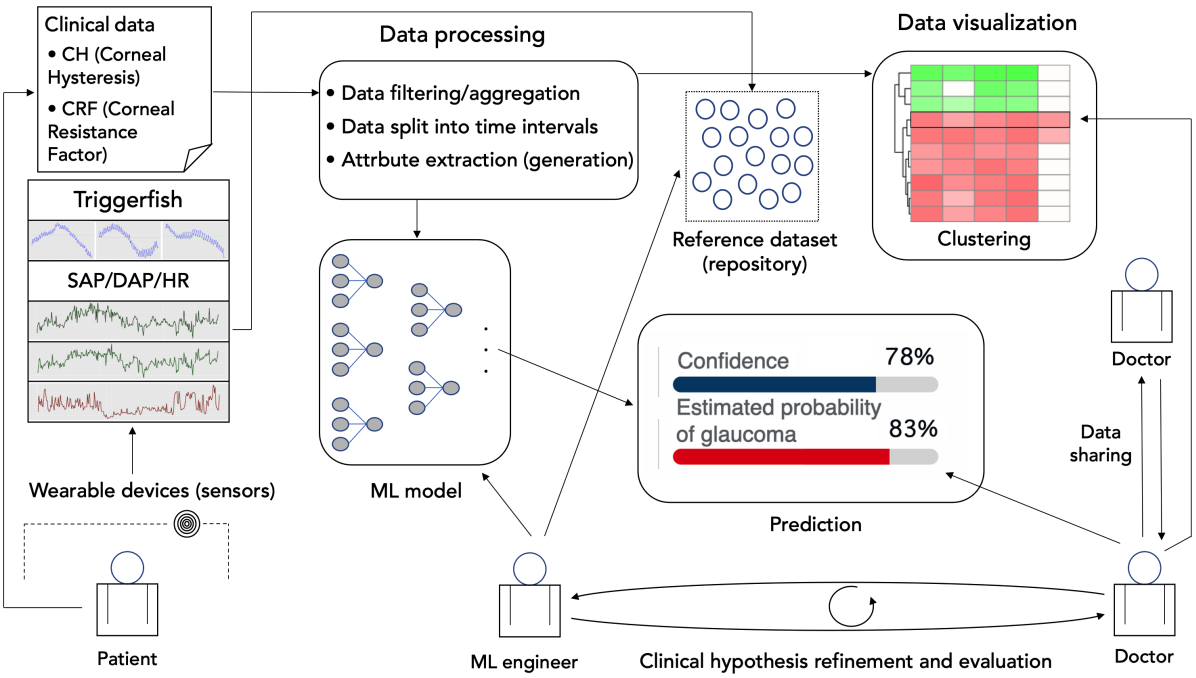


FIGURE 7.1: Data flow and interactions of users of the system. Clinical data include measurements such as age, CH, CRF, CCT (central corneal thickness), etc.

This thesis proposes comprehensive system for glaucoma diagnosis and collaborative research support. The system has the following aims:

1. Support of glaucoma detection and control using multiple ML techniques.
2. Provision of collaborative research platform for medical specialists in ophthalmology and data scientists.
3. Supply of data management services (focused on time series acquired using Triggerfish contact lens sensor and devices for continuous monitoring of cardiovascular system properties).
4. Facilitation of using various data sources in clinical hypothesis assessment and refinement.

Users of the system can be assigned with the following basic roles:

1. Ophthalmologists experienced in glaucoma diagnosis and clinical research. They can set new directions for the research and assess the practical value of the implemented solutions.
2. ML engineers or data scientists experienced in application of ML methods in biomedical data analysis. These specialists can collaborate with the doctors to design and develop new analytic components for the users.

3. Eye doctors who want to gain experience in the field of glaucoma detection and control. They can participate in collection and validation of clinical data.

General outline of the proposed approach is shown in Figure 7.1. Components of the diagram refer to data acquisition (Triggerfish and SOMNOtouch devices, clinical metadata), data management (storage, reference dataset collection, sharing), application and evaluation of ML models, data visualization, eye doctors and data science experts collaboration in clinical research related to glaucoma diagnosis and treatment.

### 7.1.1 ML-based system architecture

Increasing adoption of ML techniques has given rise to many challenges related to system development, deployment and management [22]. These challenges should be addressed along with the issues typical for standard software system engineering. ML-based system view is divided into ML subsystem and software subsystem in some high-level approaches [83]. Such view on architecture is in line with different characteristics, functions and key stakeholders of the each subsystem. Distinct nature of the each subsystem has an impact on requirements analysis and design assumptions. Development team can use different methodology and organizational principles regarding ML field with specific roles like data scientist, data engineer and domain expert. Typical concerns of the ML field include quality of data, hyperparameter tuning for algorithms, model performance assessment, visualization and explanation of the results. On the other hand standard software engineering deals with the concerns like security, availability, testing, system maintenance and update. New important role in development team may be assigned for an expert that has experience in both software engineering and application of ML. Main coordinator can more efficiently spot and manage design issues or trade-offs that arise due to the complexity of ML-based system.

Microservices architecture is getting common over the last years. This architectural paradigm assumes that system is composed of many loosely coupled components (or services) which are independently deployable [84]. Different technology stack, programming languages and data sources can be used for different components. Such design capabilities seem to be especially valuable in development of complex ML-based systems. Scalability and update flexibility of this approach outweigh the potential performance gains of monolithic architecture [85].

### 7.1.2 Data integration

Early diagnosis of glaucoma is a challenging task. Diagnostic routine includes diverse examinations and the resulting data require appropriate interpretation. The use of wearable medical devices is constantly growing in many fields of health care [86]. Continuous

monitoring of the physiological signals can provide data essential for the development of reliable diagnostic methods and management standards for the disease.

One of the important aspects of data integration is combining data from different sources and delivering unified view of them for the users of the system [87]. Notion of the case or, equivalently, set of examinations were introduced in the system to integrate series of patient examinations in time slot assigned for diagnosis. The main focus is on the analysis of sensor data therefore typical case involves Triggerfish and SOMNOtouch data from one 24-hour session. Measurements like IOP, CH, CRF, CCT (central corneal thickness) and the other anatomical readings of the eye features are also included. Any relevant examination data can be added to the case e.g. OCT, fundus images or selected optometric test results. Integration based on the notion of the case enables sharing of the sensor data and creating research data sets available for the selected users. Statistical analysis of such large data sets can lead to identification of patient subgroups that have specific characteristic related to diagnosis or management of a disease.

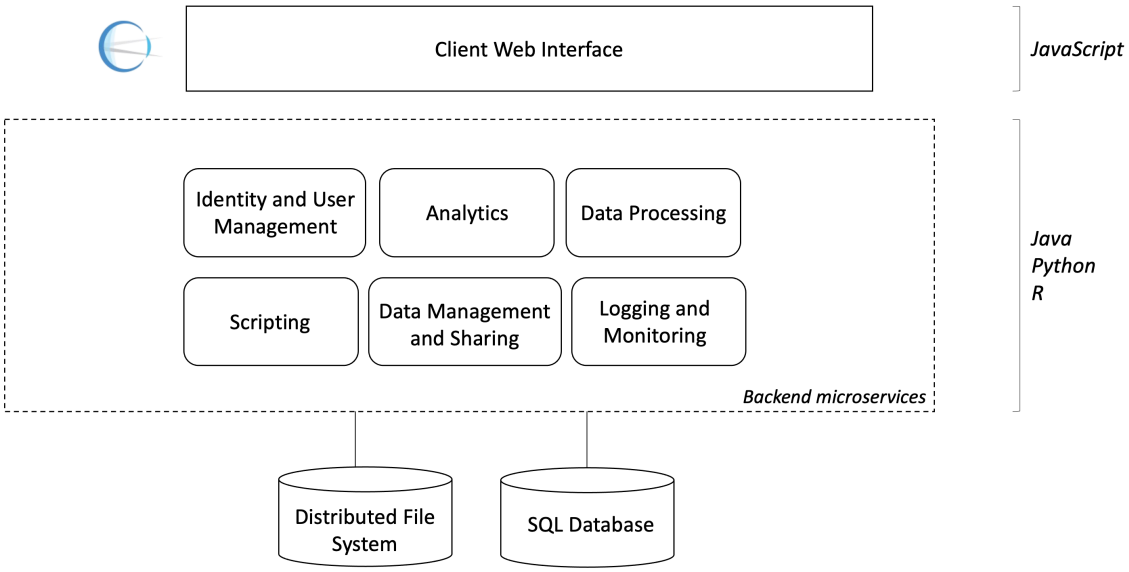


FIGURE 7.2: Overview of the basic system components.

### 7.1.3 Application services

The system is comprised of backend components, data storage and web application interface (see Figure 7.2). Core backend components are implemented in Java using Play Framework for Java. Frontend web interface is built using JavaScript libraries including Plotly Graphing Library for making interactive plots of multiple data formats. PostgreSQL is currently used as a relational database management system.

Scripting component is responsible for running Python and R code for basic processing of sensor data and application of the selected ML techniques. R language offers many libraries for data processing and the recent ML algorithms. Reference implementation of R (i.e. GNU R) works well in practice, but it is quite complex. Several attempts have been made to create an implementation of the R language that provides better performance while keeping compatibility with the original version of R [88]. Java environment is the basis for implementation of the main system therefore JVM-based interpreters for R (e.g. Renjin) seemed appropriate to examine in practice. Such approach could contribute to the closer integration of data processing components with the core system. Nevertheless, it is challenging to map semantics of scripting language like R that is significantly differ from the JVM bytecode. At this moment JVM-based interpreter for R requires additional modification/adjustment of the scripts and imposes constraints for specific R library versions. One of the objectives of the system is to run generic code in the scripting component with possibility of enhancement and fast deployment. Finally implemented solution is based on Java ProcessBuilder class to execute the scripts as operating system processes. This approach enables easy script update and use of multiple libraries.

### 7.1.4 ML environment

H2O has been chosen as the main ML framework for the system. It is open source, scalable platform providing implementation of a wide range of ML algorithms and tools. H2O can be connected to the multiple data sources such as relational databases, plain files, Apache Hadoop, Spark or Amazon S3 clusters.

ML models created and evaluated in H2O can be deployed into production environment using POJO (Plain Old Java Object) or MOJO (Model Object, Optimized) formats [74]. MOJO and POJO model files can be easy deployed in Java applications. Exported zip file with a model contains one dependency in h2o-genmodel.jar. Model package can be loaded using H2O Java API containing *MojoModel* class. Raw model prediction can be generated quickly as no connection to a running H2O cluster is required.

H2O provides Web Flow interface for model creation, tuning and evaluation. User can view model hyperparameters and detailed training and validation metrics. Variable importance plot, confusion matrix, ROC curve and other diagrams can be generated for a model. It is also possible to check model predictions for the selected dataset.

## 7.2 Application scenarios

### 7.2.1 Glaucoma diagnosis

Diagnosis is inherently individual assessment of a patient based on available data and clinical experience of medical doctor [89]. In the scenario focused on clinical decision support we consider supplementary information generated by ML model that can be used in evaluation of the most relevant diagnostic hypotheses.

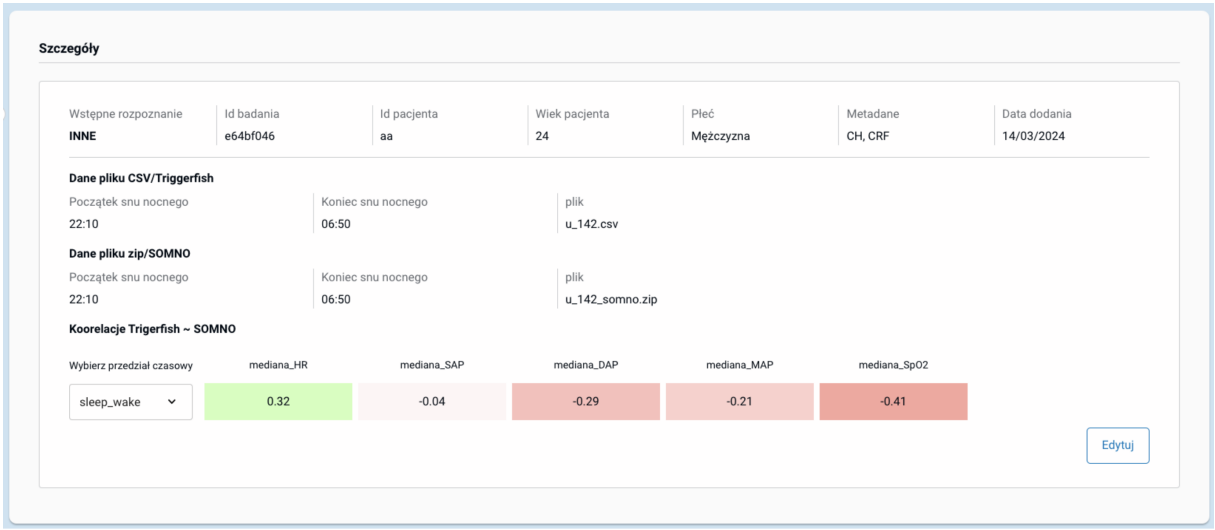


FIGURE 7.3: View of a single case data in the system. It contains Triggerfish CLS and SOMNOtouch NBIP output files (for 24-hour session), clinical measurements (e.g. CH, CRF, IOP etc.) and other data (sex, age, patient ID etc.). Correlation coefficient of Triggerfish CLS and cardiac sensor data (HR, SAP, DAP, MAP, $SpO_2$) is shown at the bottom (based on median values for bursts in the selected time interval). User can choose time interval from the list on the left (e.g. sleep_wake).

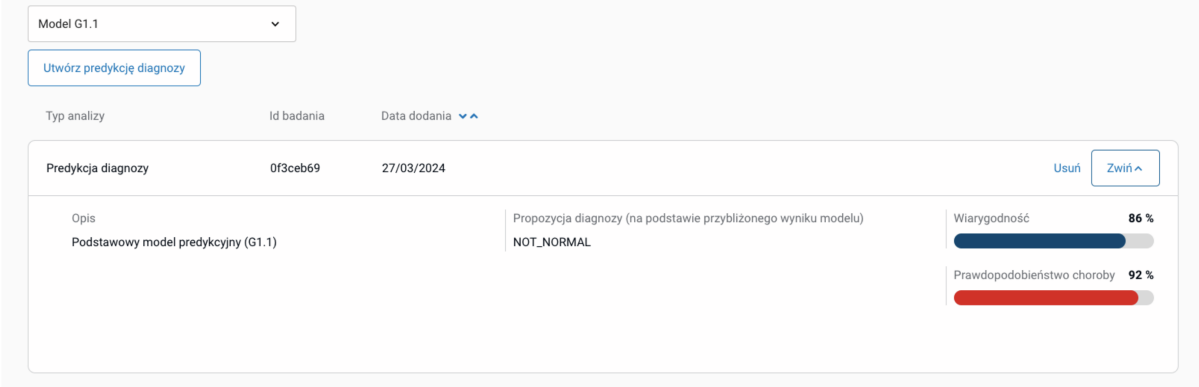


FIGURE 7.4: Prediction result for the model selected by user. Simple measure of confidence is computed for the predicted class label (blue bar). It is based on the distance between predicted probability (red bar) and reference threshold of model. If the distance is small then confidence is low (close to 0%). If the distance is large then confidence is high (close to 100%).

Diagnostic support scenario is intended for the medical professionals. They can check basic output of ML model which is predicted probability of glaucomatous neuropathy

(positive) diagnosis for a case (see Figure 7.4). The result is shown as numeric probability attached to the single bar chart. Class label (NORMAL/NOT_NORMAL) relevant to the predicted probability is also shown. It is based on a comparison with optimal classification threshold determined for the model. Diagram with explanation of the model prediction is shown to allow interpretation of the result by a user (see Figure 7.5). Understanding and comparing how a model uses the attributes to make a given prediction can provide opportunity to get insight of its properties [21]. Currently we use LIME (Local Interpretable Model-agnostic Explanations) and DALEX visualization techniques for generation of explanation of particular prediction. LIME method assumes that every complex model is linear on a local scale and can be approximated with an interpretable model [56]. DALEX break-down plots are fast approximations of Shapley values [90]. Such local explanations enable evaluation of the practical usability of the model by ophthalmologists experienced in glaucoma diagnosis. This approach can lead to refinement of the model attributes by data scientists or ML engineers working together with the doctors.

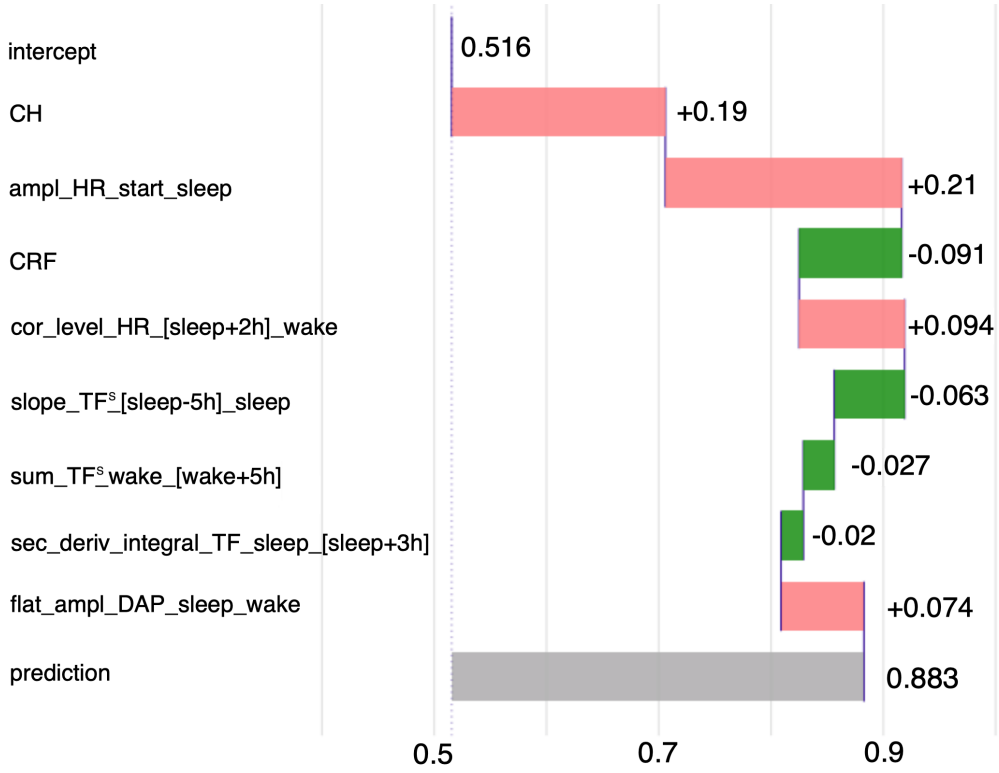


FIGURE 7.5: Prediction for particular case can be decomposed onto model attributes using DALEX break-down profile generated on the basis of conditional responses of the model. User can assess the contribution of each attribute to the prediction (0.883) for the instance. Positive attribute contributions are shown as pink bars (e.g. CH), negative as green bars (e.g. CRF). Intercept can be interpreted as mean value (an estimate of the expected value of the model's predictions for all cases).

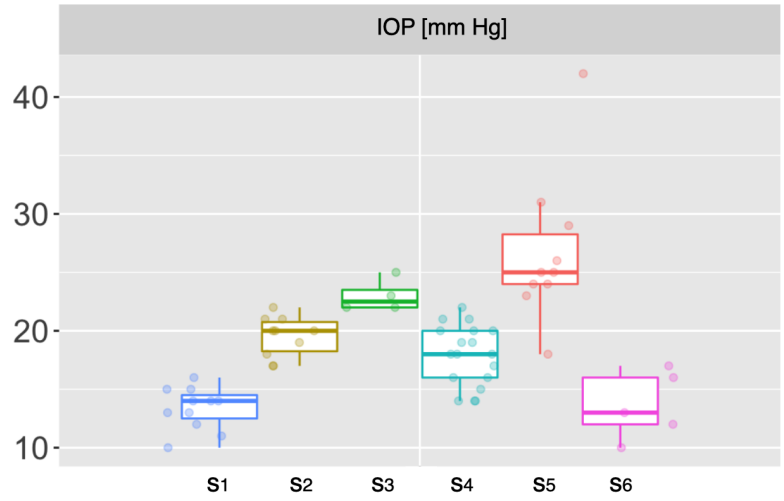


FIGURE 7.6: Comparison of IOP [mm Hg] measurements distribution in the selected sets of cases. Box plots for the 6 sets of cases defined on the basis of diagnosis: $s_1$:NORM, $s_2$:suspected OH (ocular hypertension), $s_3$:OH, $s_4$:suspected POAG/HTG, $s_5$:POAG/HTG, $s_6$:POAG/NTG.

### 7.2.2 Data visualization

Embedding model results in extended context of the available patient data can facilitate identification of specific features related to the course of disease and prognosis. Following recommendations of the domain experts we provided box plot [91] view of distribution of data included in a set of selected cases. This type of plot is common in scientific papers and usually is correctly interpreted by the medical professionals. Box plots supports visual comparison of data distribution properties across different sets of cases (see Figure 7.6). User can define set of cases using filtering by diagnosis or the range of value of the selected measurements (IOP, CH, CRF etc.). It is also possible to generate heat map view of Triggerfish and cardiac signal correlations for a set of selected cases (in time intervals chosen by the user). Rows (cases) in the heat map matrix can be ordered according to the result of hierarchical clustering for Euclidean distance (see Figure 7.7). Identification of subgroups of cases with significant positive or negative correlations and specific properties quantified by the other measurements can lead to more efficient diagnostic or treatment recommendations (related to properties of the cardiovascular system).

### 7.2.3 Collaborative research

Collaborative research is a way of tackling complex problems. Clear communication is essential for effective collaboration. Design of the system assumes creation of research projects that support activities related to development and application of personalized medicine techniques. Users working on the selected issue can add notes/comments containing multiple content in workspace of a project. Entries added by the users can contain text, images, links and embedded results of analytic functions provided by the

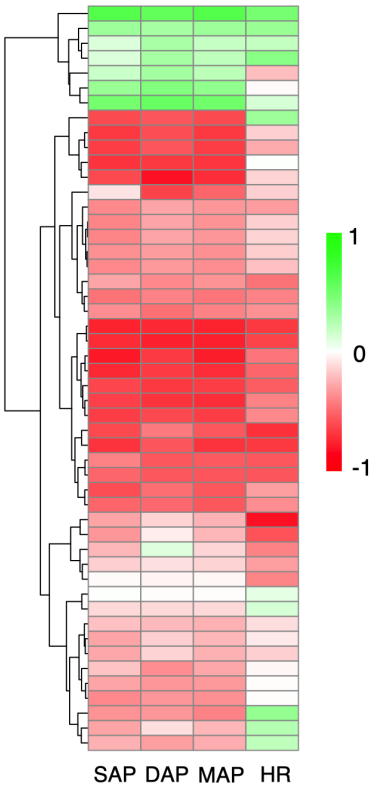


FIGURE 7.7: Heat map view of the Spearman's rank correlation coefficient values for TF and cardiac sensor signal: SAP, DAP, MAP (mean arterial pressure), HR in the first 2 hours of the main/night sleep period (SLEEP_[SLEEP+2h]). View generated for the set of 50 cases.

system. Default order of entries is aligned with the main timeline of the project. Each analytic result includes its timestamp and the details of set of cases for which it was created. Precise identification of the analytic results (such as box plots, heat maps and the other graphs) helps in tracking of discussion and development of the research conclusions.

Users can share access to selected cases from their repository (see view of a case in Figure 7.3). It encourages data collection and application of the analytic functions to assess specific properties of the data. Data sharing is one way to prompt more frequent use of sensor data in development of new glaucoma diagnosis and control standards based on continuous monitoring of eye and cardiovascular system properties. System is intended to enable flexible exploration of data and exchange of new ideas. Constraints related to research workflow in the system are assumed to be soft and primarily intended to improve coordination and reproducibility of the results.

## 7.3 Transdisciplinarity

Many complex research issues entail application of methods from multiple disciplines. Boundaries of the fields in science and technology are not ultimately determined and can change over time. Integration of knowledge and experience from many disciplines may

yield new ideas and improve credibility of research. Effective organizational structures can foster innovative science projects by introducing policies, practices and recommendations for the transdisciplinary research (addressing issues such as provision of additional resources or long-term funding) [92].

### 7.3.1 Transdisciplinary research

Transdisciplinarity concept has roots in discussions about the need for new forms of interdisciplinary collaboration [93]. It encompasses the following key characteristics:

- focus on specific, complex, real-word problems that are important for society
- transformative approach (i.e. support of action or change of the status quo)
- contemplation of broad context of the research and compatibility of its parts
- development of integrated knowledge that crosses disciplinary boundaries

Transdisciplinarity provides framework for the research projects and highlights importance of validation of the outcomes from different perspectives [94]. Many aspects of the research presented in the thesis are linked to the basic concepts of transdisciplinarity.

Glaucoma affects many millions of patients in the world and remains a major problem for health care system. As the age is one of the significant risk factors, the disease is an important issue in the aging population. One of the aims of the research is development of new diagnostic methods based on multi-sensor data. These methods can supplement current options available for patients and change diagnosis and treatment standards.

Services of the system can be extended and accommodated to handle different data formats. New system scenarios can be proposed regarding possible data sources. For example, using genotyping arrays to find single nucleotide polymorphisms (SNP) across genome will allow identification of genetic variants associated with risk of progression in particular glaucoma types [17]. Adding new system capabilities to handle such SNP data with consideration of its relationship with the other data will be consistent with transdisciplinarity assumptions related to compatibility and development of integrated knowledge. Involvement of users from many clinical and research fields is in line with the transdisciplinary attitude to the crossing of disciplinary boundaries.

### 7.3.2 Community adoption

Lack of adequate tools for management, sharing and analysis of collected data reduces productivity of the research [87]. Design of the system is focused on the services for handling Triggerfish and cardiac sensor data. Data collection workflow in the system is similar to the approach common in clinical practice. It facilitates adoption by the

eye doctors and customization of the services. We can assume evolution of the system and incremental implementation of the functions that address new requirements arising in the diagnostic and analytic scenarios. As the users of the system have different experience and knowledge it is important to introduce guidelines for application of the ML techniques to diagnosis of glaucoma. Relevant issues in this context include:

- quality of sensor data in relation to patient activity during the day
- understanding limitations of the ML models and appropriate assessment of the predicted output
- using statistical guidelines as well as clinical knowledge in interpretation of data exploration results

Appropriate organizational perspective can support setting of priorities for system development and its application. Establishment of the center of excellence for glaucoma research can provide high performance computational resources for implementation of the latest ML techniques in the field of ophthalmology, support maintenance of the system and collaborative development of personalized medicine standards.

## 7.4 Conclusions

In recent years, many new devices for continuous monitoring of patient health parameters have become available [86]. Latest ML algorithms are able to efficiently process complex data. It makes possible to develop personalized approach in many fields of medicine. 24-hour Triggerfish record joined with cardiac sensor data can be used to more accurately diagnose and track progression of glaucoma. Nevertheless, large amount of data is required to build reliable ML models [95]. At this moment Trigerfish device is not commonly used in clinical practice and software tools for sensor data processing usually offer only simple analytic functions.

While availability of the research data is limited, inclusion of new users in the system will facilitate development of the analytical services and evaluation of clinical observations. Support for scenarios involving application of ML techniques for assessment of specific cases can encourage medical professionals to collect and share more sensor data for patients. Consequently, greater availability of the data can increase interest in collaborative research scenarios for novel approaches in glaucoma diagnosis and control.

# Chapter 8

# Data analysis scenarios for collaborative research

This chapter presents exploratory data analysis scenarios that have been implemented during the research. Triggerfish and devices for continuous monitoring of cardiovascular system parameters have been introduced relatively recently into the clinical toolkit. Relation of Triggerfish CLS signal and cardiac activity with clinical data has not been extensively studied yet and there are few detailed medical knowledge sources that could facilitate interpretation of such data for specific cases. Existing publications use different data analysis methods and usually report the properties of small patient groups. These results can't be directly compared with each other or immediately used as the basis for guidelines of clinical routine.

Exploratory data analysis is an approach that uses standard data science techniques for comprehensive characterization of properties of the available data. To answer basic questions about data properties we can use statistical tests, clustering, principal component analysis (PCA) and various data visualization methods (diagrams such as box plots, scatterplots and heatmaps are commonly used in natural sciences and medicine).

Source code for the scenarios described in this chapter was created by the author using R and Python programming languages. R is scripting language designed particularly for statistical analysis and data visualization. Python is general, object-oriented, scripting language that is widely used for implementing complex and scalable ML solutions.

Current diagnostic routine in ophthalmology involves many measurements. Even imaging modalities (such as OCT) extract numerical parameters describing diagnostic output. Such data is usually saved (e.g. in CSV files/tables) and over time a considerable amount of data may be collected in clinic. Collaboration of the doctor with data scientist and ML engineer enables efficient data merging and basic inference. Clinical decisions can be made according to the complex patient data profile involving multiple modalities output. Application of statistical and ML techniques for comprehensive evaluation of

individual patient based on the available data resources can be seen as implementation of personalized medicine premises.

## 8.1 Relationship of Triggerfish CLS and cardiac sensor data derived attributes with clinical measurements

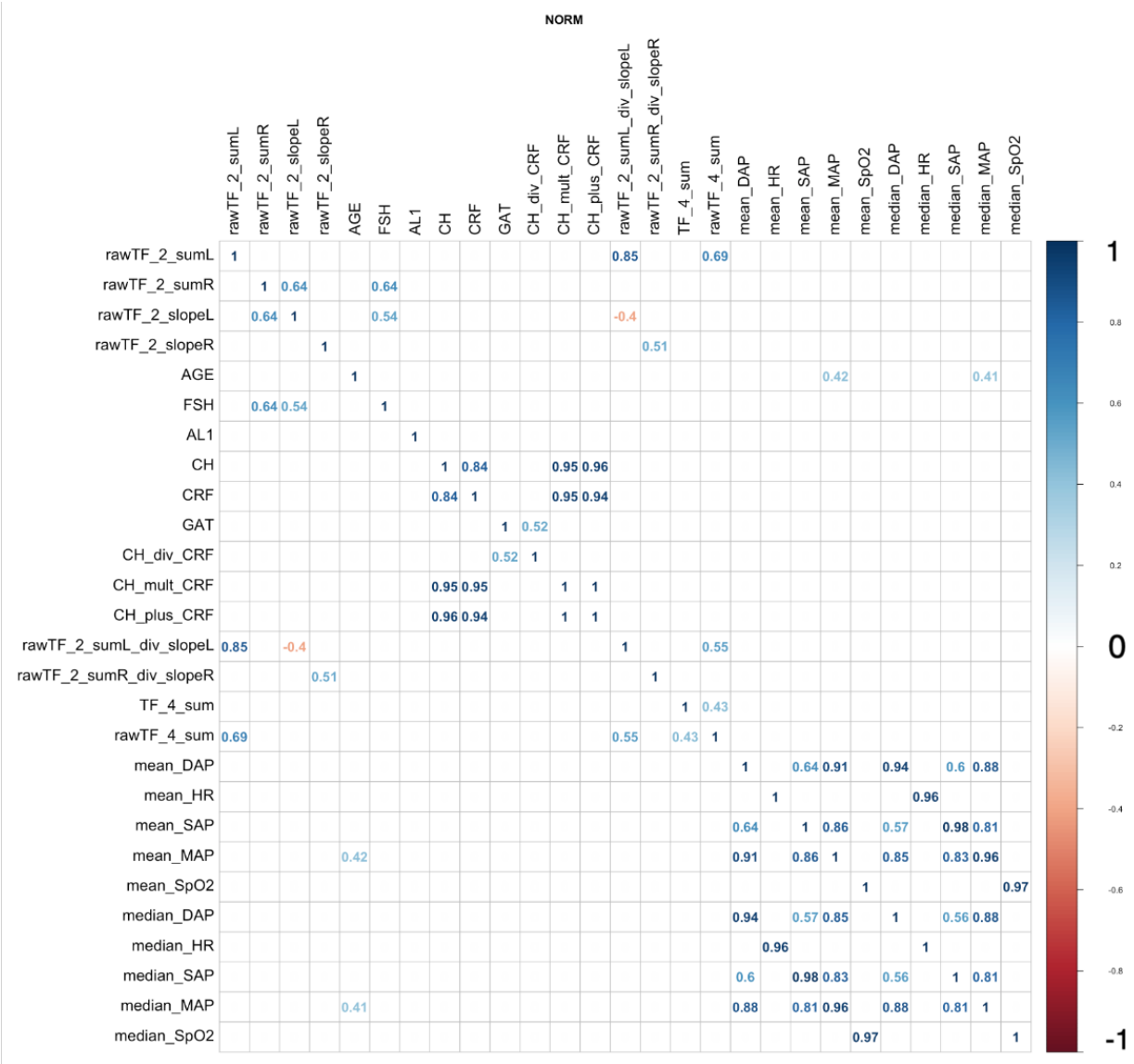


FIGURE 8.1: Correlation matrix for the normal case group (NORM).

In this scenario we investigate relation of clinical data with attributes derived for Triggerfish (TF) and SOMNOtouch data. Working hypothesis assumes that there is a relationship between clinical measurements (especially IOP or biomechanical eye properties: CH, CRF) and TF attributes derived for sleep_[sleep+3h] time interval. Additionally we focus on the increase of TF value at the beginning of the night sleep. Typically the

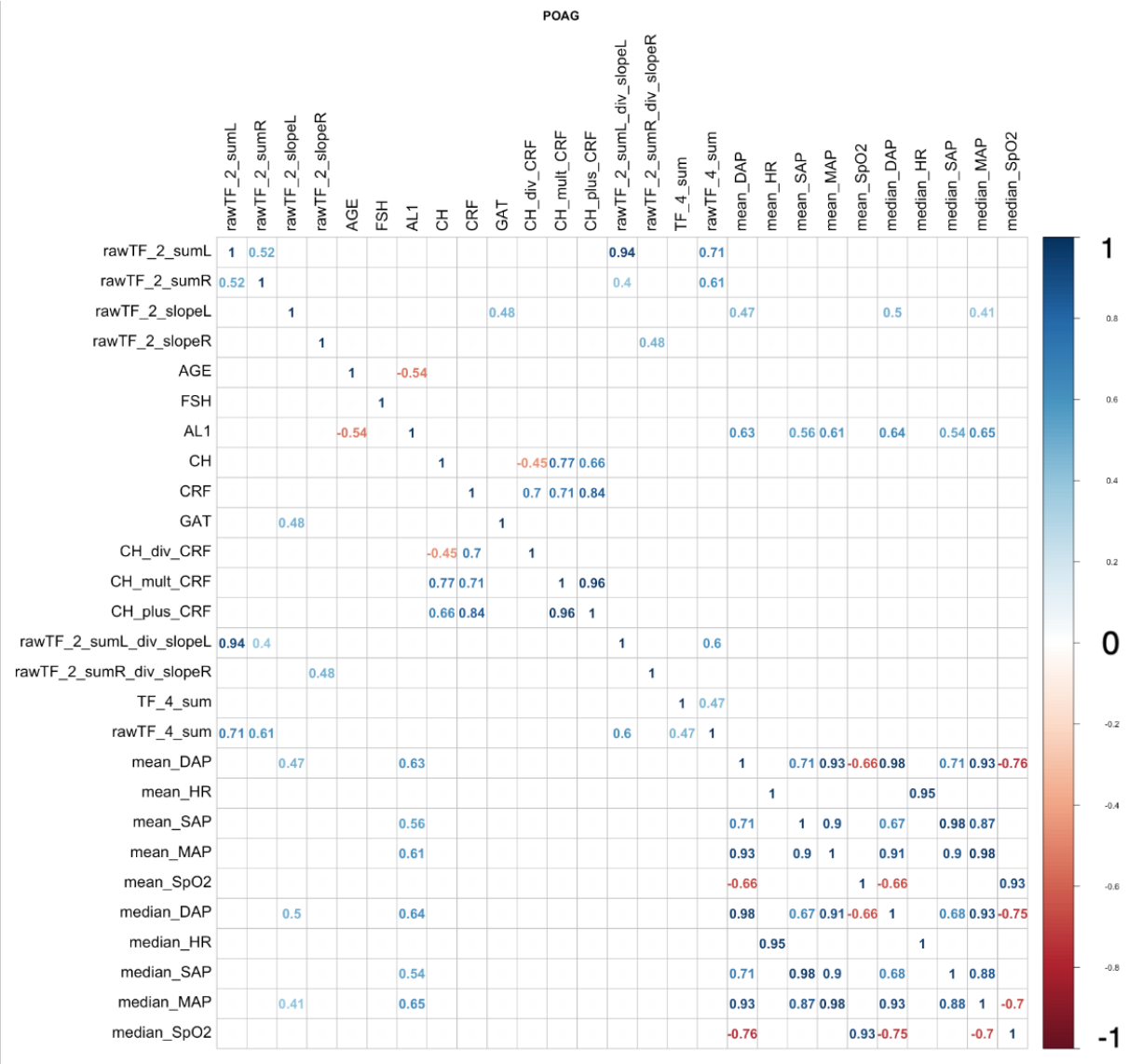


FIGURE 8.2: Correlation matrix for the glaucomatous neuropathy case group (POAG).

rate of TF increase stabilizes at some timepoint. We can mark this point as a vertex V and try to fit two straight lines (arms of the angle) to TF data to estimate exact time value for the vertex. Application of the stochastic optimization algorithm from SciPy library in Python (scipy.optimize.differential_evolution) gives correct timepoint value for most cases. Manual correction of the timepoint (within 30m range) was performed additionally for selected cases. Input dataset contains 43 normal and 46 POAG cases with the timepoint V located up to sleep+3.5h.

We divided sleep_[sleep+3h] interval using V value into the initial part (L) and final part (R). For these intervals we computed rawTF_2_sumL/R: sum under the TF curve (median values for the bursts) and rawTF_2_slopeL/R: slope (in radians) of linear regression line for TF fitted using standard least squares method. Sum, ratio and product (_plus, _div, _mult) for some of the attributes were also considered in the analysis. GAT denotes IOP measured using GAT, AL1 refers to the axial length of the eye (see details

| var1 | var2 | n | r | p.val | lower.ci | upper.ci |
|---|---|---|---|---|---|---|
| rawTF_2_sumL | rawTF_2_sumL_div_slopeL | 43 | 0.85 | <0.001 | 0.74 | 0.92 |
| rawTF_2_sumL | rawTF_4_sum | 43 | 0.69 | <0.001 | 0.49 | 0.82 |
| rawTF_2_sumR | rawTF_2_slopeL | 43 | 0.64 | <0.001 | 0.42 | 0.79 |
| rawTF_2_slopeL | rawTF_2_sumL_div_slopeL | 43 | -0.40 | 0.008 | -0.63 | -0.11 |
| rawTF_2_slopeR | rawTF_2_sumR_div_slopeR | 43 | 0.51 | <0.001 | 0.25 | 0.70 |
| AGE | mean_MAP | 38 | 0.42 | 0.008 | 0.14 | 0.64 |
| AGE | median_MAP | 38 | 0.41 | 0.010 | 0.13 | 0.64 |
| CH | CRF | 43 | 0.84 | <0.001 | 0.72 | 0.91 |
| rawTF_2_sumL_div_slopeL | rawTF_4_sum | 43 | 0.55 | <0.001 | 0.30 | 0.73 |

TABLE 8.1: Selected correlation coefficient values (r) from the matrix for the normal case group.

in 5.3), TF_4_sum refers to the (scaled/raw) sum under TF curve in wake_[wake+5h] interval. Mean/median value of SAP/DAP/MAP/HR in sleep_wake interval were included in the correlation matrix.

Squared (symmetrical) matrices contain Spearman's rank correlation coefficient values generated for the attribute pairs (see Figure 8.1 and 8.2). It is a nonparametrical statistical measure of monotonic relation between variables. Table 8.1 and 8.2 enumerate correlation coefficient values for the selected attribute pairs (r), count of non-empty, correct values in the relevant data (n), p-value of the statistical significance test (i.e. test if correlation coefficient is significantly different from 0 in population) and estimated confidence intervals (ci).

Positive correlation of rawTF_2_slopeL with IOP (GAT) and mean/median_DAP can be observed in POAG case group. But there is no significant correlation of these attributes in normal case group. Such relations are potentially valuable and can be further investigated as IOP is one of the most important factors related to development of the glaucoma.

There is no direct, significant correlation of CH and CRF with TF attributes derived for TF. Preliminary evaluation of predictive models involving CH and CRF showed that addition of the considered TF attributes (derived for the L/R interval) essentially don't improve predictive performance of the models.

## 8.2 Comparison of sensor data derived attributes in case groups based on the diagnosis and additional criteria

In this scenario we define case groups based on diagnosis (NORM, NTG and POAG) and Spearman's rank correlation coefficient value for selected intervals. We considered relation of TF (Triggerfish) and SAP/DAP/HR (SOMNOtouch) measurements. We

| var1 | var2 | n | r | p.val | lower.ci | upper.ci |
|---|---|---|---|---|---|---|
| rawTF_2_sumL | rawTF_2_sumR | 46 | 0.52 | <0.001 | 0.27 | 0.70 |
| rawTF_2_sumL | rawTF_2_sumL_div_slopeL | 46 | 0.94 | <0.001 | 0.89 | 0.97 |
| rawTF_2_sumL | rawTF_4_sum | 46 | 0.71 | <0.001 | 0.53 | 0.83 |
| rawTF_2_sumR | rawTF_2_sumL_div_slopeL | 46 | 0.40 | 0.006 | 0.12 | 0.62 |
| rawTF_2_sumR | rawTF_4_sum | 46 | 0.61 | <0.001 | 0.38 | 0.76 |
| rawTF_2_slopeL | GAT | 46 | 0.48 | 0.001 | 0.22 | 0.68 |
| rawTF_2_slopeL | mean_DAP | 42 | 0.47 | 0.002 | 0.21 | 0.67 |
| rawTF_2_slopeL | median_DAP | 42 | 0.50 | 0.001 | 0.24 | 0.69 |
| rawTF_2_slopeL | median_MAP | 42 | 0.41 | 0.008 | 0.13 | 0.62 |
| rawTF_2_slopeR | rawTF_2_sumR_div_slopeR | 45 | 0.48 | 0.001 | 0.22 | 0.68 |
| AGE | AL1 | 18 | -0.54 | 0.021 | -0.72 | -0.29 |
| rawTF_2_sumL_div_slopeL | rawTF_4_sum | 46 | 0.60 | <0.001 | 0.38 | 0.76 |

TABLE 8.2: Selected correlation coefficient values (r) from the matrix for the glaucomatous neuropathy case group.

selected results for division based on correlation of TF with HR for presentation in this section. If correlation coefficient modulus (absolute value) is greater or equal to 0.20 (custom threshold) for interval then we include a case in negative (relatively strong negative correlation) or positive group (relatively strong positive correlation).

Input dataset contains 116 cases. We focus on start_sleep, sleep_wake and tp2_wake time intervals as the day and night SAP/DAP level difference can be related to the ocular blood flow and eye function.

We compared distribution of arithmetic means of TF, SAP, DAP, MAP, HR in the intervals. Figures 8.3 and 8.4 show box plots for these attributes and additionally AGE, CH, CRF, GAT. Border color represents sign of the correlation criterion for the group (negative: red, positive: green). Fill color refers to diagnosis (NORM: blue, NTG: white, POAG: pink).

Box plots provide useful representation of variable distribution. Central horizontal bar marks the median, lower and upper edges of the box represent the first ($Q_1$) and third quartile ($Q_3$). Whiskers are usually drawn within the 1.5 interquartile range ($Q_3$-$Q_1$) from the box edges. Values outside the whiskers range are called outliers.

Nonparametric Mann–Whitney U test (known as Wilcoxon rank sum test) was used to compare attribute distributions and check if differences observed in the box plots are statistically significant (see selected comparisons in Table 8.3, 8.4, 8.5, 8.6). Output of the test depends on the shape and location parameters of the compared distributions.

Comparing NORM positive and NORM negative groups we can see difference in distribution of TF mean values in the considered time intervals. There are no significant attribute distribution differences for comparison of NORM positive with NTG positive case groups. Comparing NORM negative with NTG positive we can see difference in distribution of TF mean values in sleep_wake and tp2_wake interval.

Comparing NORM negative and NTG negative we can see difference in distribution of SAP and MAP mean values in the considered time intervals. Comparing NORM positive vs. NTG negative we can see difference in distribution of SAP and MAP mean values in the considered time intervals.

Based on the comparisons we can suppose that SAP, MAP mean values in NTG negative case group (in the considered intervals) can be important factors in the group and modification of the SAP, MAP mean levels can potentially affect ocular blood flow and eye function. Such exploratory analysis of the specific case groups may lead to detailed characterization of group properties and factors related to the condition.

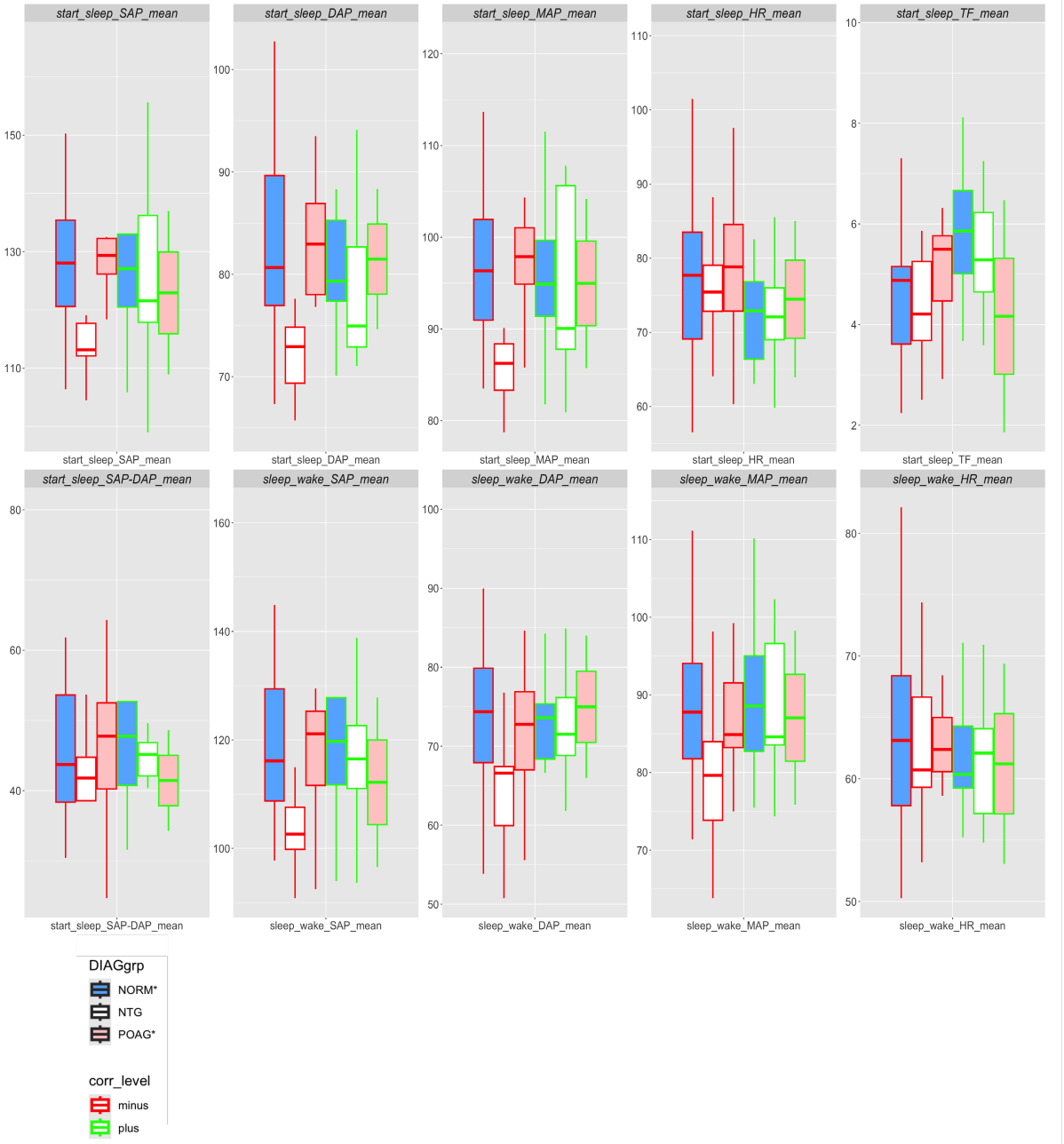


FIGURE 8.3: Side-by-side box plots for the groups based on diagnosis (fill color) and correlation of TF with HR in sleep_wake time interval (border color).

| | p.val | lower.conf.int | upper.conf.int | p.val<0.05 |
|---|---|---|---|---|
| start_sleep_SAP_mean | 0.903 | -10.86 | 13.37 | |
| start_sleep_DAP_mean | 0.542 | -8.88 | 5.16 | |
| start_sleep_MAP_mean | 0.903 | -8.39 | 6.84 | |
| start_sleep_HR_mean | 0.165 | -11.66 | 2.69 | |
| start_sleep_TF_mean | 0.018 | 0.08 | 2.53 | T |
| start_sleep_SAP-DAP_mean | 0.715 | -7.08 | 13.80 | |
| sleep_wake_SAP_mean | 0.614 | -9.33 | 17.61 | |
| sleep_wake_DAP_mean | 0.715 | -7.56 | 5.63 | |
| sleep_wake_MAP_mean | 0.821 | -7.27 | 9.42 | |
| sleep_wake_HR_mean | 0.497 | -7.52 | 2.84 | |
| sleep_wake_TF_mean | 0.015 | -3.86 | -0.26 | T |
| sleep_wake_SAP-DAP_mean | 0.302 | -4.03 | 15.88 | |
| tp2_wake_SAP_mean | 0.614 | -8.16 | 18.19 | |
| tp2_wake_DAP_mean | 0.821 | -7.01 | 6.64 | |
| tp2_wake_MAP_mean | 0.794 | -6.73 | 10.14 | |
| tp2_wake_HR_mean | 1.000 | -5.75 | 4.88 | |
| tp2_wake_TF_mean | 0.005 | -4.19 | -0.73 | T |
| tp2_wake_SAP-DAP_mean | 0.336 | -3.27 | 16.10 | |
| AGE | 1.000 | -11.00 | 12.00 | |
| CH | 0.296 | -1.50 | 0.50 | |
| CRF | 0.272 | -2.10 | 0.60 | |
| GAT | 0.943 | -2.00 | 2.00 | |

TABLE 8.3: Statistical significance of attribute distribution difference between the groups according to Mann–Whitney U test (comparison of NORM positive and NORM negative).

| | p.val | lower.conf.int | upper.conf.int | p.val<0.05 |
|---|---|---|---|---|
| start_sleep_SAP_mean | 0.001 | 6.70 | 22.24 | T |
| start_sleep_DAP_mean | 0.010 | 2.60 | 14.39 | T |
| start_sleep_MAP_mean | 0.001 | 4.02 | 15.33 | T |
| start_sleep_HR_mean | 0.589 | -6.72 | 9.26 | |
| start_sleep_TF_mean | 0.903 | -0.72 | 1.17 | |
| start_sleep_SAP-DAP_mean | 0.319 | -2.79 | 13.35 | |
| sleep_wake_SAP_mean | 0.006 | 3.59 | 22.43 | T |
| sleep_wake_DAP_mean | 0.057 | -0.32 | 16.18 | |
| sleep_wake_MAP_mean | 0.018 | 1.95 | 17.32 | T |
| sleep_wake_HR_mean | 0.794 | -5.19 | 7.53 | |
| sleep_wake_TF_mean | 0.154 | -0.24 | 3.40 | |
| sleep_wake_SAP-DAP_mean | 0.542 | -3.89 | 13.24 | |
| tp2_wake_SAP_mean | 0.002 | 3.93 | 22.14 | T |
| tp2_wake_DAP_mean | 0.053 | -0.65 | 16.83 | |
| tp2_wake_MAP_mean | 0.013 | 1.73 | 17.74 | T |
| tp2_wake_HR_mean | 0.794 | -4.49 | 7.94 | |
| tp2_wake_TF_mean | 0.154 | -0.37 | 3.80 | |
| tp2_wake_SAP-DAP_mean | 0.542 | -3.76 | 13.79 | |
| AGE | 0.235 | -19.00 | 7.00 | |
| CH | 0.010 | 0.60 | 2.80 | T |
| CRF | 0.012 | 0.50 | 4.00 | T |
| GAT | 0.020 | 0.00 | 5.00 | T |

TABLE 8.4: Statistical significance of attribute distribution difference between the groups according to Mann–Whitney U test (comparison of NORM negative and NTG negative).

| | p.val | lower.conf.int | upper.conf.int | p.val<0.05 |
|---|---|---|---|---|
| start_sleep_SAP_mean | 0.024 | 0.87 | 38.60 | T |
| start_sleep_DAP_mean | 0.040 | 0.75 | 13.01 | T |
| start_sleep_MAP_mean | 0.031 | 1.29 | 18.89 | T |
| start_sleep_HR_mean | 0.605 | -11.68 | 4.48 | |
| start_sleep_TF_mean | 0.077 | -0.07 | 2.86 | |
| start_sleep_SAP-DAP_mean | 0.190 | -4.24 | 25.86 | |
| sleep_wake_SAP_mean | 0.024 | 1.39 | 36.92 | T |
| sleep_wake_DAP_mean | 0.077 | -0.66 | 16.27 | |
| sleep_wake_MAP_mean | 0.040 | 0.69 | 21.31 | T |
| sleep_wake_HR_mean | 0.796 | -8.59 | 5.61 | |
| sleep_wake_TF_mean | 0.605 | -3.72 | 2.50 | |
| sleep_wake_SAP-DAP_mean | 0.136 | -2.59 | 29.12 | |
| tp2_wake_SAP_mean | 0.024 | 2.81 | 35.05 | T |
| tp2_wake_DAP_mean | 0.040 | 0.03 | 17.09 | T |
| tp2_wake_MAP_mean | 0.031 | 1.24 | 22.35 | T |
| tp2_wake_HR_mean | 0.730 | -7.44 | 7.89 | |
| tp2_wake_TF_mean | 0.605 | -4.70 | 2.19 | |
| tp2_wake_SAP-DAP_mean | 0.113 | -2.12 | 30.64 | |
| AGE | 0.377 | -23.00 | 10.00 | |
| CH | 0.094 | -0.30 | 2.70 | |
| CRF | 0.145 | -0.80 | 3.80 | |
| GAT | 0.083 | 0.00 | 6.00 | |

TABLE 8.5: Statistical significance of attribute distribution difference between the groups according to Mann–Whitney U test (comparison of NORM positive and NTG negative).

| | p.val | lower.conf.int | upper.conf.int | p.val<0.05 |
|---|---|---|---|---|
| start_sleep_SAP_mean | 0.412 | -6.59 | 15.28 | |
| start_sleep_DAP_mean | 0.255 | -2.99 | 8.93 | |
| start_sleep_MAP_mean | 0.286 | -4.27 | 11.36 | |
| start_sleep_HR_mean | 0.200 | -2.72 | 12.12 | |
| start_sleep_TF_mean | 0.080 | -2.28 | 0.03 | |
| start_sleep_SAP-DAP_mean | 0.876 | -7.74 | 9.02 | |
| sleep_wake_SAP_mean | 0.986 | -9.95 | 12.13 | |
| sleep_wake_DAP_mean | 0.741 | -6.82 | 7.95 | |
| sleep_wake_MAP_mean | 0.821 | -6.76 | 8.86 | |
| sleep_wake_HR_mean | 0.497 | -3.77 | 7.41 | |
| sleep_wake_TF_mean | 0.023 | 0.28 | 4.38 | T |
| sleep_wake_SAP-DAP_mean | 0.821 | -7.25 | 7.09 | |
| tp2_wake_SAP_mean | 0.986 | -10.66 | 11.29 | |
| tp2_wake_DAP_mean | 0.876 | -7.48 | 7.59 | |
| tp2_wake_MAP_mean | 0.986 | -7.73 | 8.54 | |
| tp2_wake_HR_mean | 0.958 | -4.62 | 6.00 | |
| tp2_wake_TF_mean | 0.004 | 1.42 | 5.42 | T |
| tp2_wake_SAP-DAP_mean | 0.768 | -8.00 | 6.98 | |
| AGE | 0.056 | -26.00 | 0.00 | |
| CH | 0.002 | 0.60 | 2.40 | T |
| CRF | 0.001 | 1.10 | 3.40 | T |
| GAT | 0.130 | -1.00 | 3.00 | |

TABLE 8.6: Statistical significance of attribute distribution difference between the groups according to Mann–Whitney U test (comparison of NORM negative and NTG positive).

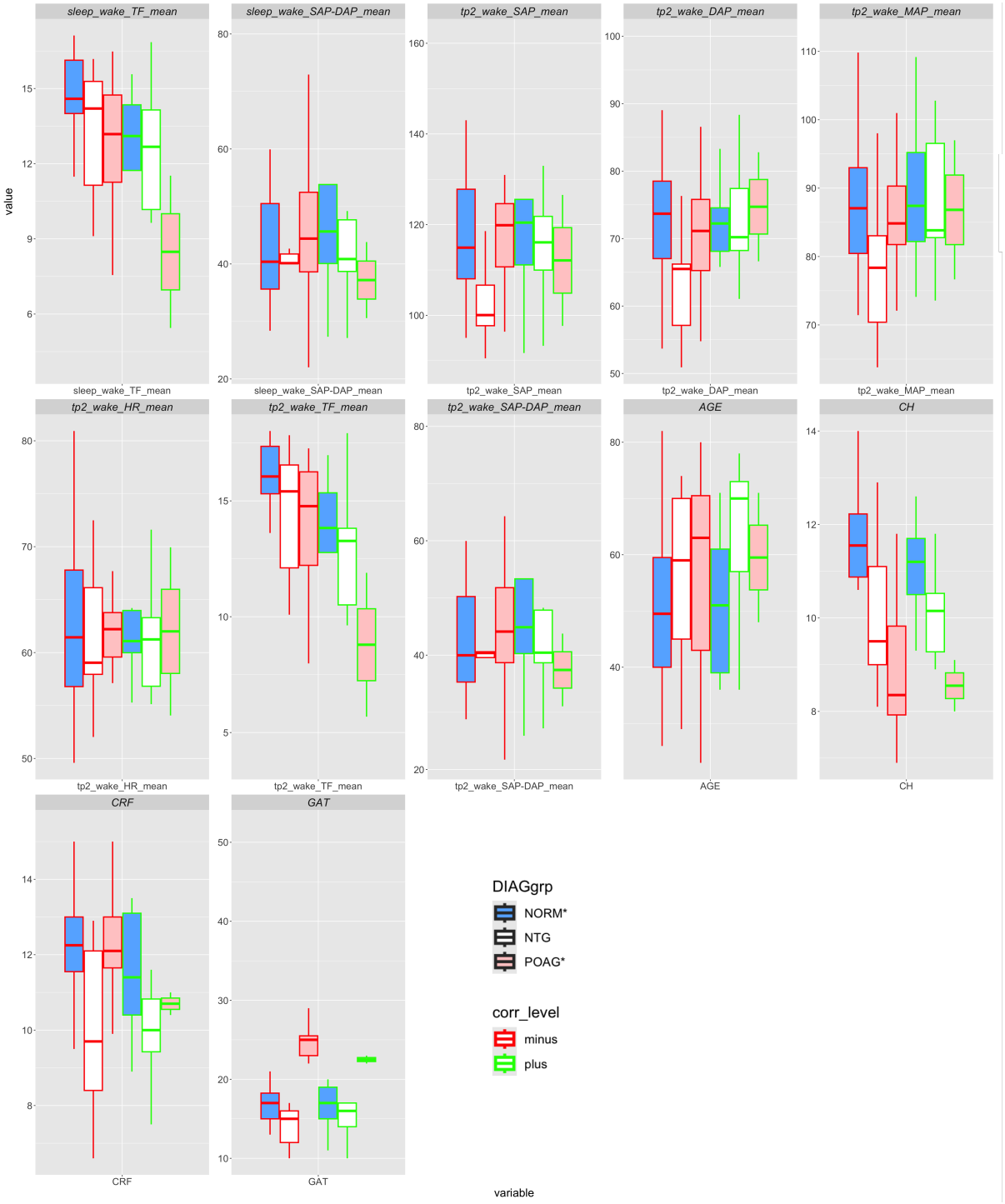


FIGURE 8.4: (continued) Side-by-side box plots for the groups based on diagnosis (fill color) and correlation of TF with HR in sleep_wake time interval (border color).

## 8.3 Clustering clinical data

In this scenario we perform clustering analysis of the set of 95 cases. We consider basic clinical measurements: GAT IOP, DCT (Dynamic Contour Tonometry IOP), OPA (ocular pulsation amplitude), IOPCC (ORA corneal compensated IOP), CH, CRF. Unsupervised ML methods such as clustering are used to characterize structure of multidimensional

datasets and identify subgroups of cases with specific properties. Hierarchical clustering is a common approach for clustering mixed-type or clinical data [96]. In this study we applied agglomerative hierarchical clustering algorithm with Ward's criterion (minimum variance method).

Dendrogram is an interpretable representation of the hierarchical clustering output. Leaf (terminal) nodes which represent single observations are plotted at zero height. The height of each node in the binary tree is proportional to the value of the intergroup dissimilarity between its two children. We receive the partition of the dataset into disjoint clusters by cutting the dendrogram horizontally at a particular height. It is equivalent to termination of the algorithm when intergroup dissimilarity exceeds a certain threshold value [46].

Principal component analysis (PCA) is a linear dimensionality reduction technique. The first principal component is a linear combination of the original variables that explains the most variance (the first PC line minimizes the sum of squared perpendicular distances of each point from the line). The second principal component is a linear combination of the original variables that is orthogonal to the first PC and explains the most (remaining) variance assuming this constraint. PCA plot is a scatter plot based on the first principal components that is commonly used to visualize output of clustering multidimensional data.

Clustering output can be used for assessment of dispersion of the IOP measured by different techniques in clusters of cases with similar biomechanical eye properties (quantified by CH and CRF) [97]. Figure 8.5 shows dendrogram tree for the input dataset. We compared 4 groups using box plots (see Figure 8.6) and PCA plot (see Figure 8.7). Group no. 2 is clearly separated from the others. It contains 8 positive cases with high IOP values. The remaining groups are larger and include positive and negative cases in different proportions. Dispersion of IOP seems to be higher for positive cases than for negative cases in most groups. Selected new cases can be compared with cases from the clusters to identify outliers. Selected individual cases can be compared with the delineated clusters to identify outliers.

Exploratory data analysis can be seen as an important part of any quantitative research. It can explain some data properties and at the same time lead to generation of new questions. It is a multi-step process involving data cleaning, imputation and transformation. Collaboration of doctors, scientists and engineers will enable selection of the most appropriate scenarios for implementation in clinical environment.

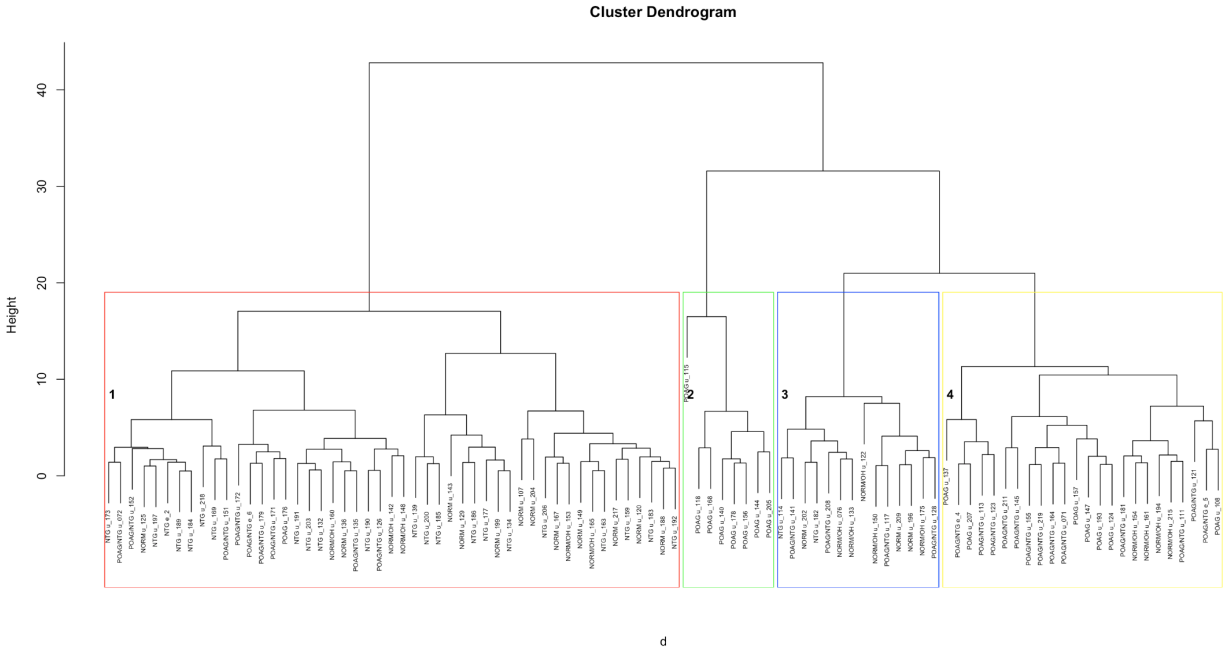


FIGURE 8.5: Dendrogram tree cut into 4 groups (color rectangles). Leaves are labeled with case ID and diagnosis.

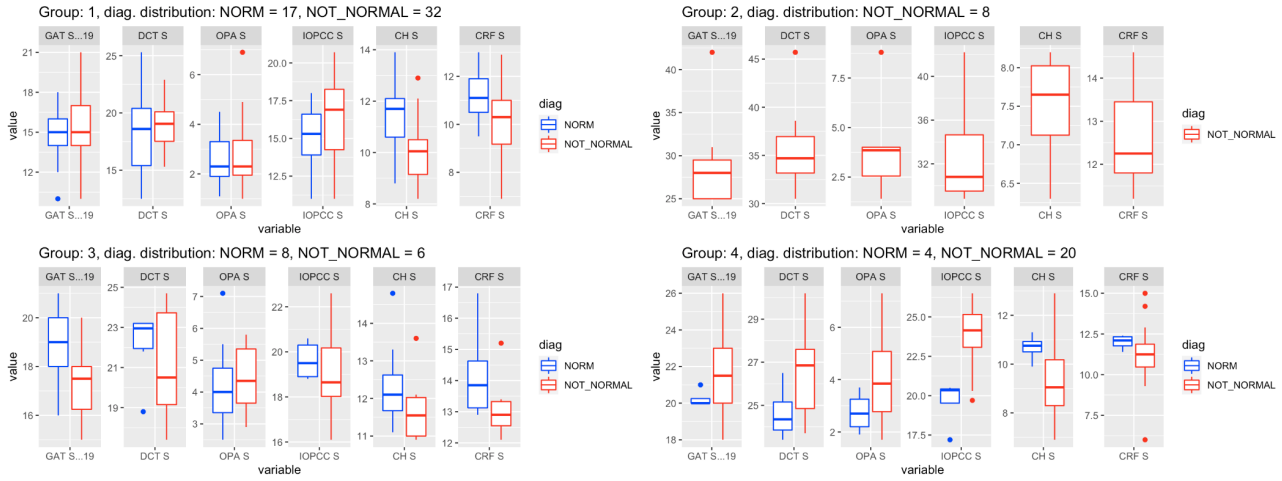


FIGURE 8.6: Comparison of 4 case groups shown in the dendrogram. Box plots are split according to the diagnosis (negative/positive).

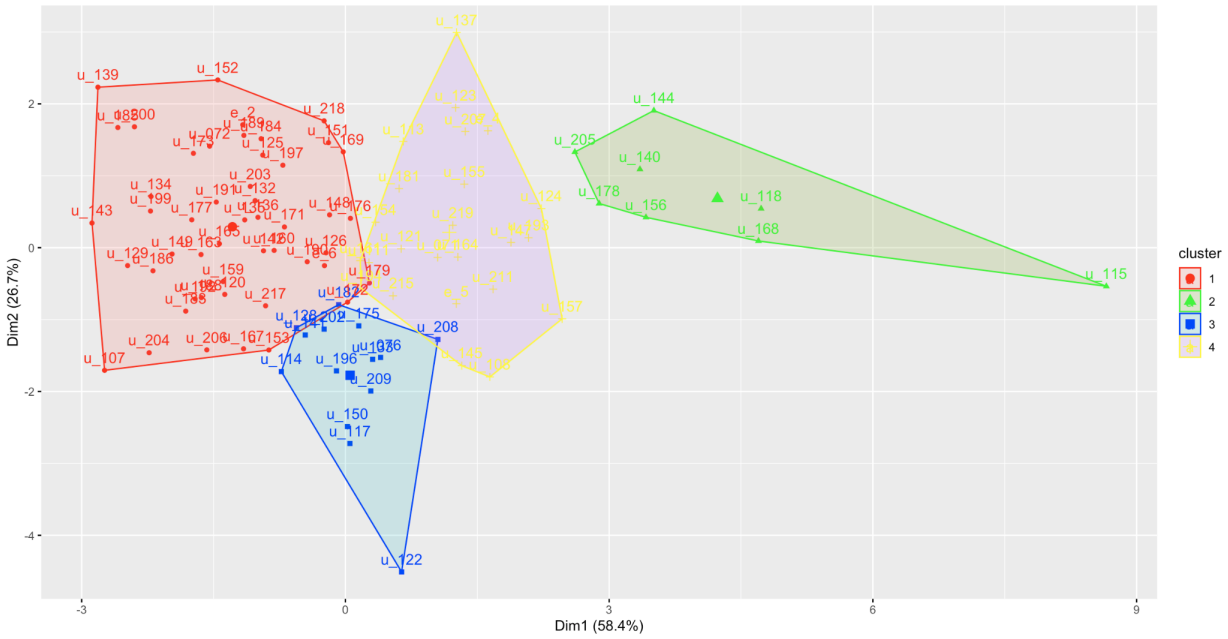


FIGURE 8.7: PCA plot with 4 clusters from the dendrogram. Cases are represented by points in the plot based on the first two principle components (dimensions).

# Chapter 9

# Summary

This thesis was focused on the development of efficient support for glaucoma diagnosis based on Trigerfish CLS and cardiac sensor data supplemented with selected clinical measurements of the eye. This approach can be considered as an innovative application of machine learning techniques for analysis of the data acquired by the devices that have recently became available to doctors.

Although modern OCT devices allow for detailed visualization of the eye morphological structures these imaging modalities don't enable tracking of dynamic changes in this complex system which is influenced by many external factors during the whole day. As for other commonly used diagnostic methods, IOP measurements can only be performed several times a day using standard tonometric techniques.

Incorporation of data from cardiovascular system monitoring device that is aligned to the Triggerfish CLS signal enables assessment of the eye during 24-hour session. In addition, the long monitoring period can be divided into time intervals according to the physiological circadian cycle properties.

The following contributions were presented in the thesis:

- Development of efficient predictive ML models for glaucoma diagnosis support that involve Triggerfish CLS and cardiac sensor data. These models don't depend on standard IOP measurements made using applanation tonometry. Instead, inclusion of the measurements of corneal biomechanical properties improves performance metrics. Furthermore, predictions of the models can be interpreted in terms related to the basic data properties and common clinical concepts.

- Implementation of raw data processing methods for assessment of relationship between Triggerfish CLS and cardiac sensor data in time intervals defined according to the physiological circadian cycle properties. This is an important relationship as the autoregulatory capacity of the eye requires adequate ocular blood flow which depends on cardiovascular system efficiency.

- Design conception and initial implementation of the software system for glaucoma diagnosis support based on ML models involving sensor data. Data management and visualization services can be used in diagnostic and collaborative research scenarios for the eye doctors and data scientists.

Solutions proposed by the author can be applied in clinical setting to support glaucoma diagnosis and development of personalized approaches for the management of the disease.

Possible future directions of the research include development of deep neural networks for processing sensor data that can provide diagnostic information supplementary to the presented solutions. Statistical analysis of the sensor data collected repeatedly over long (i.e. many years) time for the relatively large group of patients seems an important direction, once such data will be available. Comprehensive investigation of the longitudinal data may lead to more accurate assessment of the eye in early glaucoma stages and identification of specific case subgroups.

# Bibliography


[1] C. G. Campbell et al. The potential application of artificial intelligence for diagnosis and management of glaucoma in adults. *British Medical Bulletin*, 134(1):21–33, 2020. doi: 10.1093/bmb/ldaa012.

[2] K. Mansouri, A. P. Tanna, C. G. De Moraes, et al. Review of the measurement and management of 24-hour intraocular pressure in patients with glaucoma. *Survey of Ophthalmology*, 65(2):171–186, 2020. doi: https://doi.org/10.1016/j.survophthal.2019.09.004.

[3] K. R. Martin et al. Use of Machine Learning on Contact Lens Sensor-Derived Parameters for the Diagnosis of Primary Open-angle Glaucoma. *American Journal of Ophthalmology*, 194:46–53, 2018. doi: https://doi.org/10.1016/j.ajo.2018.07.005.

[4] S. Yousefi. Clinical Applications of Artificial Intelligence in Glaucoma. *Journal of Ophthalmic & Vision Research*, 18(1):97–112, 2023. doi: 10.18502/jovr.v18i1.12730.

[5] J. H. K. Liu, K. Mansouri, and R. N. Weinreb. Estimation of 24-Hour Intraocular Pressure Peak Timing and Variation Using a Contact Lens Sensor. *PLoS ONE*, 10(6), 2015. doi: 10.1371/journal.pone.0129529.

[6] N. Tojo, S. Abe, M. Ishida, T. Yagou, and A. Hayashi. The Fluctuation of Intraocular Pressure Measured by a Contact Lens Sensor in Normal-Tension Glaucoma Patients and Nonglaucoma Subjects. *Journal of Glaucoma*, 26(3):195–200, 2017. doi: https://doi.org/10.1097/ijg.0000000000000517.

[7] K. Willekens, R. Rocha, K. Van Keer, et al. Review on Dynamic Contour Tonometry and Ocular Pulse Amplitude. *Ophthalmic Research*, 55:91–98, 2016. doi: 10.1159/000441796.

[8] K. Mansouri, R. N. Weinreb, and J. H. K. Liu. Efficacy of a Contact Lens Sensor for Monitoring 24-H Intraocular Pressure Related Patterns. *PLoS ONE*, 10(5):1–14, 2015. doi: 10.1371/journal.pone.0125530.

[9] I. Pallikaris, M. K. Tsilimbaris, and A. I. Dastiridou. *Ocular Rigidity, Biomechanics and Hydrodynamics of the Eye*. Springer, 2021. doi: https://doi.org/10.1007/978-3-030-64422-2.

[10] K. R. Pillunat et al. Nocturnal blood pressure in primary open-angle glaucoma. *Acta Ophthalmologica*, 93(8):621–626, 2015. doi: 10.1111/aos.12740.

[11] Y. J. Kim, K. S. Lee, J. R. Lee, et al. Ocular pulse amplitude as a dynamic parameter and its relationship with 24-h intraocular pressure and blood pressure in glaucoma. *Experimental Eye Research*, 115:65–72, 2013. doi: https://doi.org/10.1016/j.exer.2013.06.010.

[12] T. Guergueb and M. A. Akhloufi. A review of deep learning techniques for glaucoma detection. *SN Computer Science*, 4(274):21–33, 2023. doi: https://doi.org/10.1007/s42979-023-01734-z.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv*, 2015. doi: https://doi.org/10.48550/arXiv.1512.03385.

[14] Richard M. Levenson, Yashbir Singh, et al. Advancing Precision Medicine: Algebraic Topology and Differential Geometry in Radiology and Computational Pathology. *Laboratory Investigation*, 104(6):102060, 2024. doi: doi.org/10.1016/j.labinv.2024.102060.

[15] M. Martinez-Garcia and E. Hernandez-Lemus. Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*, 8:784455, 2022. doi: 10.3389/fmed.2021.784455.

[16] A. C. Thompson, A. A. Jammal, and F. A. Medeiros. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Translational Vision Science & Technology*, 9(2):42, 2020. doi: 10.1167/tvst.9.2.42.

[17] F. Mabuchi et al. Genetic Variants Associated With the Onset and Progression of Primary Open-Angle Glaucoma. *American Journal of Ophthalmology*, 215:135–140, 2020. doi: https://doi.org/10.1016/j.ajo.2020.03.014.

[18] Y. Chen et al. Genetic Variants Associated With Different Risks for High Tension Glaucoma and Normal Tension Glaucoma in a Chinese Population. *Investigative Ophthalmology & Visual Science*, 56(4):2595–2600, 2015. doi: https://doi.org/10.1167/iovs.14-16269.

[19] Hubert Świerczyński, Juliusz Pukacki, Szymon Szczęsny, Cezary Mazurek, and Robert Wasilewicz. Sensor data analysis and development of machine learning models for detection of glaucoma. *Biomedical Signal Processing and Control*, 86:105350, 2023. doi: https://doi.org/10.1016/j.bspc.2023.105350.

[20] M. Chen, L. Kueny, and A. L. Schwartz. The role of corneal hysteresis during the evaluation of patients with possible normal-tension glaucoma. *Clinical Ophthalmology*, 12:555–559, 2018. doi: https://doi.org/10.2147/OPTH.S161675.

[21] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1 – 85, 2022. doi: 10.1214/21-SS133.

[22] R.T. Sutton, D. Pincock, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1), 2020. doi: 10.1038/s41746-020-0221-y.

[23] J. Eichner and M. Das. Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations. *Agency for Healthcare Research and Quality*, AHRQ Publication No. 10-0064-EF, 2010. URL `https://digital.ahrq.gov/sites/default/files/docs/page/CDS_challenges_and_barriers.pdf`.

[24] D.J. Park, M.W. Park, H. Lee, et al. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Nature Scientific Reports*, 11 (7567), 2021. doi: https://doi.org/10.1038/s41598-021-87171-5.

[25] Eta S. Berner. *Clinical Decision Support Systems*. Springer, 3rd edition, 2016. doi: doi.org/10.1007/978-3-319-31913-1.

[26] G.T. Berge et al. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. *BMC Medical Informatics and Decision Making*, 23(5), 2023. doi: 10.1186/s12911-023-02101-x.

[27] Tarek M. Shaarawy, Mark B. Sherwood, et al. *Glaucoma, Volume 1: Medical Diagnosis & Therapy*. Elsevier, 2nd edition, 2015. doi: https://doi.org/10.1016/C2011-1-04562-9.

[28] N. Zhang, Y. Li J. Wang, et al. Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. *Nature Scientific Reports*, 11(13762), 2021. doi: https://doi.org/10.1038/s41598-021-92971-w.

[29] M. P. Chan, D. C. Broadway, A. P. Khawaja, et al. Glaucoma and intraocular pressure in EPIC-Norfolk Eye Study: cross sectional study. *BMJ*, 358(8121), 2017. doi: 10.1136/bmj.j3889.

[30] G. Garcia et al. Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. *Artificial Intelligence in Medicine*, 118: 102–132, 2021. doi: https://doi.org/10.1016/j.artmed.2021.102132.

[31] S. Maetschke, B. Antony, H. Ishikawa, et al. A feature agnostic approach for glaucoma detection in OCT volumes. *PLoS ONE*, 14(7), 2019. doi: https://doi.org/10.1371/journal.pone.0219126.

[32] Lauren J. Coan, Bryan M. Williams, et al. Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review. *Survey of Ophthalmology*, 68(1):17–41, 2023. doi: https://doi.org/10.1016/j.survophthal.2022.08.005.

[33] L. Jones et al. CLEAR - Contact lens technologies of the future. *Contact Lens and Anterior Eye*, 44(2):398–430, 2021. doi: https://doi.org/10.1016/j.clae.2021.02.007.

[34] Yifei Niu, Junfeng Ji, et al. Regenerative treatment of ophthalmic diseases with stem cells: Principles, progress, and challenges. *Advances in Ophthalmology Practice and Research*, 4(2):52–64, 2024. doi: https://doi.org/10.1016/j.aopr.2024.02.001.

[35] Rahul M. Dhodapkar, Emily Li, et al. Deep learning for quality assessment of optical coherence tomography angiography images. *Nature Scientific Reports*, 12(13775), 2022. doi: https://doi.org/10.1038/s41598-022-17709-8.

[36] J. Crawford Downs, Michael D. Roberts, and Ian A. Sigal. Glaucomatous cupping of the lamina cribrosa: A review of the evidence for active progressive remodeling as a mechanism. *Experimental Eye Research*, 93(2):133–140, 2011. ISSN 0014-4835. doi: https://doi.org/10.1016/j.exer.2010.08.004.

[37] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. URL `https://D2L.ai`.

[38] Rita Marques et al. Automatic segmentation of the optic nerve head region in optical coherence tomography: A methodological review. *Computer Methods and Programs in Biomedicine*, 220:106801, 2022. doi: https://doi.org/10.1016/j.cmpb.2022.106801.

[39] J. Kim, L. Tran, E. Y. Chew, et al. Optic Disc and Cup Segmentation for Glaucoma Characterization Using Deep Learning. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 489–494, 2019. doi: 10.1109/CBMS.2019.00100.

[40] Ko Eun Kim et al. Development and Validation of a Deep Learning System for Diagnosing Glaucoma Using Optical Coherence Tomography. *Journal of Clinical Medicine*, 9(7), 2020. doi: 10.3390/jcm9072167.

[41] C. G. Campbell, D. S. W. Ting, et al. The potential application of artificial intelligence for diagnosis and management of glaucoma in adults. *British Medical Bulletin*, 134(1): 21–33, 2020. doi: https://doi.org/10.1093/bmb/ldaa012.

[42] Suorong Yang, Weikang Xiao, et al. Image data augmentation for deep learning: A survey. *arXiv*, 2023. doi: https://arxiv.org/abs/2204.08610.

[43] Eun Ji Lee, Tae-Woo Kim, et al. Predictive Modeling of Long-Term Glaucoma Progression Based on Initial Ophthalmic Data and Optic Nerve Head Characteristics. *Translational Vision Science & Technology*, 11(24), 2022. doi: https://doi.org/10.1167/tvst.11.10.24.

[44] Connor J. Greatbatch, Qinyi Lu, et al. Deep Learning-Based Identification of Intraocular Pressure-Associated Genes Influencing Trabecular Meshwork Cell Morphology. *Ophthalmology Science*, 4(4), 2024. doi: doi.org/10.1016/j.xops.2024.100504.

[45] A. M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433.

[46] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009. doi: https://doi.org/10.1007/978-0-387-84858-7.

[47] Steven S. Skiena. *The Data Science Design Manual*. Springer, 2017. doi: https://doi.org/10.1007/978-3-319-55444-0.

[48] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451.

[49] Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199 – 231, 2001. doi: 10.1214/ss/1009213726.

[50] Frank E. Harrell. *Biostatistics for Biomedical Research*. Vanderbilt University, 2024. URL `https://hbiostat.org/bbr/`.

[51] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874, 2006. doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[52] S. Lohr. *What Ever Happened to IBM's Watson?* The New York Times, 2021. URL `https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html`.

[53] D. Lazer et al. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343 (6176):1203–1205, 2014. doi: doi.org/10.1126/science.1248506.

[54] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591. URL `https://pbiecek.github.io/ema/`.

[55] Mateusz Staniak and Przemysław Biecek. Explanations of Model Predictions with live and breakDown Packages. *The R Journal*, 10(2):395–409, 2018. doi: 10.32614/RJ-2018-072.

[56] M. Ribeiro et al. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, USA*, 2016. doi: 10.1145/2939672.2939778.

[57] K. Gillmann, R. Wasilewicz, et al. Continuous 24-hour measurement of intraocular pressure in millimeters of mercury (mmHg) using a novel contact lens sensor: Comparison with pneumatonometry. *PLoS ONE*, 16(3): e0248211, 2021. doi: https://doi.org/10.1371/journal.pone.0248211.

[58] R. Shean, N. Yu, et al. Advances and Challenges in Wearable Glaucoma Diagnostics and Therapeutics. *Bioengineering (Basel)*, 11(2):138, 2024. doi: https://doi.org/10.3390/bioengineering11020138.

[59] Jian Li, Huiling Jia, et al. Thin, soft, wearable system for continuous wireless monitoring of artery blood pressure. *Nature Communications*, 14:5009, 2023. doi: https://doi.org/10.1038/s41467-023-40763-3.

[60] Liangqi Wang, Shuo Tian, et al. A new method of continuous blood pressure monitoring using multichannel sensing signals on the wrist. *Microsystems & Nanoengineering*, 9(117), 2023. doi: https://doi.org/10.1038/s41378-023-00590-4.

[61] G. Bilo et al. Validation of the Somnotouch-NIBP noninvasive continuous blood pressure monitor according to the European Society of Hypertension International Protocol revision 2010. *Blood Pressure Monitoring*, 20(5):291–294, 2015. doi: https://doi.org/10.1097/mbp.0000000000000124.

[62] J. Nyvad, K.L. Christensen, et al. The cuffless SOMNOtouch NIBP device shows poor agreement with a validated oscillometric device during 24-h ambulatory blood pressure monitoring. *The Journal of Clinical Hypertension*, 23(1):61–67, 2020. doi: 10.1111/jch.14135.

[63] W. Śródka. Goldmann applanation tonometry – not as good as gold. *Acta of Bioengineering and Biomechanics*, 12(2):39–47, 2010. URL `https://actabio.pwr.edu.pl/en/archive/vol12--no-2--2010`.

[64] Dianne H. Glass, Cynthia J. Roberts, et al. A Viscoelastic Biomechanical Model of the Cornea Describing the Effect of Viscosity and Elasticity on Hysteresis. *Investigative Ophthalmology & Visual Science*, 49(9):3919–3926, 2008. doi: https://doi.org/10.1167/iovs.07-1321.

[65] Hafize Gokben Ulutas, Guven Ozkaya, et al. Comparison of central corneal thickness measured by ultrasound pachymetry, corneal topography, spectral domain optical coherence tomography, and non-contact specular microscopy. *Photodiagnosis and Photodynamic Therapy*, 42:103527, 2023. doi: https://doi.org/10.1016/j.pdpdt.2023.103527.

[66] K. Balaskas J. Hopkins, P. A. Keane. Delivering personalized medicine in retinal care: from artificial intelligence algorithms to clinical application. *Current Opinion in Ophthalmology*, 31:329–336, 2020. doi: https://doi.org/10.1097/icu.0000000000000677.

[67] R. Wasilewicz, C. Mazurek, and J. Pukacki. Influence of cardiovascular system on 24 hour ocular volume changes, measured with contact lens sensor in healthy and POAG subjects. *8th World Glaucoma Congress, Melbourne, Kugler Publications on behalf of the World Glaucoma Association*, pages 198–199, 2019. URL https://wga.one/download/228/2019/163907/wgc-2019-abstract-book.pdf.

[68] R. Wasilewicz et al. Daily biorhythms of ocular volume changes and the cardiovascular system functional parameters in healthy, ocular hypertension, normal tension and primary open angle glaucoma populations. *Investigative Ophthalmology & Visual Science*, 55(13):142, 2014. URL https://iovs.arvojournals.org/article.aspx?articleid=2266649.

[69] C. N. Susanna, A. Diniz-Filho, and F. B. Daga. A Prospective Longitudinal Study to Investigate Corneal Hysteresis as a Risk Factor for Predicting Development of Glaucoma. *American Journal of Ophthalmology*, 187:148–152, 2018. doi: https://doi.org/10.1016/j.ajo.2017.12.018.

[70] Ch. Gisler, A. Ridi, M. Fauquex, et al. Towards Glaucoma Detection Using Intraocular Pressure Monitoring. *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, Tunisia, 2014*, pages 255–260, 2015. doi: https://doi.org/10.1109/SOCPAR.2014.7008015.

[71] R. Wasilewicz et al. 24 hour continuous ocular tonography Triggerfish and biorhythms of the cardiovascular system functional parameters in healthy and glaucoma populations. *Acta Ophthalmologica*, 91(s252), 2013. doi: 10.1111/j.1755-3768.2013.2721.x.

[72] K. Gillmann, R. N. Weinreb, and K. Mansouri. The effect of daily life activities on intraocular pressure related variations in open-angle glaucoma. *Nature Scientific Reports*, 11(6598), 2021. doi: doi.org/10.1038/s41598-021-85980-2.

[73] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 4th edition, 2017. doi: doi.org/10.1007/978-3-319-52452-8.

[74] M. Landry and A. Bartz. *Machine Learning with R and H2O*. H2O.ai, Inc., 7th edition, 2021. URL https://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/RBooklet.pdf.

[75] P. Teisseyre, R. A. Kłopotek, and J. Mielniczuk. Random Subspace Method (RSM) for Linear Regression. *Computational Statistics*, 31:943–972, 2016. doi: https://doi.org/10.1007/s00180-016-0658-2.

[76] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002. doi: https://doi.org/10.1613/jair.953.

[77] S. Kaushik, S. S. Pandav, and A. Banger. Relationship between corneal biomechanical properties, central corneal thickness, and intraocular pressure across the spectrum of glaucoma. *American Journal of Ophthalmology*, 153(5):840–849, 2012. doi: https://doi.org/10.1016/j.ajo.2011.10.032.

[78] Liang Liang, Ran Zhang, and Li-Ye He. Corneal hysteresis and glaucoma. *International Ophthalmology*, 39(8):1909–1916, 2019. doi: https://doi.org/10.1007/s10792-018-1011-2.

[79] S. Lundberg and S. Lee. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, page 4768–4777, 2017. URL `https://dl.acm.org/doi/10.5555/3295222.3295230`.

[80] S. L. Mansberger, F. Medeiros, and M. Gordon. Diagnostic Tools for Calculation of Glaucoma Risk. *Survey of Ophthalmology*, 53(6):11–16, 2008. doi: 10.1016/j.survophthal.2008.08.005.

[81] L. Zimprich, J. Diedrich, A. Bleeker, et al. Corneal Hysteresis as a Biomarker of Glaucoma: Current Insights. *Clinical Ophthalmology*, 14:2255–2264, 2020. doi: https://doi.org/10.2147/opth.s236114.

[82] Sejong Oh, Yuli Park, et al. Explainable Machine Learning Model for Glaucoma Diagnosis and Its Interpretation. *Diagnostics (Basel)*, 11(3):510, 2021. doi: 10.3390/diagnostics11030510.

[83] H. Muccini and K. Vaidhyanathan. Software Architecture for ML-based Systems: What Exists and What Lies Ahead. *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN), Madrid, Spain*, pages 121–128, 2021. doi: 10.1109/WAIN52551.2021.00026.

[84] B. Huang G. Liu, Z. Liang, et al. Microservices: architecture, container, and challenges. *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Macau, China*, pages 629–635, 2020. doi: 10.1109/QRS-C51114.2020.00107.

[85] O. Al-Debagy and P. Martinek. A Comparative Review of Microservices and Monolithic Architectures. *2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary*, pages 000149–000154, 2018. doi: 10.1109/CINTI.2018.8928192.

[86] C. V. Anikwe et al. Mobile and wearable sensors for data-driven health monitoring system: State-of-the-art and future prospect. *Expert Systems with Applications*, 202 (117362), 2022. doi: 10.1016/j.eswa.2022.117362.

[87] V. L. Patel S. Myneni. Organization of biomedical data for collaborative scientific research: A research information management system. *International Journal of Information Management*, 30(3):256–264, 2010. doi: https://doi.org/10.1016/j.ijinfomgt.2009.09.005.

[88] L. Stadler and A. Welc. Optimizing R language execution via aggressive speculation. *SIGPLAN Notices*, 52(2):84–95, 2017. doi: https://doi.org/10.1145/3093334.2989236.

[89] D. Hausmann, C. Zulian, et al. Tracing the decision-making process of physicians with a Decision Process Matrix. *BMC Medical Informatics and Decision Making*, 16(133), 2016. doi: 10.1186/s12911-016-0369-1.

[90] Przemyslaw Biecek. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84):1–5, 2018. URL `http://jmlr.org/papers/v19/18-416.html`.

[91] Gottfried E. Noether. *Introduction to Statistics*. Springer, 1991. doi: https://doi.org/10.1007/978-1-4612-0943-0.

[92] National Research Council. *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond*. The National Academies Press, 2014. doi: https://doi.org/10.17226/18722.

[93] M. G. Lawrence et al. Characteristics, potentials, and challenges of transdisciplinary research. *One Earth*, 5(1):44–61, 2022. doi: 10.1016/j.oneear.2021.12.010.

[94] W. L. Morton et al. Architectures of adaptive integration in large collaborative projects. *Ecology and Society*, 20(4):5, 2015. doi: http://dx.doi.org/10.5751/ES-07788-200405.

[95] A. Arora et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29:2929–2938, 2023. doi: 10.1038/s41591-023-02608-w.

[96] Caitlin E. Coombes, Xin Liu, et al. Simulation-derived best practices for clustering clinical data. *Journal of Biomedical Informatics*, 118:103788, 2021. doi: doi.org/10.1016/j.jbi.2021.103788.

[97] Aachal Kotecha, Richard A. Russell, et al. Biomechanical parameters of the cornea measured with the Ocular Response Analyzer in normal eyes. *BMC Ophthalmology*, 14 (11), 2014. doi: doi.org/10.1186/1471-2415-14-11.



Poznan University of Technology
Faculty of Computing and Telecommunications
Institute of Computing Science

This document was typeset using LaTeX.